Analysis of Performance Benefits of Multitier Gate-Level Monolithic 3-D Integrated Circuits

Inki Hong and Dae Hyun Kim, *Member, IEEE*

***Abstract*—Vertical interconnects used in monolithic 3-D integrated circuits (3-D ICs), so-called monolithic interlayer vias (MIVs), are as small as local vias. Thus, redesigning an existing 2-D IC layout in a monolithic 3-D IC generally results in shorter wire length than the 2-D IC layout. In addition, MIVs have almost negligible resistance and capacitance, so their impact on signal delay is very small. Thus, redesigning a 2-D IC layout in a monolithic 3-D IC is expected to improve its performance significantly. Some researchers designed several monolithic 3-D IC layouts and showed their timing benefits in the literature. In this paper, we present analytical models for performance (timing) benefits of multitier gate-level monolithic 3-D ICs. The analytical models we develop in this paper can be used to quickly estimate the performance benefits multitier gate-level monolithic 3-D integration provides without physically redesigning 2-D IC layouts in 3-D.**

***Index Terms*—Monolithic 3-D integration, performance analysis, prediction, reversal of critical paths (RCP).**

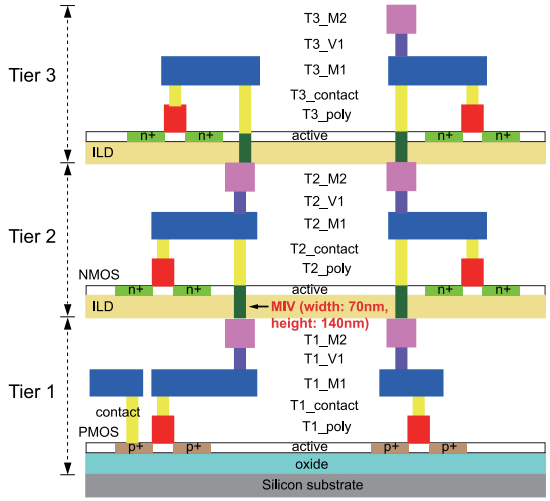


Fig. 1. Multitier monolithic 3-D integration.

I. Introduction

MULTITIER monolithic 3-D integration shown in Fig. 1 is a 3-D stacking technology by which multiple thin device layers are stacked and transistors in different tiers are connected through ultrasmall vertical interconnects, so-called monolithic interlayer vias (MIVs) [1]–[4]. MIVs are similar to local vias in terms of the size (width and z-directional height). Thus, MIV insertion is expected to cause almost no area, resistance, and capacitance overhead, whereas through-silicon-via (TSV) insertion causes non-negligible area and capacitance overheads in TSV-based integrated circuits (3-D ICs) [5], [6]. Monolithic 3-D integration, therefore, is expected to achieve the highest degree of wire length reduction and performance improvement [7].

Design of gate-level monolithic 3-D IC layouts requires almost all the design steps used for the design of 2-D IC layouts, such as placement and routing, among which placement has been actively investigated in several papers recently [5], [8]–[10]. Especially, redesigning a gate-level 2-D IC layout in a monolithic 3-D IC by uniform scaling proposed in [8] reduces the wire length ideally by $(1/\sqrt{N_T})$, where N_T is the number of tiers. Due to this property, the uniform-scaling-based placement algorithm achieves the shortest total wire length if the 2-D layout has the shortest total wire length and enough white space exists among the instances in the 2-D layout [11]. The shortened wire length is converted into delay reduction, which is one of the most important benefits monolithic 3-D integration provides. As shown by physical design and analysis in various works on monolithic 3-D IC design, monolithic 3-D integration provides performance benefits for gate-level layouts [5], [10], [12]. In this paper, we develop analytical models for fast estimation of the performance benefits of gate-level monolithic 3-D ICs.

Providing a short overview of the models, if a monolithic 3-D IC layout is designed by uniformly scaling a 2-D layout, the length of each wire in the 2-D layout is shortened by $(1/\sqrt{N_T})$. Since the delay of a wire of length l is proportional to l^2, researchers expect to achieve $N_T\times$ performance improvement (clock frequency increase) by monolithic 3-D integration. However, this analysis has three major problems as follows.

1) It ignores instance (gate) delays.
2) It does not take interconnect optimization into the performance analysis. If a path is properly buffered, the delay of the path is proportional to the path length.

Manuscript received March 4, 2017; revised July 27, 2017; accepted October 25, 2017. Date of publication October 31, 2017; date of current version July 17, 2018. This work was supported by the Defense Advanced Research Projects Agency Young Faculty Award under Grant D16AP00119. This paper was recommended by Associate Editor T.-Y. Ho. *(Corresponding author: Dae Hyun Kim.)*

I. Hong is with Cadence Design Systems, San Jose, CA 95134 USA (e-mail: inki@cadence.com).

D. H. Kim is with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164 USA (e-mail: daehyun@eecs.wsu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2017.2768427

3) It focuses only on the performance of a single path, but the performance analysis should consider multiple signal paths. If the delays of the first and the second critical paths in a given 2-D layout are dominated by wire and instance delays, respectively, the second critical path might become the most critical path in the 3-D layout, which we call reversal of critical paths (RCPs).

If we take all of them listed above into the analysis of performance benefits of multitier gate-level monolithic 3-D integration, the performance improvement is less than $\sqrt{N_T}\times$. Especially, if an RCP occurs, the performance benefit is much less than $\sqrt{N_T}\times$ as will be shown later.

In this paper, we analyze performance benefits of multitier gate-level monolithic 3-D ICs based on optimal analytical buffer insertion taking instance delay, logic depth, process technology parameters, routing congestion, and the RCP into account. The analytical models can be used to estimate the performance benefits we can obtain from multitier gate-level monolithic 3-D integration for given 2-D IC layouts without physical design. Our contributions in this paper are as follows.

1) We mathematically analyze the performance benefits multitier gate-level monolithic 3-D integration provides.
2) We mathematically analyze the RCP.
3) We qualitatively and quantitatively analyze the performance benefits using four sets of process technologies.
4) We validate the models by designing, optimizing, and analyzing 2-D and monolithic 3-D layouts.

The rest of this paper is organized as follows. We review previous work on path delay calculation, optimal buffer insertion, and monolithic 3-D ICs in Section II. In Section III, we analyze performance benefits of a single path in multitier gate-level monolithic 3-D ICs. In Section IV, we investigate the RCP, which analyzes the performance benefits of multiple signal paths. Then, we apply the performance models to two benchmarks and validate the models in Section V. Finally, we conclude in Section VI.

II. Preliminaries and Related Work

In this section, we review path delay calculation, optimal buffer insertion, and related work. Table I shows the variables used in this paper.

A. Path Delay Calculation

Suppose a single-fanout path p consists of a driver c_0, a sink c_{n+1}, n instances $c_1, \ldots, c_n$ between the driver and the sink, and wires connecting them. Then, the delay of the path can be expressed as follows:

$$d_p = \sum_{i=0}^{n} d_{c_i} + \sum_{i=0}^{n} d_{w_{i,i+1}} \tag{1}$$

where the first and the second sums are from the instance and wire delays, respectively. The above equation can be rewritten using the Elmore delay model as follows:

$$d_p = \sum_{i=0}^{n} d_{c_i} + \sum_{i=0}^{n} \left\{ R_{c_i}\left(c_{w_{i,i+1}} + C_{c_{i+1}}\right) + r_{w_{i,i+1}} \cdot C_{c_{i+1}} + \frac{1}{2} r_{w_{i,i+1}} \cdot c_{w_{i,i+1}} \right\} \tag{2}$$

where $c_{w_{i,i+1}}$ is $c_w \cdot l_{w_{i,i+1}}$ and $r_{w_{i,i+1}}$ is $r_w \cdot l_{w_{i,i+1}}$. Notice that (2) is applicable to paths of logic depth 0, where the logic depth of a path is defined as the number of combinational logic instances between the driver and the sink of the path.

TABLE I
Variables Used in This Paper

Variable	Meaning
N_{T}	# tiers
$w_{i,j}$	Wire connecting instances c_i and c_j
$l_{w_{i,j}}$	Length of wire $w_{i,j}$
$d_{w_{i,j}}$	Delay of wire $w_{i,j}$
d_g	Internal delay of instance g
R_g	Output resistance of instance g
C_g	Input capacitance of instance g
r_w	Unit wire resistance
c_w	Unit wire capacitance
τ_w	$r_w \cdot c_w$
$r_{w_{i,j}}$	Total resistance of wire $w_{i,j}$ $(= r_w \cdot l_{w_{i,j}})$
$c_{w_{i,j}}$	Total capacitance of wire $w_{i,j}$ $(= c_w \cdot l_{w_{i,j}})$
d_B	Internal delay of a buffer
R_B	Output resistance of a buffer
C_B	Input capacitance of a buffer
D_{c_i}	$d_{c_i} + R_{c_i} C_{c_i}$ (for instance c_i)
D_B	$d_B + R_B C_B$ (for buffer B)
K_B	$R_B c_w + r_w C_B$
q_{p_i}	Congestion parameter for path p_i
Q_i	$q_i \cdot L_i$

B. Optimal Buffer Insertion

In this paper, we assume the followings for simplification.

1) *Assumption 1:* All nets are single-fanout nets.
2) *Assumption 2:* All buffers are of the same type, i.e., they have the same input capacitance, internal delay, and output resistance.
3) *Assumption 3:* Any locations in a net are bufferable.
4) *Assumption 4:* The output resistance and the internal delay of the driver of a net are equal to those of a buffer.
5) *Assumption 5:* The input capacitance of the sink of a net is equal to that of a buffer.

Under these assumptions, we can optimally insert buffers into a net as follows. Suppose instance c_0 drives instance c_{n+1} through a wire of length L and we insert n buffers between them. In this case, the delay of the path is expressed as follows using (2):

$$\begin{aligned} d_p &= (n+1)d_B + R_B c_w \sum_{i=0}^{n} l_{w_{i,i+1}} + (n+1)R_B C_B \\ &\quad + r_w C_B \sum_{i=0}^{n} l_{w_{i,i+1}} + \frac{1}{2} r_w c_w \sum_{i=0}^{n} {l_{w_{i,i+1}}}^2 \\ &= (n+1)D_B + L \cdot K_B + \frac{1}{2} r_w c_w \sum_{i=0}^{n} {l_{w_{i,i+1}}}^2 \end{aligned} \tag{3}$$

for which we use $\sum_{i=0}^{n} l_{w_{i,i+1}} = L$. To find optimal locations of the buffers, we differentiate (3) with respect to $l_{w_{i,i+1}}$ for $i = 0$ to $i = n-1$ and set them to zero. Then, we obtain the following equation for each i:

$$\frac{\partial d_p}{\partial l_{w_{i,i+1}}} = \frac{1}{2} r_w c_w \left(2l_{w_{i,i+1}} - 2l_{w_{n,n+1}}\right) = 0. \tag{4}$$

Thus, we obtain $l_{w_{0,1}} = l_{w_{1,2}} = \cdots = l_{w_{n,n+1}} = L/(n+1)$, i.e., the buffers are evenly distributed along the path. In this case, the path delay in (3) is expressed as follows:

$$d_p = (n+1)D_B + L \cdot K_B + \frac{r_w c_w L^2}{2(n+1)}. \tag{5}$$

Differentiating the above equation with respect to n and setting it to zero produces the following equation for the optimal number of buffers:

$$n = \sqrt{\frac{r_w c_w L^2}{2D_B}} - 1 \tag{6}$$

and the path delay in (5) becomes as follows:

$$d_{p,\text{opt}} = L\left(K_B + \sqrt{2 r_w c_w D_B}\right). \tag{7}$$

Notice that the delay of a properly buffered net is *linearly proportional* to the length of the net as shown in (7). In addition, if n in (6) is close to zero, the net does not need any buffers. The following equation is the boundary condition for buffer insertion for a net of length L[1]:

$$L = \sqrt{\frac{2D_B}{r_w c_w}}. \tag{8}$$

If the net length is shorter than $\sqrt{(2D_B/r_w c_w)}$, we do not insert any buffers. In this case, the net delay is expressed as follows:

$$d_p = D_B + L \cdot K_B + \frac{1}{2} r_w c_w L^2. \tag{9}$$

C. Uniform-Scaling-Based Monolithic 3-D IC Design

Uniform-scaling-based monolithic 3-D IC design begins the design process with a 2-D layout [8]. For the given 2-D layout, it linear scales the location of each instance by a constant scaling ratio. In general, $(1/\sqrt{N_T})$ is used for the scaling ratio to preserve the total silicon area. The uniform scaling results in many overlaps among the instances. In addition, some of the instances need to be placed in other tiers. Thus, the uniform scaling is followed by 3-D legalization to remove overlaps, satisfy density constraints, and snap the instances [11]. According to [11], the uniform-scaling-based monolithic 3-D IC design reduces the length of each wire by $(1/\sqrt{N_T})$ if a given 2-D layout has well-distributed white space. In this paper, we assume that multitier gate-level monolithic 3-D IC layouts are generated by the uniform scaling from given 2-D layouts.

[1]The boundary condition can be slightly modified as $L = (3/2)\sqrt{(2D_B/r_w c_w)}$ and $L = 2\sqrt{(2D_B/r_w c_w)}$ if the condition is applied for $n = (1/2)$ and $n = 1$, respectively.

D. Routing Congestion

Ideally, the length of a net is shortened by $(1/\sqrt{N_T})$ by uniform scaling. If routing congestion exists in a given design, however, it is shortened by $(q/\sqrt{N_T})$, where q is greater than 1.0. To take routing congestion into account in our performance analysis models, we introduce a routing congestion parameter for each net or path as follows. Suppose the lengths of path p before and after uniform scaling are L and L', respectively. Let q_p be the congestion parameter for path p. Then, we relate L and L' as follows:

$$L' = \frac{q_p \cdot L}{\sqrt{N_T}}. \tag{10}$$

If routing congestion does not affect the length of the path, q_p is set to 1.0. If routing congestion affects the length of p, however, q_p is set to the detour overhead ratio of the path length, which is greater than 1.0. We can estimate q_p by analyzing routing congestion using a commercial tool or congestion estimation algorithms [13], [14]. For instance, if a path goes through congested grids, we can find a shortest detour path and set q_p to $(1 + t/100)$, where $t/100$ accounts for the detour overhead. In this paper, we assume that q_p is always less than $\sqrt{N_T}$ for simplification.

E. Performance Improvement by 3-D Integration

3-D integration is expected to provide performance benefits in two representative ways. First, redesigning an existing 2-D layout in 3-D reduces wire length, which leads to wire delay reduction resulting in performance improvement [7], [12], [15]. Second, restructuring 2-D architectures in 3-D can exploit the very wide intertier bandwidth 3-D integration provides, thereby improving the system performance significantly [16], [17]. Especially, the former requires only 3-D design tools, so various algorithms for physical design of 3-D layouts have been actively researched [5], [6], [18].

Regarding monolithic 3-D ICs, Panth *et al.* [5] developed a gate-level monolithic 3-D IC design methodology named Shrunk2D based on technology file scaling and showed 16.82% power saving compared to 2-D ICs. Chang *et al.* [10] developed a new, flexible design methodology named Cascade2D. Cascade2D enables design and optimization of a 3-D IC in a single 2-D design using commercial tools. They showed up to 25% higher performance than 2-D designs at the same power level. Chan *et al.* [19] developed a machine-learning-based power consumption estimation tool for 3-D ICs. The prediction accuracy of the methodology is ±5%, but it is unclear whether it could be applied to the estimation of performance benefits of 3-D ICs. In addition, generation of many data points for the machine learning training takes a large amount of time. Kim *et al.* developed TSV-aware interconnect prediction models in [7] to predict wire length, performance, and power benefits of 3-D ICs, but they predict the performance of a single net based on an abstracted set of some design parameters of a design.

In this paper, we present an analytical model to estimate performance benefits obtainable from redesigning 2-D ICs in multitier gate-level monolithic 3-D ICs by the uniform-scaling-based design methodology. Regarding the multitier monolithic

TABLE II
PARAMETERS USED FOR CASE STUDIES OBTAINED FROM [7]. THE VALUES ARE FOR 20× BUFFERS

Tech	$r_w(\Omega/um)$	$c_w(fF/um)$	$R_{c_0}(\Omega)$	$C_{c_0}(fF)$	d_{c_0} (ps)
45nm	3.31	0.171	305	1.55	70
32nm	4.14		360	1.24	56
22nm	5.17		425	1.00	45
16nm	6.46		500	0.80	36

3-D integration technology, it is possible to fabricate multitier monolithic 3-D ICs [2], but how many tiers can be stacked in a monolithic 3-D IC is not known. Thus, we simulate up to 16-tier monolithic 3-D ICs in the performance benefit analysis in this paper. We think the simulation result could motivate development of many-tier monolithic 3-D IC fabrication technology if it is not possible as of now.

III. PERFORMANCE BENEFITS OF SINGLE PATH

In this section, we analyze performance benefits obtainable from redesigning a gate-level 2-D IC in an N_T-tier gate-level monolithic 3-D IC by uniform scaling. We use the path delay for performance evaluation. Table I shows the variables used in the performance analysis. For case studies, we use the parameters shown in Table II obtained from [7].

A. Delay of Path of Logic Depth n_c

Suppose the logic depth of a path in a given 2-D layout is n_c. Assuming the path is properly buffered, the delay of the path is expressed as follows:

$$d_{p,\text{2-D}} = \sum_{i=0}^{n_c} d_{c_i} + \sum_{i=1}^{n_b} d_{b_i} + \sum_{i=0}^{n_c+n_b} d_{w_{i,i+1}} \tag{11}$$

where c_0 is the driver, c_{n_c+1} is the sink, c_1 to c_{n_c} are the logic instances, n_b is the total number of buffers, and b_i is the ith buffer. If there is no buffer in the path, n_b is set to zero and the second summation term in the right-hand side is removed. Then, the path delay computation before and after uniform scaling can be categorized into three cases as follows.

- Case 1: There is no buffer in the given path of the 2-D layout because the path length is sufficiently short. In this case, the length of the path after uniform scaling becomes shorter than that before uniform scaling, so we do not need buffer insertion after uniform scaling.
- Case 2: There are some buffers in the path of the 2-D layout. However, the length of the path becomes sufficiently short after uniform scaling, so the path does not need any buffers after uniform scaling.
- Case 3: There are some buffers in the path of the 2-D layout. The path is still sufficiently long after uniform scaling, so it needs some buffers even after uniform scaling.

In the following sections, we analyze the three cases.

B. Case 1: No Buffers in the 2-D and 3-D Paths

When the logic depth n_c of a given 2-D path is greater than zero, the logic instances act as buffers for delay reduction. Assuming the logic instances are also of the same type as buffers and can be evenly distributed along the path, we do not need buffer insertion if n_c is greater than n_b, which is the number of buffers we need for the path when the logic depth of the path is zero. In other words, we do not need buffer insertion in the 2-D path if the following inequality holds:

$$n_c > \sqrt{\frac{r_w c_w L^2}{2D_B}} - 1 \tag{12}$$

where the right-hand-side term comes from (6). A similar inequality should hold for the condition that the path does not need buffer insertion after uniform scaling. However, we assume that q_p is less than $\sqrt{N_T}$ as mentioned in Section II-D, so the inequality always holds if (12) holds. Equation (12) is rewritten for the path length as follows:

$$L < (n_c + 1)\sqrt{\frac{2D_B}{r_w c_w}} \tag{13}$$

which is a boundary condition for case 1.

When a 2-D path does not need buffer insertion, the delay of the path is just a sum of its logic instance and wire delays. Assuming the instances are evenly distributed along the path, the path delay is expressed using (5) as follows:

$$d_{p,\text{2-D}} = (n_c + 1)D_B + L \cdot K_B + \frac{r_w c_w L^2}{2(n_c + 1)}. \tag{14}$$

Similarly, the delay of this path after uniform scaling becomes as follows:

$$d_{p,\text{3-D}} = (n_c + 1)D_B + \frac{q_p \cdot L}{\sqrt{N_T}} \cdot K_B + \frac{r_w c_w (q_p \cdot L)^2}{2(n_c + 1)N_T}. \tag{15}$$

Then, the path delay ratio (performance benefit of a single path) becomes as follows:

$$\frac{d_{p,\text{2-D}}}{d_{p,\text{3-D}}} = \frac{(n_c + 1)D_B + L \cdot K_B + \frac{r_w c_w L^2}{2(n_c+1)}}{(n_c + 1)D_B + \frac{q_p \cdot L}{\sqrt{N_T}} \cdot K_B + \frac{r_w c_w (q_p \cdot L)^2}{2(n_c+1)N_T}}. \tag{16}$$

Notice that the path delay ratio is the ratio between the 2-D and the 3-D path delays, so it is greater than 1.0 if there is a performance benefit in the 3-D path.

C. Case 2: Buffers Only in the 2-D Path

In this case, the path length L in the 2-D layout is longer than $(n_c + 1)\sqrt{(2D_B/r_w c_w)}$, so it needs buffer insertion. After scaling, its length becomes $[(q_p \cdot L)/\sqrt{N_T}]$, which should be shorter than $(n_c + 1)\sqrt{(2D_B/r_w c_w)}$ because it does not require buffer insertion. Thus, case 2 occurs when the following inequalities hold:

$$(n_c + 1)\sqrt{\frac{2D_B}{r_w c_w}} \le L \le \frac{(n_c + 1)}{q_p} \cdot \sqrt{\frac{2D_B N_T}{r_w c_w}} \tag{17}$$

which is a boundary condition for case 2. In this case, we can estimate the number of buffers, n_b in (11), inserted into the 2-D path as follows assuming the logic instances and the buffers

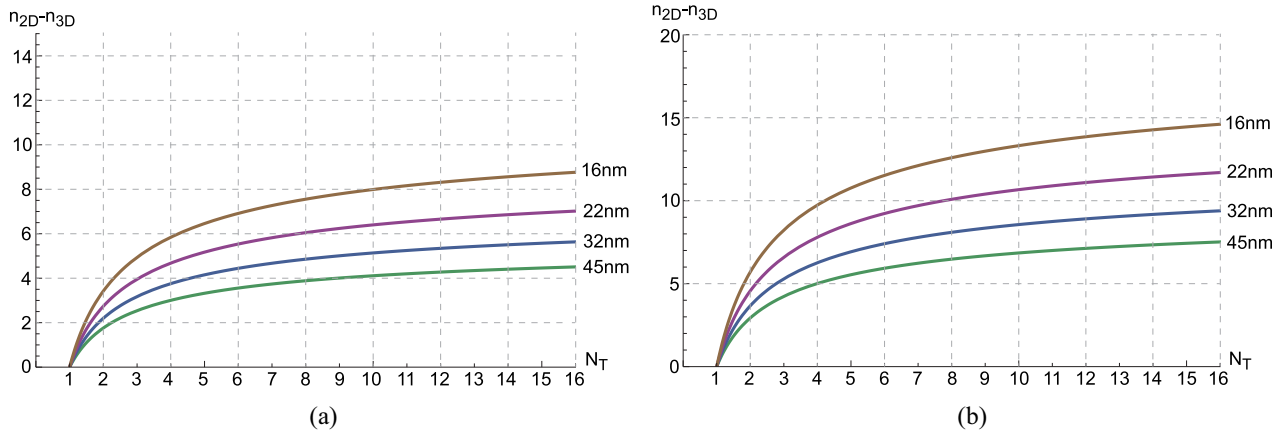


Fig. 2. Buffer count differences ($n_{2\text{-D}} - n_{3\text{-D}}$) in (27) for four process technologies. (a) $L = 3000$ um and (b) $L = 5000$ um.

are evenly distributed along the path and all the instances are of the same type:

$$n_{b,2\text{-D}} = \left(\sqrt{\frac{r_w c_w L^2}{2D_B}} - 1\right) - n_c. \tag{18}$$

Then, the delay of the 2-D path is expressed as follows using (7):

$$d_{p,2\text{-D}} = L\left(K_B + \sqrt{2r_w c_w D_B}\right). \tag{19}$$

Assuming all the buffers are removed and the logic instances are redistributed evenly along the path after uniform scaling, the delay of the 3-D path is expressed exactly the same as (15). Therefore, the path delay ratio in this case becomes as follows:

$$\frac{d_{p,2\text{-D}}}{d_{p,3\text{-D}}} = \frac{L\left(K_B + \sqrt{2r_w c_w D_B}\right)}{(n_c+1)D_B + \frac{q_p \cdot L}{\sqrt{N_T}} \cdot K_B + \frac{r_w c_w (q_p \cdot L)^2}{2(n_c+1)N_T}}. \tag{20}$$

The buffer count decreases by $n_{b,2\text{-D}}$ in (18).

D. Case 3: Buffers in the 2-D and 3-D Paths

In this case, the path in the 3-D layout is longer than $[(n_c+1)/q_p]\cdot\sqrt{(2D_B N_T / r_w c_w)}$, so both the 2-D and 3-D paths need buffer insertion. Thus, case 3 occurs when the following inequality is satisfied:

$$L > \frac{(n_c+1)}{q_p} \cdot \sqrt{\frac{2D_B N_T}{r_w c_w}} \tag{21}$$

which is a boundary condition for case 3. In this case, the number of buffers before and after uniform scaling are expressed as follows using (18):

$$n_{b,2\text{-D}} = \left(\sqrt{\frac{r_w c_w L^2}{2D_B}} - 1\right) - n_c \tag{22}$$

$$n_{b,3\text{-D}} = \left(\sqrt{\frac{r_w c_w (q_p \cdot L)^2}{2D_B N_T}} - 1\right) - n_c \tag{23}$$

where $n_{b,3\text{-D}}$ is derived by substituting $q_p \cdot L/\sqrt{N_T}$ into L in (22). The delays of the 2-D and 3-D paths in this case are expressed as follows:

$$d_{p,2\text{-D}} = L\left(K_B + \sqrt{2r_w c_w D_B}\right) \tag{24}$$

$$d_{p,3\text{-D}} = \frac{(q_p \cdot L)}{\sqrt{N_T}}\left(K_B + \sqrt{2r_w c_w D_B}\right) \tag{25}$$

so the path delay ratio becomes as follows:

$$\frac{d_{p,2\text{-D}}}{d_{p,3\text{-D}}} = \frac{\sqrt{N_T}}{q_p}. \tag{26}$$

The buffer count decreases by $n_{b,2\text{-D}} - n_{b,3\text{-D}}$ as follows:

$$n_{b,2\text{-D}} - n_{b,3\text{-D}} = L\sqrt{\frac{r_w c_w}{2D_B}}\left(1 - \frac{q_p}{\sqrt{N_T}}\right). \tag{27}$$

Fig. 2 shows (27) for the four process technologies in Table II and two path length values (3000 um and 5000 um) assuming q_p is 1.0. Notice that the buffer count difference goes up as the technology node advances because the path length (L) is a constant.

E. Analysis and Comparison of the Path Delay Ratios

Fig. 3 shows the boundary conditions for the three cases and four different process technologies assuming q_p is 1.0. Each n_c value has a horizontal line and a curve above it. The horizontal line shows the boundary condition between cases 1 and 2. If the path length is shorter than the value (i.e., the path length is below the constant line), the path delay ratio is computed by (16) for case 1. The curve above the constant line shows the boundary condition between cases 2 and 3. If the path length is longer than the value (i.e., the path length is above the curve), the path delay ratio is computed by (26) for case 3. Otherwise, the path delay ratio is computed by (20) for case 2.

When L is close to zero, the path delay ratio is calculated by (16). The equation is a strictly increasing function of L, so the path delay ratio goes up as L increases, which means the delay benefit obtainable from monolithic 3-D integration goes up. When L is greater than $(n_c+1)\sqrt{(2D_B/r_w c_w)}$, the path delay ratio is computed by (20). The equation is also a strictly increasing function of L. When L goes up

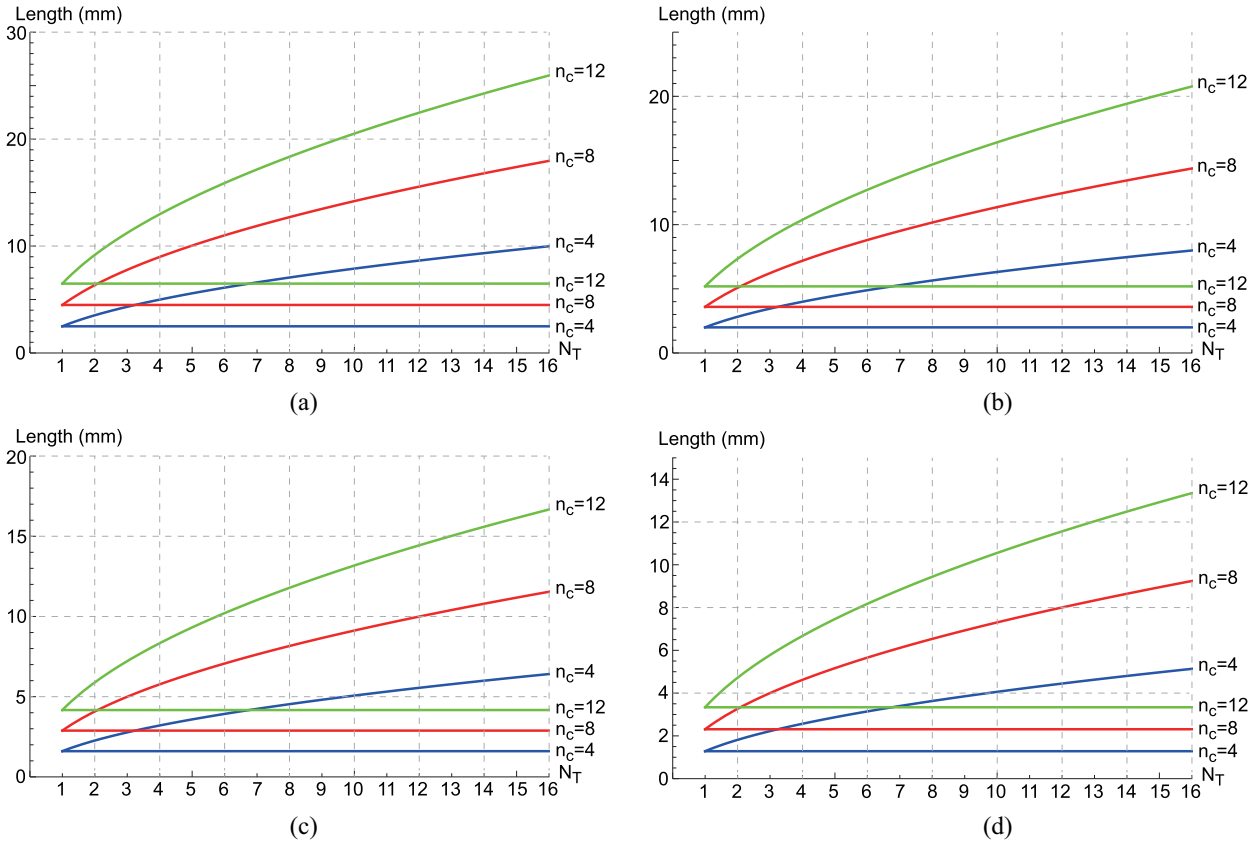


Fig. 3. Boundary conditions composed of (13), (17), and (21). (a) 45 nm, (b) 32 nm, (c) 22 nm, and (d) 16 nm.

over $[(n_c+1)/q_p]\cdot\sqrt{(2D_BN_T/r_wc_w)}$, the path delay ratio is computed by (26) and saturates at $(\sqrt{N_T}/q_p)$.

Fig. 4 shows the path delay ratios (performance benefits) for the four process technologies, two path length values (3000 um and 5000 um), and three logic depth values (4, 8, and 12) assuming q_p is 1.0. We observe the followings from the figure.

1) *L versus ($d_{p,2\text{-}D}/d_{p,3\text{-}D}$):* As L goes up, the path delay ratio goes up for a given logic depth n_c. This is because the impact of the wire delay on the path delay increases as the wire length goes up. In other words, the path delay is dominated by the total instance delay when L is small. As L goes up, however, the path delay is dominated by the wire delay, so monolithic 3-D integration helps reduce the path delay.
2) *n_c versus ($d_{p,2\text{-}D}/d_{p,3\text{-}D}$):* As n_c goes up, the path delay ratio goes down for a given path length L. This is because the impact of the total instance delay on the path delay increases as the instance count in the path goes up.
3) *N_T versus ($d_{p,2\text{-}D}/d_{p,3\text{-}D}$):* As N_T goes up, the path delay ratio goes up for a given logic depth n_c and a given path length L. This is because the wires get shorter as more tiers are stacked, so the wire delay goes down, which results in the decrease of $d_{p,3\text{-}D}$. However, the path delay ratio approaches $1+([L\cdot K_B+(r_wc_wL^2/[2(n_c+1)])]/[(n_c+1)D_B])$ and $([L(K_B+\sqrt{2r_wc_wD_B})]/[(n_c+1)D_B])$ as N_T goes to infinity for Cases 1 and 2, respectively.
4) *Technology Node Versus ($d_{p,2\text{-}D}/d_{p,3\text{-}D}$):* The path delay ratio goes up as the technology node moves from old ones (e.g., 45 nm) to new ones (e.g., 16 nm) for a given logic depth n_c, a given path length L, and a given tier count N_T. This is primarily because the wire unit resistance and buffer output resistance go up as the technology advances. Notice that this is for a fixed path length. The path delay ratio will go down if the path length is also scaled down in the new technology nodes.

Fig. 5 shows the impact of the routing congestion parameter q_p on the path delay ratios for a 5 mm-long path of logic depth 8 built with the 22 nm technology. When q_p increases from 1.00 to 1.05 and 1.10, the path delay ratios reduce from $[1.41\times, 3.10\times]$ to $[1.35\times, 3.09\times]$ and $[1.29\times, 2.97\times]$, respectively. Thus, the impact of routing congestion on the performance benefit is not negligible at all, so minimizing routing congestion is very crucial for obtaining the largest amount of performance benefits in the design of monolithic 3-D ICs.

IV. Reversal of Critical Paths

Section III shows how much we can benefit from multitier gate-level monolithic 3-D integration *for a single signal path*. In this section, we analyze performance benefits of a monolithic 3-D IC taking multiple signal paths into account.

A. Reversal of Critical Paths

Suppose a 2-D design has m signal paths $P_s = \{p_1, p_2, \ldots, p_m\}$ and the paths are sorted in the increasing

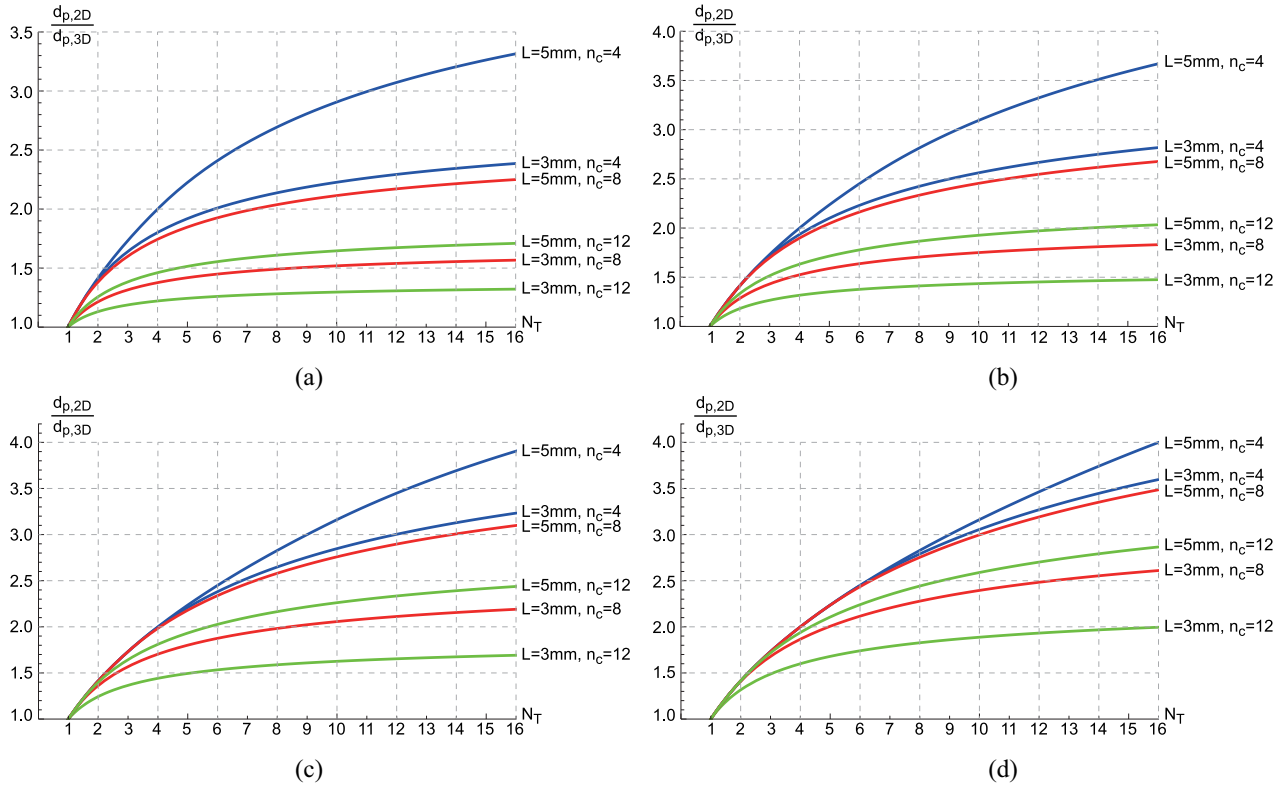


Fig. 4. Path delay ratios (performance benefits). (a) 45 nm, (b) 32 nm, (c) 22 nm, and (d) 16 nm.

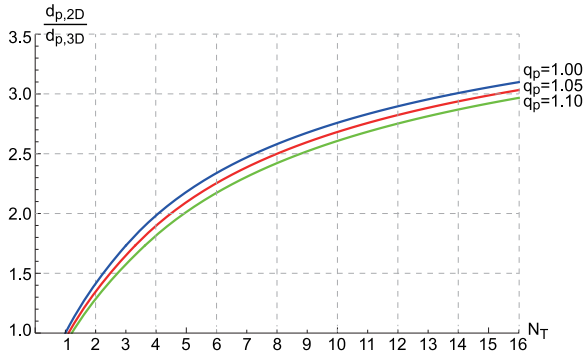


Fig. 5. Routing congestion parameter (q_p) versus path delay ratios (performance benefits). Tech: 22 nm. Path length: 5000 um. $n_c = 8$.

order of their negative slack values, so p_1 is the most critical path and p_m is the least critical path. If the delays of p_1 and p_2 are dominated by wire and instance delays, respectively, p_2 could be more critical than p_1 after uniform scaling. We call this phenomenon RCP. We define RCP more formally as follows.

Definition 1: Suppose a given 2-D design is redesigned in a monolithic 3-D IC. If the critical path in the 2-D design is not the critical path in the 3-D design, we say that an RCP occurs.

B. Necessary and Sufficient Conditions for RCP

We find necessary and sufficient conditions for the RCP as follows. If p_1 and p_2 are the most critical paths in a given 2-D and its monolithic 3-D designs ($p_1 \neq p_2$), respectively, an RCP occurs between p_1 and p_2 when the following inequalities hold:

$$d_{p_1,\text{2-D}} > d_{p_2,\text{2-D}} \tag{28}$$

$$d_{p_1,\text{3-D}} < d_{p_2,\text{3-D}} \tag{29}$$

where $d_{p_i,\text{2-D}}$ and $d_{p_i,\text{3-D}}$ are the delays of path p_i in the 2-D and 3-D layouts, respectively. $d_{p_i,\text{2-D}}$ is computed by (14) or (19) and $d_{p_i,\text{3-D}}$ is computed by (15) or (25) depending on their buffer insertion status.

Each path can be categorized into one of the three cases, Cases 1, 2, or 3 depending on its length L and logic depth n_c. For two paths, therefore, there are total nine combinations of the cases. Table III shows the boundary conditions for p_1 and p_2 and the path delay conditions shown in inequalities (28) and (29) for all the combinations. One thing to notice is that some of the nine cases do not occur depending on q_1 and q_2. For example, if q_1 is equal to q_2, case 3 $\times$ 3 cannot occur because $d_{p_1,\text{2-D}} > d_{p_2,\text{2-D}}$ leads to $L_1 > L_2$, but $d_{p_1,\text{3-D}} < d_{p_2,\text{3-D}}$ leads to $L_1 < L_2$. Assuming q_1 equals q_2, if p_1 and p_2 have the same logic depth, satisfying inequality (28) requires that L_1 is greater than L_2. In this case, the delay of p_1 will still be greater than that of p_2 after uniform scaling. Thus, two paths having the same logic depth will not lead to an RCP.[2]

Suppose the logic depth of p_1 is greater than that of p_2. If L_1 is greater than or equal to L_2, the delay of p_1 will still be greater than that of p_2 after uniform scaling, so

[2]Notice that this statement is true only under the assumptions in Section II.

TABLE III

CONDITIONS FOR AN RCP AND PERFORMANCE BENEFITS. p_1 AND p_2 ARE THE MOST CRITICAL PATHS IN THE 2-D AND 3-D LAYOUTS, RESPECTIVELY. L_i: THE LENGTH OF PATH p_i. n_i: THE LOGIC DEPTH OF p_i. B.C. DENOTES A BOUNDARY CONDITION. $Q_i = q_i \cdot L_i$

Case 1x1 p_1: Case 1 p_2: Case 1	**B.C.**(p_1): $L_1 < (n_1+1)\sqrt{\frac{2D_B}{\tau_w}}$	**B.C.**(p_2): $L_2 < (n_2+1)\sqrt{\frac{2D_B}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$(n_1-n_2)D_B + (L_1-L_2)K_B + \frac{\tau_w}{2}\left\{\frac{{L_1}^2}{n_1+1} - \frac{{L_2}^2}{n_2+1}\right\} > 0$	$(n_1-n_2)D_B + \frac{Q_1-Q_2}{\sqrt{N_T}}K_B + \frac{\tau_w}{2N_T}\left\{\frac{{Q_1}^2}{n_1+1} - \frac{{Q_2}^2}{n_2+1}\right\} < 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \left\{(n_1+1)D_B + L_1 \cdot K_B + \frac{\tau_w {L_1}^2}{2(n_1+1)}\right\} / \left\{(n_2+1)D_B + \frac{Q_2}{\sqrt{N_T}} \cdot K_B + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T}\right\}$	
Case 1x2 p_1: Case 1 p_2: Case 2	**B.C.**(p_1): $L_1 < (n_1+1)\sqrt{\frac{2D_B}{\tau_w}}$	**B.C.**(p_2): $(n_2+1)\sqrt{\frac{2D_B}{\tau_w}} \le L_2 \le \frac{(n_2+1)}{q_2} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$(n_1+1)D_B + (L_1-L_2)K_B - L_2\sqrt{2\tau_w D_B} + \frac{\tau_w {L_1}^2}{2(n_1+1)} > 0$	$(n_1-n_2)D_B + \frac{Q_1-Q_2}{\sqrt{N_T}}K_B + \frac{\tau_w}{2N_T}\left\{\frac{{Q_1}^2}{n_1+1} - \frac{{Q_2}^2}{n_2+1}\right\} < 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \left\{(n_1+1)D_B + L_1 \cdot K_B + \frac{\tau_w {L_1}^2}{2(n_1+1)}\right\} / \left\{(n_2+1)D_B + \frac{Q_2}{\sqrt{N_T}} \cdot K_B + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T}\right\}$	
Case 1x3 p_1: Case 1 p_2: Case 3	**B.C.**(p_1): $L_1 < (n_1+1)\sqrt{\frac{2D_B}{\tau_w}}$	**B.C.**(p_2): $L_2 > \frac{(n_2+1)}{q_2} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$(n_1+1)D_B + (L_1-L_2)K_B - L_2\sqrt{2\tau_w D_B} + \frac{\tau_w {L_1}^2}{2(n_1+1)} > 0$	$(n_1+1)D_B + \frac{Q_1-Q_2}{\sqrt{N_T}}K_B - \frac{Q_2}{\sqrt{N_T}}\sqrt{2\tau_w D_B} + \frac{\tau_w {Q_1}^2}{2(n_1+1)N_T} < 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \left\{(n_1+1)D_B + L_1 \cdot K_B + \frac{\tau_w {L_1}^2}{2(n_1+1)}\right\} / \left\{\frac{Q_2}{\sqrt{N_T}}(K_B + \sqrt{2\tau_w D_B})\right\}$	
Case 2x1 p_1: Case 2 p_2: Case 1	**B.C.**(p_1): $(n_1+1)\sqrt{\frac{2D_B}{\tau_w}} \le L_1 \le \frac{(n_1+1)}{q_1} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$	**B.C.**(p_2): $L_2 < (n_2+1)\sqrt{\frac{2D_B}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$(n_2+1)D_B + (L_2-L_1)K_B - L_1\sqrt{2\tau_w D_B} + \frac{\tau_w {L_2}^2}{2(n_2+1)} < 0$	$(n_2-n_1)D_B + \frac{Q_2-Q_1}{\sqrt{N_T}}K_B + \frac{\tau_w}{2N_T}\left\{\frac{{Q_2}^2}{n_2+1} - \frac{{Q_1}^2}{n_1+1}\right\} > 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \left\{L_1(K_B + \sqrt{2\tau_w D_B})\right\} / \left\{(n_2+1)D_B + \frac{Q_2}{\sqrt{N_T}} \cdot K_B + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T}\right\}$	
Case 2x2 p_1: Case 2 p_2: Case 2	**B.C.**(p_1): $(n_1+1)\sqrt{\frac{2D_B}{\tau_w}} \le L_1 \le \frac{(n_1+1)}{q_1} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$	**B.C.**(p_2): $(n_2+1)\sqrt{\frac{2D_B}{\tau_w}} \le L_2 \le \frac{(n_2+1)}{q_2} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$L_1 > L_2$	$(n_1-n_2)D_B + \frac{Q_1-Q_2}{\sqrt{N_T}}K_B + \frac{\tau_w}{2N_T}\left\{\frac{{Q_1}^2}{n_1+1} - \frac{{Q_2}^2}{n_2+1}\right\} < 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \left\{L_1(K_B + \sqrt{2\tau_w D_B})\right\} / \left\{(n_2+1)D_B + \frac{Q_2}{\sqrt{N_T}} \cdot K_B + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T}\right\}$	
Case 2x3 p_1: Case 2 p_2: Case 3	**B.C.**(p_1): $(n_1+1)\sqrt{\frac{2D_B}{\tau_w}} \le L_1 \le \frac{(n_1+1)}{q_1} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$	**B.C.**(p_2): $L_2 > \frac{(n_2+1)}{q_2} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$L_1 > L_2$	$(n_1+1)D_B + \frac{Q_1-Q_2}{\sqrt{N_T}}K_B - \frac{Q_2}{\sqrt{N_T}}\sqrt{2\tau_w D_B} + \frac{\tau_w {Q_1}^2}{2(n_1+1)N_T} < 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \frac{L_1}{Q_2} \cdot \sqrt{N_T}$	
Case 3x1 p_1: Case 3 p_2: Case 1	**B.C.**(p_1): $L_1 > \frac{(n_1+1)}{q_1} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$	**B.C.**(p_2): $L_2 < (n_2+1)\sqrt{\frac{2D_B}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$(n_2+1)D_B + (L_2-L_1)K_B - L_1\sqrt{2\tau_w D_B} + \frac{\tau_w {L_2}^2}{2(n_2+1)} < 0$	$(n_2+1)D_B + \frac{Q_2-Q_1}{\sqrt{N_T}}K_B - \frac{Q_1}{\sqrt{N_T}}\sqrt{2\tau_w D_B} + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T} > 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \left\{L_1(K_B + \sqrt{2\tau_w D_B})\right\} / \left\{(n_2+1)D_B + \frac{Q_2}{\sqrt{N_T}} \cdot K_B + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T}\right\}$	
Case 3x2 p_1: Case 3 p_2: Case 2	**B.C.**(p_1): $L_1 > \frac{(n_1+1)}{q_1} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$	**B.C.**(p_2): $(n_2+1)\sqrt{\frac{2D_B}{\tau_w}} \le L_2 \le \frac{(n_2+1)}{q_2} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$L_1 > L_2$	$(n_2+1)D_B + \frac{Q_2-Q_1}{\sqrt{N_T}}K_B - \frac{Q_1}{\sqrt{N_T}}\sqrt{2\tau_w D_B} + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T} > 0$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \left\{L_1(K_B + \sqrt{2\tau_w D_B})\right\} / \left\{(n_2+1)D_B + \frac{Q_2}{\sqrt{N_T}} \cdot K_B + \frac{\tau_w {Q_2}^2}{2(n_2+1)N_T}\right\}$	
Case 3x3 p_1: Case 3 p_2: Case 3	**B.C.**(p_1): $L_1 > \frac{(n_1+1)}{q_1} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$	**B.C.**(p_2): $L_2 > \frac{(n_2+1)}{q_2} \cdot \sqrt{\frac{2D_B N_T}{\tau_w}}$
	$\mathbf{d_{p_1,2D} > d_{p_2,2D}}$	$\mathbf{d_{p_1,3D} < d_{p_2,3D}}$
	$L_1 > L_2$	$Q_1 < Q_2$
	Performance benefit: $d_{p,2D}/d_{p,3D} = \frac{L_1}{Q_2} \cdot \sqrt{N_T}$	

inequality (29) is not satisfied. If L_1 is less than L_2 and inequality (28) is satisfied, the impact of delay reduction caused by path length reduction by uniform scaling is more significant in p_2 than in p_1, so the delay of p_1 will still be greater than that of p_2 after uniform scaling in this case. All these analyses lead to the following statement: if q_1 equals q_2, an RCP occurs only when the logic depth of p_1 is less than that of p_2. Satisfying inequality (28) in this case leads to $L_1 > L_2$.

If q_1 is greater than q_2, L_1 is reduced less effectively than L_2 by 3-D integration, so an RCP is less likely to occur. On the contrary, if q_2 is greater than q_1, an RCP is more likely to occur. In this case, however, the performance benefit will decrease if an RCP occurs.

C. Qualitative Analysis of Reversal of Critical Paths

In this section, we qualitatively analyze the RCP using Table III. We assume that q_1 equals q_2, unless mentioned otherwise.

1) Impact of N_T on the RCP Conditions: If N_T increases, the probability that an RCP occurs also increases because the wire delay is reduced further. In case 1×1, for example, $(n_1 - n_2)D_B$ is always less than zero because $n_1 < n_2$ is

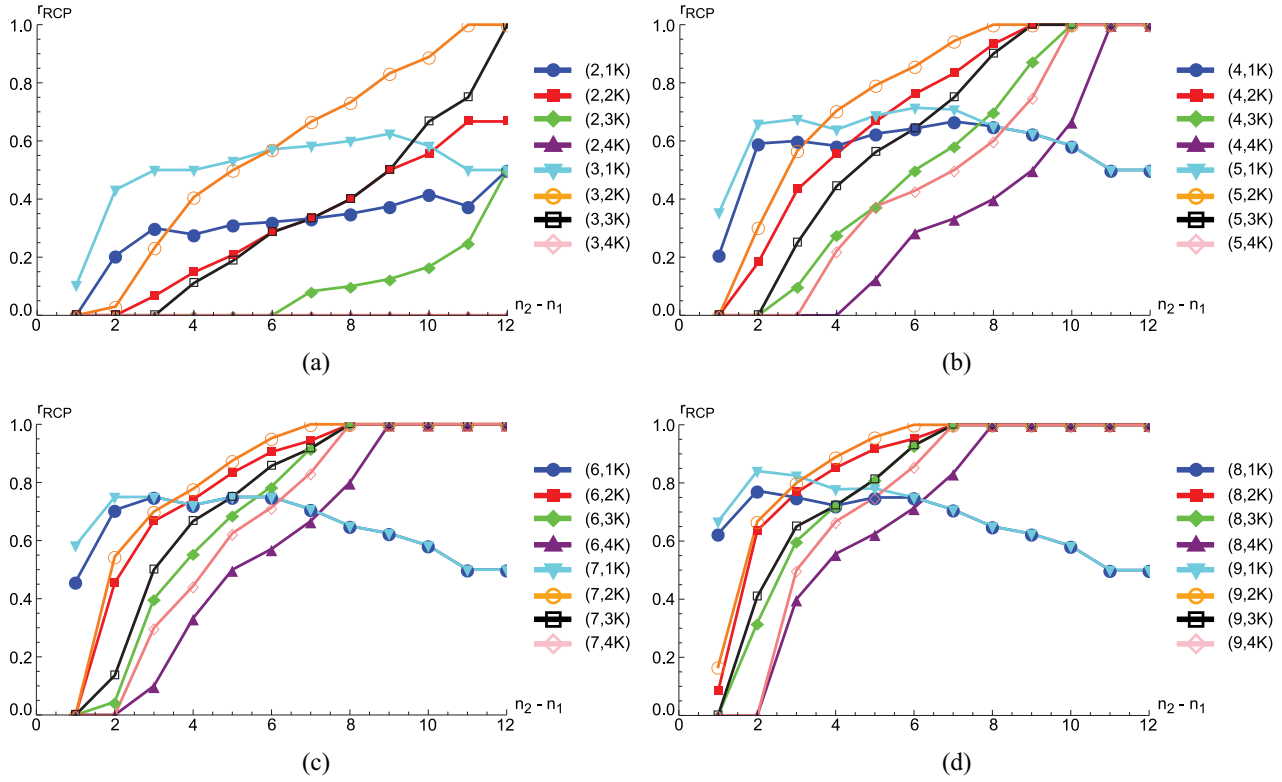


Fig. 6. Ratios (r_{RCP}) between the number of RCP occurrences and the total simulation data points for the 45 nm technology. x and y in (x, y) in the figures denote N_T and $L_1{-}L_2$ (in um), respectively. $q_1 = q_2 = 1.0$. (a) $N_T = 2$ and 3, (b) $N_T = 4$ and 5, (c) $N_T = 6$ and 7, and (d) $N_T = 8$ and 9.

included in the necessary and sufficient condition for the RCP. However, L_1 is greater than L_2, so the second and the third terms for $d_{p_1,\text{2-D}} > d_{p_2,\text{2-D}}$ for the case is positive in Table III. Since the whole sum is greater than zero, $-(n_1-n_2)D_B$ is less than the sum of the second and the third terms. If the paths are uniformly scaled in 3-D, the second and the third terms decrease by $(1/\sqrt{N_T})$ and $(1/N_T)$, respectively, so the sum of the two terms is likely to be less than $-(n_1 - n_2)D_B$. The delay decrement goes up as N_T increases, so the probability that an RCP occurs increases as N_T goes up. Other cases can be analyzed in a similar way.

2) Impact of Path Lengths on the RCP Conditions: If the path length difference $L_1{-}L_2$ goes up, the probability that an RCP occurs increases until $L_1{-}L_2$ reaches a certain value because the amount of wire delay reduction of p_1 increases or that of p_2 decreases (or both of them occur at the same time). If $L_1{-}L_2$ keeps increasing beyond a certain value, however, the probability that an RCP occurs starts decreasing. This is because the length of p_1 is still too long even after uniform scaling if $L_1{-}L_2$ keeps increasing, so the delay of p_1 is still greater than that of p_2. For example, suppose the delay of p_1 is equal to that of p_2 when L_1 is L'. If L_1 goes up over L', the delay of p_1 is always greater than that of p_2. The second and the third terms in the $d_{p_1,\text{3-D}} < d_{p_2,\text{3-D}}$ condition for case 1×1 are scaled by $(1/\sqrt{N_T})$ and $(1/N_T)$, respectively, so an RCP could still occur even if $L_1{-}L_2$ increases. If $L_1{-}L_2$ increases further, however, the wire delay dominates the total delay of p_1, so an RCP does not occur anymore.

3) Impact of Logic Depths on the RCP Conditions: If the logic depth difference $n_2{-}n_1$ goes up, L_1 should be much greater than L_2 to satisfy inequality (28). In this case, $L_1{-}L_2$ goes up, so the probability that an RCP occurs increases as explained. If the logic depth difference increases beyond a certain value, however, the probability that an RCP occurs starts decreasing. This is because the large logic depth difference between p_2 and p_1 requires a large path length difference $L_1{-}L_2$ between p_1 and p_2 for an RCP, but the large path length difference reduces the probability of the RCP as explained.

4) Impact of Routing Congestion on the RCP Conditions: If q_1 goes up, $q_1 \cdot L_1$ increases, so the probability that an RCP occurs decreases. In case 1×1, for example, $d_{p_1,\text{3-D}}$ increases as q_1 goes up, so $d_{p_1,\text{3-D}} < d_{p_2,\text{3-D}}$ would less likely occur. However, the increase of q_1 does not affect the performance benefit $(d_{p_1,\text{2-D}}/d_{p_2,\text{3-D}})$ if an RCP occurs because $d_{p_1,\text{2-D}}$ and $d_{p_2,\text{3-D}}$ are not affected by q_1. If an RCP does not occur, the performance benefit is computed by $d_{p_1,\text{2-D}}/d_{p_1,\text{3-D}}$ using (16), so the performance benefit decreases as q_1 goes up. Other cases can be analyzed in a similar way.

If q_2 goes up, $q_2 \cdot L_2$ increases, so the probability that an RCP occurs increases. In addition, $d_{p_2,\text{3-D}}$ also goes up, so the performance benefit goes down if an RCP occurs as shown in Table III. If an RCP does not occur, the performance benefit is independent of q_2.

D. Quantitative Analysis of Reversal of Critical Paths

In this section, we quantitatively analyze the RCP using the inequalities in Table III and the process technology parameters in Table II for $q_1 = q_2 = 1.0$. Our analysis methodology is as follows. Since the occurrence of an RCP is more tightly

TABLE IV
MINIMUM AND MAXIMUM PERFORMANCE BENEFITS WHEN AN RCP OCCURS (DENOTED BY RCP) AND DOES NOT OCCUR (DENOTED BY NORCP). TECHNOLOGY: 45 nm. $q_1 = q_2 = 1.0$

		N_T														
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Case 1x1	NoRCP (min)	1.029	1.040	1.047	1.052	1.055	1.058	1.060	1.061	1.063	1.064	1.065	1.066	1.067	1.067	1.068
	NoRCP (max)	1.312	1.476	1.580	1.652	1.706	1.748	1.782	1.810	1.833	1.853	1.870	1.886	1.899	1.911	1.922
	RCP (min)	1.039	1.052	1.059	1.064	1.068	1.071	1.073	1.075	1.077	1.078	1.079	1.080	1.081	1.081	1.082
	RCP (max)	1.273	1.464	1.552	1.652	1.701	1.729	1.762	1.769	1.821	1.823	1.867	1.874	1.880	1.885	1.890
Case 1x2 and 1x3	NoRCP (min)	1.029	1.040	1.047	1.052	1.055	1.058	1.060	1.061	1.063	1.064	1.065	1.066	1.067	1.067	1.068
	NoRCP (max)	1.312	1.476	1.580	1.652	1.706	1.748	1.782	1.810	1.833	1.853	1.870	1.886	1.899	1.911	1.922
	RCP	RCP does not occur.														
Case 2x1	NoRCP (min)	1.347	1.536	1.658	1.743	1.808	1.858	1.899	1.933	1.962	1.986	2.007	2.026	2.042	2.057	2.070
	NoRCP (max)	1.412	1.731	1.973	2.225	2.409	2.642	2.825	2.983	3.159	3.303	3.462	3.597	3.722	3.838	3.947
	RCP (min)	1.053	1.066	1.073	1.077	1.081	1.083	1.085	1.087	1.088	1.090	1.091	1.092	1.092	1.093	1.094
	RCP (max)	1.405	1.719	1.942	2.213	2.397	2.572	2.791	2.895	3.049	3.131	3.361	3.384	3.638	3.653	3.860
Case 2x2	NoRCP (min)	1.347	1.536	1.658	1.743	1.808	1.858	1.899	1.933	1.961	1.986	2.007	2.026	2.042	2.057	2.070
	NoRCP (max)	1.412	1.731	1.973	2.225	2.409	2.642	2.825	2.983	3.159	3.303	3.462	3.597	3.722	3.838	3.947
	RCP (min)	-	-	-	2.179	2.260	2.323	2.374	2.416	2.452	2.482	2.509	2.532	2.553	2.571	2.588
	RCP (max)	-	-	-	2.179	2.260	2.527	2.593	2.915	2.976	3.030	3.439	3.498	3.552	3.600	3.866
Case 2x3	NoRCP (min)	1.347	1.536	1.658	1.743	1.808	1.858	1.899	1.933	1.961	1.986	2.007	2.026	2.042	2.057	2.070
	NoRCP (max)	1.412	1.731	1.973	2.225	2.409	2.642	2.825	2.983	3.159	3.303	3.462	3.597	3.722	3.838	3.947
	RCP	RCP does not occur.														
Case 3x1	NoRCP (min) NoRCP (max)	1.414	1.732	2.000	2.236	2.449	2.646	2.828	3.000	3.162	3.317	3.464	3.606	3.742	3.873	4.000
	RCP (min)	1.053	1.066	1.073	1.077	1.081	1.083	1.085	1.087	1.088	1.090	1.091	1.092	1.092	1.093	1.094
	RCP (max)	1.404	1.719	1.988	2.226	2.420	2.610	2.791	2.895	3.113	3.303	3.361	3.384	3.656	3.854	3.860
Case 3x2	NoRCP (min) NoRCP (max)	1.414	1.732	2.000	2.236	2.449	2.646	2.828	3.000	3.162	3.317	3.464	3.606	3.742	3.873	4.000
	RCP (min)	-	-	-	2.179	2.260	2.323	2.374	2.416	2.452	2.482	2.509	2.532	2.553	2.571	2.588
	RCP (max)	-	-	-	2.179	2.445	2.527	2.593	2.915	2.976	3.310	3.439	3.590	3.640	3.829	3.977
Case 3x3	NoRCP (min) NoRCP (max)	1.414	1.732	2.000	2.236	2.449	2.646	2.828	3.000	3.162	3.317	3.464	3.606	3.742	3.873	4.000
	RCP	RCP does not occur.														

related to the difference of the lengths and logic depths of two given paths than their individual values, we focus on the relationships among L_1-L_2, n_2-n_1, N_T, and RCP occurrences, where L_i is the length of path p_i and n_i is the logic depth of p_i. We sweep L_1 and L_2 from 1000 um to 5000 um by a step size of 1000 um with the condition $L_1 > L_2$, n_1 and n_2 from 0 to 12 with the condition $n_2 > n_1$, and N_T from 2 to 16 and compute whether an RCP occurs between p_1 and p_2. Fig. 6 shows ratios r_{RCP} between the number of RCP occurrences and the total number of simulation data points sweeped for each (N_T, L_1-L_2) for each n_2-n_1. We plot them only for the 45 nm technology due to page limit, but other technology nodes show similar trends.

In general, r_{RCP} goes up and saturates at 1.0 as the logic depth difference (n_2-n_1) increases. If p_1 is more critical than p_2 in a given 2-D design even when the logic depth difference is large, p_1 should be much longer than p_2, so an RCP is more likely to occur as the logic depth difference increases. In addition, r_{RCP} saturates faster as the number of tiers goes up. This is because larger N_T reduces the path length further, so an RCP occurs for smaller logic depth differences as N_T increases.

However, there are a few exceptional cases, especially when the path length difference is less than 1000 um. In general, an RCP is more likely to occur, so r_{RCP} goes up as n_2-n_1 increases. If the logic depth difference increases beyond a certain point and L_1-L_2 is less than 1000 um, however, r_{RCP} starts going down for the following reason. If p_1 has longer delay than p_2, L_1-L_2 is small, and n_2-n_1 is large, their path lengths are so long (e.g., 5000 um for p_1 and 4000 um for p_2) that the path delays are dominated by their wire and buffer delays instead of their instance delays. In this case, an RCP does not occur after uniform scaling because the path delays will still be dominated by the wire and buffer delays, so r_{RCP} goes down.

E. Performance Benefits

In this section, we predict performance benefits ($d_{p,\text{2-D}}/d_{p,\text{3-D}}$) using the performance benefit equations in Table III assuming $q_1 = q_2 = 1.0$. We use the 45 nm technology parameters to quantify and investigate the performance benefits. The quantification methodology is as follows. We sweep 1) L_1 and L_2 from 1000 um to 5000 um by a step size of 1000 um; 2) n_1 and n_2 from 0 to 12; and 3) N_T from 2 to 16 and compute the performance benefits. If an RCP does not occur, we compute the performance benefit by $d_{p_1,\text{2-D}}/d_{p_1,\text{3-D}}$, and if an RCP occurs, we compute the performance benefit by $d_{p_1,\text{2-D}}/d_{p_2,\text{3-D}}$. Table IV shows the minimum and maximum performance benefits when an RCP occurs (denoted by RCP) and does not occur (denoted by NoRCP).

The maximum performance benefits of RCP are less than those of NoRCP as expected. However, the differences between the maximum performance benefits of RCP and NoRCP are less than 3%, 6%, 9%, 7%, and 9% for case 1 × 1, case 2 × 1, case 2 × 2, case 3 × 1, and case 3 × 2, respectively. In addition, the RCP and NoRCP cases have similar minimum performance benefits, which are much smaller than

TABLE V
TRENDS OF PERFORMANCE BENEFITS OBTAINABLE FROM MULTITIER GATE-LEVEL MONOLITHIC 3-D INTEGRATION. ↑ AND ↓ MEAN AN INCREASE AND A DECREASE, RESPECTIVELY

N_T	$\lvert L_1 - L_2 \rvert$	$\lvert n_2 - n_1 \rvert$	q_1	q_2	Performance benefits			
					NoRCP		RCP	
					min	max	min	max
↑	-	-	-	-	↑	↑	↑	↑
-	↑	-	-	-	↑	↑	↑	↑
-	-	↑	-	-	↓	↓	↓	↗ ↘
-	-	-	↑	-	↓	↓	-	-
-	-	-	-	↑	-	-	↓	↓

the maximum performance benefits. For instance, if all the nets in a given 2-D layout belong to case 1 because they are relatively short and/or their logic depths are high, the performance benefit we obtain is 1.082× even if 16 tiers are stacked. This case occurred when (N_T, n_1, n_2, L_1, L_2) = (16, 9, 11, 2000 um, 1000 um) in our simulation. In this case, both p_1 and p_2 belonged to case 1 and an RCP occurred. Similarly, the minimum performance benefit in case 3 × 1 when N_T is 16 is only 1.094×, which occurred when (N_T, n_1, n_2, L_1, L_2) = (16, 0, 12, 3000 um, 1000 um). In this case, an RCP occurred again. On the other hand, if an RCP does not occur, the minimum performance benefit is equal to the maximum performance benefit when the critical path belongs to case 3. In this case, the performance benefit when N_T is 16 is 4×. Thus, we observe that the performance benefits we can obtain from multitier gate-level monolithic 3-D integration are highly dependent on all the design and technology parameters, such as the number of tiers, path lengths, instance delays, and whether an RCP occurs or not.

We also observe some trends of performance benefits in Table V, which we obtained by a thorough investigation of all the computation results. As the table shows, if N_T goes up, the range ([min, max]) of the benefits goes up regardless of the RCP occurrences. This is because increasing the number of tiers always helps reduce the length of each net. If $|L_1-L_2|$ goes up, the range of the benefits also goes up because the decrement of a wire length increases as its length goes up. If $|n_2-n_1|$ goes up, the range of NoRCP goes down. If an RCP does not occur and the logic depth of the most critical path in the 2-D layout goes up, the impact of monolithic 3-D integration on the path delay reduction decreases, so the performance benefit goes down. On the other hand, if $|n_2 - n_1|$ goes up, the minimum performance benefit of RCP goes down, but the maximum performance benefit of RCP increases and then decreases. If q_1 goes up, the performance benefit goes down if an RCP does not occur, but the benefit does not change if an RCP occurs. Similarly, if q_2 goes up, the performance benefit goes down if an RCP occurs, but the benefit does not change if an RCP does not occur.

V. VALIDATION AND CASE STUDY

In this section, we design 2-D and monolithic 3-D IC layouts for two benchmarks, analyze their timing characteristics, and demonstrate how we can use the performance benefit models to estimate the performance benefits of the circuits.

A. Simulation Setup

We use low-density parity-check (LDPC) and fast Fourier transform (FFT) circuits for benchmarks. The reason we chose them for the case study is because they are relatively large (LDPC is 0.36 mm^2 and FFT is 1.11 mm^2), so they can show the benefits of multitier gate-level monolithic 3-D integration. We use the Nangate 45 nm library for a standard cell library [20], Synopsys Design Compiler for netlist synthesis, and Cadence Innovus for physical design and timing analysis. We also built an in-house software for custom analyses of the physical layouts. To generate monolithic 3-D IC layouts, we used the uniform-scaling-based design methodology in [8] for 3-D placement and the routing methodology in [5] for 3-D routing. We also performed timing optimization using Cadence Innovus on the 3-D IC layouts with modified buffer cells. The modified buffer cells have the same characteristics as the buffers defined in the original standard cell library, but they have much smaller area than the regular buffers. We use the small buffers to allow the optimization tool to insert as many buffers as it wants. We use this workaround layout generation technique because no commercial tool supports timing optimization for 3-D IC layouts. We also use 5Ω for the resistance of an MIV.

B. Validation and Performance Benefits

We estimate performance benefits of designing a 2-D layout in a monolithic 3-D IC as follows. First, we decompose the delays of top 1000 critical paths in the 2-D layout into instance delays (d_i), buffer delays (d_b), and net delays (d_n). Since all the top 1000 critical paths in the two benchmarks contain many long wires, all of them belong to case 3. The delay benefit ($d_{p,2\text{-D}}/d_{p,3\text{-D}}$) of a path belonging to case 3 is ($\sqrt{N_T}/q_p$), so we predict the delay of each path in the 3-D layout by $d_i+([q_p \cdot (d_b+d_n)]/[\sqrt{N_T}])$. Then, we compute the performance benefit assuming an RCP does not occur or an RCP occurs between the most critical path and the kth critical path ($k \geq 2$). We compare this performance benefit with the actual performance benefit we obtain from the 3-D layouts we build. The LDPC design does not have routing congestion, so we set q_p to 1.0 for LDPC, but we set q_p to 1.1 for the FFT design because it has some routing congestion.

Table VI shows the actual and predicted performance benefits of the two benchmarks. The delay of the most critical path (p_1) of the 2-D LDPC design is 5.574 ns, which consists of 1.866 ns instance delay, 1.548 ns buffer delay, and 2.140 ns net delay. If it is redesigned in a two-tier monolithic 3-D IC layout, the predicted performance benefit is 1.24× (i.e., the critical path delay will decrease by 1.24×) if an RCP does not occur. If an RCP occurs, the predicted performance benefit is dependent on with which path the RCP occurs. Thus, we show in Table VI the predicted performance benefits assuming an RCP occurs with the kth critical path. If k is in the range of [2, 50], the predicted performance benefit is between 1.27× and 1.41×. If the range of k increases, the predicted performance benefit also increases.

The critical path delay in the two-tier LDPC design is 4.071 ns, so the actual performance benefit is 1.37×. The most

TABLE VI
ACTUAL AND PREDICTED PERFORMANCE BENEFITS ($[d_{p,2\text{-D}}/d_{p,3\text{-D}}]$) OF LDPC (# GATES: 50 753, # NETS: 54 785, LAYOUT AREA: 0.36 mm^2) AND FFT (# GATES: 255 710, # NETS: 259 429, LAYOUT AREA: 1.11 mm^2). CPD DENOTES THE CRITICAL PATH DELAY. THE PATH CAUSING AN RCP IS DENOTED BY "*." "[aTH, bTH] : [$c\times$, $d\times$]" MEANS THAT IF AN RCP OCCURS BETWEEN THE FIRST AND THE kTH CRITICAL PATHS, WHERE $a \leq k \leq b$, THE RANGE OF THE PERFORMANCE BENEFIT IS PREDICTED TO BE BETWEEN $c\times$ AND $d\times$

Benchmark	CPD (ns)	Design	Actual	Predicted
LDPC	5.574	2-tier	**1.37**×	No RCP: 1.24×
				(39th) 1.36× *
				[2nd – 50th]: [1.27×, 1.41×]
				[2nd – 100th]: [1.27×, 1.45×]
				[2nd – 200th]: [1.27×, 1.50×]
				[2nd – 500th]: [1.27×, 1.56×]
				[2nd – 1,000th]: [1.27×, 1.63×]
		3-tier	**1.57**×	No RCP: 1.39×
				(21st) 1.46× *
				[2nd – 50th]: [1.39×, 1.59×]
				[2nd – 100th]: [1.39×, 1.63×]
				[2nd – 200th]: [1.39×, 1.68×]
				[2nd – 500th]: [1.39×, 1.74×]
				[2nd – 1,000th]: [1.39×, 1.83×]
FFT	3.490	2-tier	**1.22**×	No RCP: 1.17×
				(13th) 1.24× *
				[2nd – 50th]: [1.19×, 1.41×]
				[2nd – 100th]: [1.19×, 1.45×]
				[2nd – 200th]: [1.19×, 1.54×]
				[2nd – 500th]: [1.19×, 1.75×]
				[2nd – 1,000th]: [1.19×, 1.81×]
		3-tier	**1.35**×	No RCP: 1.32×
				(67th) 1.39× *
				[2nd – 50th]: [1.28×, 1.57×]
				[2nd – 100th]: [1.28×, 1.61×]
				[2nd – 200th]: [1.28×, 1.71×]
				[2nd – 500th]: [1.28×, 1.97×]
				[2nd – 1,000th]: [1.28×, 2.01×]

critical path in the two-tier 3-D design is the 39th critical path in the 2-D design, so an RCP occurred between the first and the 39th critical paths. The predicted performance benefit when an RCP occurs with the 39th critical path is 1.36×, so it is very close to the actual performance benefit. However, notice that it would require more detailed timing and physical design analysis to predict whether an RCP would occur or not and with which path an RCP would occur if an RCP would occur.

The critical path delay in the three-tier LDPC design is 3.548 ns, so the actual performance benefit is 1.57×. The most critical path in the three-tier design is the 21st critical path in the 2-D design, so an RCP occurred between the first and the 21st critical paths. Assuming $k_1 > k_2$, notice that an RCP can occur between the first and the k_2th critical paths in the three-tier design when an RCP occurs between the first and the k_1th critical paths in the two-tier design. The predicted performance benefit when an RCP occurs with the 21st path is 1.46×.

The FFT designs show similar results. The delay of the most critical path in the 2-D design is 3.490 ns, which consists of 1.145 ns instance delay, 1.216 ns buffer delay, and 1.129 ns net delay. If it is redesigned in a two-tier monolithic 3-D IC layout, the predicted performance benefit is 1.17× if an RCP does not occur. If an RCP occurs, the predicted performance benefit varies from [1.19×, 1.41 ×] to [1.19×, 1.81 ×] for an RCP occurring with one of the top 50 paths and one of the top 1000 paths, respectively. The most critical path in the two-tier 3-D design is the 13th critical path in the 2-D design and the predicted performance benefit when an RCP occurs with the 13th critical path is 1.24×. The actual performance benefit for the three-tier 3-D design is 1.35×, which lies between [1.28×, 1.61×]. An RCP occurs in this design between the most and the 67th critical paths and the predicted performance benefit for this case is 1.39×, which is very close to the actual performance benefit 1.35×.

C. Discussion

1) Pin Capacitance: Input pin capacitance is not negligible in the delay computation depending on the path length and the process technology. If pin capacitance is not negligible in a path, it is because the path has many sinks and/or the nets in the path are too short. If the path is optimized effectively, however, the sinks not on the critical paths can be grouped and driven by off-loading buffers. Thus, if pin capacitance is not negligible in a critical path, it is generally because the nets in the path are short. In this case, the path will belong to case 1 and redesigning this circuit in a monolithic 3-D IC would not improve the performance of the circuit.

2) MIV Resistance and Capacitance: In our simulation, most of the nets in the critical paths are longer than 2000 um, so the total wire *RC* is much greater than the total MIV *RC* in the paths. Thus, we ignored MIV resistance and capacitance in all the delay computations in this paper although we used 5 Ω for the MIV resistance in the actual design and analysis.

If the MIV *RC* is not negligible, we can take the MIV *RC* into the analytical models as follows. Suppose r_m and c_m are the resistance and capacitance of an MIV, respectively, and n_m is the total number of MIVs in a net of length l. Assuming the MIVs are evenly distributed throughout the net, the total resistance and capacitance of the net are $l \cdot r_w + n_m \cdot r_m$ and $l \cdot c_w + n_m \cdot c_m$, respectively. We can consider this net as a net of length l whose unit resistance and capacitance are $(l \cdot r_w + n_m \cdot r_m)/l$ and $(l \cdot c_w + n_m \cdot c_m)/l$, respectively, and insert buffers using (3).

3) Multifanout Nets: If a net is a multifanout net, the source-to-sink length has a bigger impact than the total net length on the net delay. This is because the Elmore delay of a multifanout net varies significantly depending on the locations of the sinks and the routing topology of the net. If we can assume that the routing topology of a net in a given 2-D design is preserved in its monolithic 3-D design, we would be able to estimate the performance benefit of the net more accurately by a more sophisticated analytical model.

VI. CONCLUSION

In this paper, we mathematically analyzed the performance benefits of multitier gate-level monolithic 3-D ICs designed by the uniform-scaling-based 3-D placement and analytical optimal buffer insertion. The delay benefit of a single-fanout net obtainable from N_T-tier monolithic 3-D integration is $(\sqrt{N_T}/q)\times$ when the net is sufficiently long and optimally buffered. In a multitier gate-level monolithic 3-D IC, however, multiple nets are scaled down at the same time and the delays

of some paths are dominated by instance delays, whereas those of some other paths are dominated by net delays. Thus, RCP could occur. We also analyzed the RCP in detail in this paper to investigate under what conditions an RCP would occur. Finally, we validated the performance benefit models by building and comparing 2-D and monolithic 3-D IC layouts. The proposed analytical models for performance benefits of multitier gate-level monolithic 3-D ICs can be used for fast estimation of performance benefits without physical designing a 2-D IC in 3-D.

REFERENCES

[1] P. Batude *et al.*, "Advances in 3D CMOS sequential integration," in *Proc. IEEE Int. Electron Devices Meeting*, Baltimore, MD, USA, Sep. 2009, pp. 1–4.

[2] S. Bobba *et al.*, "CELONCEL: Effective design technique for 3-D monolithic integration targeting high performance integrated circuits," in *Proc. Asia South Pac. Design Autom. Conf.*, Yokohama, Japan, Jan. 2011, pp. 336–343.

[3] P. Batude *et al.*, "Demonstration of low temperature 3D sequential FDSOI integration down to 50 nm gate length," in *Proc. IEEE Symp. VLSI Technol.*, Honolulu, HI, USA, Jun. 2011, pp. 158–159.

[4] C.-H. Shen *et al.*, "Monolithic 3D chip integrated with 500ns NVM, 3ps logic circuits and SRAM," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, USA, Dec. 2013, pp. 9.3.1–9.3.4.

[5] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD methodologies for low power gate-level monolithic 3D ICs," in *Proc. Int. Symp. Low Power Electron. Design*, La Jolla, CA, USA, Aug. 2014, pp. 171–176.

[6] D. H. Kim, K. Athikulwongse, and S. K. Lim, "Study of through-silicon-via impact on the 3-D stacked IC layout," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 21, no. 5, pp. 862–874, May 2013.

[7] D. H. Kim, S. Mukhopadhyay, and S. K. Lim, "TSV-aware interconnect distribution models for prediction of delay and power consumption of 3-D stacked ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 9, pp. 1384–1395, Sep. 2014.

[8] S.-E. D. Lin, P. P. Pande, and D. H. Kim, "Optimization of dynamic power consumption in multi-tier gate-level monolithic 3D ICs," in *Proc. Int. Symp. Qual. Electron. Design*, Santa Clara, CA, USA, Mar. 2016, pp. 29–34.

[9] K. M. Kim, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "Four-tier monolithic 3D ICs: Tier partitioning methodology and power benefit study," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2016, pp. 70–75.

[10] K. Chang *et al.*, "Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, Austin, TX, USA, Nov. 2016, pp. 1–8.

[11] Y. Chen and D. H. Kim, "A legalization algorithm for multi-tier gate-level monolithic three-dimensional integrated circuits," in *Proc. Int. Symp. Qual. Electron. Design*, Santa Clara, CA, USA, Mar. 2017, pp. 1–6.

[12] K. Acharya *et al.*, "Monolithic 3D IC design: Power, performance, and area impact at 7nm," in *Proc. Int. Symp. Qual. Electron. Design*, Santa Clara, CA, USA, Mar. 2016, pp. 41–48.

[13] J. Lou, S. Thakur, S. Krishnamoorthy, and H. S. Sheng, "Estimating routing congestion using probabilistic analysis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 1, pp. 32–41, Jan. 2002.

[14] R. S. Shelar, S. S. Sapatnekar, P. Saxena, and X. Wang, "A predictive distributed congestion metric with application to technology mapping," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 5, pp. 696–710, May 2005.

[15] T. Thorolfsson, K. Gonsalves, and P. D. Franzon, "Design automation for a 3DIC FFT processor for synthetic aperture radar: A case study," in *Proc. ACM Design Autom. Conf.*, San Francisco, CA, USA, 2009, pp. 51–56.

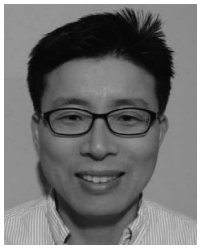
[16] D. Fick *et al.*, "Centip3De: A cluster-based NTC architecture with 64 ARM cortex-M3 cores in 3D stacked 130 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 104–117, Jan. 2013.

[17] D. H. Kim *et al.*, "Design and analysis of 3D-MAPS (3D massively parallel processor with stacked memory)," *IEEE Trans. Comput.*, vol. 64, no. 1, pp. 112–125, Jan. 2015.

[18] M. Pathak, Y.-J. Lee, T. Moon, and S. K. Lim, "Through-silicon-via management during 3D physical design: When to add and how many?" in *Proc. IEEE Int. Conf. Comput.-Aided Design*, San Jose, CA, USA, 2010, pp. 387–394.

[19] W.-T. J. Chan, Y. Du, A. B. Khang, S. Nath, and K. Samadi, "3DIC benefit estimation and implementation guidance from 2DIC implementation," in *Proc. ACM Design Autom. Conf.*, San Francisco, CA, USA, Jun. 2015, pp. 1–6.

[20] *Nangate 45nm Open Cell Library*, Nangate, Santa Clara, CA, USA, 2011. [Online]. Available: http://www.nangate.com



Inki Hong received the B.S. degree in management science from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, the M.S. degree in computer science from Stanford University, Stanford, CA, USA, and the Ph.D. degree in computer science from the University of California at Los Angeles, Los Angeles, CA, USA.

He is a Research and Development Senior Group Director of Digital Signoff Group of Cadence, San Jose, CA, USA. Since 2005, he has been researching on logic and physical design technologies for Cadence's IC implementation tools. He is currently leading a global team of Research and Development engineers to develop the next generation physical optimization technology for Innovus physical implementation system. From 1998 to 2005, he was with Synopsys, Mountain View, CA, USA, developing physical synthesis technology. He has published over 25 articles on power optimization and physical design in premier journals and conferences.

Dr. Hong was a recipient of the Design Automation Conference Graduate Student Scholarship in 1997.

Dae Hyun Kim (S'08–M'12) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2002, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2007 and 2012, respectively.

He is an Assistant Professor with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA. He researched on physical layout optimization with Cadence Design Systems, Inc., San Jose, CA, USA, from 2012 to 2014. His current research interests include electronic design automation and computer-aided design for very large scale integration (VLSI), high-performance, low-power VLSI and computer systems, and 3-D ICs and systems.

Dr. Kim was a recipient of the Cadence Excellence in Innovation Award in 2014, the Defense Advanced Research Projects Agency Young Faculty Award in 2016, and the EECS Early Career Award from the School of Electrical Engineering and Computer Science at Washington State University in 2017.