Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs

Lennart Bamberg, Alberto García-Ortiz
ITEM, University of Bremen
{bamberg, agarcia}@uni-bremen.de

Lingjun Zhu, Sai Pentapati, Da Eun Shim, Sung Kyu Lim
GTCAD, Georgia Institute of Technology
{lingjun, sai.pentapati, daeun, limsk}@ece.gatech.edu

***Abstract*—Memory-on-logic and sensor-on-logic face-to-face stacking are emerging design approaches that promise a significant increase in the performance of modern systems-on-chip at reasonable costs. In this work, a netlist-to-layout design flow for such heterogeneous 3D systems is proposed. The proposed technique overcomes the severe limitations of existing 3D physical design methodologies. A RISC-V-based multi-core system, implemented in a commercial technology, is used as a case study to evaluate the proposed design flow. The case study is performed for modern/large and small cache sizes to show the superiority of the proposed methodology for a broad set of systems. While previous 3D design flows do not show to optimize performance against 2D baseline designs for processor systems with a significant memory area occupation, the proposed flow shows a performance and power improvement by 20.4–28.2 % and 3.2–3.8 %, respectively.**

I. Introduction

Integration into the third dimension enables scaling beyond the end of Moore's law. Three main variants of 3D integrated circuits (ICs) exist: Through-silicon-via (TSV) based, monolithic, and face-to-face stacked [1]. The relatively large geometrical dimensions of today's TSVs and their low manufacturing yield limit TSV-based 3D stacking to designs with a small number of inter-die connections [2]. Monolithic 3D integration is an emerging alternative where the 3D system is fabricated sequentially, instead of stacking prefabricated 2D dies, which enables fine-grain 3D interconnects, overcoming problems caused by the TSV area occupation. However, the manufacturing yield and cost of monolithic 3D ICs are currently even worse than that of TSV-based 3D ICs. While sophisticated manufacturing techniques—developed over decades for 2D ICs—are used to form transistors and metal layers in stacked 3D ICs, new process steps are required for monolithic 3D integration. After producing the first tier of a monolithic 3D IC, only relatively low temperatures can be used to form subsequent tiers to prevent degradation of already manufactured metal layers [3].

The preferred approach until technology advances is face-to-face (F2F) stacking. A F2F stack is made up of two prefabricated 2D dies connected through the topmost metal layers of both dies using a face-to-face bonding technology. Since F2F bonding bumps are much smaller and easier to manufacture than TSVs, F2F stacking enables a high 3D interconnect density at a low cost [4].

Like all 3D-integration techniques, F2F stacking enables heterogeneous integration. While one die can be manufactured in an aggressively scaled technology node to integrate semi-custom digital components made up of standard cells, the other die can be used to integrate specific types of full-custom components. This heterogeneity brings numerous advantages as the technology of the second die can be optimized solely for the needs of the full-custom components. A well-known example of heterogeneous integration is memory-on-logic stacking, where the second die is dedicated exclusively to memory blocks. Thereby, the memory is no longer constrained by being process compatible with logic as the two dies are fabricated separately. Another example is sensor-on-logic integration. In contrast to logic components, sensors, and other analog/mixed-signal components typically do not benefit from using ultimately scaled technology nodes. Thus, in a sensor-on-logic stack, the sensing die can be integrated into a larger technology node than the logic die. Hence, heterogeneous 3D integration promises significantly better power, performance, area, and cost than homogeneous 3D integration.

State-of-the-art physical design methodologies for homogeneous F2F-stacked 3D ICs are based on commercial 2D placement, routing, and sign-off tools to produce commercial quality 3D IC layouts [5], [6]. However, the design methodologies are not suited for heterogeneous memory-on-logic or sensor-on-logic 3D integration. To overcome this problem, this work presents a physical design methodology for this specific kind of F2F-stacked 3D ICs, named *Macro-3D*. The flow enables to build high-performance memory-on-logic and sensor-on-logic 3D ICs. Moreover, the proposed design methodology is the first one that uses commercial 2D electronic design automation (EDA) tools without extending them substantially to produce a valid 3D placement and routing. This fact significantly improves the layout quality for the proposed flow.

A case study on memory-on-logic stacking for a tape-out-proven RISC-V multi-core system shows the strong superiority of the proposed flow and the design style. While the previous 3D physical design methodologies do not allow to increase the performance against 2D baseline designs, the proposed flow shows a performance/timing improvement by up to 28.2 %.

II. Macro-on-Logic 3D Integration

In this section, a specific type of heterogeneous 3D integration is introduced: Macro-on-logic (MoL) F2F stacking. The floorplan, as well as the cross-view of a MoL 3D IC and a logically equivalent 2D IC are illustrated in Fig. 1. In an F2F-stacked IC, two prefabricated 2D dies are bonded in a face-to-face manner. Thereby the electrical connections between the dies are established through F2F bumps. With F2F stacking, the same substrate area is available for active circuit elements as in

978-3-9819263-4-7/DATE20/©2020 EDAA

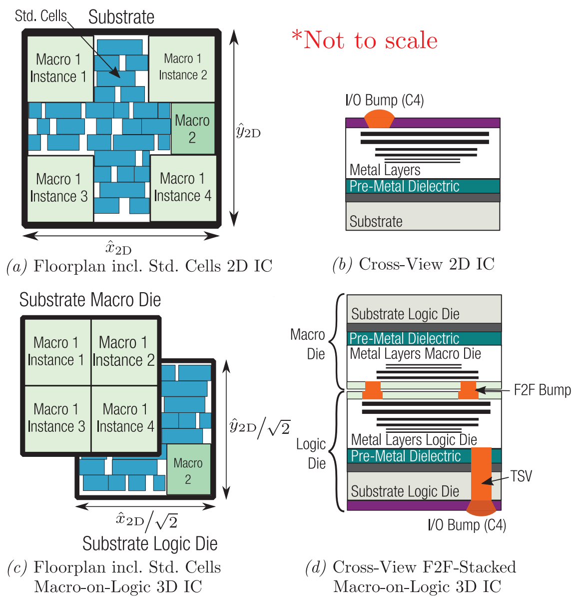


Fig. 1. 2D IC and a logically equivalent F2F-stacked MoL 3D IC.

a traditional 2D system with a footprint that is $2\times$ larger. Thus, without increasing the substrate integration density, the x and y size of a system can be reduced by a factor of $\sqrt{2}\times$ through F2F stacking. Hence, F2F stacking reduces the maximum half-perimeter wire length (HPWL) by almost 30 % and thus shows vast potential to improve the system performance and energy dissipation for aggressively scaled technologies, in which interconnects are a bottleneck. However, this requires a minimum F2F-bump pitch in the range of the wire spacing. One promising solution is hybrid wafer-to-wafer bonding, which enables direct metal-to-metal/dielectric-to-dielectric bonding between the back end of lines (BEOLs) of two fabricated wafers. With this technique, F2F-bump pitches below 1 μm are possible due to the precise wafer-level integration [1]. This enables drastic 3D integration gains due to reduced wire lengths for a wide range of products.

An additional driver for 3D integration is heterogeneous integration. In a 2D system-on-chip (SoC), every component needs to be integrated within the same substrate/technology, and only one global BEOL structure exists. Stacking two dies provides the design-flexibility to have two different substrates and BEOLs. For the integration of semi-custom digital components, constructed from standard cells, using the most aggressively scaled technology is advantageous, but it demands a large number of metal layers. However, for full-custom components (*e.g.*, memories, analog-digital converters), it can be advantageous to use a different technology or a substrate with different physical characteristics. Furthermore, the routing of most full-custom components requires fewer metal layers, even if integrated into the same technology node, due to the more regular wiring, which has the potential to reduce manufacturing cost. Thus, adding the heterogeneity that semi-custom digital components made up of standard cells are only integrated into one of the dies has the potential to boost the gain of F2F stacking. Afterward, the second die can have a different BEOL and is used only to integrate full-custom components that appear in the EDA flow as regular macro blocks. This specific approach of heterogeneous integration is denoted as macro-on-logic (MoL) stacking throughout this work, as, in the resulting layout, standard-logic cells are only placed into the bottom die while the top one only includes macro blocks. The structure of an MoL stack is illustrated in Fig. 1*(c)–(d)*. Note, that in the die where the standard cells are placed (logic die) macros can still be placed (an analysis of modern multi-core systems shows that even with relatively small cache sizes macros occupy more than 50 % of the area).

In an MoL stack, designers can change the technology for the design of the full-custom components in the macro die as long as the interface (*e.g.*, power supply voltage) is compatible with the standard cells in the other die. However, such changes do not affect the physical design of digital systems, during which all blocks (*i.e.*, macros and standard cells) are treated as black boxes. Thus, only a heterogeneity between the BEOLs of the two dies is considered in the following.

III. Motivation & Fundamental Idea

Two physical design methodologies for F2F-stacked 3D ICs have been proposed[1]: *Shrunk-2D* (*S2D*) [5] and *Compact-2D* (*C2D*) [6]. Both are based on 2D EDA tools to overcome the lack of true commercial-grade 3D EDA tools. In *S2D*, all standard cells and interconnect dimensions are initially shrunk by 50 %. Floorplanned macros are replaced by placement blockages. At an *(x,y)* location of the full stack, where a macro is placed in one of the two dies, the 2D place-and-route (P&R) tool considers a blockage of 50 %. Where macros are placed in both dies, the tool considers a full blockage. This allows the complete design to fit into a 2D floorplan with a footprint of the target two-die design. Afterward, the shrunk cells are placed and routed in this so-called *S2D Design*—having the BEOL of one die—while taking care of the macro blockages. Thereby, the shrunk cells are placed and routed with the same HPWL and metal-layer utilization as the target F2F design. However, this requires equal BEOLs in both dies of the F2F stack. After 2D P&R, tier partitioning is performed to determine the (*z*)/die location of the cells resized to their original size. Afterward, F2F-via planning decides the actual F2F-bump locations and the true inter-die routing based on the *(x,y,z)* placements.

Compact-2D overcomes the issue of *S2D* that shrinking cells and routing geometries is not possible for ultimately scaled technology nodes as it requires 2D P&R engines for a future technology node. Furthermore, post-tier-partitioning optimization is added. The core idea of *C2D* is to increase the floorplan footprint by a factor of $2\times$ compared to the final F2F stack. For a good estimate of the wire parasitics of the target 3D design, despite the increased floorplan, the interconnect parasitics per unit length are reduced by a factor of $\sqrt{2}\times$. Again, partial and full blockages are added to take macros into

[1]Note that die-by-die routing methodologies are not suitable for the F2F-bump counts hybrid wafer-to-wafer bonding enables.

account. However, here, the blockage areas are increased by a factor of 2× to account for the increased floorplan footprint. After P&R of the standard cells in the increased floorplan is done, the cell locations are linearly mapped to the 3D design to obtain the *(x,y)* locations of the cells in the F2F design. Afterward, tier partitioning is performed, followed by F2F-via planning like in *S2D*. Finally, post-tier-partitioning optimization and incremental routing are done.

Although both flows have shown promising results, they have significant drawbacks. At first, shrinking the interconnect dimensions or parasitics leads to inaccuracies in the estimated wire parasitics for the intermediate *S2D*/*C2D Design*. Many routes are included in the designs that do not exist in the final F2F design. Furthermore, the nature of the full, double metal stack, including the F2F bumps, is not considered. If interconnected cells are located next to each other in the *C2D*/*S2D Design*, but in different dies in the final F2F design, the parasitics change drastically due to the large interconnects parasitics in modern technologies. Generally, the whole initial P&R of the 2D tools does not represent reality. Therefore, timing can be heavily mispredicted, resulting in many paths being over-optimized (*e.g.*, too large buffers) or under-optimized (*e.g.*, too small buffers). This also requires a new routing after tier partitioning, degrading the performance compared to the *S2D*/*C2D Design*, as this second routing cannot be fully co-optimized with the placement. Furthermore, existing tools do not optimize for different BEOLs in the dies.

For macros in the design, the situation becomes worse as 50 % blockages lead to large errors in the predicted (*x,y*) cell locations compared to the final F2F design. Our experiments with the tools showed that the spatial resolution used by commercial 2D P&R tools to take care of partial blockages is not fine enough, resulting in many overlaps after tier partitioning. Fixing these overlaps again showed huge performance degradations. Also, the macro routing parasitics are estimated rather poorly. For example, for MoL integration, during the *S2D* or *C2D* stage, the P&R tool considers the macro pins as located within the same BEOL as the standard cell pins. However, in the final 3D design, the macro pins will be located in the other die, which increases the parasitics due to added vias and wires. Furthermore, this results in routing congestions that are not predictable at earlier stages. In fact, our experiments show that neither *S2D* nor *C2D* for a homogeneous or heterogeneous 3D integration scheme can improve the maximum performance of modern multi-cache-level processor systems compared to baseline 2D designs due to the large area occupation of macro cells (see later Sec. V). This is contrary to homogeneous designs mainly made up of standard cells where both flows, *S2D* and *C2D*, showed good performance improvements over 2D designs [5], [6].

To overcome the previously outlined issues, this work proposes to exploit the specific structure of an MoL stack such that nothing is modified (*e.g.*, shrunk) that still has to be placed or routed, and no partial blockages are used while still only standard 2D EDA tools are required for P&R and power-performance-area (PPA) analysis. Furthermore, the 2D EDA tools will be given a combined BEOL that represents the full metal stack of the two stacked dies, including the F2F vias. In other words, for the first time, the 2D EDA tools are tricked to seeing the physical reality for P&R despite only allowing for one substrate. Thereby, in contrast to all previous 3D flows, the P&R and PPA results of the 2D tools are directly equal to the final ones for the target 3D stack, and no further steps such as partitioning, F2F-via planing, or inter-die/incremental routing are required. Furthermore, the highly-optimized 2D routing engines take care of the F2F-via planning, which also enables routing paths starting and ending in the same die but still traversing the other die to avoid congestions. This increases routability compared to previous approaches.

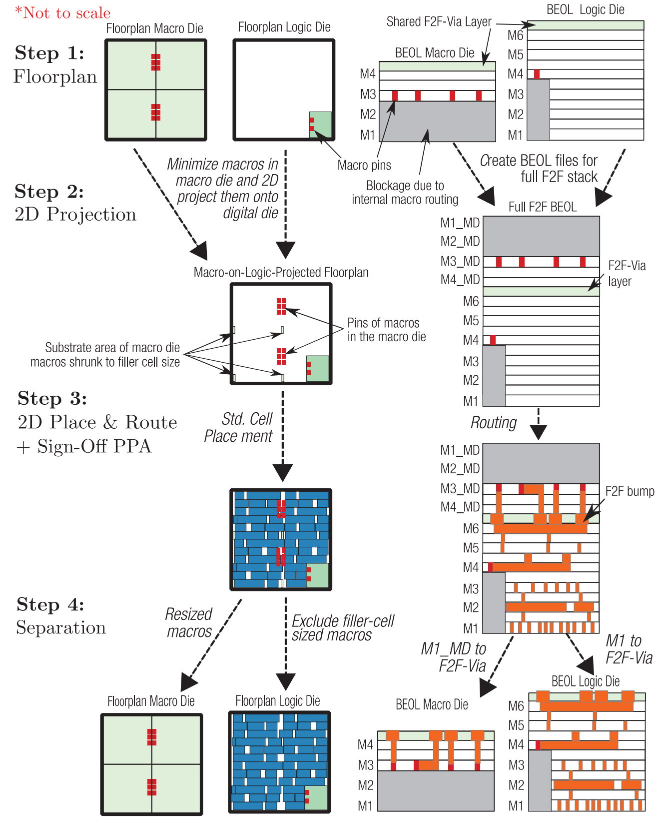


Fig. 2. Proposed design methodology to design F2F-stacked MoL 3D ICs.

IV. Proposed Design Methodology

This section presents *Macro-3D*, a physical design methodology for commercial-quality MoL-stacked 3D ICs. The flow consists of four main steps, illustrated in Fig. 2. First, two separate 2D floorplans with the same footprint as the final F2F stack are generated: One for the pure macro die, and one for the logic die (which can include macros as well). Afterward, the macro blocks are placed in these floorplans.

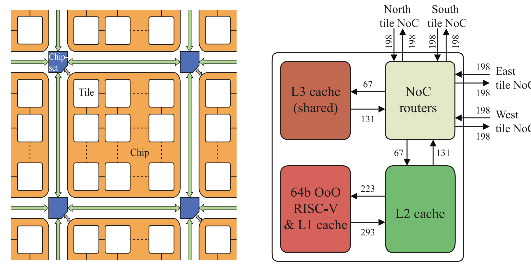
Second, a memory-on-logic-projected 2D floorplan is generated from the perspective of the logic die. To obtain a 2D P&R result that remains valid for the final F2F design and considers the correct net parasitics, the BEOL of the full metal stack is generated in the form of *tch* files for parasitic extraction (one for each corner) and a *techlef* file for the abstract view of the layers. Since layers need unique names, the layers of the macro die are extended by the suffix "*_MD*". For example, if the logic die has six metal layers (*M1* to *M6*) and the macro die four (*M1* to *M4*), the layer order is: *M1*→*VIA12* … →*M6* →*F2F_VIA*→*M1_MD*→*VIA12_MD*…→*M4_MD*. Since the macros in the macro die occupy no space in the logic die, their substrate area is shrunk to the minimum possible size, which is the size of a filler cell (note that commercial tools do not allow a substrate area of 0). The pin layers of the macro-die macros are edited to represent the new naming of the metal layers (*e.g.*, *M3* macro-pin layer definitions are modified to *M3_MD*). The same modification is done for the layers of the routing blockages due to the internal routing within the macros in the macro die. The (*x,y*) boundaries of the macro pins and routing blockages are left unmodified. All changes are done by simple scripted modifications in the *lef* files of the related macros. Afterward, the floorplan representing the logic die, and the one with the shrunk macros and edited layers, are superimposed to a single 2D floorplan.

In the third step, this floorplan and the combined double-stack BEOL are fed into a standard 2D P&R engine. Since this engine sees all macro pins at the correct positions, has the full BEOL of the whole F2F stack for routing and parasitic extraction, and the correct area to place standard cells, it produces a standard-cell placement and routing that remains valid for the target MoL 3D structure. This further implies that the placed and routed *Macro-3D* design can be used to obtain PPA values with standard 2D sign-off tools that are valid for the final F2F-stacked 3D design.

Finally, the design is separated again into two parts to generate the individual *GDSII* files required for production. The logic die contains all substrate objects, but the filler-cell-sized macros, physically located in the macro die. Furthermore, it includes the metal layers of the logic die (layers *M1* to *M6* in the previous example), and the F2F bumps (*F2F_VIA* layer). The macro die includes the related macros rescaled to their original size, the metal layers of the macro die (*M1_MD* to *M4_MD* in the previous example) and the F2F bumps. Thus, the *F2F_VIA* layer is included in both parts.

V. Evaluation

In this work, *OpenPiton* [7], a tape-out-proven multi-core system, is used as the benchmark architecture. It is highly configurable, as the core count, cache sizes, *etc.* can be defined arbitrarily. The *OpenPiton* system is shown in Fig. 3(a). A full system consists of at least one chip, made up of multiple tiles. Thus, a tile is an atomic piece out of which systems with arbitrary core counts can be constructed. Hence, the tile design is analyzed while ensuring a correct functionality when multiple tiles are instantiated to create large systems (more details in Sec. V-1). Thereby, the reported results are valid for systems with arbitrary core counts. The tile architecture is illustrated in Fig. 3(b). It consists of a 64-bit out-of-order (OoO) RISC-V Ariane core and three cache levels (L1–L3). The first two levels are private to the individual cores, while the



(a) Full System *(b)* Tile Architecture

Fig. 3. OpenPiton architecture (adopted from [7]).

third level cache is coherently shared among all cores. Three parallel on-chip networks (NoCs) are used as the scaleable inter-tile communication architecture.

Two tile architectures with different cache sizes are analyzed: A tile with a modern/large cache system with 16 kB of L1 instruction and data cache, 128 kB of L2 cache, and 1 MB of L3 cache per tile; and a small-cache tile, including 8 kB of L1 instruction cache, 16 kB of L1 data cache, 16 kB of L2 cache, and 256 kB of L3 cache per tile. Gate-level syntheses show that, even for the small cache sizes, memory macros occupy more than 50 % of the substrate area, showing the suitability of MoL stacking for a wide range of standard products.

1) Design Setup: In the tile designs, the inter-tile interconnects must be captured through constraints, because those paths start in one tile and end in another. Consider an exemplary NoC path starting in one tile instance and ending in the north adjacent tile instance. This path is represented in the tile design by a path starting at an NoC register and ending at a north output pin, combined with a path starting at a south input pin ending at another NoC register. Thus, both paths together have to finish in one clock cycle, and the north output pin and the associated south input pin locations have to be aligned such that the tile instances can be connected without additional routing. Thus, in the tile design, all pins are located in *M6*, input and output pins of inter-tile paths are constrained with a half-cycle delay, and associated output-input pin pairs have the same x location at the north-south edges or the same y location at the east-west edges. This ensures timing closure for systems with arbitrary tile counts. After P&R of a design, full-chip static timing/power analysis is performed to obtain the PPA values. Thereby, the toggle ratio per clock cycle for inputs and registers is set to 0.2. The maximum achievable clock frequency for a design is here used as the performance metric.

2) Tool and Technology Setup: A commercial 28-nm, high-κ metal-gate, planar technology is used for the physical design performed with *Cadence* tools. Multiple process corners are considered while timing closure is done at the slowest corner, and power is reported at the typical corner. In the full/double BEOL of the whole F2F stack, the F2F bumps are included as vias. The minimum-pitch, size, and height of these F2F vias are chosen as 1 um, 0.5 um $\times$ 0.5 um, and 0.17 um, respectively, based on [1] and the BEOL of the used 28-nm technology. According to extraction results for the typical corner, the mean resistance and capacitance of a F2F via/bump is 44 mΩ and 1.0 fF, respectively.

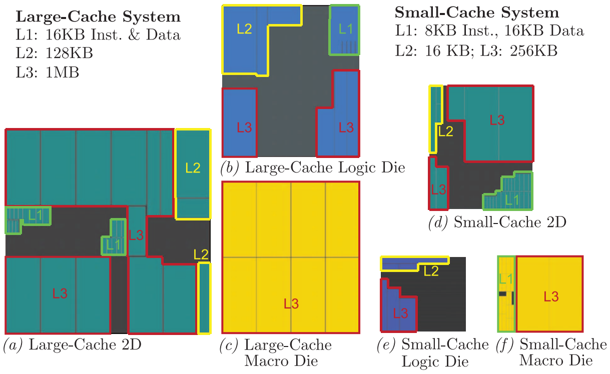


Fig. 4. Memory-macro floorplans of the 2D and the MoL 3D designs.

Table I
MAX-PERFORMANCE PPA AND COST COMPARISON OF THE 2D AND THE 3D DESIGNS FOR THE SMALL-CACHE SYSTEM.

	2D	MoL *S2D* [5]	BF *S2D* [5]	*Macro-3D*
f_{clk} [MHz]	390	227	260	470
E_{mean}* [pJ/cycle]	116.7	123.1	112.9	117.6
$A_{footprint}$ [$(mm)^2$]	1.20	0.60	0.60	0.60
F2F bumps	0	5405	8703	4740

* Equivalent to power-per-megahertz.

Floorplans with placed memory macros are created. For a fair comparison, the area ratio between the footprints of a 2D and the related 3D floorplan is $2\times$, so the same silicon area is available in 2D and 3D. The macro floorplans for the 2D and MoL designs are shown in Fig. 4. To reach the maximum performance, all 2D and 3D floorplans are highly optimized by considering the tile architecture and P&R results for multiple floorplan alternatives.

A. Analysis and Results

First, a max-performance comparison between the proposed 3D physical design flow, the state-of-the-art 3D physical design flows (*S2D*/*C2D*), and the standard 2D design flow is drawn. Therefore, P&R is executed for all designs with six metal layers (per die for 3D designs) to have the same metal capacities in all designs, ensuring a fair comparison. The results reveal that, for designs with a significant amount of macros, *S2D* performs significantly better than *C2D*. As expected, previous 3D physical design flows perform better for the small-cache system, due to the lower macro-over-standard-cell area ratio. Thus, only results for *S2D* and the small-cache architecture are reported here. As outlined in Sec. III, *S2D* and *C2D* perform better if the number of partial macro blockages is minimized. Thus, a second balanced floorplan (BF) is created for the *S2D* flow in which memory blocks overlap as much as possible (resulting in more full blockages).[2] Thereby, best-case values are reported for the previous 3D design methodologies. In Table I, the resulting max-performance PPA and manufacturing-cost metrics are presented. Even in the best-case scenario (BF *S2D*), the previous 3D flows results in a 33.3 % lower maximum clock frequency, f_{clk}, than the baseline 2D design, due to the limitations outlined in Sec. III. If the process advantages of MoL stacking are wanted, BF *S2D* cannot be used, and the performance degradation is as high as 41.8 %. This dramatically lower max-performance does not even result in a significantly improved energy dissipation per instruction, quantified by E_{mean}. In contrast, the proposed *Macro-3D* flow results in a max-performance increase of 20.5 %, against the 2D design, without increasing the energy consumption noticeably. Furthermore, this is obtained at lower manufacturing cost than previous 3D flows, not only due to the intrinsic advantages of MoL stacking but also due to the reduced number of F2F bumps (−45.5 %) with the same die footprint area, $A_{footprint}$.

[2]Note that for a balanced floorplan, the manufacturing/design advantages of MoL stacking are lost.

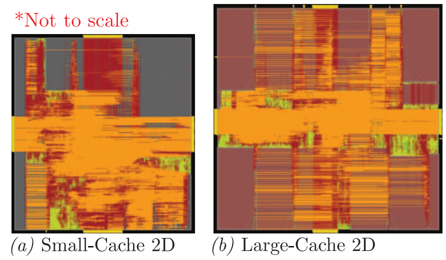


Fig. 5. Final placed and routed 2D-design layouts

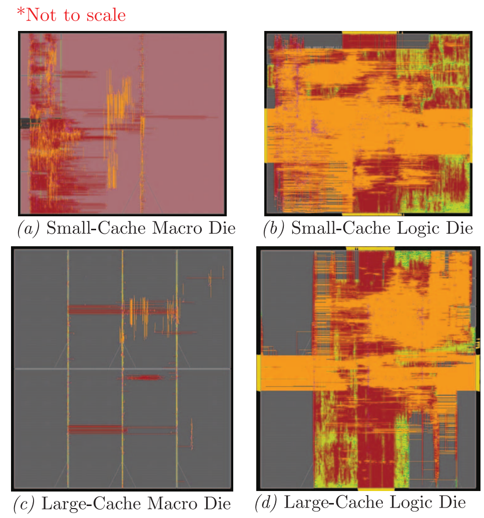


Fig. 6. Final placed and routed MoL layouts of the macro die and the logic die resulting from the proposed *Macro-3D* flow (red dots indicate F2F-bump).

Due to the previously outlined performance degradation of existing 3D flows for macro-heavy designs, the following in-depth analysis and discussion are limited to comparing the *Macro-3D* designs with the related 2D designs. The layouts of the baseline 2D designs and the *Macro-3D* designs are shown in Fig. 5 and 6, respectively. In Table II, the values for the in-depth comparison are reported. The results show that the maximum frequency of the *Macro-3D* MoL design is even 28.2 % higher than the 2D baseline design for the large-cache tile. The huge performance improvements of the proposed flow for both tile structures are due to the smaller footprints, resulting in lower wire lengths/parasitics and better clock-tree characteristics. This also has a huge impact on the wirelength of the critical path, especially for the small-cache system,

Table II
IN-DEPTH COMPARISON OF 2D AND THE PROPOSED MACRO-3D DESIGNS.

	Small-Cache		Large-Cache	
	2D	*Macro-3D*	2D	*Macro-3D*
f_{clk} [MHz]	390	470 (**+20.5%**)	328	421 (**+28.2%**)
E_{mean} [nJ/cycle]	116.7	117.6 (**+0.8%**)	369.3	366.1 (**-0.9%**)
$A_{\text{footprint}}$ [$(mm)^2$]	1.20	0.60 (**-50.0%**)	3.88	1.94 (**-50.1%**)
$A_{\text{logic-cells}}$ [$(mm)^2$]	0.29	0.30 (**+1.6%**)	0.47	0.47 (**+1.2%**)
Total wirelength [m]	6.3	5.6 (**-11.8%**)	12.2	10.4 (**-14.8%**)
F2F bumps	0	4740	0	1215
$C_{\text{pin,total}}$ [nF]	0.36	0.38 (**+5.6%**)	0.52	0.56 (**+7.4%**)
$C_{\text{wire,total}}$ [nF]	0.89	0.83 (**-7.2%**)	1.61	1.44 (**-10.2%**)
Max. clk.-tree depth	13	14 (**+7.7%**)	20	16 (**-20.0%**)
Crit.-path wirelength [mm]	1.49	0.55 (**-63.0%**)	2.21	1.50 (**-32.0%**)

Table III
IMPACT OF REMOVING TWO METAL LAYERS OF THE MACRO DIE ON THE MAX-PERFORMANCE PPA AND COST METRICS.

	Small-cache		Large-cache	
	Macro-3D M6–M6	*Macro-3D M6–M4*	*Macro-3D M6–M6*	*Macro-3D M6–M4*
f_{clk} [MHz]	470	462 (**-1.8%**)	421	423 (**+0.5%**)
E_{mean} [nJ/cycle]	117.6	119.0 (**+1.3%**)	366.1	362.5 (**-1.0%**)
A_{metal} [$(mm)^2$]	7.20	6.0 (**-16.7%**)	23.3	19.4 (**-16.7%**)
F2F bumps	4740	3866 (**-18.4%**)	1215	922 (**-24.1%**)

where the critical path starts at a flip-flop and ends at a memory block in 2D. Such critical paths do not occur for MoL stacking, where memory blocks can be placed above standard logic. The standard-cell areas, $A_{\text{logic-cells}}$, and total pin capacitances, $C_{\text{pin,total}}$, are slightly increased for the *Macro-3D* designs when compared to the 2D designs, due to the higher drive strengths of some cells required for the increased clock frequencies. Despite the increased performance, *Macro-3D* designs, on average, do not consume more energy per cycle. When the *Macro-3D* designs are re-implemented for the same target frequency (iso-performance) as the 2D designs (328 MHz), the proposed flow shows power consumption reductions by 3.2 % and 3.8 % for the small and the large-cache system, respectively, due to reduced wirelengths.

1) Unbalanced/Heterogeneous Metal Stack: The 2D designs must be routed with at least six metal-layers. The internal routing of a memory block fully occupies the first four layers, making it impossible to route over memory blocks in the horizontal and vertical direction with less than six metal layers (*e.g.*, to reach I/O pins). However, for the MoL/*Macro-3D* designs, the number of metal layers in the macro die can be reduced from six to four without losing routability. An additional experiment shows that, despite the implied lower manufacturing cost, such a modification in the BEOL has no significant impact on the performance. As shown in Table III, the maximum frequency of the small-cache system only decreases by 1.8% after removing the two metal layers, while the maximum frequency of the large-cache system even increases by 0.5%. That is because, in the original *Macro-3D* design with six metal layers per die (*M6–M6*), most of the signal routing is done inside the logic die. Inter-die interconnects (*i.e.*, F2F vias) are mainly used to access memory pins located on the top die. Therefore, removing the two layers reduces the routing space, but it does not significantly increase routing congestions. Furthermore, by minimizing the metal layers in the upper die, the number of required F2F bumps is reduced by 18.4 % and 24.1 % as the top BEOL is, in this case, exclusively used for accessing memory pins and not for inter-standard-cell routing. This reduction further facilitates manufacturing. Thus, the proposed *Macro-3D* flow enables us to build F2F stacked 3D ICs with heterogeneous BEOLs, saving manufacturing costs, without degrading the performance noticeably.

VI. CONCLUSION

This work presented a novel physical design methodology for heterogeneous F2F-stacked 3D ICs. The proposed flow is completely based on commercial 2D EDA tools, which results in commercial-quality 3D IC layouts. While previous flows that are also based on 2D EDA tools initially perform a pseudo placement and routing of the standard cells, the proposed technique directly performs a true/valid 3D placement and routing without additional steps or tools. To achieve this, the flow exploits the specific nature of memory or sensor-on-logic 3D stacking, making it exclusively usable for such systems. However, both design styles promise an improvement in the system performance and manufacturing cost, compared to other 3D design styles, as it allows to add technological heterogeneity between the two dies. In contrast, previous 3D design methodologies perform particularly poorly for memory/sensor-on-logic 3D integration. Thus, the proposed technique extends previous works in an ideal way. While the previous 3D design methodologies do not show to optimize performance against 2D baseline designs for a RISC-V processor system and memory-on-logic stacking, the proposed technique shows to improve performance by up to 28.2 %, while it still facilitates manufacturing. The considered design style enables to design memory or sensor blocks of an SoC without the need to be process compatible with standard logic. Exploiting this feature to boost the 3D-integration gains further is left for future work.

REFERENCES

[1] E. Beyne, “The 3-D interconnect technology landscape,” *IEEE Design & Test*, vol. 33, no. 3, pp. 8–20, 2016.
[2] X. Dong, J. Zhao, and Y. Xie, “Fabrication cost analysis and cost-aware design space exploration for 3-D ICs,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 29, no. 12, pp. 1959–1972, 2010.
[3] P. Batude, C. Fenouillet-Beranger, L. Pasini, V. Lu, F. Deprat, L. Brunet, B. Sklenard, F. Piegas-Luce, M. Cassé, B. Mathieu *et al.*, “3DVLSI with CoolCube process: An alternative path to scaling,” in *2015 Symposium on VLSI Technology*. IEEE, 2015, pp. T48–T49.
[4] C. S. Tan, L. Peng, J. Fan, H. Li, and S. Gao, “Three-dimensional wafer stacking using Cu–Cu bonding for simultaneous formation of electrical, mechanical, and hermetic bonds,” *IEEE Trans. Device Mater. Rel.*, vol. 12, no. 2, pp. 194–200, 2012.
[5] 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, 2017.
[6] 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 *International Symposium on Physical Design*. ACM, 2018, pp. 90–97.
[7] J. Balkind, M. McKeown, Y. Fu, T. Nguyen, Y. Zhou, A. Lavrov, M. Shahrad, A. Fuchs, S. Payne, X. Liang *et al.*, “Openpiton: An open source manycore research framework,” in *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 2. ACM, 2016, pp. 217–232.