Compact-2D: A Physical Design Methodology to Build Two-Tier Gate-Level 3-D ICs

Bon Woong Ku[®], Kyungwook Chang[®], Student Member, IEEE, and Sung Kyu Lim, Senior Member, IEEE

Abstract-The recent advancement of wafer bonding and monolithic integration technology offers fine-grained 3-D interconnections to face-to-face (F2F) and monolithic 3-D (M3D) ICs. In this article, we propose a full-chip RTL-to-GDSII physical design solution to build commercial-quality two-tier gate-level F2F and M3D ICs. The state-of-the-art flow named shrunk-2D (S2D) requires shrinking of standard cells and interconnects by a factor of 50% to fit into the target 3-D footprint of a two-tier design. This, unfortunately, necessitates commercial place/route engines that handle one node smaller geometries, which can be challenging and costly. Our flow named compact-2D (C2D) does not require any geometry shrinking. Instead, C2D implements a 2-D IC with scaled interconnect RC parasitics and contracts the layout to the 3-D integrated circuit footprint. In addition, C2D offers post-tier-partitioning optimization (post-TP opt) which is completely missing in S2D. This additional optimization step is shown to be effective in fixing timing violations caused by intertier 3-D routing overhead. Lastly, we present a methodology to reuse the routing result of post-TP opt for the final GDSII generation. Our experimental results show that at iso-performance, C2D offers up to 28.0% power reduction and 15.6% silicon area savings over commercial 2-D ICs without any routing resource overhead.

Index Terms—Compact-2D (C2D), face-to-face (F2F)-bonded 3-D integrated circuit (3-D IC), monolithic 3-D (M3D) IC, physical design methodology.

I. INTRODUCTION

I N THE looming end of Moore's law, 3-D integration technology has been drawing huge attention in the hope of inheriting the transistor scaling benefit for the system-level improvement. However, 3-D integrated circuits (3-D ICs) call for a tighter 3-D interconnect pitch to further improve the functional density and power-performance-area benefits as 2-D interconnects become denser in the advanced technology nodes. Among several notable achievements, hybrid wafer-to-wafer (W2W) bonding technology [2], [3] and monolithic 3-D (M3D) technology [4], [5] have emerged as promising solutions for the tight 3-D contact pitch in the advanced 3-D

Manuscript received June 14, 2018; revised December 13, 2018, April 21, 2019, and August 14, 2019; accepted October 21, 2019. Date of publication November 8, 2019; date of current version May 22, 2020. This work was supported by DARPA ERI 3DSOC Program under Award HR001118C0096. This article was recommended by Associate Editor Bustany_Chu. (*Corresponding author: Bon Woong Ku.*)

Bon Woong Ku is with Synopsys, Inc., Mountain View, CA 94043 USA (e-mail: bon.ku@synopsys.com).

Kyungwook Chang and Sung Kyu Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: k.chang@gatech.edu; limsk@ece.gatech.edu).

Digital Object Identifier 10.1109/TCAD.2019.2952542

Hybrid W2W Bonding Technology





Fig. 1. (1) Hybrid W2W bonding technology [2], [3]. (a) Both wafers are fabricated in parallel before the bonding. The wafer surface is flattened by chemical-mechanical planarization. (b) Top wafer is flipped, and wafers are aligned, bonded at room temperature, and annealed at < 250 °C. (c) F2F via is naturally formed at the location of a direct metal-to-metal bonding. (2) M3D technology [4], [5]. (a) Empty top wafer is sheared off from the bulk carrier by the H+ ion implant cut and bonded at room temperature to the BEOL of a prefabricated bottom wafer. (b) After planarization on the top wafer, MIVs are created for 3-D interconnections using lithography technology. (c) Top FEOL and BEOL are fabricated within a low thermal budget to keep the integrity of the bottom tier.

integration. Fig. 1 illustrates the two-tier integration process of each technology.

Hybrid W2W bonding enables direct metal-to-metal (damascene-pad) and dielectric-to-dielectric bonding between the back-end-of-lines (BEOLs) of prefabricated wafers in a face-to-face (F2F) fashion. Wafers are fabricated in parallel with the conventional process before the bonding. After the wafer surface is flattened by chemical-mechanical planarization, wafers are aligned, bonded at room temperature,

0278-0070 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information. and annealed at <250 °C to strengthen the interfacial bonding. As a result, an F2F via is naturally formed at the location of a direct metal-to-metal bonding. Recently, a 1.8- μ m F2F via pitch has been demonstrated, and the minimum pitch is projected down to 0.8- μ m in the near future thanks to the advancement of wafer-alignment precision [6], [7].

On the other hand, M3D technology enables sequential fabrication on top of a prefabricated bottom wafer. An empty top wafer is sheared off from the bulk carrier by H+ ion implant cut and bonded at room temperature to the BEOL of the bottom wafer. After planarization on the top wafer, monolithic intertier vias (MIVs) are created for 3-D interconnections using lithography technology, and top frontend-of-line (FEOL) and metal stacks are fabricated within a low thermal budget to keep the integrity of the bottom FEOL and BEOL. Because of this intertier process variation, M3D integration is farther from commercialization than F2F integration,¹ but the minimum pitch of MIVs is projected down to less than 0.1 μ m, which offers more fine-grained 3-D interconnections.

Compared to the traditional through-silicon-via (TSV)based 3-D integration where the minimum pitch of 3-D interconnections is larger than 10 μ m, both F2F and M3D integration offer higher degrees of freedom in 3-D placement and routing (P&R) to designers. Therefore, while TSV-based 3-D ICs are suitable for the block-level 3-D integration where coarse-grained 3-D interconnects are utilized for the interblock communication, F2F and M3D integration open up the gate-level 3-D integration era enabling the massive number of intergate 3-D connections. However, commercial electronic design automation tools are not available to support gate-level F2F and M3D IC implementation yet, and new design and CAD solutions are required to fully benefit from the advanced 3-D integration technologies. In this article, we present an RTL-to-GDSII physical design methodology named compact-2D (C2D) to build areaoptimal and commercial-quality two-tier gate-level F2F and M3D ICs.

II. MOTIVATION

The state-of-the-art full-chip design flow named shrunk-2D (S2D) [1], [8] proposed a unique method to build gate-level 3-D ICs using 2-D P&R tools. Assuming no silicon-area overhead in a two-tier 3-D IC, the 3-D design footprint is 50% of the 2-D IC footprint. To fit all the synthesized gates into this small 3-D footprint, S2D shrinks the standard cells and interconnects by 50% to double the chip capacity and uses them to implement an S2D design with the conventional 2-D P&R steps. This approach is based on the assumption that the Z dimension is so small and thus negligible, and the half perimeter wirelength (HPWL) of the S2D design is the same as that of the flattened 3-D design, where cells in 3-D space are projected onto the single placement layer.

As a result of S2D design, the (X, Y) location of a cell is determined, and tier partitioning is subsequently performed to

decide the Z location of the cell. At this moment, reducing the perturbation from legalization is the key objective for the tier partitioning. For example, consider that there are two adjacent cells at a placement row in the S2D design, and they are assigned at the same tier. Then legalization with the restored cell size causes displacement from their optimal (X, Y) locations. However, if they are assigned to different tiers, they can keep their original (X, Y) locations even after the size restoration. To address this issue, S2D dissects the S2D design into partitioning bins in a regular fashion and performs Fiduccia– Mattheyses (FM) [9] min-cut partitioning while keeping the area balance within each local bin [8], [10], [11].²

After tier partitioning and placement legalization at each tier, S2D conducts detailed routing for intertier 3-D nets with full metal stacks for both tiers and finds the resulting (X, Y) location of a 3-D via as the implicit position of a 3-D connection between two tiers. This is called 3-D via planning. 3-D via planning results in netlists containing those 3-D connections as primary I/O ports for each tier, and detailed routing at separate tiers produces the final GDSII layouts. Although S2D presents how to use commercial 2-D P&R engines to build two-tier gate-level F2F and M3D ICs, it introduces the following new issues, especially, in the advanced technology nodes.

- To handle shrunk geometries, S2D requires P&R engines and design rule checkers that target one node smaller technology, which is both challenging and costly.³
- The shrunk dimension of interconnects leads to inaccuracy in *RC* parasitics of the S2D design unless the parasitic database is rebuilt for the shrunk geometries.⁴
- S2D does not support any optimization after tier partitioning. Therefore, it is prone to timing failure caused by intertier 3-D routing overhead.

Besides S2D flow, [13]–[15] also proposed the physical design flow for the advanced 3-D ICs utilizing commercial 2-D P&R engines. Ku et al. [13] presented derated-2D flow, which supports timing closure after tier-partitioning to address the intertier variations in M3D ICs. However, the optimization was performed only for the bottom tier cells, and the buffer insertion was not based on the legal placement. Ku et al. [14] presented projected-2D flow and studied the power-performance-cost tradeoffs for M3D ICs. Projected-2D transforms a 2-D IC into an M3D IC directly. This might help minimize the performance degradation caused by intertier 3-D routing overhead, but the power saving from the reduced wirelength in the M3D IC was not accounted in the analysis. Chang et al. [15] presented cascade-2D flow for M3D ICs as well, which utilizes the hierarchical design methodology in 2-D IC implementation tools. However, cascade-2D requires

¹Recently, one of the top semiconductor manufacturers has officially announced their F2F wafer bonding technology dubbed wafer-onwafer (WoW) targeting advanced 3-D-integrated graphics processing units.

²The tier partitioning algorithm is not limited to the bin-based FM min-cut partitioning, but various strategies have been proposed for the given design and technology constraints [12], [13].

³Our conversations with S2D flow users at industry design houses revealed an exponential increase in design rule violations (DRVs) at the 7 nm node. Designers reported that an excessive number of violations may cause commercial engines to terminate abruptly or produce low-quality layouts.

⁴The resistance of a shrunk wire segment in the S2D design is not the same as that of the final 3-D design. Similarly, the shrunk width and spacing of interconnects lead to inaccurate capacitance values in the S2D design.

 TABLE I

 Terminologies in Our C2D Flow. The Target 3-D Design Footprint is N% of the 2-D Design Footprint

Compact-2D	An initial 2D design with unit length RC scaled by
Design	a factor of $\sqrt{N\%}$
Memory	Expanding the pin locations and memory macro
Expansion	boundaries by a factor of $1/\sqrt{N\%}$
Placement	Linearly contracting the placement solution of the
Contraction	Compact-2D design by a factor of $\sqrt{N\%}$
Compact 3D	Performing timing, power, and 3D via location co-
Via Planning	optimization to address inter-tier routing overhead
Incremental	Recycling the routing result from Compact 3D Via
Routing	Planning for final GDSII generation



Fig. 2. Our C2D flow. C2D is a full-chip RTL-to-GDSII physical design solution to build commercial-quality two-tier gate-level 3-D ICs. C2D utilizes 2-D IC implementation tools and finds the optimal (X, Y) location of a cell through a C2D design and placement contraction. Then, C2D determines the Z location of a cell by tier partitioning. Compact 3-D via planning performs post-TP opt using commercial 2-D P&R engines to compensate intertier 3-D routing overhead. Incremental routing is a CAD solution to preserve the routing result from compact 3-D via planning for the final GDSII file generation of each tier. In color are the new steps proposed in this article which do not exist in the state-of-the-art 3-D IC flow [8].

tier-partitioning solution from the beginning and neglects the parasitics of 3-D vias during the design closure.

III. DESIGN METHODOLOGY

This section presents our design methodology named C2D flow to build commercial-quality two-tier gate-level F2F and M3D ICs. C2D flow finds the (X, Y) placement solution of a 3-D design using the original geometries of standard cells and interconnects. It also introduces an optimization capability to take intertier 3-D routing overhead into account correctly. Table I explains new terminologies used in this article, and the overall design methodology is shown in Fig. 2. Although we assume that the form-factor of a two-tier 3-D design is 50% of the 2-D design footprint in this section for the ease of explanation, C2D flow is not limited to that assumption but scalable to implement area-optimal gate-level F2F and M3D ICs.

A. Compact-2D Design

A C2D design is a pseudo-3D design in C2D flow to produce the optimal (X, Y) location of cells and macros in a two-tier 3-D design. While the corresponding pseudo-3D

design in S2D flow is implemented with the shrunk layout objects in the target 3-D footprint, the C2D design is made of original layout objects from the target technology node in the 2-D footprint. Then, the resultant placement solution is linearly contracted to the target 3-D footprint. While implementing the C2D design, C2D introduces unique ways to handle memory macros and interconnect *RC* parasitics.

1) Handling Memory Macros: In the conventional 2-D IC design, memory macros do not have any overlap between each other, and none of standard cells is placed inside the memories. However, the C2D design needs to allow overlaps between memories when they are placed at the same (X, Y) location but at different tiers. Moreover, P&R engines need to place standard cells inside memory regions unless the memories fully occupy the same (X, Y) regions in both tiers. These interesting needs are applicable to S2D as well. S2D has proposed shrinking the footprint of a memory macro down to the minimum placement unit and using placement blockages to cover its original boundary. Note that at most two memories may overlap at any particular location in a two-tier 3-D design. Full placement blockages are used in the fully overlapped regions and restrict the standard cell placement. To enable the standard cell placement at partially vacated regions, 50% partial placement blockages are used. The pin locations of a memory macro are retained, which serve as anchors for the standard cell placement regardless of the shrunk memory size.

C2D also uses full/partial blockages to cover memory macro boundaries to implement the pseudo-3D design. However, C2D requires an additional treatment for the size of a blockage at the C2D design. To handle the memory macros in the C2D design correctly, it is important to understand the correlation between placement contraction and the standard cell density. Recall that once the C2D design is implemented, it is linearly contracted to the target 3-D footprint given the lowerleft corner of the core as the contraction origin. As a result, this placement contraction makes the standard cell density at any particular region become twice because the distance between cells is contracted by a factor of $\sqrt{50\%}$ while the original footprint of a cell is retained.

Assuming that only two memories exist in the netlist and they are placed on different tiers in the final two-tier 3-D design, Fig. 3 depicts the impact of placement contraction on the standard cell density in the C2D design. Note that blockage boundaries are contracted in proportion to the core size. Focusing on partial blockage regions, the standard cell density at those regions becomes twice after placement contraction $(D'_{r2} = 2D_{r2} \text{ and } D'_{r3} = 2D_{r3})$. Given that we use 50% partial blockages at those regions in the C2D design, D'_{r2} and D'_{r3} are equal to the target density of the 3-D design (T_d) . It allows cells located at each partial blockage region to move to the vacated region on the tier where a memory does not occupy without any placement density overhead $(D'_{r2} = D_{b2} = T_d$ and $D'_{r3} = D_{t2} = T_d$). This implies that, after placement contraction, the size of the vacated region on a tier should be matched with a part of the original memory macro placed on the other tier.

As a result, the blockage boundaries need to be expanded by a factor of $1/\sqrt{50\%} = 1.414$ in the C2D design as shown



Fig. 3. (a) Assume that two memories (Mem_t, Mem_b) exist in the netlist and they are placed on different tiers in the final two-tier 3-D design. Then, the C2D design is divided into four different regions. T_d is the target placement density in the 3-D design. D_{r1} is the standard cell density on the white space region. D_{r2} and D_{r3} is for each partial blockage region, and D_{r4} for the full blockage region. (b) During placement contraction, the entire blockage floorplan of the C2D design is linearly contracted to the target 3-D footprint by a factor of 50%. This allows that standard cell density at a partial blockage region becomes twice $(D'_{r2} = 2D_{r2}$ and $D'_{r3} = 2D_{r3})$, which is equal to T_d . (c) It allows cells at each partial blockage region on a tier to move to the vacated region on the other tier where a memory does not occupy without any placement density overhead $(D'_{r2} = D_{b2} = T_d)$ and $D'_{r3} = D_{t2} = T_d)$. This implies that, after placement contraction, the size of the vacated region on a tier should be matched with a part of the original memory macro placed on the other tier.

in Fig. 4. The pin locations of a memory macro also should be expanded to correctly anchor the standard cells around the placement blockages, otherwise unwanted routing change occurs in the final 3-D design. To summarize how we handle memory macros in the C2D design, we first prepare for the expanded memory macro library exchange format (LEF) files (memory expansion) that scale the pin locations by a factor of 1.414 while shrinking the footprint down to the minimum placement unit.⁵ Then we assign the tier location of a memory macro and preplace the memories manually considering the intermodule connectivity (memory preplacement). Lastly, we generate placement blockages on the expanded memory region while flattening the tier location of a memory macro (memory flattening). As a result, we obtain the initial C2D design floorplan, in which memory modules are replaced with expanded pins and full/partial placement blockages.

2) Interconnect RC Scaling: The HPWL of a net in the C2D design is $1/\sqrt{50\%} = 1.414 \times$ longer than the corresponding net in the final 3-D design when both are projected on the X-Y plane. To match the electrical length between the C2D and the 3-D design despite the difference in geometrical length, Fig. 5 illustrates the need for interconnect RC scaling in the C2D design. By scaling the unit RC per length by a factor of $\sqrt{50\%} = 0.707$, we avoid the redundant buffer insertions caused by increased geometrical length in the C2D design while still using the original geometries of standard cells and interconnects.⁶ Then, we perform all the required conventional



Fig. 4. (a) Blockage boundary need to be expanded by a factor of $1/\sqrt{N\%}$ in the C2D design because the expanded macro boundary naturally becomes the original macro size $(H \times W)_{macro}$. Also, the pin locations of a memory macro should be expanded to correctly anchor the standard cells around the expanded placement blockage, otherwise unwanted routing change occurs after placement contraction.



Fig. 5. Need for interconnect *RC* scaling in the C2D design. The length of interconnects will be reduced to $0.707 \times$ in the final 3-D layout by placement contraction. In order to reflect this in advance, we scale the unit length *RC* by a factor of 0.707 in the C2D design. The red line in the most left figure indicates an interconnect with reduced parasitics.

implementation steps using commercial 2-D P&R tools to build the C2D design. As a result, we obtain the fully legalized (X, Y) placement solution to deliver at placement contraction.

B. Placement Contraction

Placement contraction linearly maps the placement of the C2D design to the 3-D design footprint to determine the (X, Y) locations of cells and macros in the 3-D design. Assume that the ratio of the target 3-D design footprint to the C2D design footprint is N%. Given the lower-left chip-to-core offset in the C2D design (d_x, d_y) , the center point of a cell/macro (X, Y) in the C2D design is mapped to the center point of the cell/macro (X', Y') in the final 3-D footprint by the equation

$$(X', Y') - (d_x, d_y) = ((X, Y) - (d_x, d_y)) \times \sqrt{N\%}.$$
 (1)

Note that we subtract the chip-to-core offset to calibrate the contraction origin.

Since standard cells retain their footprints during contraction, it is obvious that they have overlaps on the final 3-D design footprint. However, they will be legalized and snapped to the placement rows at each tier after their Z locations are decided. It is worth noting that the placement contraction

⁵If the target 3-D design footprint is N% of the 2-D footprint, the scaling factor for placement blockage and pin location becomes $1/\sqrt{N\%}$.

⁶If the target 3-D design footprint is N% of the 2-D footprint, we scale the unit length *RC* by a factor of $\sqrt{N\%}$.



Fig. 6. Our C2D flow demonstrated with OpenSparc T2 [16] single core design: memory expansion and preplacement, memory flattening, C2D design, and placement contraction. Tier partitioning and compact 3-D via planning follow next.

applied to a memory macro is distinct from that applied to a standard cell in that the center point of the macro is manually defined at memory preplacement step while the center point of the standard cell is determined once all the P&R steps are performed in the C2D design. Recall that an expanded macro boundary in the C2D design naturally becomes the original macro size on the 3-D footprint after placement contraction. This allows us to instantiate a memory macro at the location derived from (1) without creating any large halo of dead spaces. Fig. 6 demonstrates C2D flow from memory preplacement to placement contraction using OpenSparc T2 [16] single core design. It is important to note that the placement solution based on the shrinking idea from S2D and that based on interconnect RC scaling/placement contraction ideas from our C2D are ideally the same. However, C2D necessitates the P&R engines that handle the target technology node only while S2D relies on the CAD engines for the next technology node.

C. Tier Partitioning

Since the tier location of a memory macro is preassigned manually, standard cells within a memory boundary move to the tier where the memory does not occupy. To determine the Z location of each standard cell outside memory boundaries, C2D adopts bin-based FM min-cut tier partitioning algorithm [8], [10], [11]. Given that (X, Y) placement solution is confined to the final 3-D footprint after placement contraction, we define partitioning bins that dissect the entire 3-D footprint in a regular fashion. A partitioning bin is assigned to the cells located under the region of it, and we apply two-way FM min-cut partitioning algorithm with the area skew constraint enforced at individual local partitioning bin. For the initial solution, the partition or the tier of a cell is randomly chosen while balancing the area of two partitions within a local bin. Then, a cell move across the partitions is found legal when it meets the area skew defined at a local partitioning bin.

It is important to note that one cutsize objective is imposed over the entire circuit, while the area balance is enforced locally in each bin. Several passes are run until the cutsize is not decreased anymore, and the tier location of a cell is determined based on the partitioning result. The number of cutsize, which turns into the minimum number of intertier connections, is controlled by the size of a partitioning bin. Larger bin size or less stringent local area skew constraint allows the tier partitioning algorithm to find a lower cut size solution leading to less routing congestion between the two tiers. However, this also leads to larger local area skew or displacement from the original (X, Y) solution determined by placement contraction. Therefore, a sweet spot exists along the partitioning bin size. Once tier partitioning determines the Z location of each cell, a design exchange format (DEF) file for each tier is created, and the placement engine legalizes the overlaps between standard cells caused by placement contraction at each tier (tier-by-tier legalization).

D. Compact 3-D via Planning

Once we decide the (X, Y, Z) placement solution, the optimal (X, Y) location of a 3-D via needs to be determined based on the given 3-D placement. This is called 3-D via planning. In this step, intertier 3-D routing overhead, which has not been accounted by the C2D design, starts to affect the timing closure. S2D is not only susceptible to this degradation but none of 3-D-routing-aware optimization is introduced after tier partitioning. In order to support post-tier-partitioning optimization (post-TP opt) to compensate the intertier 3-D routing overhead, C2D presents a unique stage named compact 3-D via planning. Compact 3-D via planning consists of two steps, and the following sections describe them in detail.

1) Compact Placement: After tier partitioning and tierby-tier legalization, cells have been separated into two tiers and snapped to the placement rows at each tier. To support power, timing, and 3-D via location co-optimization by applying a commercial 2-D P&R tool that only works on a single tier, post-TP opt needs to compress the two-tier placement information into the single tier, while still keeping the tier information of a cell. To address the needs, C2D creates a synthetic placement called compact placement.

Given the two-tier placement information, our in-house program creates a DEF file that flattens the two-tier placements. However, cells are instantiated based on their tier locations in the DEF file, so that post-TP opt keeps the tier information of a cell. In order to do this, 3-D technology LEF is required, which includes all the metal stacks from both tiers. Next, 3-D macro LEF is required to instantiate a cell based on its tier location as depicted in Fig. 7. This allows commercial 2-D



Fig. 7. 3-D technology and macro LEF files prepared for compact 3-D via planning. 3-D technology LEF defines the 3-D metal stack and its design rules. 3-D macro LEF defines the cells based on their tier locations. It is worth to note that an MIV blockage is defined in M3D macro LEF to prevent MIVs from penetrating the top tier cells.



Fig. 8. (a) S2D flow [1] does not support post-TP opt because of the placement overlaps. (b) Placement row splitting in our C2D flow enables the optimization by fully legalizing the placement overlaps.

P&R engines to distinguish the pin layers of a top cell from those of a bottom cell even though both cells are placed at the same (X, Y) location. It is noteworthy that F2F and M3D ICs have different 3-D via definition and the metal stack configuration for the top tier. To build M3D ICs, an MIV must not penetrate the device region of top tier cells. Therefore, 3-D macro LEF for M3D ICs contains the blockage (OBS) at the MIV layer covering the entire macro boundary of a top tier cell.

In the compact placement, 3-D footprint is too small to accommodate all the cells from both tiers. Nevertheless, we should not have placement overlaps because the commercial optimization engine requires the fully legalized placement solution. To find the legal (X, Y) placement solution on the final 3-D IC footprint, we split a placement row into the top and bottom rows, and change the height of standard cells in 3-D macro LEF to the half of the original to fit into the split rows. In Fig. 8, Row0 and Row1 are two adjacent placement rows to be split. Row1 is vertically flipped over to share the power rail with Row0. Now, placement row splitting turns each row into two horizontally split rows. In Row0, the bottom half is reserved for the bottom tier placement, and the top half for the top tier. However, in Row1, the bottom half is reserved for the top tier placement, and the top half for the bottom tier due to the flipped orientation of Row1. As a result, the placement

overlap is fully legalized while accommodating every cell in the design on the final 3-D footprint. It is worth noting that the pin locations of a standard cell are still preserved regardless of splitting placement rows. Therefore, pins may be found outside the boundary of the top/bottom half of a cell in the compact placement. The bottom-tier cell pins are assigned to metal layers MXB and the top-tier cell pins are assigned to MXT (see Fig. 7).

Placement row splitting allows us to perform legalization between the top-tier and bottom-tier standard cells, but memory macros are large and occupy a number of placement rows. To handle memory macros in the compact placement, we transform the physical size of a memory macro into the minimum placement unit, and we create a full placement blockage at its original boundary. The (X, Y) memory pin locations are retained despite the change of memory size, while memory pins are assigned to metal layers based on the tier of the memory. Then, we keep the standard cells inside memory regions fixed at their given (X, Y) locations to avoid legalization by full placement blockages. Note that full placement blockages limit the additional placement other than cells that already exist, and we preserve the placement inside memory regions intact.

The artificial natures of the compact placement are worthy of note because the compact placement is a special representation of cells from two tiers using a single chip. The standard cells are compressed to the half-height to be legally placed in the small 3-D footprint. Since the original (X, Y) location of pins in a cell is still preserved, pins are partially outside the boundary of the half-height cell, which is a significant deviation from the classical macro definition. The tier information of a cell is turned into the metal layer of pins by 3-D macro LEF, so that commercial P&R engines can fully access to the pins of either top or bottom cells while only a single placement layer exists. This synthetic placement result is then fed into a commercial 2-D IC tool to perform post-TP opt.

2) Post-Tier-Partitioning Optimization: C2D performs post-TP opt for timing, power, and 3-D via location cooptimization to close the design under intertier 3-D routing overhead. On top of the compact placement, we conduct detailed routing with 3-D technology LEF and utilize the post-route optimization capability of a commercial 2-D optimization engine for post-TP opt. Post-TP opt requires RC corners for the full 3-D metal stack (3-D interconnect technology file) and the timing corners for both top and bottom tiers (3-D liberty file). Full optimization capabilities, including resizing, moving, insertion, and deletion, are employed for both top and bottom cells to address the intertier 3-D routing overhead. Recall that we have compressed the cell height into the half but preserved the width of standard cells in the compact placement. Therefore, placement is kept legal although there are updates while post-TP opt proceeds. To be specific, when a cell gets resized, the width of the cell is updated, and this affects the location of neighboring cells by legalization. When a cell is inserted, the P&R engine finds the best macro instance (either a top or a bottom instance) based on the detailed routing as well as the (X, Y) location of the cell for design closure. The macro instance used turns into the tier

location of the cell naturally. Once the optimization is done, we save a DEF file to deliver at incremental routing.

E. Incremental Routing

Incremental routing is a CAD solution to preserve the routing result from post-TP opt for the final GDSII file generation of each tier. Our in-house tool parses the DEF file delivered from post-TP opt and constructs a graph for each net to tract the routing structure. Each graph consists of vertices and edges representing individual routing objects and their connectivity, respectively. Routing objects include wires, vias, I/O ports, and cell pins. We store the tier location (based on 3-D technology LEF) and size of the routing objects. The (X, Y) locations where those routing objects cross each other are kept along with their edge definitions.

For a 2-D graph that contains all vertices located at a single tier, we simply deliver the graph information when creating a DEF and a Verilog file for the tier. However, for a 3-D graph that contains a 3-D via as one of its vertices, care must be taken since we need to divide the graph into the set of 2-D subgraphs located at two different tiers. We convert a 3-D via into two 3-D I/O ports for the top and the bottom tier each. When there exist a driver cell, external input ports, or other 3-D input ports in its connected subgraph, the direction of a 3-D I/O port becomes the output, otherwise the input. As a result, the 3-D graph turns into a group of 2-D subgraphs, and each subgraph information is delivered when creating a DEF and a Verilog file for each tier. Throughout this process, we do not reduce the number of 3-D vias, and 2-D subnets are assigned to the original 3-D net at the top Verilog file to be identified as the same net.

Based on the collected graph information, our in-house tool creates a DEF file and a Verilog file for each tier. The DEF file introduces the 3-D vias as external I/O ports at their (X, Y)locations (located in the middle of the layout). Note that the (X, Y) location of a 3-D via becomes the implicit location of a 3-D connection. Also the DEF file contains the final cell locations after restoring the cell height to the original. Since the pin locations of a cell are preserved regardless of halving the cell height, we can easily retrieve the correct (X, Y) locations of a cell based on the original cell height. Also, we reproduce the routing structure for each graph based on the actual connection points defined in each edge. As a result, I/O port locations, cell placement, routing information of each graph are stored in the DEF file for a corresponding tier. For the Verilog file for each tier, the connectivity among 3-D I/O ports and the cells within the same tier are provided. Lastly, a top Verilog and standard parasitic exchange format (SPEF) files are created, which define the connections between 3-D I/O ports in separate tiers and RC parasitics of 3-D vias, respectively.

For final GDSII file generation, we load the DEF and the Verilog file at each tier and perform sign-off physical DRV fixing. Since the routing information of a net is retained in the DEF file, C2D minimize the perturbation between post-TP opt and final tier-by-tier GDSII layout. The reason why this additional DRV fixing is necessary is that tools built for 2-D ICs do not support full DRV fixing for the pins outside the



Fig. 9. Final 2-D and two-tier gate-level F2F and M3D IC GDSII layouts of OpenSpare T2 [16] single core implemented using our C2D flow. The layout of bottom die of the F2F IC is mirrored for the flipped F2F bonding.

macro boundary while employing compact placement. When the sign-off DRV fixing is done, *RC* parasitics of each tier are extracted, and we proceed the final 3-D timing & power analysis.

IV. STATE-OF-THE-ART COMPARISON

In Table II, we compare the timing & power savings of C2D with those of S2D based on the OpenSparc T2 [16] single core (SPC) design at 1.0-GHz clock frequency. We find the maximum frequency in that the worst negative slack (WNS) is less than 5% of the clock period. This is to ensure the performance is tightly optimized. 3-D ICs are built at the maximum frequency of the 2-D IC for iso-performance comparison. We use dual-Vt cell libraries in 28-nm commercial-grade technology process design kit (28-nm PDK). Six metal layers are used for 2-D, and both top and bottom tiers for 3-D implementations (Fig. 9). MIV diameter, pitch, resistance, and capacitance are assumed to be 0.1 μ m, 0.2 μ m, 16 Ω , and 0.1 fF, respectively. The F2F via diameter, pitch, resistance, and capacitance are assumed to be 0.5 μ m, 1.0 μ m, 0.5 Ω , and 0.2 fF, respectively. For the static power analysis, we set the switching activity as 0.1 for primary input ports and register output pins, and 2.0 for a clock port.

We observe that both C2D and S2D designs significantly decrease the net switching power thanks to the huge wirelength saving in 3-D designs. Following buffer reduction contributes to the combinational cell and clock network power savings. The maximum total power reduction of C2D is 12.56% while S2D offers 12.16% saving over the 2-D IC at iso-performance. In addition, it is remarkable that C2D manipulates more 3-D interconnections than S2D and reduces the total negative slack (TNS) violations by 74.97% while S2D worsens the timing. This is because C2D supports post-TP opt to compensate intertier 3-D routing overhead. This result not only shows that C2D offers comparable power reduction as the state-of-the-art S2D but also proves that C2D builds timing-robust 3-D ICs. Most of all, C2D is more scalable than S2D in that our C2D flow performs with P&R engines, technology files, and design rules for the target technology and does not require handling of the next smaller node.

TABLE II

ISO-PERFORMANCE TIMING AND POWER COMPARISON AMONG 2-D, S2D [8], AND C2D USING OPENSPARC T2 [16] SINGLE CORE (28 nm). Δ% SHOWS % IMPROVEMENT OVER 2-D. TARGET CLOCK FREQUENCY IS 1.0 GHz. WNS STANDS FOR THE WNS, TNS FOR THE TNS, AND TPS FOR THE TOTAL POSITIVE SLACK. IN THIS ARTICLE, WE CLAIM THAT A DESIGN MEETS THE TIMING WHEN THE WNS IS LESS THAN 5% OF THE CLOCK PERIOD TO ENSURE THAT THE PERFORMANCE IS TIGHTLY OPTIMIZED. TWO UTILIZATION NUMBERS OF 3-D DESIGNS ARE FOR THE TOP AND THE BOTTOM TIER, RESPECTIVELY, AND THE UTILIZATION SAVING OF 3-D ICS ARE CALCULATED BASED ON THE AVERAGE UTILIZATION OF BOTH TIERS. C2D OFFERS COMPARABLE POWER REDUCTION AND SIGNIFICANT SLACK IMPROVEMENT COMPARED WITH S2D

	2D		F2F	ICs		M3D ICs			
	20	S2D	$ \Delta\% $	C2D	$\Delta\%$	S2D	$\Delta\%$	C2D	$ \Delta\% $
Footprint (mm^2)	2.53	1.26	50.02%	1.26	50.02%	1.26	50.02%	1.26	50.02%
Place. Util. (%)	73.59	70.03 / 71.84	3.64%	70.93 / 70.23	4.10%	70.00 / 71.84	3.64%	71.15 / 70.10	4.05%
3D Via #	-	157,415	-	193,224	-	195,973	-	261,032	-
Total WL (m)	15.69	11.59	26.10%	11.41	27.25%	11.80	24.78%	11.81	24.68%
Total Pwr (mW)	334.90	295.77	11.68%	293.84	12.26%	294.17	12.16%	292.84	12.56%
Net Pwr (mW)	183.81	151.04	17.83%	148.54	19.19%	149.37	18.74%	147.03	20.01%
Cell Pwr (mW)	80.75	78.14	3.23%	77.98	3.43%	78.22	3.13%	78.12	3.25%
Leak. Pwr (mW)	70.34	66.58	5.34%	67.32	4.29%	66.58	5.34%	67.69	3.76%
WNS (ps)	-24.47	-45.68	-86.68%	-34.53	-41.11%	-58.29	-138.21%	-29.96	-22.44%
TNS (ps)	-1721.23	-1656.10	3.78%	-480.16	72.10%	-2050.68	-19.14%	-430.85	74.97%
TPS (ns)	36077.60	38061.70	5.50%	39538.00	9.59%	37859.10	4.94%	38721.70	7.33%

V. EXPERIMENTAL RESULTS

In this section, we analyze the impact of each design step in C2D flow with LDPC, AES-128, and JPEG from OpenCore benchmark suites [17]. Assumptions on the technology and analysis are the same as we made in Section IV. The initial utilization density for both AES-128 and JPEG is 60%, while 40% for wire-dominated LDPC. The maximum clock frequency for each benchmark is decided when the design closes the timing with WNS less than 5% of the clock period, and it is 2.0-GHz for LDPC, 5.4-GHz for AES-128, and 2.16-GHz for JPEG. Fig. 10 shows the GDSII layouts of 28 nm 2-D and C2D-based two-tier gate-level F2F and M3D IC implementations for each benchmark.

A. Impact of Interconnect RC Scaling

In the C2D design, we scale interconnect *RC* parasitics by a factor of $\sqrt{50\%} = 0.707$ to imitate the parasitics of the final 3-D design based on the assumption that the 3-D design footprint is 50% of the 2-D design footprint. However, the *RC* scaling factor can be generalized and set to be $\sqrt{40\%} = 0.632$ in case the 3-D design footprint is 40% of the 2-D design footprint. Table III shows C2D design results with various 3-D/2-D footprint ratios.

With a low *RC* scaling factor, such as under 0.6, all benchmarks have huge power and standard cell area savings because of the reduced interconnect parasitics and the less number and lower drive-strength of buffers. However, since the target footprint saving is much larger than the standard cell area saving, it results in the impractical placement utilization per each tier in the 3-D design. Assuming placement utilization in [70%, 80%] range is allowed, 3-D footprint savings reach up to 65% for LDPC and 56% for both AES-128 and JPEG. In case of wiredominated LDPC design, since the 2-D footprint is determined by the routability, huge wirelength reduction in the 3-D design helps increase the footprint saving more than other circuits.

When the same placement utilization in both 2-D and 3-D ICs should be considered, we observe that $53\% \sim 57\%$ 3-D footprint savings are good target for all designs. With exactly

TABLE III IMPACT OF INTERCONNECT *RC* SCALING BASED ON THE TARGET 3-D FOOTPRINT. ASSUMING PLACEMENT UTILIZATION IN [70%, 80%] RANGE IS ALLOWED, 3-D FOOTPRINT SAVINGS REACH UP TO 65% FOR LDPC AND 56% FOR BOTH AES-128 AND JPEG

Footprint $(3D/2D)$	50%	45%	40%	35%					
Silicon Area $(3D/2D)$	100%	90%	80%	70%					
RC Scaling	0.707	0.671	0.632	0.592					
LDPC									
Std. Cell Area (mm^2)	0.180	0.178	0.177	0.172					
Avg. Place. Util. / Die	58.31%	63.92%	72.03%	79.69%					
Place. Util. $(3D/2D)$	87.83%	96.30%	108.50%	120.04%					
Total Power (mW)	179.23	174.48	167.70	158.03					
Footprint $(3D/2D)$	50%	47%	44%	41%					
Silicon Area $(3D/2D)$	100%	94%	88%	82%					
RC Scaling	0.707	0.686	0.663	0.640					
	AES-	128							
Std. Cell Area (mm^2)	0.359	0.356	0.355	0.355					
Avg. Place. Util. / Die	70.10%	73.88%	78.99%	84.58%					
Place. Util. $(3D/2D)$	95.09%	100.22%	107.15%	116.15%					
Total Power (mW)	331.68	330.49	324.54	323.39					
	JPE	G							
Std. Cell Area (mm^2)	0.943	0.941	0.939	0.936					
Avg. Place. Util. / Die	70.71%	70.71%	80.07%	85.65%					
Place. Util. $(3D/2D)$	96.03%	101.78%	108.73%	116.32%					
Total Power (mW)	579.17	573.52	565.84	563.80					

50% footprint saving constraint, we find $4\% \sim 12\%$ placement utilization savings in 3-D designs. This shows that sweeping the interconnect *RC* scaling factor helps to set the practical and optimal target footprint of a 3-D design. For the rest of experiments, we keep the 50% 3-D footprint saving constraint for all benchmarks to factorize the impact of other steps clearly.

B. Impact of Tier Partitioning

While placement contraction is deterministic in that the (X, Y) location of a cell is scaled by 0.707, bin-based tier partitioning is heuristic with regard to the size of partitioning bins. Depending on the partitioning bin size, the number of cells applied to the local area balance constraint varies, resulting in different cutsizes between the tiers. Table IV shows



Fig. 10. 28-nm GDSII die images of 2-D and two-tier gate-level F2F and M3D IC implementations using our C2D flow. (a) LDPC (2.0 GHz). (b) AES-128 (5.4 GHz). (c) JPEG (2.16 GHz).

TABLE IV Depending on the Partitioning Bin Size, the Number of Cells Applied to the Local Area Balance Constraint Varies, Resulting in Different Cutsizes Between the Tiers. The Large Bin Size Allows the Algorithm to Find the Minimal Cutsize

Bin Size (µm)	5	10	20	40	80				
LDPC									
Bin #	6,169	1,542	386	96	24				
Avg. Cell # / Bin	11	42	169	677	2,707				
Cutsize #	42,655	16,451	11,445	10,916	10,571				
AES-128									
Bin #	10,247	2,562	640	160	40				
Avg. Cell # / Bin	14	55	219	877	3,507				
Cutsize #	83,869	39,414	31,971	15,150	7,698				
JPEG									
Bin #	26,680	6,670	1,668	417	104				
Avg. Cell # / Bin	11	43	171	682	2,729				
Cutsize #	196,517	80,676	58,967	42,500	32,341				

how the different partitioning bin sizes change the number of cutsize between two tiers under 5% area skew.

1) Displacement Analysis: After tier partitioning, we create a DEF file for each tier, and the placement engine legalizes the overlaps caused by placement contraction at each tier. Therefore, tier-by-tier legalization causes displacement from the optimal (X, Y) placement solution offered by placement contraction. To minimize this unwanted displacement, two cells with an overlap need to be separated into the different tiers as much as possible. Fig. 11 shows that when we decrease the partitioning bin size to enforce the local area balance, the displacement caused by tier-by-tier legalization is minimized. Total displacement per cell is the averaged Euclidean distance between the (X, Y) location determined by placement contraction and the (X, Y) location after tier-by-tier legalization.

As the partitioning bin size increases, we observe the degree of displacement grows monotonically. In case of gate-dominant circuits (AES-128 and JPEG) where the local connectivity is primal, most of cells are placed close to each other at the C2D design. Therefore, the number of overlaps is significantly large when the placement is contracted into the



Fig. 11. Impact of partitioning bin size on the displacement caused by tier-by-tier legalization. Total displacement per cell is the averaged Euclidean distance between the (X, Y) location determined by placement contraction and the (X, Y) location after tier-by-tier legalization. As the partitioning bin size increases, we observe the degree of displacement grows monotonically.

final 3-D footprint. The overlaps are effectively removed under small partitioning bin size (<10 μ m) at the similar level of the displacement in wire-dominant circuit (LDPC). However, large partitioning bin size (>40 μ m) leads to the imbalance in local area skew, resulting in huge displacement (For AES-128, 3.55 μ m displacement per cell at 80 μ m partitioning bin size) after tier-by-tier legalization.

Note that the displacement value depends on the target process node because of the different size of cells. To understand the displacement without process node dependency, we analyze the displacement in the number of placement rows. X and Y displacement per cell means the independent movement on X and Y axes, respectively. With small partitioning bin size (5 μ m), both X and Y displacement is suppressed under 0.25 placement rows, which proves the effectiveness of local area balancing in bin-based tier partitioning. In the worst case, (AES-128 at 80 μ m partitioning bin size), we observe that Y displacement is as large as 0.90 rows, and 1.66 rows for X displacement. We find that the ratio of X to Y displacement varies from 1.42 to 2.15, and the average ratio is 1.74 out of all the tier partitioning results.

The displacement caused by tier-by-tier legalization should be minimized to preserve the optimized placement solution from placement contraction. We observe that using small partitioning bin size in bin-based tier partitioning helps reduce the displacement overhead. However, small partitioning bin size

Bin Size (μm)	5	10	20	40	80	5	10	20	40	80
LDPC			F2F IC	M3D IC						
3D Via #	55,468	26,999	20,850	19,802	19,726	66,642	38,336	32,204	31,421	31,173
Avg. WL / net (μm)	39.16	38.85	38.83	38.84	38.82	39.27	39.04	39.10	39.13	39.17
3D Net # (%)	61.41	24.71	17.73	16.89	16.75	62.01	26.73	20.12	19.57	19.30
3D Net WL Savings (%)	26.73	27.58	27.87	27.87	27.95	26.61	27.38	27.37	27.30	27.20
2D Net WL Savings (%)	26.60	26.93	26.80	26.79	26.77	26.43	26.51	26.37	26.34	26.32
Total WL Savings (%)	26.70	27.28	27.32	27.30	27.33	26.57	27.00	26.89	26.84	26.77
AES-128			F2F IC				Ν	A3D IC		
3D Via #	104,306	61,902	51,460	22,311	10,824	120,919	78,626	67,767	43,109	31,434
Avg. WL / net (μm)	16.45	16.24	16.56	18.16	18.83	17.40	17.17	17.58	19.54	20.27
3D Net # (%)	59.67	28.11	22.91	11.14	5.96	60.28	30.18	25.14	15.25	10.06
3D Net WL Savings (%)	20.57	22.10	21.50	18.45	16.73	15.16	15.22	12.55	-3.04	-10.84
2D Net WL Savings (%)	22.74	22.20	19.95	11.46	8.76	21.38	20.80	18.71	10.34	6.48
Total WL Savings (%)	21.14	22.15	20.60	12.94	9.71	16.74	17.87	15.91	6.54	3.03
JPEG			F2F IC				Ν	A3D IC		
3D Via #	247,236	153,184	127,265	106,405	86,536	240,301	120,921	94,868	71,353	53,810
Avg. WL / net (μm)	14.54	14.57	14.76	15.06	15.67	15.48	15.45	15.84	16.38	16.97
3D Net # (%)	61.36	25.19	18.42	13.27	10.10	53.42	26.96	20.73	16.70	13.34
3D Net WL Savings (%)	20.69	21.73	21.61	21.39	19.11	14.20	14.36	11.04	5.77	1.81
2D Net WL Savings (%)	19.76	19.01	17.66	15.53	12.17	17.35	16.57	15.24	13.57	10.17
Total WL Savings (%)	20.47	20.31	19.29	17.60	14.28	15.30	15.43	13.32	10.37	7.09

TABLE V

IMPACT OF PARTITIONING BIN SIZE ON WIRELENGTH SAVING. SAVING VALUES ARE BASED ON THE COMPARISON OVER THE C2D RESULT

causes large cutsize, which turns into more congestion on the intertier 3-D routing. As a result, small partitioning bin size reduces the wirelength saving in 3-D ICs. In the next section, we discuss the wirelength saving tradeoff of tier partitioning in detail.

2) Wirelength Analysis: After tier-by-tier legalization, compact placement flattens the two-tier placement solution into a single chip and allows us to perform detailed intertier 3-D routing. In Table V, we analyze the detailed 3-D routing result to explore the impact of tier partitioning on the wirelength saving. A net is defined as either a 2-D or a 3-D net based on their 3-D via usage, and we compare its wirelength with that in the C2D design. LDPC shows the best wirelength savings (27.3%) when the bin size is in the range of 20 μ m to 80 μ m. On the other hand, AES-128 and JPEG have 22.15% and 20.31% wirelength savings, respectively, at 10- μ m bin. It is noteworthy that gate-dominant circuits steeply lose the wirelength savings along with increasing the bin size over the sweet spot. M3D AES-128 shows even negative wirelength saving on 3-D nets starting from 40- μ m partitioning bin. Considering the best wirelength savings for both F2F and M3D ICs, we determine the size of partitioning bins as $40-\mu m$ for LDPC, 10- μ m for both AES-128 and JPEG for the following analysis. This shows that finding the optimal partitioning bin size is critical to maximize the wirelength savings of 3-D ICs.

We observe that the average wirelength per net based on the detailed intertier 3-D routing (Avg. WL/net in Table V) is strongly correlated to the optimal partitioning bin size. If the bin size is much smaller than Avg. WL/net, most of nets become 3-D, causing congestion and detouring to meet the design rules for 3-D vias. This is the reason that the wirelength saving of 3-D nets decreases at $5-\mu m$ bin, lowering the total wirelength saving. On the other hand, if the bin size is too large, then most of nets remain at 2-D with increased wirelength by huge displacement from tier-by-tier legalization. Therefore, embracing too many 2-D nets again degrades the total wirelength saving.

One way to estimate the Avg. WL/net before we actually perform detailed 3-D routing is to utilize the C2D design result. Recall that the HPWL of a net in the C2D design is $1/\sqrt{N\%}$ times longer than that in the final 3-D design where the target 3-D design footprint is N% of the 2-D design footprint. Therefore, we can calculate the Avg. WL/net by multiplying $\sqrt{N\%}$ to the average net wirelength calculated from the C2D design. For example, the average net wirelength based on the LDPC C2D design is 53.48 μ m and this gives us 37.82- μ m for estimated Avg. WL/net which is close to both the actual Avg. WL/net (38.84 μ m) and the optimal partitioning bin size. Note that the wirelength overhead caused by intertier 3-D routing and tier-by-tier legalization are not addressed in this estimation. This makes the estimated Avg. WL/net a little smaller than the actual value. However, it gives us a good reference point for bin size optimization.

C. Impact of Intertier 3-D Routing Overhead

1) Routed Versus Predicted Wirelength: In the C2D design, the interconnect *RC* parasitics are scaled by a factor of 0.707 given that every net in the C2D design takes 29.3% HPWL saving by placement contraction. This is true when we only consider the HPWL based on the flattened two-tier placement information (compact placement). However, after we perform detailed intertier 3-D routing on top of the compact placement, the actual routed length of a net can be different from what was predicted in the C2D design. Fig. 12 shows the net count distribution based on the ratio of the routed wirelength at the compact placement for F2F ICs to the predicted wirelength at the C2D design. Percentage numbers indicate the proportion



Fig. 12. Net count distribution based on the ratio of the routed wirelength at the compact placement for F2F ICs to the predicted wirelength at the C2D design. In the C2D design, we predicted that the wirelength of a net becomes 29.3% shorter in the compact placement. However, the actual routing at the compact placement shows the distribution centered at the predicted wirelength, justifying the necessity of post-TP opt.

out of the total net counts in the specified range of wirelength ratio.

We find that the shape of distribution depends on the circuit characteristics. For the wire-dominant circuit (LDPC), 38.3% of nets have routed with the predicted wirelength, and the distribution shows the steep bell-shaped curve, which implies that intertier 3-D routing overhead does not affect the timing closure made at C2D design too seriously. However, lower than 30% of nets are routed with the predicted wirelength in gate-dominant circuits (26.9% for AES-128, and 19.8% for JPEG), and the deviation of distribution is larger than that of LDPC. When averaging the ratio between routed and predicted wirelength out of the entire nets, we observe that only 2.8% mismatch from the predicted wirelength for LDPC, while 10.1% and 12.7% for AES-128 and JPEG, respectively. Note that the difference between routed and predicted wirelength causes the net delay difference, and this intertier 3-D routing overhead will be addressed by post-TP opt after all. When the timing path slack becomes positive due to the positive wirelength mismatch (Routed WL/Predicted WL < 1.0), post-TP opt has a margin to improve the power consumption.

2) F2F Versus M3D: In Table V, we found that the 3-D nets in F2F ICs have more wirelength savings than those of M3D ICs. Ideally thinking, F2F ICs go through both top and bottom BEOLs for the intertier 3-D routing, while M3D ICs need only bottom BEOL. Also, F2F via pitch is assumed to be as $5 \times$ large as MIV pitch in this article, which is supposed to cause additional routing detour in F2F ICs. To analyze this observation in detail, we dissect the wirelength of 3-D nets with running metal layers and tabulate the wirelength distribution per each layer in Table VI.

For the intertier 3-D routing in M3D ICs, MIVs must not penetrate the top tier cells. Since our tier partitioning is areabalanced locally, generally 50% of cells should be placed on the top tier (under 5% area skew in this article), resulting in significant reduction in the available area for MIV insertion as shown in Fig. 13. Because of this unique constraint in M3D ICs, 3D routing turns out to be congested and detoured.

TABLE VI

DIFFERENT INTERTIER 3-D ROUTING OVERHEADS IN F2F AND M3D ICS. THE WIRELENGTH DISSECTION OF 3-D NETS IS PRESENTED BASED ON RUNNING METAL LAYERS. WIRELENGTH UNIT IS mm. IN M3D ICS, MIVS MUST NOT PENETRATE THE TOP TIER CELLS. BECAUSE OF THIS UNIQUE CONSTRAINT, 3-D ROUTING TURNS OUT TO BE CONGESTED AND DETOURED IN M3D ICS. ON THE OTHER HAND, INTERTIER 3-D ROUTING IN F2F ICS MUST GO THROUGH BOTH TOP AND BOTTOM BEOLS, WHICH LENGTHENS THE WIRELENGTH

F2F	LDPC		AES	5-128	JP	JPEG		
3D Net WL	Тор	Bottom	Тор	Bottom	Тор	Bottom		
M1	2.54	1.22	1.30	0.91	3.89	1.42		
M2	93.80	100.48	159.29	169.26	329.65	338.12		
M3	135.53	129.28	213.67	213.68	445.02	444.85		
M4	139.77	135.99	119.11	113.28	223.84	214.53		
M5	167.11	143.08	58.59	59.01	71.77	92.02		
M6	59.18	108.46	8.26	14.26	7.40	17.87		
Die WL	597.93	618.51	560.22	570.41	1081.56	1108.81		
Total WL	1216.44		113	0.62	2190.38			
M3D	LD	PC	AES	5-128	JP	EG		
M3D 3D Net WL	LD Top	PC Bottom	AES Top	-128 Bottom	JP Top	EG Bottom		
M3D 3D Net WL M1	LD Top 23.20	DPC Bottom 1.60	AES Top 29.34	-128 Bottom 0.94	JP Top 63.14	EG Bottom 1.51		
M3D 3D Net WL M1 M2	LD Top 23.20 155.50	DPC Bottom 1.60 101.95	AES Top 29.34 248.98	-128 Bottom 0.94 183.51	JP Top 63.14 551.79	EG Bottom 1.51 356.98		
M3D 3D Net WL M1 M2 M3	LE Top 23.20 155.50 162.17	DPC Bottom 1.60 101.95 151.00	AES Top 29.34 248.98 183.54	-128 Bottom 0.94 183.51 228.95	JP Top 63.14 551.79 343.17	EG Bottom 1.51 356.98 460.48		
M3D 3D Net WL M1 M2 M3 M4	LE Top 23.20 155.50 162.17 138.22	DPC Bottom 1.60 101.95 151.00 147.23	AES Top 29.34 248.98 183.54 34.02	-128 Bottom 0.94 183.51 228.95 161.40	JP Top 63.14 551.79 343.17 12.71	EG Bottom 1.51 356.98 460.48 324.64		
M3D 3D Net WL M1 M2 M3 M4 M5	LD Top 23.20 155.50 162.17 138.22 121.83	PPC Bottom 1.60 101.95 151.00 147.23 189.14	AES Top 29.34 248.98 183.54 34.02 0.54	-128 Bottom 0.94 183.51 228.95 161.40 124.56	JP Top 63.14 551.79 343.17 12.71 0.12	EG Bottom 1.51 356.98 460.48 324.64 262.04		
M3D 3D Net WL M1 M2 M3 M4 M5 M6	LD Top 23.20 155.50 162.17 138.22 121.83 47.95	PC Bottom 1.60 101.95 151.00 147.23 189.14 160.41	AES Top 29.34 248.98 183.54 34.02 0.54 0.53	-128 Bottom 0.94 183.51 228.95 161.40 124.56 107.89	JP) Top 63.14 551.79 343.17 12.71 0.12 1.14	EG Bottom 1.51 356.98 460.48 324.64 262.04 212.06		
M3D 3D Net WL M1 M2 M3 M4 M5 M6 Die WL	LD Top 23.20 155.50 162.17 138.22 121.83 47.95 648.88	PPC Bottom 1.60 101.95 151.00 147.23 189.14 160.41 751.34	AES Top 29.34 248.98 183.54 34.02 0.54 0.53 496.95	5-128 Bottom 0.94 183.51 228.95 161.40 124.56 107.89 807.24	JP) Top 63.14 551.79 343.17 12.71 0.12 1.14 972.08	EG Bottom 1.51 356.98 460.48 324.64 262.04 212.06 1617.71		



Fig. 13. Compact 3-D via planning snapshots for (a) F2F and (b) M3D ICs. A top tier cell in M3D ICs contains an MIV blockage to prevent MIVs from penetrating itself.

To find the optimal white space to insert an MIV, M3D ICs use a maximum $11.8 \times$ more wirelength on the M6 layer of the bottom tier for intertier 3-D routing than F2F ICs. LDPC is a wire-dominant circuit so it has lower placement utilization than other gate-dominant benchmarks. Therefore, LDPC allows more white spaces to insert MIVs while having only $1.5 \times$ M6 usage for detouring. On the other hand, the intertier 3-D routing overhead in F2F ICs is captured at M5 and M6 layers usage on the top tier. While M3D ICs have very little usage on the M5 and M6 layers on the top tier, F2F ICs use a maximum 576.5 \times more wirelength on the top tier M5 layer for intertier 3-D routing. However, when comparing the difference in the total wirelength of F2F and M3D ICs, we observe that detouring for MIV insertion has worse impact on the 3-D net wirelength saving than lengthened intertier 3D routing in F2F ICs.

Fig. 14 compares M3D ICs with F2F ICs on the mismatch between the wirelength in detailed intertier 3-D routing and



Fig. 14. Comparison between M3D IC and F2F IC on the mismatch between the wirelength in detailed intertier 3-D routing and the predicted wirelength in the C2D design. We observe that the overall distribution of M3D ICs is pushed to the right, implying that the mismatch from the predicted wirelength becomes worse due to the unique 3-D routing constraint in M3D ICs.

the predicted wirelength from the C2D design. We observe that the overall distribution of M3D ICs is pushed to the right, implying that the mismatch from the predicted wirelength increases due to the unique 3-D routing constraint in M3D ICs. When averaging the ratio between routed and predicted wirelength out of the entire nets, we observe that 3.5% mismatch for LDPC, while 16.2% and 19.6% for AES-128 and JPEG, respectively. This shows that intertier 3-D routing overhead negatively impacts on M3D ICs more than F2F ICs with regard to the wirelength.

D. Impact of Post-Tier-Partitioning Optimization

Using LDPC, Table VII shows how effectively post-TP opt fixes the timing violations caused by the intertier 3-D routing overhead. When intertier 3-D routing is done (Before Opt), WNS and TNS are degraded, and both F2F and M3D designs fail to meet the timing. Note that the M3D IC has more timing degradation than the F2F IC with regard to TNS and the number of violated paths. These timing violations are fixed after we perform post-TP opt (After Opt). WNS is improved by 44.4%, and TNS is restored by 91.5% with the negligible power overhead (0.1%) in the F2F IC. The timing restoration in the M3D IC is more drastic in that WNS and TNS values become significantly better than those of the F2F IC, respectively. This proves that post-TP opt is critical to implement timing-robust 3-D ICs.

While fixing the timing violation, post-TP opt indeed optimizes the positive slack paths and shows the standard cell area and pin capacitance savings. This cell-level saving is translated into the power saving in the M3D IC but not in the F2F IC. The interesting fact in the F2F IC is that, the wirelength is actually shorter than the M3D IC as discussed in Section V-C2, but the wire capacitance is larger than that of the M3D IC. This is because the F2F IC utilizes M6 layers on both tiers. Although intertier 3-D routings are made with short wirelength, using more M6 layers create larger wire capacitance, resulting in larger power consumption than the M3D IC. In general, post-TP opt restores the timing by inserting or up-sizing the buffers

TABLE VII Impact of Post-Tier-Partitioning Optimization. Intertier 3-D Routing Overhead Introduces Huge Timing Violation (Before Opt), and Our Optimization Fixes the Timing Violation With Negligible Power Overhead (After Opt)

LDPC F2F IC			
Before Opt	After Opt	$\Delta\%$	
65,187	65,271	-0.1	
179,815	179,645	0.1	
200.96	199.81	0.5	
2718.52	2720.88	-0.1	
303.48	306.27	-0.9	
-43.57	-24.23	44.4	
-2637.13	-222.99	91.5	
383	27	93.0	
178.25	178.49	-0.1	
LDPC	C M3D IC		
Before Opt	After Opt	$\Delta\%$	
65,187	65,149	0.1	
179,815	179,302	0.3	
200.96	199.14	0.9	
2735.01	2736.04	-0.1	
2735.01 299.34	2736.04 300.72	-0.1 -0.5	
2735.01 299.34 -40.05	2736.04 300.72 -7.15	-0.1 -0.5 82.1	
2735.01 299.34 -40.05 -6426.20	2736.04 300.72 -7.15 -18.53	-0.1 -0.5 82.1 99.7	
2735.01 299.34 -40.05 -6426.20 737	2736.04 300.72 -7.15 -18.53 11	-0.1 -0.5 82.1 99.7 98.5	
	LDP4 Before Opt 65,187 179,815 200.96 2718.52 303.48 -43.57 -2637.13 383 178.25 LDPC Before Opt 65,187 179,815 200.96	LDPC F2F IC Before Opt After Opt 65,187 65,271 179,815 179,645 200.96 199.81 2718.52 2720.88 303.48 306.27 -43.57 -24.23 -2637.13 -222.99 383 27 178.25 178.49 LDPC M3D IC Before Opt After Opt 65,187 65,149 179,815 179,302 200.96 199.14	

while minimizing the power increase. However, if the power overhead becomes the issue, then post-TP opt can start to delete or down-size the buffers at the expense of the timing margin.

E. Commercial 2D Versus C2D

Based on the optimal footprint derived from Section V-A, Table VIII compares the design results of 2-D IC with those of C2D-based F2F and M3D ICs. The total footprint savings of 3-D designs over the 2-D design is 57.8% for LDPC, and 53.0% for both AES-128 and JPEG while having similar placement utilization as that of 2-D IC. Our C2D flow offers $20\% \sim 34\%$ wirelength savings and $4\% \sim 13\%$ standard cell area savings. Therefore, the wire-dominated LDPC, which shows the highest wirelength to standard cell area ratio, benefits most from C2D in terms of the total power saving at iso-performance (28.0% in M3D IC), whereas the lowest wirelength to standard cell area ratio benchmark, JPEG, gains the lowest total power savings (5.7% in F2F IC).

F. Summary and Future Directions

To sum up the strengths of our C2D flow, first C2D does not shrink the standard cell and interconnect geometries, so we can utilize the 2-D P&R engines for the target technology node. Second C2D offers strong post-TP opt that enables timing, power, and 3-D via location co-optimization further. This makes C2D flow more favorable and adaptable in the advanced technology node. On the other hand, C2D requires the accurate parasitic database of the full 3-D metal stack for the decent post-TP opt, which is challenging due to the limited support from tools and commercial PDKs built for 2-D ICs. Also, regarding placement row splitting, supporting full DRV

1105

TABLE	VIII
-------	------

2-D VERSUS C2D ISO-PERFORMANCE POWER COMPARISON. Δ% INDICATES THE SAVINGS OVER THE 2-D DESIGN

	20		A 07		A 07
Design	2D	C2D-M3D	$\Delta\%$	C2D-F2F	$\Delta\%$
	LDPC, 2.06	GHz			
Footprint $(\mu m \times \mu m)$	555.7×555.1	361.2×360.8	57.8	361.2×360.8	57.8
3D Via Count	-	35,923	-	21,575	-
Standard Cell Area (μm^2)	204,782	178,700	12.7	178,876	12.7
Placement Utilization (%)	66.4	68.6	-3.3	68.6	-3.3
Total Wirelength (m)	3.8	2.5	33.6	2.5	33.6
Total WL / Cell Area (m^{-1})	18.7	14.0	25.1	14.2	24.1
Switching Power (mW)	193.9	134.2	30.8	136.9	29.4
Cell Internal Power (mW)	33.0	28.9	12.4	28.8	12.7
Leakage Power (mW)	11.1	8.2	26.1	8.2	26.1
Total Power (mW)	237.8	171.3	28.0	174.0	26.8
Normalized Power-Area Product	1.000	0.608	39.1	0.618	38.2
	AES-128, 5.4	GHz			
Footprint $(\mu m \times \mu m)$	716×715.6	490.9×490.6	53.0	490.9×490.6	53.0
3D Via Count	-	75,439	-	63,211	-
Standard Cell Area (μm^2)	377,702	356,021	5.7	361,096	4.4
Placement Utilization (%)	73.7	73.8	-0.1	75.0	-1.8
Total Wirelength (m)	2.9	2.4	17.2	2.2	22.9
Total WL / Cell Area (m^{-1})	7.7	6.7	13.0	6.2	19.5
Switching Power (mW)	250.8	221.8	11.6	223.7	10.8
Cell Internal Power (mW)	113.6	107.1	5.7	108.4	4.6
Leakage Power (mW)	17.5	15.7	10.3	16.1	8.0
Total Power (mW)	381.9	344.6	9.8	348.2	8.8
Normalized Power-Area Product	1.000	0.848	15.2	0.857	14.3
	JPEG, 2.160	GHz		- -	
Footprint $(\mu m \times \mu m)$	1156.3 × 1153.7	792.8×791.0	53.0	792.8×791.0	53.0
3D Via Count	-	148,943	-	121,357	-
Standard Cell Area (μm^2)	982,231	941,791	4.1	943,812	3.9
Placement Utilization (%)	73.6	75.1	0.2	75.3	0.2
Total Wirelength (m)	5.8	4.8	17.2	4.6	20.2
Total WL / Cell Area (m^{-1})	5.9	5.1	13.6	4.9	16.9
Switching Power (mW)	415.8	385.4	7.2	385.9	7.2
Cell Internal Power (mW)	195.1	189.7	2.7	189.9	2.7
Leakage Power (mW)	30.2	28.7	5.0	28.5	5.6
Total Power (mW)	641.1	603.7	5.8	604.4	5.7
Normalized Power-Area Product	1.000	0.885	11.5	0.886	11.4

fixing on the macro pins outside the boundaries will make post-TP opt more precise and eventually remove the sign-off DRV fixing at the incremental routing stage.

This article reveals the unique characteristics of both F2F and M3D ICs in the C2D flow. We observe that the intertier 3-D routing overhead negatively impacts M3D ICs more than F2F ICs with regard to the total wirelength. This is because M3D integration has the unique constraint on intertier 3-D routing in that 3-D vias should not penetrate top-tier cells. This leads to the congestion and detouring on intertier 3-D routings. However, we observe that longer wirelength in M3D ICs does not lead to more power consumption than that of F2F ICs. This is because F2F ICs utilize more M6 layers on both tiers for 3-D net routing, resulting in larger wire capacitance.

For the future directions, we can generalize C2D flow to handle more than two tiers for the advanced 3-D integration technology. For example, we can adjust the scaling factors in interconnect RC scaling/placement contraction for multitier designs. Various multiway balanced partitioning schemes can be applied to the tier partitioning. Also, placement row splitting in the compact 3-D via planning can be performed based on the number of dies. In addition, we can study the thermal and power integrity issues in 3-D ICs using our C2D flow, and lastly, we can apply C2D to build the variety of commercial-quality 3-D systems, including neuromorphics and heterogeneous 3-D ICs. All these challenges are the future works for C2D flow.

VI. CONCLUSION

To maximize the utilization of 3-D interconnect and the power-performance-area benefit of advanced 3-D ICs, in this article, we proposed a full-chip RTL-to-GDSII physical design solution named C2D that offers commercial-quality F2F and M3D IC physical layouts. We presented interconnect *RC* scaling and placement contraction, which allow us to utilize the original technology files and design rules for the target technology node for 3-D IC implementation. We also introduced compact placement to enable post-TP opt. With our extensive experiments and analysis, we evaluated the impact of those ideas step-by-step, and showed that using 28-nm process design kit, F2F, and M3D ICs implemented by our C2D flow offers a maximum 28.0% of total power reduction as well as 15.6% silicon area saving at its maximum compared to the 2-D IC at iso-performance.

REFERENCES

- [1] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Shrunk-2-D: A physical design methodology to build commercial-quality monolithic 3-D ICs," IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 36, no. 10, pp. 1716-1724, Oct. 2017.
- [2] T. Suga, R. He, G. Vakanas, and A. La Manna, "Direct Cu to Cu bonding and other alternative bonding techniques in 3D packaging," in 3D Microelectronic Packaging. Cham, Switzerland: Springer, 2017, pp. 129-155.
- [3] P. R. Morrow, C.-M. Park, S. Ramanathan, M. J. Kobrinsky, and M. Harmes, "Three-dimensional wafer stacking via Cu-Cu bonding integrated with 65-nm strained-Si/low-k CMOS technology," IEEE Electron Device Lett., vol. 27, no. 5, pp. 335-337, May 2006.
- [4] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain, and P.-E. Gaillardon, "3-D sequential integration: A key enabling technology for heterogeneous co-integration of new function with CMOS," IEEE J. Emerg. Sel. Topics Circuits Syst., vol. 2, no. 4, pp. 714-722, Dec. 2012.
- [5] J. Franco et al., "Gate stack thermal stability and PBTI reliability challenges for 3D sequential integration: Demonstration of a suitable gate stack for top and bottom tier nMOS," in Proc. IEEE Int. Rel. Phys. Symp. (IRPS), Monterey, CA, USA, 2017, pp. 2B-3.1-2B-3.5.
- [6] S.-W. Kim et al., "Ultra-fine pitch 3D integration using face-to-face hybrid wafer bonding combined with a via-middle through-silicon-via process," in Proc. Electron. Compon. Technol. Conf., Las Vegas, NV, USA, 2016, pp. 1179–1185. [7] E. Beyne, "The 3-D interconnect technology landscape," *IEEE Des.*
- Test., vol. 33, no. 3, pp. 8-20, Jun. 2016.
- [8] 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, 2014, pp. 171-176.
- [9] C. M. Fiduccia and R. M. Mattheyses, "A linear-time heuristic for improving network partitions," in Proc. 19th Design Autom. Conf., Las Vegas, NV, USA, 1982, pp. 175-181.
- [10] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs," in Proc. Int. Symp. Phys. Design (ISPD), 2014, pp. 47-54.
- [11] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs," IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 34, no. 4, pp. 540-553, Apr. 2015.
- [12] S. K. Samal, D. Nayak, M. Ichihashi, S. Banna, and S. K. Lim, "Tier partitioning strategy to mitigate BEOL degradation and cost issues in monolithic 3D ICs," in Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD), Austin, TX, USA, 2016, pp. 1-7.
- [13] B. W. Ku et al., "Physical design solutions to tackle FEOL/BEOL degradation in gate-level monolithic 3D ICs," in Proc. Int. Symp. Low Power Electron. Design, 2016, pp. 76-81.
- [14] B. W. Ku, P. Debacker, D. Milojevic, P. Raghavan, and S. K. Lim, "How much cost reduction justifies the adoption of monolithic 3D ICs at 7nm node?" in Proc. IEEE Int. Conf. Comput.-Aided Design, Austin, TX, USA, 2016, pp. 1-7.
- [15] 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, 2016, pp. 1-8.
- [16] OpenSPARC. Accessed: Feb. 23, 2020. [Online]. Available: http://www.opensparc.net
- [17] OpenCores. Accessed: Feb. 23, 2020. [Online]. Available: http://www.opensparc.net



Bon Woong Ku received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2014, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2017 and 2019, respectively.

His research interests include emerging device modeling, and physical design and CAD solutions for the next generation 3-D ICs in the advanced technology and neuromorphic computing.



Kyungwook Chang (Student Member, IEEE) received the B.S. and M.S. degrees in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2007 and 2010, respectively, and the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2019.

His current research interests include computeraided design solutions for 3-D ICs, designtechnology co-optimization, physical/logic design

and analysis, power delivery network, and parallel

computing/memory architecture.



Sung Kyu Lim (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the University of California at Los Angeles, Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001, where he is currently the Dan Fielder Endowed Chair Professor. His research on 3-D IC reliability is featured as Research Highlight in the Communication of the ACM in 2014. His 3-D IC test chip published in

the IEEE International Solid-State Circuits Conference in 2012 is generally considered the first multicore 3-D processor ever developed in academia. He has authored the book entitled Practical Problems in VLSI Physical Design Automation (Springer, 2008). His current research interests include modeling, architecture, and electronic design automation (EDA) for 3-D ICs.

Dr. Lim is a recipient of the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006, the Advisory Board of the ACM Special Interest Group on Design Automation from 2003 to 2008 and awarded a Distinguished Service Award in 2008, the Best Paper Awards from the IEEE Asian Test Symposium in 2012, the IEEE International Interconnect Technology Conference in 2014, and the Class of 1940 Course Survey Teaching Effectiveness Award from Georgia Institute of Technology in 2016. He was an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS from 2007 to 2009. He has been an Associate Editor of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS since 2013. He has served on the Technical Program Committee of several premier conferences in EDA.