Design-Aware Partitioning-Based 3-D IC Design Flow With 2-D Commercial Tools

Kyungwook Chang, Saurabh Sinha, Brian Cline, Greg Yeric, and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*—3-D ICs can continue to improve power, performance, area, and cost beyond traditional Moore's law scaling limitations by leveraging the third dimension and short vertical interconnects. Several recent studies present methodologies to implement 3-D ICs, but most of these studies implement each tier separately after partitioning a design into multiple tiers, resulting in inaccurate buffer insertion, which becomes more severe in advanced technology nodes. In this article, we present a new methodology called "Cascade2D flow" which utilizes design and microarchitecture insight for tier partitioning and implements 3-D ICs using 2-D commercial tools. By modeling vertical interconnects with sets of anchor cells and dummy wires, Cascade2D flow places, and routes and optimizes multiple tiers simultaneously in the 2-D version of a 3-D IC called "cascade2D design," which enables accurate buffer insertion. Two flavors of 3-D ICs—monolithic 3-D (M3D) and face-to-face-bonded (F2F-bonded) 3-D ICs—of a commercial in-order, 32-bit application processor at foundry 28 nm, 14/16 nm, and predictive 7-nm technology nodes are implemented using this new methodology. We investigate the power, performance and area improvements of 3-D ICs over the 2-D counterparts to examine the efficacy of the methodology. Our new methodology outperforms the state-of-the-art 3-D IC design flows in the both flavors of 3-D ICs with up to 4× better power savings. In the best case, 3-D ICs from Cascade2D flow show 25% better performance at iso-power and 20% lower power at iso-performance.**

***Index Terms*—3-D IC design flow, face-to-face-bonded 3-D IC, monolithic 3-D IC.**

I. Introduction

As 2-D scaling faces its limitations due to the physical limits of channel-length scaling, lithography limitation along with increased parasitics and costs, 3-D IC technology has emerged as a promising solution to extend Moore's law.

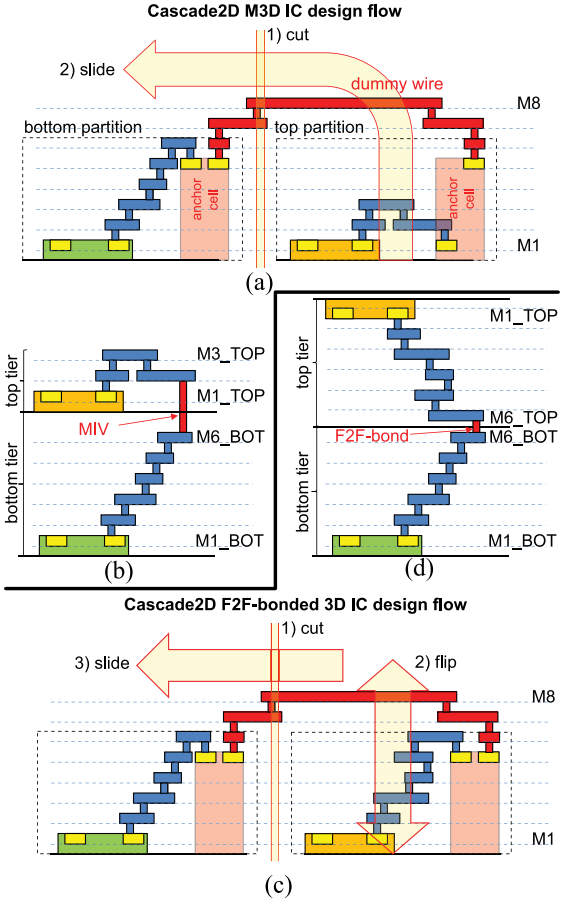

Fig. 1. 3-D IC implementation scheme of Cascade2D flow. (a) *Cascade2D design* with a set of anchor cells and dummy wires, which models an MIV and (b) equivalent M3D IC. (c) *Cascade2D design* and (d) corresponding F2F-bonded 3-D IC.

In 2018, TSMC announced wafer-on-wafer (WoW) silicon wafer stacking technology, which bonds the topmost metal layers of two fabricated dies with 10 μm×10 μm vertical interconnects, establishing a face-to-face (F2F) die interface as shown in Fig. 1(d). Unlike through-silicon via (TSV)-based 3-D ICs, which stack the back side of a fabricated die on the face side of another die, F2F-bonded 3-D ICs achieve finer vertical interconnects between dies owing to their smaller size.

On the other hand, in monolithic 3-D (M3D) ICs, transistors in multiple tiers are fabricated sequentially. Compared to F2F-bonded 3-D ICs, the sequential fabrication allows two adjacent tiers to have much finer vertical interconnects using fine-pitched monolithic intertier vias (MIVs), which connect

Manuscript received August 24, 2019; revised December 25, 2019, June 10, 2020, and November 5, 2020; accepted February 5, 2021. Date of publication March 9, 2021; date of current version February 21, 2022. This article was recommended by Associate Editor C. Zhuo. *(Corresponding author: Kyungwook Chang.)*

Kyungwook Chang is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea (e-mail: k.chang@skku.edu).

Saurabh Sinha and Brian Cline are with the Research Department, ARM Inc., Austin, TX 78735 USA (e-mail: saurabhr8here@gmail.com; brian.cline@arm.com).

Greg Yeric is with Cerfe Labs, Inc., Austin, TX 78759 USA (e-mail: greg.yeric@cerfelabs.com).

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

Digital Object Identifier 10.1109/TCAD.2021.3065005

the topmost metal layer of the bottom tier and the bottom-most metal layer of the top tier as shown in Fig. 1(b) [1], [2]. Owing to the very small sizes and parasitics of MIVs, and recent researches on manufacturing technology involving the ability to process thinner dies, M3D ICs harness true benefits of 3-D ICs with fine-grained vertical integration.

In 3-D ICs, standard cells and hard macros are partitioned into multiple tiers, and vertical interconnects (i.e., F2F-bonds in F2F-bonded 3-D ICs; MIVs in M3D ICs) are utilized for intertier connections. We reduce wire length by utilizing the short vertical interconnects instead of using long wires in 2-D space. 3-D ICs also save standard cell area because a smaller number of buffers and lower drive-strength cells are needed to drive the reduced wire load. Power savings in 3-D ICs are attributed to the reduced wire length and buffer area.

Currently, commercial electronic design automation (EDA) tools do not support 3-D ICs and, hence, previous studies have explored 3-D IC implementation approaches using 2-D commercial EDA tools. In Shrunk2D flow [3], in order to estimate cell placement of a 3-D IC, the dimensions of the cells and wires are shrunk, and the "shrunk2D design" is implemented in the half footprint of the 2-D IC counterpart. However, using a *shrunk2D design* is prone to inaccurate buffer insertion because of inaccurate wire-load estimation and wire-load increase caused by tier partitioning [4].

In order to resolve timing violations caused by the inaccuracy and wire-load increases, Compact2D flow [5] deploys post tier-partitioning optimization. In addition, by utilizing scaled wire parasitics instead of shrunk cells and wires for wire-load estimation, it does not require one technology-node smaller place-and-route engines and design rule checkers while implementing pseudo3D designs before partitioning, which are called "compact2D designs." However, obtaining the optimal parasitic scaling factor requires several physical design iterations, and the scope of the post tier-partitioning optimization is limited because it attempts to close timing on top of the predefined placement from the *compact2D design*. In addition, the optimization requires to create 3-D technology files using raw 2-D technology files (e.g., ICT in Cadence tools, ITF in Synopsys tools), which are usually foundry confidential and not available in most of the foundry process design kits (PDKs). Moreover, the above two flows are design-agnostic, utilizing unnecessarily a large number of vertical interconnects, which can degrade power and performance.

Other 3-D IC implementation methodologies are proposed in [6] and [7], which use only the locations of cells to partition them into two tiers. Billoint *et al.* [6] used a design folding technique but shows marginal wire-length savings with no power savings, whereas [7] utilizes a standard cell row alternating technique, but showing performance degradation.

In order to relieve worsening electrostatics associated with scaling planar transistors, industry has transitioned to 3-D FinFETs. However, FinFETs have higher parasitic capacitance owing to their 3-D structure and the introduction of local interconnects to contact transistors. Therefore, to reduce power consumption in FinFET-based technology nodes, it is crucial to reduce standard cell area effectively.

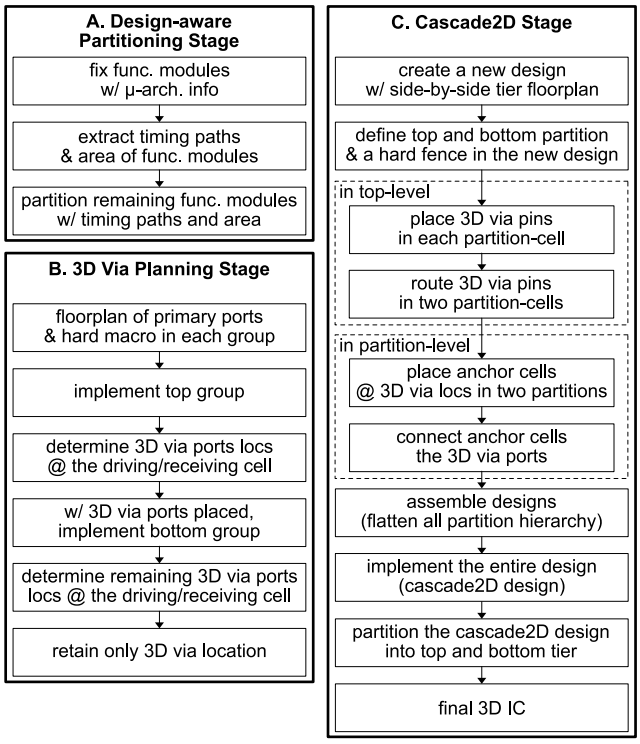

Fig. 2. Flow diagram of the proposed methodology, Cascade2D flow.

Fig. 1 shows our "Cut-and-Slide" methodology of Cascade2D flow with sets of anchor cells and dummy wires. As can be clearly seen, the anchor cells and dummy wires model vertical interconnects, and the *cascade2D design* in Fig. 1(a) is functionally equivalent to the M3D IC in Fig. 1(b), whereas the *cascade2D design* in Fig. 1(c) is equivalent to the F2F-bonded 3-D IC in Fig. 1(d).

The main contributions of this work are as follows: 1) we present a novel 3-D IC design methodology that implements M3D and F2F-bonded 3-D ICs while giving designers the ability to incorporate design and microarchitecture insight to guide the partitioning scheme; 2) our methodology is partition-scheme agnostic, hence making it an ideal platform to evaluate different partitioning schemes; 3) it effectively reduces the standard cell area as well as wire length compared to 2-D ICs, resulting in significant power savings; 4) the proposed Cascade2D flow shows better power savings compared to state-of-the-art 3-D IC implementation methodologies; and 5) the efficacy of the proposed design flow on M3D and F2F-bonded 3-D ICs are examined.

II. Implementation Methodology

This section presents our RTL-to-GDSII design methodology, Cascade2D flow, to implement sign-off quality 3-D ICs. The inputs and outputs of the proposed method are as follows.

1) *Input:* Design RTL and constraints, 2-D libraries.
2) *Output:* GDSII layouts, timing/power analysis results.

Fig. 2 shows the flow diagram of the proposed methodology. First, the functional modules in a design are partitioned into two groups, the top and bottom group, creating signals crossing the two groups, which eventually become 3-D vias (i.e.,

TABLE I
QUALITATIVE COMPARISON OF CASCADE2D FLOW AND OTHER 3-D IC IMPLEMENTATION FLOWS

	Cascade2D	Shrunk2D	Compact2D	[6][7]
partitioning granularity	block-level gate-level	gate-level	gate-level	gate-level
RTL-level constraints	supported	not supported	not supported	not supported
partitioning algorithm	flexible	bin-based min-cut	bin-based min-cut	geometry -based
designer's control on partitioning	complete control	only on bin-size	only on bin-size	no control
tier optimization	both tier simultaneous optimization	tier-by-tier optimization	post tier-part. optimization	no optimization
technology parameters for optimization	actual tech. param.	shrunk tech. param.	actual tech. param. + 3D techfiles	actual tech. param.

MIVs or F2F-bonds) in the 3-D IC (design-aware partitioning stage). Then, the locations of the 3-D vias are determined (3-D via planning stage) and, finally, the *cascade2D design* is implemented with sets of anchor cells and dummy wires in 2-D space, which is equivalent to the 3-D IC (Cascade2D stage).

Table I presents a qualitative comparison of Cascade2D flow with other 3-D IC implementation flows. As a design is placed and routed after partitioning, Cascade2D flow offers more flexibility on design partitioning compared to Shrunk2D and Compact2D flow, but at the cost of restricted cell placement as cell placement depends on design partitioning.

A. Design-Aware Partitioning Stage

In this stage, we partition the RTL of a design into two groups: 1) the top and 2) bottom groups, which represent the top and bottom tiers of the 3-D IC, respectively. The partitioning is performed in two steps: 1) first, manually fix some of the functional modules based on the microarchitecture of the design, giving designers the ability to control over the partitioning and then, 2) automatically partition the remaining functional modules with the design information from the 2-D IC.

In the first step, some of the functional modules in a design can be manually fixed into separate tiers using a detailed information of microarchitecture organization from front-end designers. This step is similar to the floorplanning step in the regular 2-D IC implementation flow, but, in here, we determine only the tier assignment of modules, not the exact location of modules in each tier. The architectural information used for regular 2-D IC floorplanning can be used in this step, which includes power-domain specifications, the data flow and timing budgets among functional modules, etc. For example, consider two functional modules having a large number of data paths with tight timing budgets (e.g., an ALU and its register bank). Placing these modules onto separate tiers and connecting them with 3-D vias help reduce the wire length.

In the second step, the remaining functional modules, whose microarchitectural organizations are not available or important, are partitioned automatically by utilizing the design

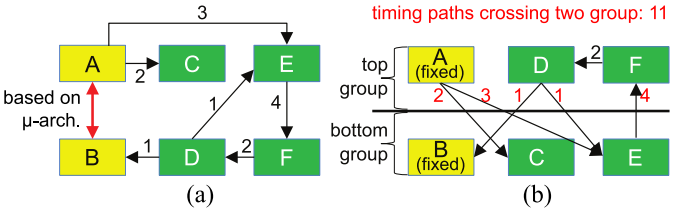

Fig. 3. Example of our design-aware partitioning scheme showing (a) functional modules manually fixed (yellow box) based on microarchitectural organizations and the degree of connectivity (numbers on arrows) of the rest of modules (green box). (b) Result of the design-aware partitioning stage after the nonfixed functional modules are automatically partitioned.

information from the synthesized netlists or the 2-D ICs. Because the depth (z-dimension) of an IC is extremely thin compared to its width (x-dimension) and height (y-dimension), the previous studies have shown that 3-D ICs reduce the distance between cells using vertical interconnects and, hence, achieve power and performance improvements [3], [4]. Similarly, we reduce the distance between modules, by placing heavily communicating functional modules on top of each other in separate tiers.

By extracting the timing paths from a 2-D IC, we can quantify the number of the timing paths crossing each pair of its functional modules. We call this number the "degree of connectivity" between functional modules. Timing paths crossing the module boundary indicate that the paths have timing constraints with respect to a clock, and they need to be optimized in order to close timing. Therefore, if two functional modules have a large degree of connectivity, EDA tools need to optimize a large number of timing paths by placing instances in the timing paths closely and minimizing wire length between those instances. We also extract the standard cell area of each functional module from the synthesized netlist or the 2-D IC to balance the cell area between the tiers.

After obtaining the degree of connectivity of the functional modules and their cell area, the design is partitioned into two groups based on the following criteria.

1) Balance cell area of the top and bottom groups.
2) Maximize the number of the timing paths crossing two groups.

These criteria help: 1) the functional modules, which have high degree of connectivity, be placed into separate tiers, hence minimizing the distance between them and 2) balance the standard cell area of the two tiers.

Fig. 3 shows an example of our partitioning methodology. Functional modules A and B are manually fixed on two different groups based on the microarchitecture of the design, and modules C, D, E, and F are automatically partitioned by the partitioner, maximizing the number of the timing paths crossing the two groups and balancing the cell area of the two groups. It should be emphasized, however, that Cascade2D flow is extremely flexible, and it can incorporate any number of constraints for partitioning cells or modules into tiers. Depending on the type of a design, the designer may wish to employ different partitioning criteria than presented here, and the subsequent stages (3-D via planning stage and Cascade2D stage) would remain the same. Hence, this flow is an ideal platform to evaluate different partitioning schemes for 3-D ICs.

At this stage, it is important to understand that there are two types of IO ports in a partitioned design. There are a set of IO ports, which is created because of "design-aware partitioning stage." These IO ports connect the top and bottom groups of the design and are referred to as "3-D via ports/pins" in the rest of this article since they eventually become 3-D vias in a 3-D IC. Additionally, we have a set of IO ports for the top level of a design before partitioning. These are same as the conventional IO ports of a 2-D IC, and referred to as "primary ports/pins."

B. 3-D Via Planning Stage

After partitioning the RTL of a design into the top and bottom groups, the locations of 3-D vias are determined. We first manually floorplan macro blocks (e.g., memory blocks) on the tiers which the macro blocks are assigned to in the previous stage. Next, we implement the top group and then place 3-D via ports above their driving or receiving cells on the top routing metal layer, so that the wire length between 3-D via ports and the relevant cells is minimized. Note that 3-D via ports are placed above the standard cells, instead of the edge of a die as would be done in a conventional 2-D IC. As explained in the previous section, 3-D via ports are actually IO ports that connect the top and bottom groups. We leverage the fact that all cell placement algorithms in commercial EDA tools tend to place the relevant cells close to IO ports to ease timing closure. Hence, we implement the bottom group using the locations of the 3-D vias determined from the top group implementation, so that the cell placement of the top group guides that of the bottom group using prefixed 3-D via ports.

We assume that primary ports of designs are connected only to the top tier in the 3-D ICs, and they can be either manually placed before the top group implementation or automatically placed by EDA tools during the top group implementation. However, it is possible that some primary ports need to be directly connected to the functional modules in the bottom group. These "feed-through" signals will not traverse any cells on the top group. Therefore, the 3-D via ports for those signals cannot be placed with the top group implementation and are determined during the bottom group implementation.

Since the spacing of MIVs is a few tens of nanometer, which is smaller than the size of cells, the spacing is automatically ensured while placing 3-D via ports above their relevant cells. On the other hand, in F2F-bonded 3-D ICs, although the recent improvements on high alignment precision of wafer-level integration help reduce the pitch of F2F-bonds in a 1-μm range [8], [9], the spacing is still too large to be automatically ensured by cells. Therefore, to ensure the spacing of F2F-bonds, we set the minimum distance among the driving and receiving cells connected to 3-D via ports to the minimum spacing of F2F-bonds during the top and bottom group implementations.

Fig. 4(a) and (b) shows the locations of 3-D vias after implementing the bottom group of an M3D IC and an F2F-bonded 3-D IC, respectively. The figures clearly show that the 3-D vias in the F2F-bonded 3-D IC are larger and more spread out compared to those in the M3D IC. After obtaining a complete set of 3-D via locations, the standard cell placement in the top and bottom group implementations is discarded, and only the locations of 3-D via ports are retained.

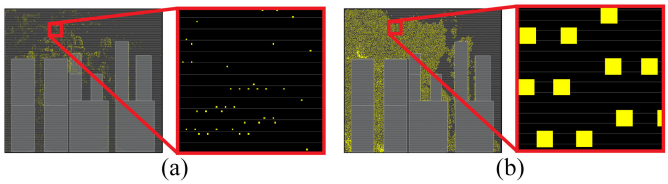

Fig. 4. Locations of 3-D vias (yellow dots) after completing the 3-D via planning stage: (a) MIVs in an M3D IC and (b) F2F-bonds in an F2F-bonded 3-D IC.

C. Cascade2D Stage

In this step, we simultaneously implement the top and bottom tier of a 3-D IC in a single design, a *cascade2D design*, so that all the cells in the 3-D IC are placed, routed and optimized with real cell location and accurate parasitics. The main advantages of this method are as follows: 1) it performs accurate buffer insertion as it does not involve parasitic estimation with shrunk geometry (Shrunk2D flow) or scaled wire parasitic (Compact2D flow); 2) it can better optimize the timing paths crossing the tiers as they are in a single design (i.e., all the timing information of both tiers are available) from the beginning of the implementation (i.e., without prefixed placement as Compact2D flow); and 3) it does not require 3-D technology files that are usually not available for foundry PDKs. A *cascade2D design* is modeled with sets of anchor cells and dummy wires, using a partitioning technique supported in Cadence Innovus. Note that the partitioning technique in here is a 2-D design technique and different from that in design-aware partitioning stage.

We first create a new floorplan with both tiers placed side-by-side with the same total area as the 2-D IC counterpart. We define the top and bottom partitions in the floorplan and set a hard fence for placement, so that cells in the top partition are placed only on the top half of the floorplan, and cells in the bottom partition only on the bottom half of the floorplan. Then, two hierarchies of the design are created as follows.

1) *First Level of Hierarchy:* Top view, which contains only two cells: a) the top-partition cell and b) the bottom-partition cell. These two cells contain pins that represent 3-D vias (i.e., 3-D via pins) for the top and bottom tiers, respectively.
2) *Second Level of Hierarchy:* Top-partition cell, which contains the top-partition view, where cells from the top group are placed and routed.
3) *Second Level of Hierarchy:* Bottom-partition cell, which contains the bottom-partition view, where cells from the bottom group are placed and routed.

In the top view, we place 3-D via pins in the top-partition cell and the bottom-partition cell on the topmost routing metal layer [e.g., M6 in Fig. 1(a) and (c)]. We use the 3-D via port locations derived in Section II-B for the pins in each of the top and the bottom-partition cells. Fig. 5(a) shows the placed pins for MIVs in an M3D IC in the top view. The locations of MIV pins in the top half of the floorplan (i.e., the top-partition cell) are identical to those in the bottom half of the floorplan (i.e., the bottom-partition cell).

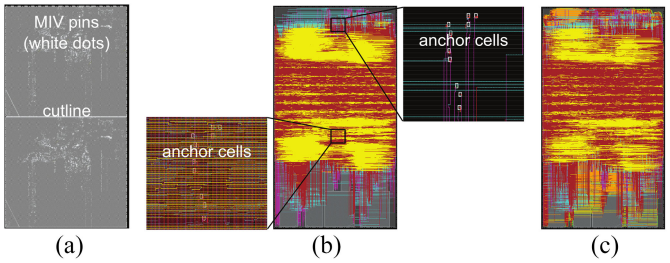

Fig. 5. Die images in different steps in the Cascade2D stage described in Section II-C while implementing an M3D IC: (a) top view after placing 3-D via pins for MIVs, (b) after assembling the top view with the top-partition and the bottom-partition views, and (c) after implementing the *cascade2D design*.

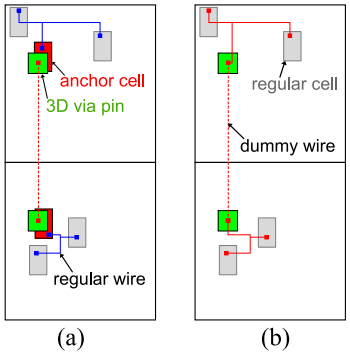

Fig. 6. Example of utilizing anchor cells to distinguish regular wires (solid lines) and dummy wires (dashed line) during delay and parasitic calculation: (a) with anchor cells and (b) without anchor cells. Red lines represent the nets whose parasitic and delay are 0, whereas blue lines are nonzero.

Then, using 3–4 additional metal layers above the topmost routing metal layer used in actual design, [e.g., M7 and M8 in Fig. 1(a) and (c)], the 3-D via pins in the top-partition cell and the bottom-partition cell are routed and connected. As the locations of the pins are identical in the x-axis in the top-partition and bottom-partition cells, routing tools create long vertical wires crossing the two partition cells. These wires on the additional 3–4 metal layers used to connect the 3-D via pins of the top-partition and bottom-partition cells are called "dummy wires" because their only function is establishing logical connections between the two tiers in physical designs (i.e., *cascade2D designs*). The delay and parasitics associated with these wires will not be considered in final 3-D ICs.

In an M3D IC, an MIV connects a wire in the topmost routing metal layer of the bottom tier and that in the bottommost routing metal layer of the top tier. We wish to emulate this connectivity in our *cascade2D designs*. Hence, we need a mechanism to connect the bottommost routing metal layer in the top-partition view with the topmost routing metal layer in the bottom-partition view. This is achieved with what we call as "anchor cells." An anchor cell is a dummy cell that implements buffer logic. Anchor cells model zero-delay virtual connection between a dummy wire and one of metal layers. After connecting the two partition cells with dummy wires, anchor cells are placed below the 3-D via pins in each partition view. In this step, only anchor cells are placed, not logic cells.

On the other hand, although F2F-bonds connect the topmost routing metal layer of the top and bottom tier, anchor cells

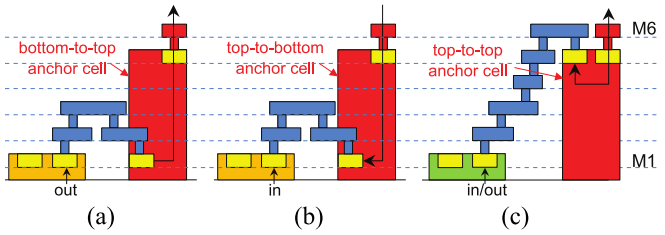

Fig. 7. Three types of anchor cells: (a) bottom-to-top anchor cell; (b) top-to-bottom anchor cell; and (c) top-to-top anchor cell.

should be still utilized as they are used by EDA tools to explicitly distinguish regular wires and dummy wires. Commercial EDA tools perform parasitics extraction and delay calculation based on nets, not wires (Note that a net is composed of multiple wires and vias). Fig. 6 compares the *cascade2D design* for an F2F-bonded 3-D IC with and without anchor cells. The dummy wire (dashed line) in Fig. 6(a) is bounded by anchor cells and separated from regular wires (solid wires), resulting in three different nets. This makes EDA tools possible to set the delay and parasitics of only dummy wires to zero (the red wire), while keeping regular wires intact (the blue wires). However, in case anchor cells are not deployed as shown in Fig. 6(b), EDA tools treat two regular wires and a dummy wire as one net, making the delay and parasitic to 0 not only for the dummy wires but also for the regular wires connected to the dummy wires, which is not desirable.

Depending on the flavor of a 3-D IC, the partition and the signal direction of a 3-D via, three flavors of anchor cells are utilized: 1) bottom-to-top anchor cells [Fig. 7(a)], which receive signals from the bottommost routing metal layer and drive a dummy wires; 2) top-to-bottom anchor cells [Fig. 7(b)], which send the signal in the reverse direction; and 3) top-to-top anchor cells [Fig. 7(c)], which connect a dummy wire to the topmost routing metal layer. In M3D ICs, bottom-to-top and top-to-bottom anchor cells are utilized for the top-partition cell to drive and receive signals of dummy wires, respectively, whereas top-to-top anchor cells are deployed for the bottom-partition cell. In F2F-bonded 3-D ICs, only top-to-top anchor cells are used to connect dummy wires to the topmost routing metal layer of both tiers. After placement, anchor cells and the corresponding 3-D via pins are connected.

Next, all hierarchies are flattened (i.e., the top view and the two partition views are assembled), projecting all anchor cells and dummy wires in the two partition views into the top view and creating a single assembled design shown in Fig. 5(b).

With the assembled design, we set the delay and parasitic of dummy wires to 0, and anchor cells and dummy wires are set to be fixed, so that they cannot be modified. These sets of anchor cells and dummy wires effectively act as "wormholes." In M3D ICs, these wormholes connect the topmost routing metal layer of the bottom partition to the bottommost routing metal layer of the top partition without delay, emulating the behavior of MIVs. On the other hand, in F2F-bonded 3-D ICs, the wormholes connect the topmost routing metal layers of the top and bottom partition, emulating F2F-bonds. Note that the parasitics of 3-D vias are added in the final timing stage.

Then, a regular place-and-route flow is performed, which involves the placement of logic cells in the design,

clock-tree-synthesis (CTS), post-CTS-hold, route, post-route, and post-route-hold stages. Owing to: 1) wormholes, which provide virtual connections between the top and bottom partitions and 2) the hard fence, which sets the boundary for the top and bottom partition, a commercial EDA tool places cells on the separate 2-D partitioned space where the cells belong to, with virtual connections (i.e., dummy wires) being the only wires crossing the separated 2-D partitioned space. At this stage, we call the resulting design the "cascade2D design."

CTS in the Cascade2D flow is performed as the regular 2-D implementation flow. As described in Section II-B, a clock signal from the corresponding primary input port first goes to the top partition, and is divided into two branches inside the partition. One of branches is used to generate a clock tree in the top partition, and the other branch is connected to the bottom partition through a set of anchor cells and a dummy wire and used to generate a clock tree in the bottom partition.

Fig. 5(c) shows a *cascade2D design* after the implementation. Although we set the delay and parasitics of dummy wires to 0 in the implementation stage, commercial EDA tools calculate power consumption with the actual RC parasitics for dummy wires. Therefore, *cascade2D designs* are again partitioned into the top and bottom partitions, pushing all cells and wires to the corresponding partitions except dummy wires. Then, the RC parasitics of each partition are extracted. The final 3-D IC is created by connecting the two extracted designs with the parasitics of 3-D vias. The final timing and power analysis is performed on the 3-D IC.

III. Experimental Setup

A. Cascade2D Versus Shrunk2D and M3D IC Versus F2F-Bonded 3-D IC

The experimental setup for the comparison between Cascade2D and Shrunk2D flow, as well as between M3D ICs and F2F-bonded 3-D ICs is the same as that described in [10] and is reproduced here for the sake of clarity and completion. Table II shows the representative metrics for each technology node used in our study, based on the previous publications [11]–[15]. The foundry 28-nm process is planar transistor-based while the foundry 14/16 nm is the first-generation foundry FinFET process. For these nodes, we have used production-level standard cell libraries containing over 1000 cells and memory macros that were designed, verified, and characterized using foundry PDKs.

We utilize a predictive 7-nm PDK to generate the required views for this study. We have developed the predictive 7-nm PDK containing electrical models (BSIM-CMG), design rule check (DRC), logic versus schematic (LVS), extraction, and technology library exchange format (LEF) files. The transistor models incorporate scaled channel lengths, fin-pitches, and increased fin-heights compared to the previous technology nodes to improve the performance at lower supply voltages. Multiple threshold voltages (V_T) and variation corners are supported in the predictive 7-nm PDK. Process metrics, such as gate pitch and metal pitches are linearly scaled from previous technology nodes [14], and design rules are created considering lithography challenges associated with printing these pitches. The interconnect stack is modeled based on similar scaling assumptions. A 7-nm standard cell library and memory macros are designed and characterized using this PDK.

TABLE II
Key Metrics for the Foundry 28, 14/16, and Predictive 7-nm Technology Node Used in This Study

parameters	28 nm [11]	14/16 nm [12][13]	7 nm scaled from [14]
transistor type	planar	FinFET	FinFET
supply voltage (V)	0.9	0.8	0.7
contacted poly-pitch (nm)	110∼120	78∼90	50
M1 pitch (nm)	90	64	36
MIV size (nm^2)	80×80	40×40	32×32
MIV spacing (nm)	80	40	32
MIV height (nm)	140	170	170
F2F-bond size (μm^2)	0.5×0.5 [15]		
F2F-bond spacing (μm)	0.5 [15]		
F2F-bond height (μm)	0.9 [15]		

Our 3-D ICs require six routing metal layers on both the top and the bottom tiers. The MIVs connect M6 of the bottom tier with *M*1 of the top tier, whereas F2F-bonds connect *M*6 of the top and bottom tiers. We limit the size of the MIVs to be $2\times$ the minimum via size allowed in the technology node to reduce the MIV resistance. The MIV heights take into account the fact that MIVs need to traverse through intertier dielectrics and transistor substrates to contact to M1 on the top tier. The MIV height increases from 28 to 14/16 and 7-nm technology nodes because of the introduction of local interconnect middle-of-line (MOL) layer in the sub-20-nm nodes. The MIV resistance is estimated based on the dimension of the vias and we used the previously published values for the MIV capacitance from [3].

Since the fabrication of M3D ICs is done sequentially, high-temperature processing for the front-end device on the top-tier can adversely affect the interconnects in the bottom tier while low-temperature processing will result in inferior top-tier devices. Recent works reporting low-temperature processings that achieve similar device behavior across both tiers have been presented in [16] and, hence, all our studies are done with the assumption of similar device characteristics in both the tiers.

As fabricated dies are bonded together to build F2F-bonded 3-D ICs, the size, spacing, and height of F2F-bonds depend on wafer-level integration technology rather than process technology. Therefore, the size and spacing of F2F-bonds are set based on [15] and remain the same for all technology nodes.

The standard cell libraries and memory macros for the foundry 28, 14/16 and the predictive 7-nm technology nodes are used to synthesize, place, and route the full-chip designs. The 2-D, M3D, and F2F-bonded 3-D ICs of a commercial, in-order, 32-bit application processor are implemented sweeping the target frequency from 500 MHz to 1.2 GHz in 100-MHz increments across the three technology nodes. Full-chip timing is closed at the slowest corner for setup fix and the fastest corner for hold fix. Power is reported at the typical corner. The memory block floorplans are customized within each tier for each technology node to close timing but kept constant during frequency sweeps. The chip area is fixed such that the final cell utilization is similar across technology nodes.

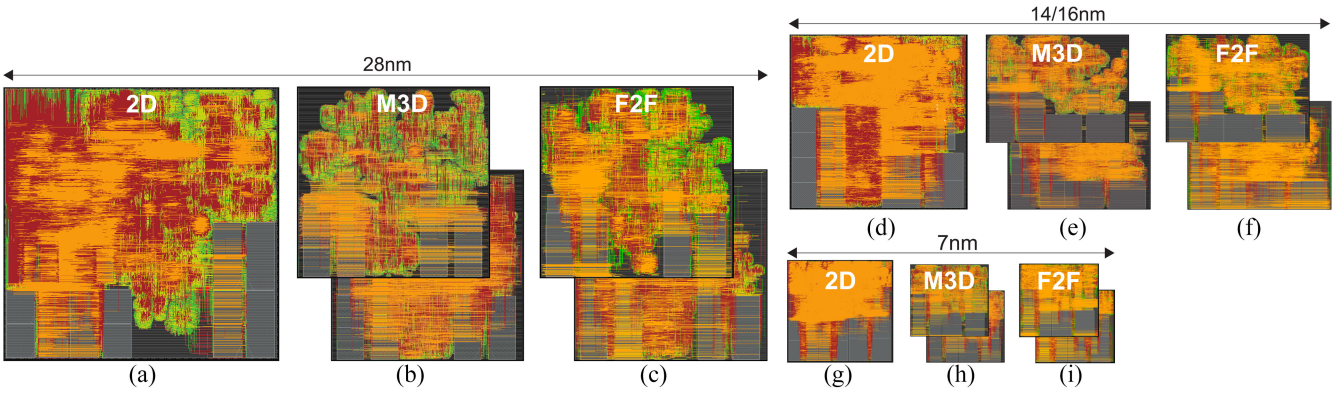

Fig. 8. Layouts of (a) 28-nm 2-D. (b) 28-nm Cascade2D M3D. (c) 28-nm Cascade2D F2F-bonded 3-D. (d) 14/16-nm 2-D. (e) 14/16-nm Cascade2D M3D. (f) 14/16-nm Cascade2D F2F-bonded 3-D. (g) 7-nm 2-D. (h) 7-nm Cascade2D M3D. (i) 7-nm Cascade2D F2F-bonded IC of the application processor at 1.0 GHz.

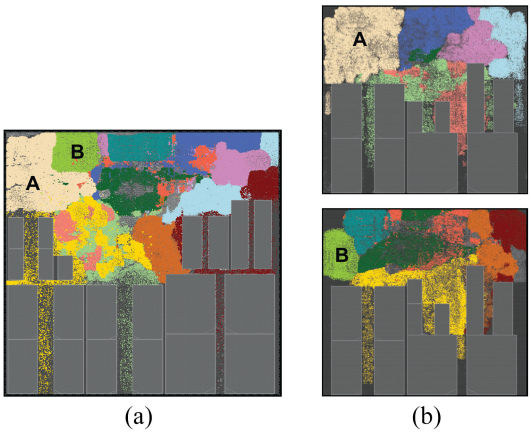

Fig. 9. Color map of functional modules in 7 nm (a) 2-D IC and (b) Cascade2D M3D IC of a commercial processor at 1.0 GHz.

B. Cascade2D Versus Compact2D

In order to compare the proposed methodology to Compact2D flow, the 2-D, M3D, and F2F-bonded 3-D ICs are implemented with Nangate 45-nm Open Cell Library based on FreePDK45 because of its full accessibility to the raw technology file, which is essential to create a 3-D technology file for post tier-partitioning optimization in Compact2D flow.

AES-128 and JPEG encoder from OpenCores are used as benchmarks. The target frequency for AES-128 is set to 1500 MHz, which is the maximum frequency at which its 2-D IC is able to close timing. In order to investigate the timing optimization behavior of the two flows during timing closure, we intentionally set the target frequency of the JPEG encoder to 1000 MHz, which is ∼100 MHz higher than the maximum frequency of its 2-D IC.

Functional modules in the designs are partitioned using the methodology presented in Section II-A with the following constraints from their microarchitecture. In AES-128, shift-rows, mix-columns, and add-round-key modules in each encryption round, which form computation intensive data paths, are assigned to alternating tiers (e.g., shift-rows and add-round-key modules in an encryption round are assigned to the bottom tier, and the mix-columns module is assigned to the top tier). In the JPEG encoder, quantizer and entropy encoder modules of each *Y*, Cb, and Cr components are assigned to different tiers. The rest of the functional modules are automatically partitioned based on the methodology in Section II-A.

IV. Results and Analysis: Cascade2D versus Shrunk2D

A. Power and Performance Benefit

Fig. 8 shows the die images of 2-D, Cascade2D M3D ICs (i.e., M3D ICs implemented with Cascade2D flow), and Cascade2D F2F-bonded 3-D ICs of the commercial, in-order, 32-bit application processor at 1.0 GHz in the foundry 28, 14/16, and the predictive 7-nm technology nodes. Since the 28 and 14/16-nm designs are unable to meet timing at 1.2 GHz, the designs of the target frequency up to 1.1 GHz are presented. For the 7 nm, we report the results up to 1.2 GHz.

From the timing path extraction from the 2-D ICs, we find that the functional modules *A* and *B* in Fig. 9(a) have a large number of timing paths crossing them. Those functional modules are placed side-by-side by EDA tools in the 2-D IC, resulting in long wire length and large standard cell area in the timing paths between the modules. On the other hand, in the Cascade2D M3D IC, those functional modules are automatically partitioned with the two criteria (i.e., balancing the cell area and maximizing the number of the timing paths) in the second step of the design-aware partitioning stage, minimizing the distance between them using MIVs. These MIVs reduce the wire length of the timing paths crossing the modules, as well as the standard cell area of the modules because of the reduced wire load.

The normalized total power consumption of the 2-D and Cascade2D M3D ICs across technology nodes are presented in Fig. 10. We observe that the Cascade2D M3D ICs consume less power in all cases. Hence, at iso-power, the M3D ICs run at higher frequencies compared to the 2-D counterparts. For example, considering the 14/16-nm technology node, we see that the M3D ICs show 25 % higher performance with the same total power compared to the 2-D ICs. Fig. 11 shows the power saving comparison between the Cascade2D M3D ICs and the Shrunk2D M3D ICs from their 2-D counterparts. The

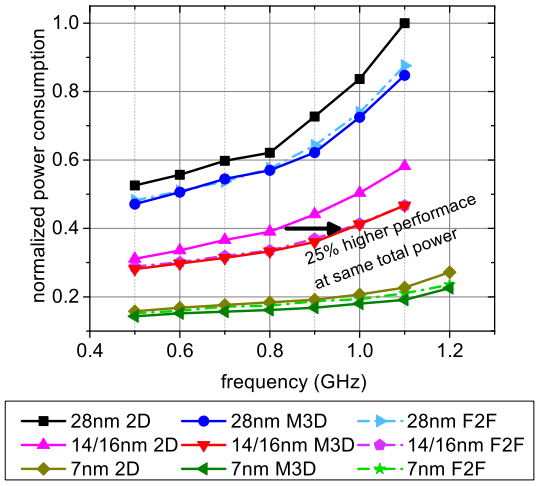

Fig. 10. Normalized power consumption of 2-D, Cascade2D M3D and F2F-bonded 3-D ICs across technology nodes.

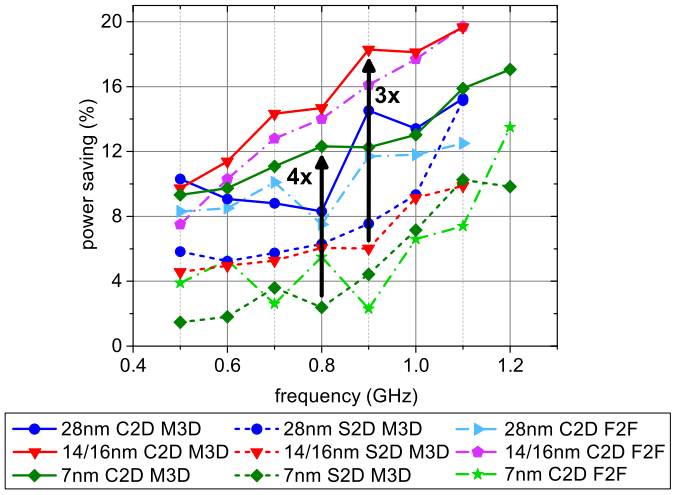

Fig. 11. Power savings of Cascade2D M3D (C2D M3D; solid lines), Shrunk2D M3D (S2D M3D; dotted lines), and Cascade2D F2F-bonded 3D (C2D F2F; dashed-dotted lines) ICs over 2D ICs.

Cascade2D M3D ICs show up to 3×–4× better power savings than the Shrunk2D M3D ICs depending on the technology node and the target frequency. In the best case scenario, the Cascade2D M3D IC consumes 20 % less power than the 2-D IC at iso-peformance (14/16 nm technology node at 1.1 GHz).

B. Comparison to Shrunk2D Flow

To analyze the difference in the power savings between Cascade2D M3D and Shrunk2D M3D ICs, we use the following equation explaining dynamic power consumption:

$$P_{dyn} = P_{INT} + \alpha \cdot \left(C_{\text{pin}} + C_{\text{wire}}\right) \cdot {V_{DD}}^2 \cdot f_{clk}. \tag{1}$$

The first term P_{INT} is the internal power of cells. The second term describes the switching power where C_{pin} is the pin capacitance of the cells, and C_{wire} is the wire capacitance in the design. α is the activity factor, and f_{clk} is the design target clock frequency. Since internal power and pin capacitance depend on the standard cell area, and wire capacitance is correlated to the wire length, we extend (1) to (2), which describes

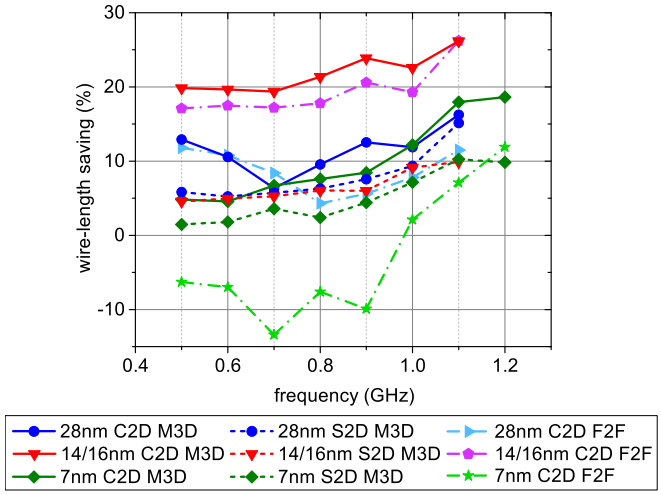

Fig. 12. Wire-length reduction comparison between Cascade2D M3D, Shrunk2D M3D, and Cascade2D F2F-bonded 3-D ICs over 2D ICs.

TABLE III
NUMBER OF MIVS IN 28, 14/16, AND 7-NM M3D ICS OF THE APPLICATION PROCESSOR AT 1.0 GHz

MIV count	28 nm	14/16 nm	7 nm
Cascade2D M3D IC	7,545	7,545	7,545
Shrunk2D M3D IC	164,553	120,770	99,587

the factors affect the power saving of M3D ICs

$$\Delta P_{dyn} = \Delta_{\text{cell}} \cdot \left(P_{INT} + \alpha \cdot C_{\text{pin}} \cdot {V_{DD}}^2 \cdot f_{clk}\right) + \Delta_{\text{wire}} \cdot \alpha \cdot C_{\text{wire}} \cdot {V_{DD}}^2 \cdot f_{clk} \tag{2}$$

where Δ_{cell} and Δ_{wire} are the differences in the standard cell area and wire length between 2-D and M3D ICs, respectively.

The primary advantage of Shrunk2D M3D ICs comes from reduced wire length, which results in reduced wire switching power dissipation [10]. As shown in Fig. 12, Shrunk2D M3D ICs reduce wire length by 20 % to 25 % consistently across technology nodes and target frequencies. Its wire-length reduction is mainly attributed to a large number of MIVs between the tiers. Table III compares the number of MIVs in Shrunk2D M3D and Cascade2D M3D ICs. Shrunk2D flow partitions "cells" into two tiers and utilizes multiple MIVs for a net, whereas Cascade2D flow partitions "functional modules" and allows one MIV per net. For these reasons, the number of MIVs in the Shrunk2D M3D ICs is an order of magnitude higher than that in the Cascade2D M3D ICs.

The large number of MIVs in Shrunk2D M3D ICs helps reduce the wire length, but it also increases the total capacitance of MIVs, limiting the wire capacitance reduction. As shown in Table IV, although the Shrunk2D M3D ICs reduce more wire length than the Cascade2D M3D ICs in the 14/16 and 7-nm technology node, the wire capacitance reduction of the Cascade2D M3D ICs is higher than the Shrunk2D M3D ICs. The negative impact of the large number of MIVs on wire capacitance stems from the bin-based partitioning scheme of Shrunk2D flow [3]. While bin-based partitioning helps distribute cells evenly on both tiers, it has a tendency to partition neighboring cells into separate tiers, which are

TABLE IV
NORMALIZED ISO-PERFORMANCE COMPARISON OF 2-D IC, SHRUNK2D M3D IC, CASCADE2D M3D, AND F2F-BONDED 3-D ICS OF THE APPLICATION PROCESSOR ACROSS TECHNOLOGY NODES AT 1.0 GHz. ALL VALUES ARE NORMALIZED TO CORRESPONDING 28-nm 2-D IC PARAMETERS. CAPACITANCE AND POWER VALUES ARE NORMALIZED TO 28-nm 2-D IC TOTAL CAPACITANCE AND TOTAL POWER, RESPECTIVELY

	normalized 2D IC			Shrunk2D M3D IC %Δ from 2D IC			Cascade2D M3D IC %Δ from 2D IC			Cascade2D F2F-bonded 3D IC %Δ from 2D IC		
parameters	28 nm	14/16 nm	7 nm	28 nm	14/16 nm	7 nm	28 nm	14/16 nm	7 nm	28 nm	14/16 nm	7 nm
std. cell area	1	0.331	0.077	−7.6 %	−6.8 %	−7.5 %	−9.5 %	−11.9 %	−8.8 %	−9.0 %	−13.4 %	−4.1 %
wire-length	1	0.728	0.404	−19.3 %	−24.1 %	−24.6 %	−11.9 %	−22.6 %	−12.2 %	−7.8 %	−19.3 %	−2.1 %
wire cap.	0.531	0.375	0.205	−18.1 %	−14.2 %	−13.7 %	−9.5 %	−19.7 %	−19.2 %	−5.9 %	−16.4 %	−5.1 %
pin cap.	0.469	0.422	0.203	−12.1 %	−6.3 %	−9.7 %	−11.1 %	−13.2 %	−7.9 %	−10.3 %	−14.5 %	−1.4 %
total cap.	**1**	**0.797**	**0.408**	**−15.5 %**	**−10.1 %**	**−11.7 %**	**−9.6 %**	**−15.2 %**	**−12.9 %**	**−7.3 %**	**−14.3 %**	**−2.6 %**
internal power	0.428	0.282	0.128	−4.8 %	−7.6 %	−4.7 %	−14.5 %	−15.2 %	−11.1 %	−13.4 %	−16.4 %	−7.9 %
switching power	0.505	0.318	0.119	−13.4 %	−10.6 %	−10.1 %	−13.0 %	−20.8 %	−15.1 %	−10.8 %	−18.9 %	−5.0 %
leakage power	0.066	0.002	0.000	−7.7 %	−4.0 %	−2.0 %	−9.5 %	−7.7 %	−2.8 %	−8.9 %	−8.3 %	−1.4 %
total power	**1**	**0.602**	**0.247**	**−9.3 %**	**−9.1 %**	**−7.2 %**	**−13.4 %**	**−18.1 %**	**−13 %**	**−11.8 %**	**−17.7 %**	**−6.6 %**

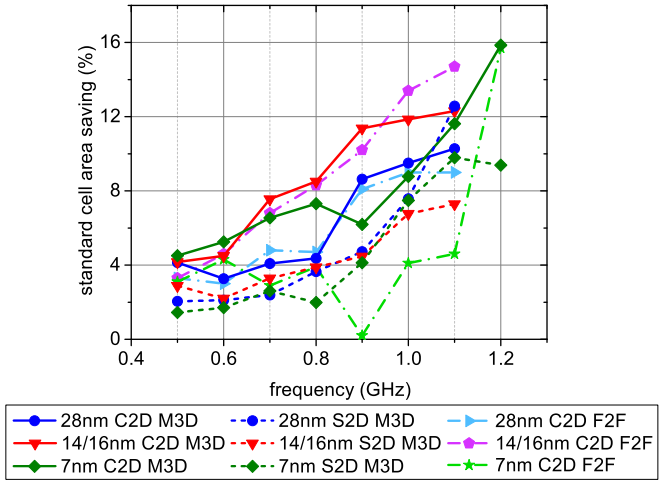

Fig. 13. Standard cell area savings in Cascade2D M3D, Shrunk2D M3D, and Cascade2D F2F-bonded 3-D ICs over 2-D ICs.

originally connected using short local wires in the 2-D counterpart, increasing the wire capacitance. On the other hand, in Cascade2D flow, as the partitioning is based on functional modules, cells in a functional module are connected with short local wires on the xy-plane, and only intermodule connections, which tend to be longer, can be split onto tiers using MIVs.

Most importantly, Cascade2D M3D ICs save their power mainly through reducing standard cell area. Shrunk2D flow uses a *shrunk2D design* to estimate the wire length and wire load of the resulting M3D IC. However, while shrinking technology geometries, the minimum width of each metal layer is also scaled, and extrapolation is performed by commercial EDA tools during wire parasitic extraction. This extrapolation tends to overestimate wire parasitics, especially in scaled technology nodes, resulting in an unnecessarily large number of buffers inserted in *shrunk2D designs* to close timing [4]. In Cascade2D flow, buffers are inserted while implementing and optimizing the top and bottom partitions simultaneously (in the Cascade2D stage described in Section II-C) with actual technology geometries. Therefore, Cascade2D flow achieves more accurate parasitic extraction and, hence, more standard cell area savings than Shrunk2D flow, as shown in Fig. 13. This standard cell area savings also help reduce the wire length of Cascade M3D ICs.

TABLE V
NORMALIZED ISO-PERFORMANCE COMPARISON OF 2-D IC, CASCADE2D M3D ICS WITH SAME DIE-AREA, AND 10 % REDUCED DIE-AREA AT 1.1 GHz IN PREDICTIVE 7-nm TECHNOLOGY NODE

parameters	2D	Cascade2D M3D IC	
die-area	1	1	0.9
density	69.7 %	63.2 %	71.1 %
total power	1	0.841	0.871

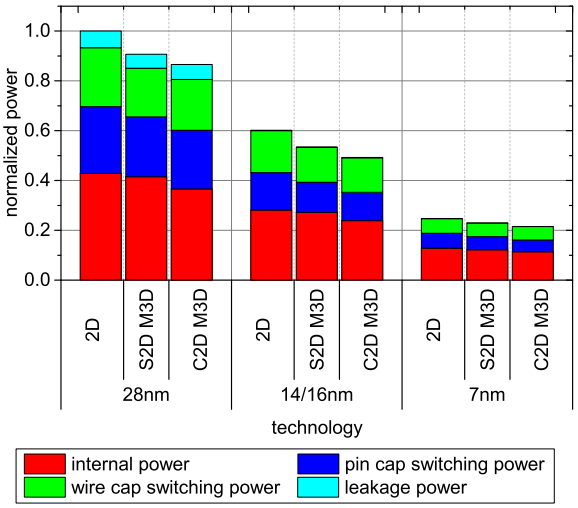

Fig. 14. Power breakdown into internal power, pin capacitance switching power, wire capacitance switching power and leakage power for 2-D, Shrunk2D M3D (S2D M3D), and Cascade2D M3D (C2D M3D) ICs at 1.0 GHz in foundry 28, 14/16, and predictive 7-nm technology nodes.

With more reduction in the standard cell area, the cell density of a Cascade2D M3D IC reduces as well. Hence, we leverage this feature to reduce the die-area while increasing the cell density. We implement two separate M3D ICs using Cascade2D flow, one with the same total die-area as the 2-D counterpart and another with a 10 % reduced area. Table V shows that we can maintain similar power savings with the reduced die-area of the M3D IC. The ability of reducing die-area makes M3D ICs extremely attractive for main-stream adoption because less area directly translates to reduced costs.

As shown in (2), the standard cell area reduction affects both internal power and pin capacitance switching power reduction, whereas wire-length reduction reduces only wire capacitance switching power. Fig. 14 shows the power breakdown of the

TABLE VI
RUNTIME COMPARISON BETWEEN SHRUNK2D AND CASCADE2D FLOW WITH THE APPLICATION PROCESSOR AT 1.0 GHz IN 7-NM TECHNOLOGY NODE

Shrunk2D flow		Cascade2D flow	
step	run-time	step	run-time
1. shrunk2D impl.	5 h	1. design-aware part.	0.5 h
2. gate-level part.	0.5 h	2. 3D via plan	2 h
3. MIV plan	0.5 h	3. cascade2D impl.	4.5 h
4. top/bottom tier impl.	1.5 h	-	-
total	7.5 h	total	7 h

TABLE VII
DESIGN AND POWER METRIC COMPARISON OF CASCADE2D F2F-BONDED 3-D ICS USING 0.5 μm AND 0.25-UM F2F-BONDS SIZE AND SPACING IN 7-NM TECHNOLOGY NODE AT 1.2 GHz

F2F-bond spacing	0.5 µm	0.25 µm
std. cell area	−15.7 %	−17.7 %
wire-length	−11.9 %	−16.8 %
wire cap.	−14.7 %	−20.1 %
pin cap.	−9.3 %	−12.2 %
total cap.	**−11.4 %**	**−15.5 %**
internal power	−16.8 %	−18.4 %
switching power	−10.1 %	−15.1 %
leakage power	−5.8 %	−5.8 %
total power	**−13.5 %**	**−16.7 %**

2-D, Cascade2D M3D, and Shrunk2D M3D ICs. As shown in the figure, the internal power and pin capacitance switching power, which depend on the standard cell area, account for over 70 % of the total power, and they contribute even more in the 14/16 and 7-nm ICs. This is because FinFETs have larger pin capacitance, which comes from 3-D fin structure and MOL layers, which do not exist in 28-nm technology node. Therefore, reducing standard cell area becomes more important as technology advances. As the Cascade2D M3D ICs reduce more standard cell area compared to the Shrunk2D M3D ICs by attacking 70 % of the total power; they achieve better power savings consistently, even though the wire-length reduction of the Cascade2D M3D ICs is less than the Shrunk2D M3D ICs.

Table VI compares the runtime of Cascade2D flow and Shrunk2D flow. For Shrunk2D flow, we assume that the design library with shrunk geometry is available. The total runtime for each flow is comparable. It is important to note that both flows need a reference 2-D IC. A 2-D IC is needed in Shrunk2D flow to evaluate the quality of the final M3D IC, while it is useful in Cascade2D flow to extract timing and standard cell area information for the design-aware partitioning stage.

V. RESULTS AND ANALYSIS: F2F-BONDED 3-D IC VERSUS M3D IC

A. Power and Performance Benefit of F2F-Bonded 3-D IC

Fig. 10 also presents the power consumption of the Cascade2D F2F-bonded 3-D ICs of the application processor across the technology nodes and target frequencies, which is normalized to the 28-nm 2-D IC at 1.1 GHz. The Cascade2D F2F-bonded 3-D ICs also outperform the 2-D counterparts showing up to 19.7 % power savings in the best case scenario (14/16-nm technology node at 1.1 GHz).

However, as presented in Fig. 11 which compares the power savings of 2-D and 3-D ICs, the power savings of the F2F-bonded 3-D ICs are less than the M3D ICs in almost all cases. We observe a significant difference in the 7-nm technology node, showing up to 9.9 % less power savings at 900 MHz.

B. Cascade2D F2F-Bonded 3-D ICs Versus M3D ICs

The less power savings of the F2F-bonded 3-D ICs compared to the M3D ICs are first attributed to the less wire-length savings as shown in Fig. 12 and Table IV.

As shown in Fig. 1(d), to route a net from a bottom-tier cell to a top-tier cell in an F2F-bonded 3-D IC, the net needs to traverse through all the metal layers of the bottom tier, an F2F-bond and then all the metal layers of the top tier. On the other hand, the number of metal layers utilized to connect cells in different tiers is less in an M3D IC [Fig. 1(b)]. This increases the wire length of F2F-bonded 3-D ICs, which, in turn, reduces Δ_{wire} in the second term of (2).

Another and the most important factor that worsens the wire-length savings of F2F-bonded 3-D ICs is the larger size and spacing of F2F-bonds compared to MIVs. As multiple tiers of M3D ICs are fabricated sequentially, the size and spacing of MIVs are in a few tens of nanometer and scale down as technology advances, enabling ultra fine-grained vertical integration. However, in F2F-bonded 3-D ICs, as fabricated wafers are bonded, they depend on wafer-level integration technologies involving high precision wafer alignment and bonding technologies [8], [9], [15]. Therefore, the size and spacing of F2F-bonds are significantly larger than MIVs, which are in micron-scale and consistent across technology nodes.

The larger size and spacing of F2F-bonds make their vertical integration more spread out compared to M3D ICs as shown in Fig. 4. This prevents F2F-bonds from being placed on their optimal locations and from minimizing the wire length of nets crossing the tiers. The negative impact becomes worse as technology advances as the size and spacing of F2F-bonds do not scale while cell sizes shrink, even showing negative wire-length saving compared to 2-D ICs as shown in Fig. 12.

As the wire length of the nets crossing the tiers in F2F-bonded 3-D ICs are longer than M3D ICs, the wire parasitics of the nets are also larger. The larger wire parasitics worsen the timing closure of F2F-bonded 3-D ICs, which makes them utilize more buffers while closing timing as shown in the similar trends in Figs. 12 and 13. The less standard cell area savings decrease the first term of (2), hence reducing the total power savings of F2F-bonded 3-D ICs.

In order to investigate the impact of the size and spacing of F2F-bonds on the power savings of F2F-bonded 3-D ICs, we compare the design and power metrics of F2F-bonded 3-D ICs in the 7-nm technology node at the maximum achievable frequency (i.e., 1.2 GHz) when the size and spacing are 0.5 and 0.25 μm in Table VII. It shows that using F2F-bonds with the smaller size and spacing, the F2F-bonded 3-D IC achieves a significantly higher wire-length saving (5.1 % higher) by placing its F2F-bonds on more optimized locations, which also contributes more standard cell area savings (2.0 % higher) due to reduced wire loads. The higher savings are translated to

TABLE VIII
ISO-PERFORMANCE COMPARISON OF 2-D IC AND THEIR 3-D COUNTERPARTS OF AES-128 AND JPEG ENCODER IMPLEMENTED USING COMPACT2D AND CASCADE2D FLOW WITH THE NANGATE 45-nm OPEN CELL LIBRARY. NOTE THAT THE TARGET FREQUENCY OF JPEG ENCODER IS INTENTIONALLY SET HIGHER THAN ITS MAXIMUM FREQUENCY. ALL PERCENTAGE VALUES ARE COMPUTED WITH RESPECT TO ITS 2-D IC VALUE. WNS, TNS, AND TPS STAND FOR WORST NEGATIVE SLACK, TNS, AND TOTAL POSITIVE SLACK, RESPECTIVELY

	AES-128@1,500MHz					JPEG encoder@1,000MHz				
	2D	M3D %Δ from 2D		F2F %Δ from 2D		2D	M3D %Δ from 2D		F2F %Δ from 2D	
		Compact2D	Cascade2D	Compact2D	Cascade2D		Compact2D	Cascade2D	Compact2D	Cascade2D
std. cell area (mm^2)	0.205	2.1 %	−3.4 %	2.0 %	−2.8 %	0.639	−8.7 %	−7.7 %	−8.5 %	−8.0 %
wire-length (m)	2.078	−17.8 %	−7.6 %	−17.0 %	−6.7 %	4.003	−12.8 %	8.6 %	−12.0 %	8.6 %
wire cap. (pF)	237.0	−9.1 %	−6.9 %	−5.6 %	−5.8 %	384.4	1.4 %	12.0 %	2.7 %	11.8 %
pin cap. (pF)	491.7	3.6 %	−12.9 %	3.5 %	−11.7 %	1,301.7	−20.2 %	−19.8 %	−19.9 %	−20.3 %
total cap. (pF)	**728.7**	**−0.5 %**	**−10.9 %**	**0.5 %**	**−9.8 %**	**1,686.1**	**−15.3 %**	**−12.5 %**	**−14.8 %**	**−13.0 %**
total res. (MΩ)	**10.9**	**6.1 %**	**−5.0 %**	**7.5 %**	**−4.3 %**	**23.4**	**9.4 %**	**1.3 %**	**9.7 %**	**1.2 %**
internal power (mW)	131.9	−1.2 %	−6.6 %	−1.0 %	−5.4 %	535.0	−13.6 %	−15.2 %	−13.4 %	−15.3 %
switching power (mW)	128.3	−16.0 %	−18.0 %	−15.1 %	−16.2 %	467.4	−14.2 %	−16.2 %	−13.9 %	−16.5 %
leakage power (mW)	3.5	3.4 %	−11.1 %	3.2 %	−10.2 %	7.8	−19.9 %	−18.5 %	−19.6 %	−18.9 %
total power (mW)	**263.7**	**−8.4 %**	**−12.2 %**	**−7.8 %**	**−10.7 %**	**1,010.2**	**−14.0 %**	**−15.7 %**	**−13.7 %**	**−15.8 %**
WNS (ns)	**-0.056**	**-0.106**	**-0.056**	**-0.104**	**-0.047**	**-0.105**	**-0.232**	**-0.091**	**-0.241**	**-0.099**
TNS (ns)	-28.0	-86.7	-53.2	-88.3	-46.8	-120.2	-64.9	-53.5	-94.5	-55.7
TPS (ns)	2,351	2,500	2,342	2,495	2,342	5,514	4,883	5,110	4,854	4,933

the wire and pin capacitance reduction, enhancing the power savings of the F2F-bonded 3-D IC.

VI. RESULTS AND ANALYSIS: CASCADE2D VERSUS COMPACT2D

A. Power and Performance Benefit

Table VIII shows the power and performance comparison of 3-D ICs implemented with Cascade2D and Compact2D flow. We use the maximum achievable frequency of the 2-D IC for AES-128, and set the target frequency for JPEG encoder 10 % higher than its maximum on purpose to examine the timing closure in the flows. We set the default parasitic scaling factor of *compact2D designs* (i.e., pseudo3D designs before partitioning the designs into two tiers; corresponds to *shrunk2D designs* in Shrunk2D flow) to 0.7.

In AES-128, we observe that both Cascade2D M3D and F2F-bonded 3-D IC present 12.2 % and 10.7 % power savings, respectively, while achieving similar worst negative slack (WNS) with their 2-D counterpart, which is similar to what is discussed in Section IV-A. We also observe slightly less power savings in Cascade2D F2F-bonded 3-D IC compared to its M3D counterpart because of the larger number of metal layers to traverse to connect top- and bottom-tier cells, and the larger size and spacing of F2F-bonds as discussed in Section V-A. For the over-clocked design, JPEG encoder, the Cascade2D M3D and F2F-bonded 3-D IC achieve even higher power savings, 15.7 % and 15.8 %, respectively, with the improved timing closure [i.e., improved WNS and total negative slack (TNS)] than its 2-D counterpart.

B. Comparison to Compact2D Flow

By comparing Cascade2D with Compact2D flow, we observe a similar trend to the comparison to Shrunk2D flow: Cascade2D M3D and F2F-bonded 3-D IC show worse wire-length savings, but better standard cell area savings. Despite of the much larger wire-length savings in the Compact2D 3-D ICs, the difference in the wire capacitance savings of the Compact2D and Cascade2D 3-D ICs is marginal. As

TABLE IX
COMPARISON OF THE NUMBER OF 3-D VIAS IN 3-D ICS IMPLEMENTED WITH COMPACT2D AND CASCADE2D FLOW

	AES-128		JPEG encoder	
	Compact2D	Cascade2D	Compact2D	Cascade2D
M3D IC	54,669	1,794	100,371	4,330
F2F-bonded 3D IC	45,188	1,794	74,134	4,330

Compact2D and Shrunk2D flow share the same partitioning algorithm (i.e., the gate-level bin-based min-cut partitioning algorithm) whereas Cascade2D flow uses block-level partitioning, the Compact2D 3-D ICs also utilize an order of larger number of 3-D vias compared to the Cascade2D 3-D ICs as shown in Table IX.

The excessive number of 3-D vias also increases the total resistance of the Compact2D 3-D ICs as well, which degrades both power consumption and timing closure. Higher resistance of wires adversely impacts the slew of the wires and, in turn, increases the internal power consumption of the subsequent cells, resulting in higher internal power in the Compact2D 3-D ICs compared to the Cascade2D counterparts. Furthermore, the larger standard cell area savings powered by accurate buffer insertion in Cascade2D flow help reduce the internal power, as well as pin capacitance switching power, which widens the internal power saving gap of the two flows even further.

The higher resistance of the Compact2D 3-D ICs is detrimental also to their timing closure because it increases the latency as well as the slew, which degrades timing closure. In Compact2D flow, post tier-partitioning optimization tries to fix timing violations after tier partitioning, but the scope is limited as it is performed on predefined placement (i.e., incremental optimization), leaving unresolved timing violations as shown in the relatively high WNS and TNS in Table VIII.

Fig. 15 compares the slack of the timing paths in JPEG encoder between 2-D and the corresponding M3D ICs. As shown in the red dots in Fig. 15(a) and (b), the Compact2D M3D IC has much more number of timing paths utilizing 3-D vias compared to the Cascade2D M3D IC, which aligns the observation in Table IX. The excessive number of 3-D vias

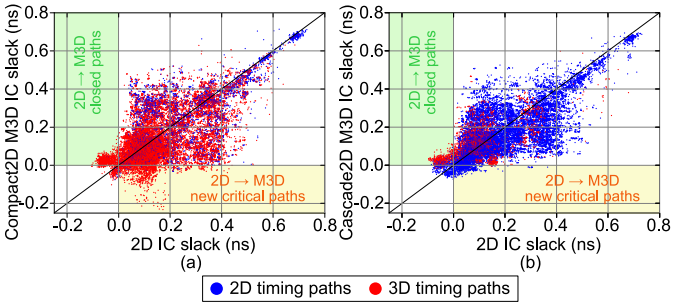

Fig. 15. Comparison of the slack of the timing paths in JPEG encoder: (a) 2-D versus Compact2D M3D IC and (b) 2-D versus Cascade2D M3D IC. Timing paths in the design are represented as dots. Blue dots indicate the timing paths which do not cross the tier boundary (i.e., 2-D timing paths), while red dots represent the timing paths crossing the tiers (i.e., 3-D timing paths).

prevents the Compact2D M3D IC from using short local wires on the xy-plane during optimization. As shown in the red dots in the yellow region in Fig. 15(a), it increases the latency and degrades the timing closure of some timing paths, especially in the designs with tight timing budgets (note that we intentionally set the target frequency of JPEG encoder ∼100 MHz higher than the maximum frequency of its 2-D IC to investigate the timing optimization behavior). The slack of those timing paths are positive (closed) in the 2-D IC, but negative (violated) in the Compact2D M3D IC.

On the other hand, in the Cascade2D M3D IC, only the timing paths that cross the functional module boundary can utilize 3-D vias, which tend to be long timing paths. As shown in the highly dense red dots in the green area in Fig. 15(b), the timing paths that cross the functional module boundary tend to be critical paths in 2-D IC (i.e., negative slack), but their timings are closed in the Cascade2D M3D IC using 3-D vias.

Although timing violations due to 3-D vias are minimized in the Cascade2D M3D IC as shown only a few red dots in the yellow region in Fig. 15(b), the proposed partitioning scheme has drawbacks.

1) As shown in the red dots with the positive 2-D IC slack in Fig. 15(b), those timing paths are noncritical but also utilize 3-D vias because they just cross the module boundaries, which is unnecessary for the timing closure (i.e., unnecessary 3-D vias).
2) There are also timing paths which do not utilize 3-D vias because they do not cross the module boundaries even though they are critical paths in the 3-D IC as shown in the blue dots with the negative Cascade2D M3D IC slack (i.e., missing D vias).

These problems stem from the lack of timing information during partitioning. Unlike Shrunk2D and Compact2D flow, which implement *shrunk2D and compact2D designs*, respectively, the functional modules in a design are partitioned in RTL before the design is implemented in the Cascade2D flow. Therefore, the timing information is not yet available during partitioning, resulting in unnecessary and missing 3-D vias.

Actually, the two-step implementation scheme of Compact2D flow (i.e., first, perform placement with *a single global* parasitic scaling factor before partitioning and then perform the incremental optimization after partitioning) has two major problems especially in cell dominated circuits: 1) failing to close timing with a low parasitic scaling factor and 2) unnecessary buffer insertion to noncritical timing path with a high parasitic scaling factor. These problems become more evident with tighter timing budgets as shown in the JPEG encoder case. In that case, the Compact2D 3-D ICs fail to close timing even with higher power than the Cascade2D 3-D ICs.

TABLE X
COMPARISON OF POWER AND TIMING METRICS BETWEEN CASCADE2D M3D IC AND COMPACT2D M3D IC WITH DIFFERENT PARASITIC SCALING FACTORS IN JPEG ENCODER

	Compact2D				Cascade2D
(parasitic scaling factor)	0.6	0.7	0.8	0.9	
std. cell area (mm)	0.573	0.584	0.580	0.588	0.590
wire-length (m)	3.419	3.489	3.258	3.327	4.347
total cap. (pF)	1,359.4	1,428.5	1,395.5	1,433.2	1,474.9
total res. (MΩ)	25.2	25.6	24.6	25.1	23.7
total power (mW)	828.4	869.1	860.8	882.2	851.8
WNS (ns)	-0.160	-0.232	-0.147	-0.148	-0.091
TNS (ns)	-90.1	-64.9	-24.8	-23.3	-53.5
TPS (ns)	4,030	4,883	5,458	5,228	5,110

Table X presents the key timing and power metrics of the Compact2D M3D ICs with different parasitic scaling factors, comparing them with the Cascade2D M3D IC. With the low parasitic scaling factor (i.e., 0.6), the Compact2D M3D IC fails to close timing as EDA tools underestimate the parasitics of the final 3-D IC, inserting insufficient buffer in the *compact2d design*. Later, post tier-partitioning optimization tries to fix the timing violations with actual parasitics, but as the optimization is performed on top of the predefined placement, it leaves unresolved timing violations as shown in the high WNS and TNS. In this case, higher parasitic scaling factor should be used, so that EDA tools can utilize enough buffers during *compact2d design* implementation before partitioning. However, the higher parasitic scaling factor makes EDA tools overestimate the parasitics even for noncritical timing paths, inserting unnecessarily large number of buffers on those paths and, in turn, resulting in high power consumption. We observe that the timing closure is improved with increasing parasitic scaling factor in *compact2D designs*, but the power consumption is drastically increasing.

On the other hand, as Cascade2D flow simultaneously implements the top and bottom tier of 3-D ICs for all the steps of the physical design from the placement to post-route optimization, it can harness all the optimization features that EDA tools offer, inserting buffers only on the timing paths with tight timing constraints. The yellow and orange bars in Fig. 16 show the slack histogram of the timing paths in the Compact2D M3D ICs implemented with the parasitic scale factor of 0.7 and 0.9. We observe that increasing the parasitic scale factor not only reduces the number of timing paths with negative slack but also increases the number of timing paths with high positive slack. (i.e., all the positive slack bins are increased with the increased parasitic scale factor). This is due to the increased parasitic scale factor is applied to all the timing paths in the design (i.e., global), making EDA tools insert more buffers even in high positive slack timing paths.

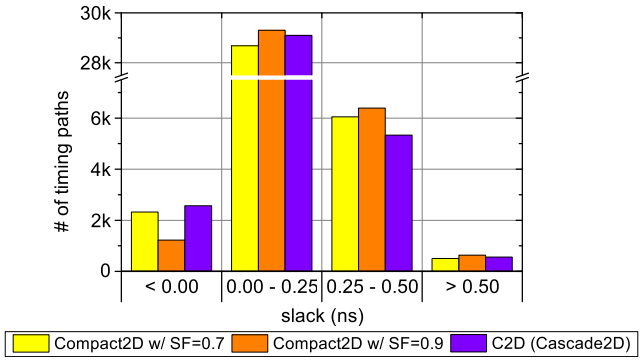

Fig. 16. Histogram of the slack of the timing paths of Cascade2D M3D IC and Compact2D M3D IC with parasitic scaling factor = 0.7 and 0.9 in Table X. Note that WNS (Compact2D w/ SF = 0.7) = − 0.232 ns, WNS(Compact2D w/ SF = 0.9) = −0.148 ns, and WNS (Cascade2D) = −0.091 ns.

In contrast, we observe less number of timing paths in high positive slack timing paths while achieving less WNS in the Cascade2D M3D IC (−0.091 ns) compared to the Compact2D M3D ICs (−0.232 ns when SF = 0.7 and −0.148 ns when SF = 0.9), which indicates that the Cascade2D flow resolves the timing closure by inserting buffers only on the critical timing paths.

VII. Future Works

As a future direction of our studies, we can first improve the design-aware partitioning scheme to use timing information from the corresponding 2-D IC, so that unnecessary and missing 3-D vias discussed in Section VI-B are minimized.

In addition, 3-D floorplanning methodologies can be explored as a part of the design-aware partitioning stage, which would help improve the quality of tier partitioning.

Next, we can generalize Cascade2D flow to 3-D ICs with more than two tiers. The work involves devising a partitioning method, which supports more than two partitions while maximizing timing paths and balancing cell area. In addition, as the locations of 3-D vias connecting a pair of two neighboring tiers can affect those connecting another pair, a new optimization methodology is required for 3-D Via Planning Stage. Furthermore, 3-D via locations need to take into account the nets which traverse more than two tiers skipping some of the tiers in the middle. In the Cascade2D stage, an efficient way to manage dummy wires is needed as there can be significant congestion on dummy wire layers on middle tiers.

Anchor cells also require improvements to further optimize designs as the area taken by anchor cells would degrade the standard cell resizing and buffer insertion during optimization especially in 3-D ICs with more than two tiers because the number of anchor cells in middle tiers can be roughly doubled.

Finally, in order to improve the wire-length savings, the Cascade2D stage can be improved by utilizing multiple 3-D vias for a net as in Shrunk2D and Compact2D flow.

VIII. Conclusion

In this article, we presented a new methodology called "Cascade2D flow" to implement 3-D ICs using 2-D commercial tools. Cascade2D flow utilizes a design-aware partitioning scheme, where it gives designers ability to control over the partitioning functional modules while partitioning heavily communicating functional modules into separate tiers. One of the main advantages of this flow is that it is extremely flexible and is partition-scheme agnostic, making it an ideal platform to evaluate different 3-D IC partitioning algorithms. 3-D vias are modeled as sets of anchor cells and dummy wires, which enable us to implement and optimize both the top and bottom tiers simultaneously in a single 2-D design. Cascade2D flow reduces the standard cell area effectively, resulting in significantly better power savings than the state-of-the-art 3-D IC design flows developed previously. Experimental results with a commercial, in-order, and 32-bit application processors in foundry 28, 14/16, and predictive 7-nm technology nodes show that Cascade2D 3-D ICs can achieve up to 4× better power savings compared to the state-of-the-art 3-D IC design flows, while using an order of magnitude less 3-D vias. In the best case scenario, the 3-D ICs implemented using this new methodology result in 25% higher performance at iso-power and up to 20% power reduction at iso-performance compared to the 2-D ICs. By leveraging a smaller standard cell area, we demonstrate that 3-D ICs can save up to 10 % die-area, which directly translates to reduced costs. Additionally, we compared two flavors of 3-D ICs: 1) M3D and 2) F2F-bonded 3-D ICs, to examine and analyze the power benefits from the new methodology, and observe that M3D ICs benefit more especially in advanced technology nodes because of the fine granularity and scalability of MIVs. We hope that this work paves the way for more research to combat manufacturing, thermal, process variation and EDA tool challenges associated with this novel technology.

References

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

[2] O. Thomas, M. Vinet, O. Rozeau, P. Batude, and A. Valentian, "Compact 6T SRAM cell with robust read/write stabilizing design in 45 nm monolithic 3D IC technology," in *Proc. Int. Conf. IC Design Technol.*, Austin, TX, USA, 2009, pp. 195–198.

[3] S. A. 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.

[4] K. Chang, S. Pentapati, D. E. Shim, and S. K. Lim, "Road to high-performance 3D ICs: Performance optimization methodologies for monolithic 3D ICs," in *Proc. Int. Symp. Low Power Electron. Design*, 2018, pp. 1–6.

[5] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A physical design methodology to build commercial-quality face-to-face-bonded 3D ICs," in *Proc. Int. Symp. Phys. Design*, 2018, pp. 90–97.

[6] O. Billoint *et al.*, "A comprehensive study of monolithic 3D cell on cell design using commercial 2D tool," in *Proc. Design Autom. Test Eur. Conf. Exhibit.*, Grenoble, France, 2015, pp. 1192–1196.

[7] O. Billoint, M. Brocard, S. Thuries, G. Berhault, and H. Sarhan, "A partitioning-free methodology for optimized gate-level monolithic 3D designs," in *Proc. SOI 3D Subthreshold Microelectron. Technol. Unified Conf.*, Burlingame, CA, USA, 2017, pp. 1–2.

[8] 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. 66th Electron. Compon. Technol. Conf.*, Las Vegas, NV, USA, 2016, pp. 1179–1185.

[9] E. Beyne, "The 3-D interconnect technology landscape," *IEEE Design Test*, vol. 33, no. 3, pp. 8–20, Jun. 2016.

[10] K. Chang, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "Match-making for monolithic 3D IC: Finding the right technology node," in *Proc. Design Autom. Conf.*, Austin, TX, USA 2016, pp. 1–6.

[11] S. H. Yang *et al.*, "28 nm metal-gate high-K CMOS SoC technology for high-performance mobile applications," in *Proc. Custom Integr. Circuits Conf.*, San Jose, CA, USA, 2011, pp. 1–5.
[12] S.-Y. Wu *et al.*, "A 16 nm FinFET CMOS technology for mobile SoC and computing applications," in *Proc. Int. Electron Devices Meeting*, Washington, DC, USA, 2013, pp. 1–4.
[13] T. Song *et al.*, "A 14 nm FinFET 128 Mb SRAM with V_{MIN} enhancement techniques for low-power applications," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 158–169, Jan. 2015.
[14] K.-I. Seo *et al.*, "A 10 nm platform technology for low power and high performance application featuring FINFET devices with multi Workfunction gate stack on bulk and SOI," in *Proc. Symp. VLSI Technol.*, Honolulu, HI, USA, 2014, pp. 1–2.
[15] E. Beyne *et al.*, "Scalable, sub 2 μm pitch, Cu/SiCN to Cu/SiCN hybrid wafer-to-wafer bonding technology," in *Proc. Int. Electron Devices Meeting*, San Francisco, CA, USA, 2017, pp. 1–4.
[16] P. Batude *et al.*, "3DVLSI with CoolCube process: An alternative path to scaling," in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, 2015, pp. T48–T49.

Kyungwook Chang 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.

He joined the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea, in 2021, where he serves as an Assistant Professor. His current research interests include computer-aided design solutions for 3-D ICs, design-technology co-optimization, physical/logic design and analysis, power delivery network, and parallel computing/memory architecture.

Saurabh Sinha received the B.Tech. degree in electronics and instrumentation engineering from the National Institute of Technology Rourkela, Rourkela, India in 2006, and the M.S. and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2008 and 2011, respectively.

He is currently employed as a Principal Research Engineer with ARM Research, Austin, TX, USA. His research interests include predictive technology, design-technology co-optimization at advanced technology nodes, and disruptive technologies, such as 3-D-ICs and their impact on design.

Brian Cline received the B.S. degree in electrical engineering from the University of Texas, Austin, TX, USA, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA, in 2006 and 2010, respectively.

From 2006 to 2010, he was a Graduate Fellow with Semiconductor Research Corporation (SRC), Durham, NC, USA. He is a Principal Research Engineer with the Devices, Circuits, and Systems Research Group, ARM, Austin. He has published more than 25 papers, including a number of best paper award nominees, invited papers, and plenary sessions, and has filed more than 25 patents. His research interests and responsibilities include the circuits-level and system-level design impact of emerging logic devices (primarily post-CMOS), memory devices, next-generation lithography (e.g., EUV, DSA, etc.), and next-generation integration strategies (e.g., fine-grained 3-D bonding, monolithic/sequential 3-D, etc.). He also has interests in electronic design automation tools and methodologies and design-technology co-optimization, mainly with respect to high-speed, power-efficient digital logic design.

Dr. Cline was chosen as one of the recipients of the 2017 SRC Mahboob Khan Outstanding Industry Liaison Award.

Greg Yeric received the Ph.D. degree in microelectronics from the University of Texas at Austin, Austin, TX, USA, in 1993.

He joined Motorola's Advanced Products Research and Development Laboratories in embedded nonvolatile memory process integration, subsequently researching on technology development roles at TestChip Technologies, Plano, TX, USA, HPL Technologies, San Jose, CA, USA, and Synopsys, Mountain View, CA, USA. He then spent 12 years with Arm Research, Austin, eventually as a Fellow focusing on design-technology co-optimization and predictive technology. In October 2020, he co-founded Cerfe Labs, Austin, a spin-out from Arm Holdings focusing on materials and device research including novel nonvolatile memory Research and Development, where he serves as the CTO.

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 (2012) is generally considered the first multicore 3-D processor ever developed in academia. He has authored the book titled *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 Award in 2006. He was on the advisory board of the ACM Special Interest Group on Design Automation from 2003 to 2008 and awarded a Distinguished Service Award in 2008. He received the Best Paper Awards from the IEEE Asian Test Symposium in 2012 and the IEEE International Interconnect Technology Conference in 2014. He received 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.