3D-GCP: An Analytical Model for the Impact of Process Variations on the Critical Path Delay Distribution of 3D ICs*

Siddharth Garg, Diana Marculescu
Dept. of ECE, Carnegie-Mellon University, PA USA
E-mail: sgarg1@andrew.cmu.edu, dianam@ece.cmu.edu

Abstract— 3D Integrated Circuits (ICs) have been recently proposed as a solution to the increasing wire delay concerns in scaled technologies. At the same time, technology scaling leads to increasing variability in manufacturing process parameters, making it imperative to quantify the impact of these variations on performance. In this work, we take, to the best of our knowledge, **the first step towards formally modeling the impact of process variations on the clock frequency of fully-synchronous (FS) 3D ICs**. The proposed analytical models demonstrate theoretically and experimentally that **3D designs behave very differently under the impact of process variations as compared to equivalent 2D designs**. In particular, for the same number of critical paths, we show that a 3D design is always less likely to meet a pre-defined frequency target compared to its 2D counterpart. Furthermore, as opposed to models for 2D ICs, the 3D models need to accurately account for not only within-die (WID) critical paths, i.e., paths that lie entirely within one of the die layers, but also D2D critical paths that use through-silicon vias (TSVs) to span across multiple dies in the 3D stack. Finally, we show, theoretically and experimentally, that the mapping of critical paths to the die layers of a 3D IC can also affect the timing yield of a design, while the mapping issue does not arise in the 2D case since there is only a single die layer in a 2D IC. The accuracy of the proposed models is experimentally verified and found to be in excellent agreement with detailed SPICE and gate-level Monte Carlo (MC) simulations.

Keywords— Statistical timing analysis, 3D Integrated Circuits (ICs)

I. Introduction

Wire delay has been shown to be an increasing fraction of gate delay with technology scaling, making it difficult to cross the length of a die in a single clock cycle [11]. 3D IC technology is a promising solution to the growing wire delay problem and is being aggressively pursued in industry [9] and academia [6], [2], [12]. While there exist numerous flavors of 3D IC technology, in this work, we concentrate primarily on stacked-die 3D integration. In this technology, each device layer is manufactured separately using a conventional 2D fabrication process and the planar (i.e., 2D) dies are subsequently stacked and bonded vertically on top of each other to realize a 3D IC. Wires can cross from one layer in the stack to another using die-to-die interconnects called through-silicon vias (TSVs). 3D integration is therefore able to significantly reduce the average and worst-case wirelength of a conventional 2D design, thereby alleviating the performance impact of slow global interconnects.

Along with wire delay, the increase in manufacturing process variations (PVs) is another major cause for concern in scaled technologies, since it leads to variations in the performance and power characteristics of fabricated dies. Moreover, process variations manifest themselves as both die-to-die (D2D) variations that affect each die differently but all transistors on a particular die in the same way, and within-die (WID) variations that affect each transistor on every die differently. From a performance perspective, there has been significant amount of prior work to accurately and efficiently model the impact of PVs on the cycle time (or clock frequency) of fully-synchronous 2D ICs. While a number of authors have proposed so called Statistical Static Timing Analysis (SSTA) algorithms to compute the process variation driven distribution of the maximum critical path delay for *specific* designs, [3] introduced the *generic critical path (GCP)* model that encapsulates the low-level implementation details of a circuit using two high-level parameters (the number of critical paths in the circuit and the number of stages per critical path), which are then used to compute the maximum critical path delay distribution. The authors use the proposed model to analyze the impact of technology scaling and increased design complexity on the mean and variance of critical path delay. Moreover, the GCP model has since been *widely* used in a number of studies at the micro-architecture and system-level to efficiently characterize process variations without having to resort to time consuming circuit level timing statistical simulation or analysis [14], [15], [13].

Unfortunately, for a number of reasons that will be made clear later in the paper, the GCP model introduced by [3] cannot be simply extended for the case of 3D ICs. Therefore, we introduce for the first time, a generic critical path model for 3D ICs (3D-GCP) that allows us to analytically characterize the impact of process variations on the cycle time distribution for fully-synchronous 3D systems. The proposed models provide the following advantages: (1) they allow for *efficient characterization* of the impact of increased design complexity, increased integration complexity (i.e., more die layers) and increased magnitude of process variations on the maximum critical path delay distribution of 3D ICs; (2) they serve as *efficient high-level process variation models* that can be used at the micro-architecture and system level design abstraction; and (3) they serve to *theoretically motivate*

* This research was supported in part by NSF award CCR-0702451.

and explain variability-aware, design time optimization approaches that are unique to the 3D IC design flow. The first two advantages are analogous to those provided by the original GCP model to 2D circuits, while the third is specific to the case of 3D ICs and arises from gate/module to die layer mapping flexibility that exists in 3D designs.

II. Related Work

Performance analysis of 3D ICs in terms of improvement in timing characteristics has been studied extensively before. [6] and [2] study the change in the wirelength distribution, and implicitly the increase in clock frequency, obtained by moving from a 2D to a 3D implementation. On a related note, [12] studies the improvement in system performance metrics such as latency and throughput obtained by reducing the average number of hops between modules using a 3D network-on-chip (NoC). However, none of these works have investigated the role of process variations on performance, which is the focus of our work.

Over the last few years, the analysis of impact of process variations on cycle time and frequency for 2D circuits has been the focus of extensive research. We point the interested reader to a comprehensive tutorial paper on this subject [7]. While some of the previously proposed SSTA algorithms could potentially be adapted to handle the case of 3D ICs, the goal of this work is more aligned with the high-level GCP model introduced by [3]. However, compared to our work, [3] focuses exclusively on conventional 2D circuits, which as we will show, cannot be trivially extended for the case of 3D ICs. Finally, the only other work to look at variations for 3D ICs [8] makes use of Monte Carlo simulations and has a very different goal, variability-aware post fabrication assembly strategies, compared to our work.

III. Preliminaries and Assumptions

We begin by introducing the mathematical notation and assumptions about fully-synchronous 2D and 3D designs that we make in this work. Before introducing the proposed model for 3D systems, we will first briefly review the high-level GCP model proposed by [3] to study the impact of process variations for 2D systems. In particular, [3] assumes that a generic 2D circuit can be fully characterized, from a timing perspective, by N_{cp} critical paths, where each path consists of a chain of n_{cp} two-input *NAND* gates. Furthermore, the D2D component of process variations is modeled using a single random variable (RV) G and the uncorrelated WID component is modeled using a set of independent and identically distributed (*iid*) RVs L_{ij} ($1 \leq i \leq N_{cp}$,$1 \leq j \leq n_{cp}$) that represent the impact of WID random variations on gate j of critical path i. The maximum deviation in delay for a fully-synchronous 2D design, ΔT^{2D}_{max}, can therefore be written as:

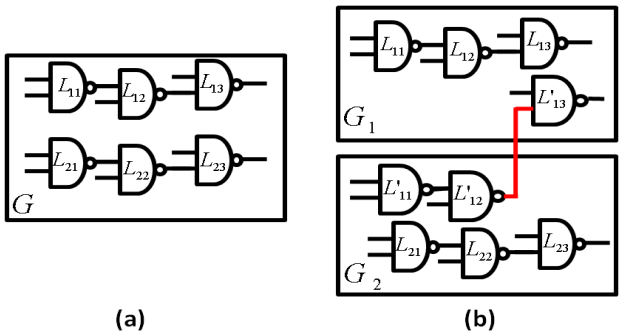

Fig. 1. (a) Conventional 2D IC with two critical paths (b) 3D IC with two WID critical paths, and one D2D critical path. The TSV is shown in red. Note that as a matter of convention, the WID and D2D critical paths are numbered separately.

$$\begin{aligned} \Delta T^{2D}_{max} &= \max_{1\leq i\leq N_{cp}} (\sum_{j=1}^{n_{cp}} \alpha(G + L_{ij})) \\ &= \max_{1\leq i\leq N_{cp}} (n_{cp}\alpha G + \sum_{j=1}^{n_{cp}} \alpha L_{ij}) \end{aligned} \quad (1)$$

where α is the sensitivity of the gate delay to process variations. Now, assuming that variations in process parameters are normally distributed [3], i.e., $G \sim N(0, \sigma_G)$ and $L_{ij} \sim N(0, \sigma_L)$, we can write the probability that $\Delta T^{2D}_{max} \leq \tau$ as:

$$Pr\{\Delta T^{2D}_{max} \leq \tau\} = F_{\Delta T^{2D}_{max}}(\tau) = f_G(\frac{\tau}{\alpha n_{cp}}) * (F_L(\frac{\tau}{\alpha\sqrt{n_{cp}}})^{N_{cp}}) \quad (2)$$

where $*$ represents the convolution operation, and $f_X()$ and $F_X()$ represent the probability density function (*pdf*) and cumulative distribution function (*cdf*) of RV X respectively.

To model 3D systems, we assume, without any loss of generality, a system consisting of L die layers[1] stacked on top of each other. Connections between die layers are accomplished using through-silicon vias (TSVs). As shown in Figure 1(b), the critical paths in such a system can be classified as either:

- **Within-die (WID) Critical Paths**, i.e., paths that lie fully in one of the L die layers. The total number of WID critical paths in the system is represented as N^{WID}_{cp}, with layer i containing N^i_{cp} paths ($\sum_{i=1}^{L} N^i_{cp} = N^{WID}_{cp}$). Each WID critical path is assumed to consist of n^{WID}_{cp} gates. Finally, without loss of generality, the mapping function $m()$ maps each WID critical path to one of the L device layers.
- **Die-to-die (D2D) Critical Paths**, i.e., critical paths that use TSVs to cross from one die layer to another. The total number of D2D critical paths in the system is represented as N^{D2D}_{cp}. Each D2D critical path consists of n^{D2D}_{cp}

[1]Not to be confused with L_{ij} that denote the uncorrelated WID RVs.

gates, where the number of gates that lie in layer i is given by n_{cp}^i ($n_{cp}^{D2D} = \sum_{i=1}^{L} n_{cp}^i$).

Figure 1(b) shows an example of a 3D system with two layers ($L = 2$), two WID critical paths ($N_{cp}^{WID} = 2$, $N_{cp}^1 = 1$, $N_{cp}^2 = 1$) consisting of three stages ($n_{cp}^{WID} = 3$) and one D2D critical path ($N_{cp}^{D2D} = 1$) consisting also of three stages ($n_{cp}^{D2D} = 3$, $n_{cp}^1 = 1$, $n_{cp}^2 = 2$). We note that while all the results that follow will be based on this canonical model for the critical paths in a 3D circuit, more complicated models (for example, one in which the number of stages for a WID critical path vary from one layer to another) can be analyzed using simple extensions of the proposed methodology.

As opposed to 2D systems where there is a single RV representing D2D variations, each layer in a 3D system has a different D2D variation component. We therefore define, for each layer i in the 3D system, a RV G_i ($1 \leq i \leq L$) that corresponds to the impact of D2D variations in that layer. Furthermore, for *homogeneous* stacked-die 3D systems in which each die is fabricated in the same process technology, the RVs G_i and G_j ($i \neq j$) can be assumed to be independent and identically distributed (*iid*) RVs. On the other hand, for *heterogeneous* integration, the D2D RVs of two layers fabricated in different processes may not be identically distributed, although they would still be independent. While the proposed framework can handle both cases, in the rest of the paper, we will focus specifically on the former case of homogeneous integration. Finally, the random WID component of process variation is modeled by RVs L_{ij} and L'_{ij} that represent variations in the process parameters of gate j in the i^{th} WID and D2D critical path respectively. For illustration, the example 3D circuit in Figure 1(b) has been annotated with its corresponding D2D and WID RVs.

Under these assumptions, we can now write the variation in the critical path delay of the i^{th} WID critical path, ΔT_i^{WID}, as:

$$\Delta T_i^{WID} = n_{cp}^{WID} \alpha G_{m(i)} + \sum_{j=1}^{n_{cp}^{WID}} \alpha L_{ij}, 1 \leq i \leq N_{cp}^{WID} \quad (3)$$

and the variation in the critical path delay of the i^{th} D2D critical path, ΔT_i^{D2D}, as:

$$\Delta T_i^{D2D} = \sum_{j=1}^{L} \beta n_{cp}^j G_j + \sum_{j=1}^{n_{cp}^{D2D}} \beta L'_{ij}, 1 \leq i \leq N_{cp}^{D2D} \quad (4)$$

where, as before, α and β represent the sensitivity of a gate in a WID and a D2D critical path to process variations respectively. Note that in general, the sensitivities of WID and D2D paths to process variations can be different due to the differences in the electrical properties of wires and TSVs. In the experimental results section, we provide a detailed SPICE based characterization of these values. Finally, the maximum deviation in critical path delay for a 3D system ΔT_{max}^{3D} can simply be written as:

$$\Delta T_{max}^{3D} = \max_{1 \leq i \leq N_{cp}^{WID}, 1 \leq j \leq N_{cp}^{D2D}} (\Delta T_i^{WID}, \Delta T_j^{D2D}) \quad (5)$$

IV. 3D-GCP Variability Model

Having described the canonical structure of a 3D circuit in terms of a few high-level parameters, we would now like to derive analytical expressions for the *cdf* of the maximum deviation in critical path delay, ΔT_{max}^{3D} in terms of these parameters. However, compared the conventional GCP model, the increase in model complexity for the 3D case is immediately clear. In particular, instead of only two parameters that were used to describe a 2D circuit, the canonical model for a 3D circuit has $2L + 3$ parameters. In addition, as opposed to a single component of D2D variations, each layer of a 3D design has its own RV corresponding to D2D variations. Finally, the introduction of D2D critical paths, which have very different timing characteristics under process variations as compared to WID paths, adds another degree of complexity.

To reduce the complexity, we adopt a constructive approach towards developing the 3D-GCP model - we begin with analyzing 3D designs that consist of *only WID critical paths* and provide exact analytical expressions for the same. Next we introduce D2D critical paths in the system, which makes the computation of an exact analytical expression for the *cdf* of ΔT_{max}^{3D} intractable. We therefore derive analytical *stochastic lower and upper bounds* on the *cdf* of ΔT_{max}^{3D}. As we will show in the experimental results, the derived bounds are, in practice, extremely tight and provide close approximations of the actual *cdf*.

A. Modeling WID Critical Paths

For a design that consists only of WID critical paths, i.e., $N_{cp}^{D2D} = 0$, we can re-write ΔT_{max}^{3D} from Equation 3 and Equation 5 as:

$$\Delta T_{max}^{3D} = \max_{1 \leq i \leq N_{cp}^{WID}} (n_{cp}^{WID} \alpha G_{m(i)} + \sum_{j=1}^{n_{cp}^{WID}} \alpha L_{ij}) \quad (6)$$

By noticing that the maximum critical path deviation for each layer in the 3D system is independent of every other layer, we can write the *cdf* of ΔT_{max}^{3D}, $F_{\Delta T_{max}^{3D}}(\tau)$, as:

$$F_{\Delta T_{max}^{3D}}(\tau) = \prod_{i=1}^{L} f_G(\frac{\tau}{\alpha n_{cp}^{WID}}) * (F_L(\frac{\tau}{\alpha \sqrt{n_{cp}^{WID}}})^{N_{cp}^i}) \quad (7)$$

Note that Equation 7 is the most general form of the analytical *cdf* for 3D ICs with only WID critical paths, i.e., it makes no assumption about how critical paths are mapped to die layers. To make the relationship between the *cdf* and the number of device layers L clearer, we examine the specific case in which the WID critical paths are *evenly divided* between the device layers, i.e., $N_{cp}^i = \frac{N_{cp}^{WID}}{L}$. Under this assumption, we can write $F_{\Delta T_{max}^{3D}}(\tau)$ as:

$$F_{\Delta T_{max}^{3D}}(\tau) = [f_G(\frac{\tau}{\alpha n_{cp}^{WID}}) * (F_L(\frac{\tau}{\alpha \sqrt{n_{cp}^{WID}}})^{\frac{N_{cp}^{WID}}{L}})]^L \quad (8)$$

From this equation, it is clear that besides the exponential dependence of the *cdf* on the number of critical paths in the system, the *cdf* is also exponentially dependent on the number of layers, L, in the system. Furthermore, while D2D variations only impact the variance of the maximum critical path delay for 2D circuits [3], in the 3D case, both the *mean and the variance* are impacted by D2D variations.

A.1 2D Vs. 3D

While the analytical results help to intuitively understand the difference between 2D and 3D circuits from a timing variability perspective, we now try to *theoretically* explore the difference between the two. Indeed, we prove that for the *same number of total critical paths*, and assuming the 3D circuit has no D2D critical paths, **a 2D circuit is always more likely to meet a given timing specification τ than a 3D circuit**, i.e., $Pr\{\Delta T^{2D}_{max} \leq \tau\} \geq Pr\{\Delta T^{3D}_{max} \leq \tau\}$. The proof is based on the following result on comparing the maximum of multivariate Gaussian random vectors:

Theorem 1: Given two Gaussian Random Vectors X and Y of cardinality N such that $E(X_i) = E(Y_i) = 0$ $(1 \leq i \leq N)$, $E(X_i^2) = E(Y_i^2)$ $(1 \leq i \leq N)$, and

$$E(X_i X_j) \geq E(Y_i Y_j), \forall i \neq j \tag{9}$$

then $max(Y) \geq_{st} max(X)$.

Proof: A detailed proof of this theorem can be found in [1], and is not repeated here for clarity of exposition. ■

We note that the $\geq_{st}$ symbol refers to a *stochastic inequality*: $A \geq_{st} B$ implies that $Pr\{A \leq \tau\} \leq Pr\{B \leq \tau\}, \forall \tau \in \mathbb{R}$. Using this result, we now prove formally the relationship between 2D and 3D circuits.

Theorem 2: If (a) the number of critical paths for a 2D circuit is the same as the total number of WID critical paths for a 3D circuit, i.e., $N^{2D}_{cp} = N^{WID}_{cp}$, (b) the number of stages for the 2D and 3D circuit are the same, i.e., $n^{2D}_{cp} = n^{WID}_{cp}$ and (c) the 3D circuit has zero D2D paths, i.e., $N^{D2D}_{cp} = 0$; then $\Delta T^{3D}_{max} \geq_{st} \Delta T^{2D}_{max}$.

Proof: Let $\Delta T^{2D}_{max} = \max_{1 \leq i \leq N}(\Delta T^{2D}_i)$ and $\Delta T^{3D}_{max} = \max_{1 \leq i \leq N}(\Delta T^{3D}_i)$. From Equation 1 and Equation 3, we can see that $E(\Delta T^{2D^2}_i) = E(\Delta T^{3D^2}_i)$. Furthermore:

$$E(\Delta T^{2D}_i \Delta T^{2D}_j) = E(\Delta T^{3D}_i \Delta T^{3D}_j), \forall i,j : m(i) = m(j) \tag{10}$$

$$E(\Delta T^{2D}_i \Delta T^{2D}_j) > E(\Delta T^{3D}_i \Delta T^{3D}_j), \forall i,j : m(i) \neq m(j) \tag{11}$$

Therefore, using the result from Theorem 1, this implies that $\Delta T^{3D}_{max} \geq_{st} \Delta T^{2D}_{max}$. ■

Significance: The number of critical paths in a design is a useful proxy for design complexity [3]. From this perspective, Theorem 2 implies that a 3D system is always worse impacted by process variations than a 2D design with the same design complexity. Therefore, while the theorem establishes process variations to be a potentially even greater cause for concern in 3D circuits, it **should not**, in general, be used to compare a specific 2D circuit with its equivalent 3D implementation, since a gate-level 3D place-and-route process may alter the number of critical paths in the design [6].

A.2 Mapping of WID Critical Paths

Mapping of gates or modules in a design to die layers is an important step in the physical design flow for 3D ICs that affects both the final wirelength histogram [6] and the temperature profile [5] of the design. From Equation 7, we can see that the mapping of critical paths to die layers also affects the *cdf* of maximum critical path delay - i.e., for the same number of total critical paths, different mappings of critical paths to die layers can produce different results. Indeed, we *prove theoretically* that, for a design with only WID critical paths, the worst case impact of process variations on critical path delay occurs when the critical paths are *evenly divided between the die layers*, i.e., $N^i_{cp} = \frac{N^{WID}_{cp}}{L}, \forall i \in [1, L]$. Before describing the proof, we first define some useful notation. To explicitly consider the relationship between the *cdf* and the mapping of critical paths, we let $F_{\Delta T^{3D}_{max}}(\tau) = R_\tau(N^1_{cp}, N^2_{cp}, \ldots, N^L_{cp})$, where from Equation 7 we can see that:

$$R_\tau(N^1_{cp}, N^2_{cp}, \ldots, N^L_{cp}) = \prod_{i=1}^{L} r_\tau(N^i_{cp}) \tag{12}$$

$$r_\tau(N) = f_G(\frac{\tau}{\alpha n^{WID}_{cp}}) * (F_L(\frac{\tau}{\alpha\sqrt{n^{WID}_{cp}}})^N) \tag{13}$$

Theorem 3: For a 3D system with L layers, N^{WID}_{cp} critical paths, and zero D2D critical paths, $Pr\{T^{3D}_{max} \leq \tau\}$ is *minimized* for any value of τ when $N^i_{cp} = \frac{N^{WID}_{cp}}{L}, \forall i \in [1, L]$.

Proof: We begin by defining a function $Q_\tau(N^1_{cp}, N^2_{cp}, \ldots, N^L_{cp}) = log_e(R_\tau(N^1_{cp}, N^2_{cp}, \ldots, N^L_{cp}))$. Now consider the following optimization problem:

$$\min Q_\tau(N^1_{cp}, N^2_{cp}, \ldots, N^L_{cp}) \tag{14}$$

subject to:

$$\sum_{i=1}^{L} N^i_{cp} = N^{WID}_{cp} \tag{15}$$

$$N^i_{cp} \geq 0, \forall i \in [1, L] \tag{16}$$

Since the *log()* function is strictly monotonic, R_τ and Q_τ achieve their minimum values at the same point. Furthermore, we observe that the function $Q_\tau(N^1_{cp}, N^2_{cp}, \ldots, N^L_{cp}) = \sum_i^L log(r_\tau(N^i_{cp}))$ is *convex* because: (a) $r_\tau(N)$ can be shown to be log-convex[2], and therefore, $log(r_\tau(N))$ is convex; and (b) the sum of convex functions is convex [4].

[2]A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is said to be log-convex if $log(f)$ is convex.

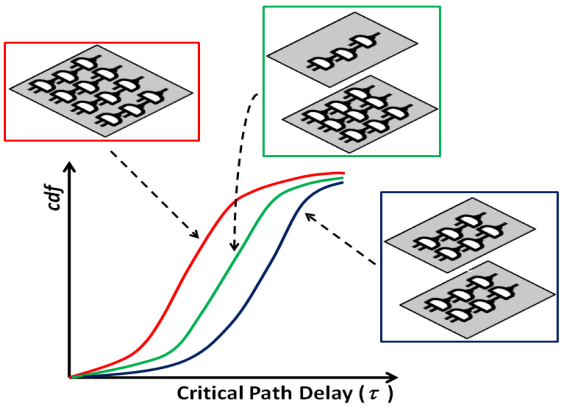

Fig. 2. Overview of results presented in Section **4-A**

Now, we pick any point $S_1 = \{N_{cp}^1, N_{cp}^2, \ldots, N_{cp}^L\}$ in the feasible region and define a set $\theta = \{S_1, S_2, \ldots, S_{L!}\}$ that consists of all possible permutations of S_1. Furthermore, we define point S_* as:

$$S_* = \frac{1}{L!}\sum_{i=1}^{L!} S_i = \{\frac{N_{cp}^{WID}}{L}, \frac{N_{cp}^{WID}}{L}, \ldots, \frac{N_{cp}^{WID}}{L}\} \quad (17)$$

Finally since $Q_\tau(N_{cp}^1, N_{cp}^2, \ldots, N_{cp}^L)$ is convex, we know that $Q_\tau(S_*) \leq \frac{\sum_{i=1}^{L!} Q_\tau(S_i)}{L!}$ [4]. Since $Q_\tau(S_i) = Q_\tau(S_j)$ $(1 \leq i, j \leq L!)$, we can restate this inequality as $Q_\tau(S_*) \leq Q_\tau(S_1)$, where S_1 can be *any* point in the feasible region of the optimization problem. In fact, barring the degenerate cases in which either $\sigma_L = 0$ or $\sigma_G = 0$, the convexity of the objective function implies that S_* is a point of *unique* minima in the feasible region, in other words, $Q_\tau(S_*) < Q_\tau(S_1)$ $\forall S_1 \neq S_*$. Therefore, $R_\tau(S_*) < R_\tau(S_1)$ $\forall S_1 \neq S_*$ ■

Significance: Theorem 3 demonstrates that a mapping solution that is unaware of the interaction between the mapping of critical paths to die layers and process variations may result in a sub-optimal design - in other words, it introduces critical path mapping as an important element of the variability-aware design space for 3D ICs.

A.3 Summary of Results

The results presented so far are summarized in Figure 2, which compares the *cdf* of maximum critical path delay for three cases: (a) a 2D design with four critical paths, (b) a 3D design with the same number of WID critical paths as the 2D design, with an uneven mapping of paths to die layers; in this case, the first layer has three paths and the second has one, and (c) a 3D design with the same number of WID critical paths as the 2D design, with critical paths evenly divided between die layers; i.e., with each layer having two critical paths. Based on the Theorem 2 and Theorem 3, we can conclude that the 3D design with an even mapping of paths, i.e., case (c), will always have the least likelihood of meeting a specified timing constraint, τ, while an equivalent 2D design, i.e., case (a), will have the highest likelihood of meeting timing constraints. This can be observed graphically in Figure 2.

B. Modeling WID+D2D Paths

We now consider the most general case in which $N_{cp}^{3D} > 0$, i.e., a 3D system with both WID and D2D critical paths. Unfortunately, in the general case, it is not possible to obtain a closed-form analytical expression for the *cdf* of ΔT_{max}^{3D} due to the correlations between the ΔT_i^{D2D} and ΔT_j^{WID} terms in Equation 5. In particular we can see that:

$$E(\Delta T_i^{WID} \Delta T_j^{D2D}) = \alpha\beta n_{cp}^{m(i)} \sigma_G^2 \geq 0 \quad (18)$$

Instead of providing exact analytical expressions for the *cdf*, we therefore concentrate on computing provable stochastic lower bounds and upper bounds on ΔT_{max}^{3D}.

B.1 Stochastic Lower Bound

The lower bound on ΔT_{max}^{3D} can be obtained by simply ignoring the impact of D2D critical paths on the maximum delay variation. Let ΔT_{max}^{LB} be the RV that stochastically lower bounds ΔT_{max}^{3D}. We therefore write:

$$\Delta T_{max}^{LB} = \max_{1 \leq i \leq N_{cp}^{WID}} (\Delta T_i^{WID}) \quad (19)$$

We now prove formally that the RV T_{max}^{LB} is indeed a guaranteed lower bound on the actual critical path delay distribution.

Theorem 4: $\Delta T_{max}^{LB} \leq_{st} \Delta T_{max}^{3D}$, i.e., $Pr\{\Delta T_{max}^{LB} \leq \tau\} \geq Pr\{\Delta T_{max}^{3D} \leq \tau\}$ $\forall \tau \in \mathbb{R}$.

Proof: For any two random variables, X and Y, $Pr\{X, Y\} = Pr\{X\}Pr\{Y/X\} \leq Pr\{X\}$. Therefore it follows that $Pr\{\max(\Delta T_i^{WID}) \leq \tau\} \geq Pr\{\max(\Delta T_i^{WID}) \leq \tau, \max(\Delta T_i^{D2D}) \leq \tau\}$. ■

While ignoring the D2D critical paths may seem to be a loose lower bound, we show experimentally that this is indeed not the case. Intuitively, this is because the path delay of a D2D critical path involves a summation over the D2D RVs G_i, which reduces the standard deviation of delay of a D2D path with respect to a WID path, hence reducing the probability that a D2D path will be the speed limiting path of the design. In fact, this intuition can be proved formally:

Theorem 5: If $n_{cp}^{D2D} = n_{cp}^{WID}$, $N_{cp}^{D2D} = N_{cp}^{WID}$ and $\alpha = \beta$, then probability that a WID critical path is the speed limiting path in the design is greater than the probability that a D2D path is the speed limiting path in the design, i.e., $Pr\{max(\Delta T_i^{WID}) \geq max(\Delta T_i^{D2D})\} \geq 0.5$

Proof: We first define new RVs $\Delta T_i^{WID,S}$ $(1 \leq i \leq N_{cp}^{WID})$ as follows:

$$\Delta T_i^{WID,S} = n_{cp}^{WID} \alpha G_1 + \sum_{j=1}^{n_{cp}^{WID}} \alpha L_{ij}, 1 \leq i \leq N_{cp}^{WID} \quad (20)$$

Comparing this definition with Equation 3, it is clear that $max(\Delta T_i^{WID,S}) \leq max(\Delta T_i^{WID})$, where the inequality is *deterministically true*, i.e., it is true for any realization of the RVs. Therefore, we can write:

$$Pr\{max(\Delta T_i^{WID}) \geq max(\Delta T_i^{D2D})\} \geq Pr\{max(\Delta T_i^{WID,S}) \geq max(\Delta T_i^{D2D})\} \quad (21)$$

Now, we can write the RV $Z = \{max(\Delta T_i^{WID,S}) - max(\Delta T_i^{D2D})\}$ as:

$$Z = \alpha\{n_{cp}^{WID}G_1 - \sum_{j=1}^{L} n_{cp}^{j}G_j + max(L_{ij}) - max(L'_{ij})\} \quad (22)$$

In the above expression, $n_{cp}^{WID}G_1 - \sum_{j=1}^{L} n_{cp}^{j}G_j$ is zero mean normal RV, and is therefore also symmetric about the ori- gin. Furthermore, $max(L_{ij})$ and $max(L'_{ij})$ are *iid* RVs and therefore $max(L_{ij}) - max(L'_{ij})$ is a zero mean RV symmetric about the origin. Finally, since the sum of two zero mean symmetric RVs is also zero mean and symmetric, we know that the RV Z is zero mean and symmetric. Therefore,

$$Pr\{max(\Delta T_i^{WID,S}) \geq max(\Delta T_i^{D2D})\} = 0.5 \quad (23)$$

Equation 21 and Equation 23 together complete the desired proof. ■

Significance: Theorem 5 shows that, assuming all other things being equal, i.e., for a system with the same number of WID and D2D paths and the same number of stages, a D2D path is always *less* likely to be the speed constraining path in the system as compared to a WID path. This is, as mentioned before, because of the averaging affect on a D2D path of crossing multiple die layers. From a design perspective, the result suggests allocating as many critical paths as possible as D2D paths as a strategy to mitigate the impact of process variations on performance, unless the sensitivity of a D2D path to PVs is significantly greater than that of a WID path ($\beta >> \alpha$). We note that our results suggest that for typical interconnect and TSV dimensions, $\beta \approx \alpha$.

B.2 Stochastic Upper Bound

Let ΔT_{max}^{UB} be the RV that corresponds to the provable upper bound on ΔT_{max}^{3D}. We will first provide the expression for ΔT_{max}^{UB} before explaining its physical significance. In particular we can write:

$$\Delta T_{max}^{UB} = max(\max_i(\Delta T_i^{WID}), \max_i(T_i^{D2D,UB})) \quad (24)$$

where

$$\Delta T_i^{D2D,UB} = \sum_{j=1}^{L} \beta n_{cp}^{j} G'_j + \sum_{j=1}^{n_{cp}^{D2D}} \beta L'_{ij} \quad (25)$$

and G'_j $(1 \leq j \leq L)$ are newly introduced RVs such that G'_j and G_j are *iid* $\forall j \in [1, L]$.

The physical significance of the upper bound can be understood by setting $\sum_{j=1}^{L} \beta n_{cp}^{j} G'_j = G_{L+1}$. In particular, the upper bound corresponds to a 3D system in which an additional *dummy* die layer $L+1$ has been introduced and to which all the D2D critical paths in the system have been allocated.

Theorem 6: $\Delta T_{max}^{UB} \geq_{st} \Delta T_{max}^{3D}$

Proof: The proof follows directly from Theorem 1 by observing that $E(\Delta T_i^{WID} \Delta T_j^{D2D}) > 0$ from Equation 18, while $E(\Delta T_i^{WID} \Delta T_j^{D2D,UB}) = 0$. Therefore, $E(\Delta T_i^{WID} \Delta T_j^{D2D,UB}) \leq E(\Delta T_i^{WID} \Delta T_j^{D2D})$ $(i \in [1, N_{cp}^{WID}], j \in [1, N_{cp}^{D2D}])$ while all other elements of the two co-variance matrices remain unchanged. ■

From Equation 24 and Equation 25, we can now write an analytical expression for the *cdf* of ΔT_{max}^{UB} as:

$$F_{\Delta T_{max}^{UB}}(\tau) = [\prod_{i=1}^{L} f_G(\frac{\tau}{s_1}) * (F_L(\frac{\tau}{s_2})^{N_{cp}^{i}})][f_G(\frac{\tau}{s_3}) * (F_L(\frac{\tau}{s_4})^{N_{cp}^{D2D}})] \quad (26)$$

where $s_1 = \alpha n_{cp}^{WID}$, $s_2 = \alpha\sqrt{n_{cp}^{WID}}$, $s_3 = \beta\sum_{i=1}^{L}(n_{cp}^{i})^2$ and $s_4 = \beta\sqrt{n_{cp}^{D2D}}$.

B.3 Summary of Results

For the generic case of 3D systems with both WID and D2D critical paths, we observed that it is difficult to obtain a precise analytical expression for the *cdf* of the maximum critical path delay. Therefore, we derive theoretically guaranteed lower and upper bounds on the desired *cdf*, in other words, for any timing constraint τ, we provide a lower and an upper bound on the probability of meeting that constraint. Finally, to compare the relative impact of D2D and WID paths on the maximum critical path delay distribution, we show in Theorem 5 that all other things being equal, WID paths are more likely to be the speed-limiting path in a system than D2D paths.

Having described in detail the proposed models for 3D ICs with only WID critical paths and then for the more general case, we now present our experimental results that validate and demonstrate the effectiveness of the proposed 3D-GCP framework.

V. Experimental Results

We now present our experimental results that first validate the proposed models against detailed SPICE-based Monte-Carlo simulation and then investigate the application of the models to characterize the impact of variations on 3D designs.

A. SPICE-based Validation

To comprehensively test the accuracy of the proposed models, we compare the *cdfs* obtained from the analyt-

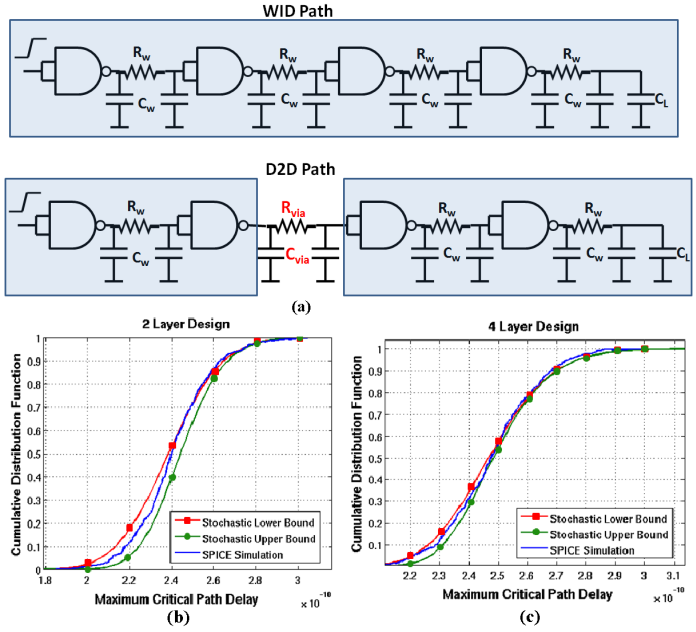

Fig. 3. (a) SPICE models for WID and D2D paths. Maximum critical path delay *cdfs* for (b) a two layer design, and (c) a four layer design.

TABLE I
PARAMETERS FOR 3D DESIGNS IN FIGURE 3

Design	$N_{cp}^{WID} : \{N_{cp}^{i}\}$	n_{cp}^{WID}	$n_{cp}^{D2D} : \{n_{cp}^{i}\}$
2 Layer	$100 : \{50, 50\}$	6	$6 : \{3, 3\}$
4 Layer	$100 : \{25, 25, 25, 25\}$	6	$6 : \{1, 2, 2, 1\}$

ical expressions with Monte Carlo simulations on input SPICE models of the canonical 3D circuits. In keeping with the methodology described by [3], each critical path in the SPICE netlist is modeled as a chain of two-input NAND gates in a *90 nm* PTM technology. Furthermore, wire (via) delays are inserted between gates that lie in the same layer (cross die layers) using a standard π model as shown in Figure 3(a). The RC parameters associated with the wire models are computed using the average dimensions for Metal 2 wires reported in [16], while vias are assumed to be $1.2\mu m \times 1.2\mu m$, with a $2.4\mu m$ pitch and $20\mu m$ length as reported in [9]. Furthermore, to equalize the nominal delay of the D2D and WID critical paths, the gate widths of the D2D critical paths are sized up. We observed that in practice, the additional delay introduced by a TSV with the given parameters is only slightly larger than the delay of a regular interconnect, requiring the gates on a D2D path to be sized up by 10% for a 2 layer design and 30% for a four layer design. Finally, variations in process parameters are modeled by introducing both D2D and random WID variations in gate length, each with a σ of 5% of the nominal value.

Under these assumptions, we modeled two 3D systems with parameters as shown in Table I and performed 500 runs of Monte Carlo simulations in SPICE to determine the maximum critical path delay distribution for each design. We note that we were restricted to relatively small circuits due to the prohibitive run-time and memory requirements of larger designs. Finally, using the models proposed in this paper, we computed the *cdfs* corresponding to the analytical lower and upper bounds for each design. The obtained *cdfs* for the two designs are shown in Figure 3(b) and (c). It is clear that in both cases, the stochastic lower and upper bounds are in excellent agreement with the SPICE results. Furthermore, the upper bound behaves ideally for both cases, i.e., always predicts a lower yield than predicted by the SPICE-based *cdf* for any cycle time constraint, while the lower bound behaves ideally for all but high yield values ($> 80\%$). The non-ideality arises because of two reasons: (1) The assumption made in the original GCP model about gate delays being linearly dependent on process variations parameters (also used in our model) is not *perfectly* accurate, although prior results and the results from Figure 3 indicate that this is still a good approximation, and (b) the *cdf* obtained from SPICE simulations itself has some inherent error since it is obtained from Monte Carlo simulations. Nonetheless, given the complexity of the SPICE transistor models, we believe that the results in Figure 3 demonstrate the effectiveness of the proposed methodology.

B. Design Space Exploration

Having validated the accuracy of the proposed models, we now use the models to explore the impact of various design decisions on the *cdf* of maximum critical path delay of 3D ICs and also verify experimentally, some of the theoretical results shown before. We begin by analyzing designs that consist of only WID critical paths and then look at the general case of designs with both D2D and WID critical paths.

B.1 WID Critical Paths Only

For all the results that follow, process variations are modeled by introducing both D2D and WID variations in the gate length parameter with a total standard deviation of 10% of the mean [7], though we note that our framework is general and can handle other sources of variation as well. The parameter $\gamma = \frac{\sigma_G^2}{\sigma_{tot}^2}$ represents the contribution of D2D variations to total variations. Finally, to reflect the design complexity in $90nm$ technology, we choose total number of critical paths in the all the designs that we evaluate to be $1,000$, as suggested by [3].

We begin by looking at the impact of increasing the number of die layers in a 3D IC for the same number of total WID critical paths. In Figure 4(a), we plot the percentage increase in the mean of maximum critical path delay with increasing number of die layers and for three values of $\gamma = \{0.25, 0.5, 0.75\}$. From the figure, we can see that for the medium and high values of γ, i.e., when D2D variations have moderate to high contributions to total variability, increasing the number of device layers has a *significant* impact on the mean critical path delay. For example, for $\gamma = 0.5$ the mean delay increases by 9.5% when going from a 2D design to a six layer design. On the other hand, when

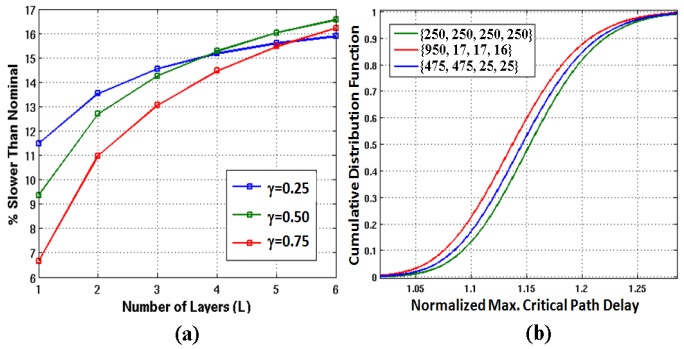

Fig. 4. (a)Impact of number of die layers on maximum critical path delay (b) *cdf* of maximum critical path delay for three different WID critical path to die mappings.

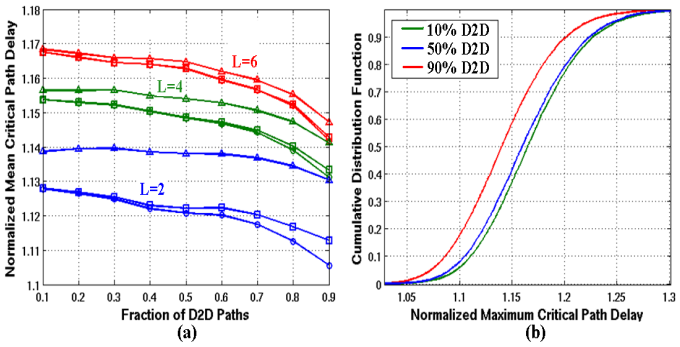

Fig. 5. (a) Mean of the maximum critical path delay distribution as a function of the fraction of D2D critical paths in the system. The lower bounds, actual values and upper bounds are given by circle, square and triangle markers respectively. (b) *cdf* of maximum critical path delay for D2D paths $\{10\%, 50\%, 90\%\}$ of total number of paths.

WID variations dominate, we observe that the impact is less pronounced. This result confirms our prediction from Theorem 2, i.e., in all cases we see that the 3D designs are worse than their equivalent 2D counterparts from a variability perspective.

Figure 4(b) shows the impact of different WID critical path to die-layer mapping on the *cdf* of maximum critical path delay. As proven in Theorem 3, we can see that the mapping in which the $1,000$ critical paths are equally split between the four layers, i.e., $\{250, 250, 250, 250\}$, stochastically upper bounds the other two mappings. In particular, the design with equally split paths provides only 36% yield at the 50% yield point of the best design, i.e., $\{950, 17, 17, 16\}$ in which 95% of the critical paths lie in a single layer. We note that this is not an arbitrary design point - [10] have shown that up to 95% of the critical paths of a typical microprocessor lie in the cache arrays - from this perspective, the best design would correspond to a case in which the caches are implemented in one layer and the logic in the rest. We point out that while the 2D and 3D GCP models do not explicitly model timing variations for memory structures, [10] have shown that from a variation perspective, memory arrays can be equivalently modeled using chains of logic gates as well.

B.2 WID+D2D Paths

We now analyze the accuracy of the proposed stochastic upper and lower bounds for the more general case of 3D ICs with both D2D and WID critical paths. While previously we used SPICE based Monte Carlo simulations to obtain the actual *cdf* for small designs, we now use gate-level MC simulations (using sensitivity information extracted from SPICE) so as to analyze larger systems. Figure 5(a) shows the impact on mean critical path delay (relative to the nominal delay) of varying the fraction of D2D paths in the system from 20% to 80% for $L = \{2, 4, 6\}$. Also shown are the corresponding mean values obtained from the computed stochastic lower and upper bounds. We observe that except for the upper bound for the two layer design, all the other bounds are extremely tight.

We note that these results confirm the intuition developed from Theorem 5, i.e., as the fraction of D2D critical paths in the design increases, the mean of the critical path delay distribution *improves*. Finally, in Figure 5(b), we plot the actual *cdfs* of maximum critical path delay for three cases, i.e., when D2D paths make up $\{10\%, 50\%, 90\%\}$ of the total critical paths in the design. While it is clear from the figure that the 90% D2D paths case has significantly higher timing yield than the 10% and 50% cases, we note that the 90% case would come with an associated overhead in terms of the number of TSVs required to implement the design.

VI. Conclusion and Future Work

In this paper, we present 3D-GCP, a high-level analytical model for the impact of process variations on stacked-die 3D ICs, that takes as input a canonical description of a 3D circuit encapsulated within a few parameters. Using this model we are able to provide an exact analytical expression for the *cdf* of maximum critical path delay for 3D circuits with only WID critical paths and guaranteed stochastic lower and upper bounds for circuits with both D2D and WID paths. The model is validated against SPICE based Monte Carlo simulations and demonstrated to be in excellent agreement with SPICE results. Furthermore, using this model, we prove theoretically that a 3D circuit is always *less likely* to meet a specified frequency target compared to a 2D design with the same number of critical paths. Finally, we show, theoretically and experimentally, that the mapping of critical paths to die layers and the ratio of WID to D2D critical paths in the design can both have an impact on the timing yield of the design.

As future work, we plan to look in to variability mitigation techniques for 3D ICs, both using variability-aware design time mapping of critical paths and variability-aware post-fabrication assembly techniques.

References

[1] R. Adler. *An Introduction to Continuity Extrema and Related Topics for General Gaussian Processes.* Institute of Mathematical Statistics, 1990.

[2] K. Banerjee, S.J. Souri, P. Kapur, and K.C. Saraswat. 3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Intercon-

nect Performance and Systems-on-Chip Integration. *Proceedings of the IEEE*, 2001.

[3] K.A. Bowman, S.G. Duvall, and J.D. Meindl. Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration. *IEEE Journal of Solid-state Circuits*, 2002.

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

[5] J. Cong, J. Wei, and Y. Zhang. A thermal-driven floorplanning algorithm for 3D ICs. In *Proceedings of ICCAD*, 2004.

[6] Y. Deng and W.P. Maly. Interconnect characteristics of 2.5-D system integration scheme. In *Proceedings of ISPD*, 2001.

[7] SG Duvall. Statistical circuit modeling and optimization. In *International Workshop on Statistical Metrology*, 2000.

[8] C. Ferri, S. Reda, and R.I. Bahar. Strategies for improving the parametric yield and profits of 3D ICs. In *Proceedings of ICCAD*, 2007.

[9] S. Gupta, M. Hilbert, S. Hong, and R. Patti. Techniques for Producing 3D ICs with High-Density Interconnect. In *Proceedings of the 21st International VLSI Multilevel Interconnection Conference*, 2004.

[10] S. Herbert, S. Garg, and D. Marculescu. Reclaiming Performance and Energy Efficiency from Variability. In *Proceedings of IBM Pac2*, 2006.

[11] R. Ho, K.W. Mai, and M.A. Horowitz. The future of wires. *Proceedings of the IEEE*, 2001.

[12] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M.S. Yousif, and C.R. Das. A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In *Proceedings of ISCA*, 2007.

[13] X. Liang and D. Brooks. Microarchitecture parameter selection to optimize system performance under process variation. In *Proceedings of ICCAD*, 2006.

[14] D. Marculescu and E. Talpes. Variability and energy awareness: a microarchitecture-level perspective. In *Proceedings of DAC*, 2005.

[15] R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas. Mitigating Parameter Variation with Dynamic Fine-Grain Body Biasing. In *IEEE/ACM International Symposium on Microarchitecture*, 2007.

[16] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45nm design exploration. In *Proceedings of ISQED*, 2006.