Fine-pitch 3D System Integration and Advanced CMOS nodes: Technology and System Design Perspective

Dragomir Milojevic[b], Giuliano Sisto[c], Geert Van der Plas[a], and Eric Beyne[a]

[a]IMEC, Kapeldreef 75, B-3001 Leuven, Belgium
[b]BEAMS, Ecole Polytechnique de Bruxelles - ULB, Bruxelles, Belgium
[c]Cadence Design Systems, San Jose, USA

ABSTRACT

Advanced CMOS SoCs with more cores and more complex memory hierarchies are hitting the memory wall, especially in intermediate cache levels (L2, L3). Managing the memory wall thus represent major challenge in the design of future systems and should include memory tech tuning, macro design & Logic-to-Memory interconnect optimization using multi-die packages & different 3D structures. To understand the benefits of 3D interconnects on Memory-on-Logic partitioning we analyze four different partitioning options of intermediate (L2) cache assuming high density CuCu hybrid bonding. We observe that the partitioning of the complete sub-system (memory macros & controller logic) is less beneficial with respect to reference 2D integration when compared to memory macro only partitioning schemes. Further, more memory macros are moved from the logic die, better the gains are (up to 40% total wirelength reduction). Such gains come at the expense of higher 3D pin count, motivating finer 3D pitches. Finally, we demonstrate design enablement of 3D aware IR-drop analysis for micro- and nano-TSVs with Buried Power Rail for Back Side power delivery.

Keywords: 3D integration, CuCu Hybrid Bonding, Cache Partitioning, 3D Place&Route, 3D IR-drop

1. INTRODUCTION

High-end CPUs execute hundreds of software threads simultaneously in a single IC package (e.g. Cascade Lake Platinum 9282). To achieve such high processing density, complex SoCs trade-off core frequency for power, cache memory configuration & capacity for total area & cost. Logical & physical implementation of such SoCs are structured using multiple levels of hierarchy. Few cores (<10) and L1 memories are tightly coupled in *core clusters* (Fig. 1). Multiple fully shared L2 sub-systems require efficient low delay/latency access to the cores (e.g. crossbars). More physical cores per cluster allow more efficient execution of fewer threads, resulting in increased performance. Multiple clusters are organized around next cache memory level (L3). Inter-cluster communication are typically implemented with Networks-on-Chip (NoCs). As the number of cores in the cluster increases, intra-cluster & inter-cluster communication contribute to the memory wall.[1] Bandwidth estimates, extrapolated from actual high-performance SoC[2] and taking CMOS scaling into account point memory wall in intermediate cache levels, especially for bigger core clusters (figure in the middle, cluster level), while current HBMs could sustain the bandwidth needs at chip level (figure on the right).

2. IMPROVING THE SYSTEM CONNECTIVITY & SYSTEM PARTITIONING

A path towards more optimized system interconnect is in *multi-dies packages* that enable *functional system partitioning*, so that different dies in the system and Die-to-Die interconnect can be optimized for a given functionality (Fig. 2). Bottommost 2.5D die serves as a hub to the outside world (e.g. Si bridge to the optical off-module interconnect) and/or to the in-package dense memory (e.g. Si bridge connection to higher and/or last level cache). Note that this 2.5D die could be active. SoC functionalities that don't require stringent performance, could be implemented there to offload more expensive real estate elsewhere in the package. 2.5D

Further author information: (Send correspondence to Dragomir Milojevic)
Dragomir Milojevic: E-mail: dragomir.milojevic@ulb.be, Telephone: +32.16.28.17.65
Geert Van der Plas: E-mail: Geert.VanderPlas@imec.be

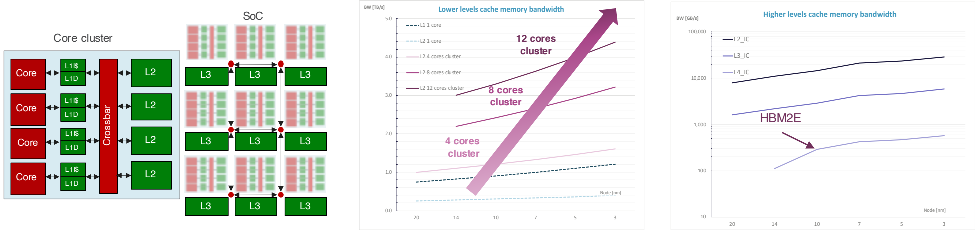


Figure 1: Complex high-performance hierarchy and bandwidth estimates – the Memory Wall

die implements *horizontal die-to-die planar interconnect*. In contrast, the chiplets, placed on top of the hub die are composed of Face-to-Face (F2F) stack that implement *area-array* interconnect using smaller 3D structures with much denser & shorter Die-to-Die interconnects. These are well suited for partitioning of intermediate and lower level cache memories. To enable practical integration of multi-dies packages we need 3D technologies with different integration densities to match bandwidth density, latency and energy per bit needs when partitioning a particular cache memory level. These structures could be also used for power delivery, from the package to the hub die and from the hub die to other dies. Two types of structures are illustrated on Fig. 3. First, ones that enable connectivity from the Back Side (BS) to the Front Side (FS) – the TSVs, and secondly those that enable Front-to-Front side connectivity (micro-bumps and Cu Pads). Standard TSV structures targeting pitches of 5μm, are scaled down to 2 and 1.4μm with micro-TSVs (μTSVs). Nano-TSVs (nTSVs)[3] target pitches in range of 0.2μ and connect to Buried Power Rails[4] (BPR). As for micro-bumps, typical 40μm pitches are scaled down to 20μm and below, embedded micro-bumps target 10μm, while direct hybrid bonding Cu pads aim $<3\mu$m.[5] Note that in case of front side 3D structures there is no area overhead, since they are processed after the topmost metal layer in the stack. In this paper we focus on a single F2F Memory-on-Logic chiplet, implemented using hybrid CuCu wafer bonding (Fig. 4). At system-level we look at partitioning of intermediate cache memory hierarchy levels (L2) of a relatively complex sub-systems composed of: a) multiple memory macros, with different functionalities (e.g. data, tag, buffers) & capacity and b) the control logic, that usually takes a non-negligible amount of logic resources. Each macro is composed of XY symmetrical bit-cell arrays and the internal controller. Having in

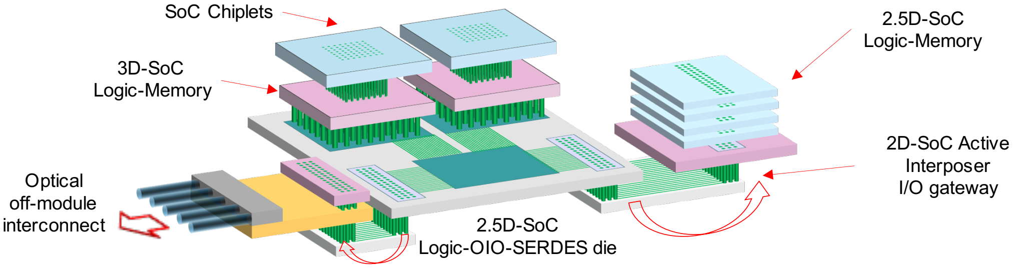


Figure 2: Complex system integration using multi-die packages and different 3D technologies

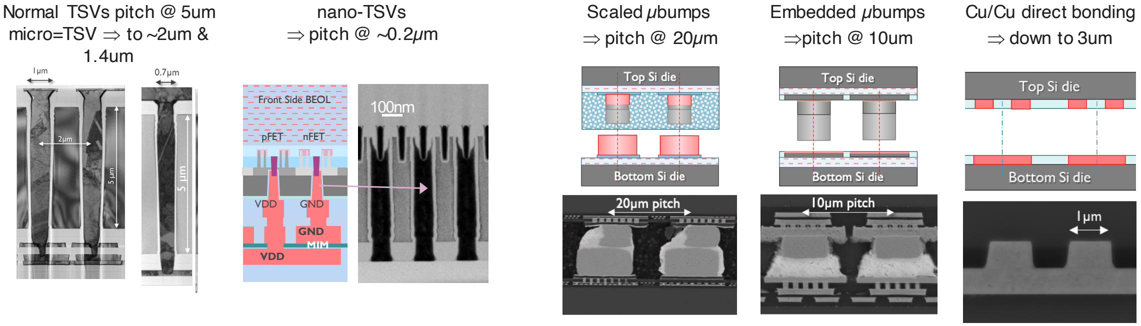


Figure 3: 3D structures for Back-to-Front and Front-to-Front connectivity and their target pitches

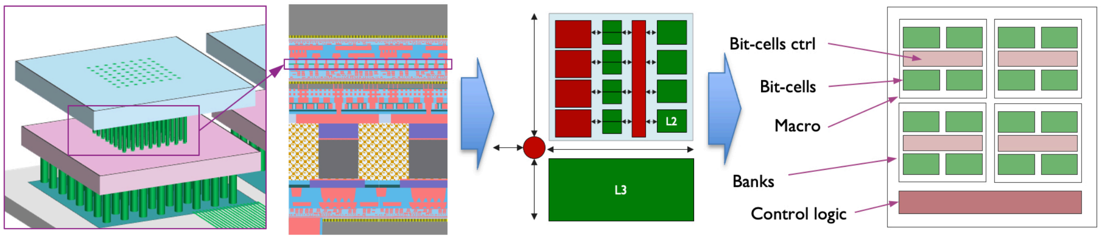


Figure 4: Dense Face-to-Face integration and system Memory-on-Logic partitioning

mind the structure of cache memories & when implementing Memory-on-Logic, different partitioning schemes could be defined depending on what is exactly partitioned where. Here we focus on two different partitioning schemes illustrated on Fig. 5.

Scheme1 - Cache sub-system partitioning – In this partitioning scheme the Memory die contains all cache memory macros and the controller logic. From implementation perspective this scheme is simple, because there are no choices to be made: simply take the complete cache sub-system (e.g. L2 or L3), and make a cut at interface level. Note that both dies now have to implement logic functionalities, so similar FEOL/BEOL configurations are expected for both dies, meaning less opportunities for potential cost optimization of wafer processing. On the other hand, since the cut is made at interface level, this scheme will a priori generate fewer 3D connections and thus stress less the required 3D interconnect pitch. Hence, this scheme could be used for both F2F and F2B configurations, with limited area overhead in F2B due to the limited number of 3D structures required.

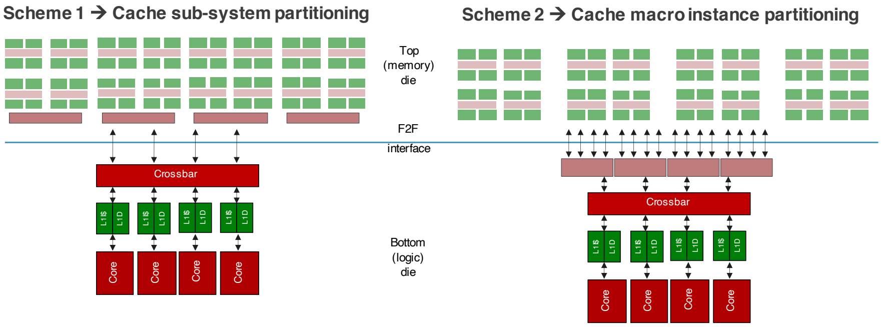

Scheme 1 → Cache sub-system partitioning
Scheme 2 → Cache macro instance partitioning
Top (memory) die
F2F interface
Crossbar
L1S
L1D
Core
Bottom (logic) die
Crossbar
L1S
L1D
Core

Figure 5: Different partitioning scenarios for Memory-on-Logic

Scheme2 - Cache macro instance partitioning – Here, the Memory die includes only cache memory macros. Since cache sub-system contain different macros, both function & capacity wise, multiple partitioning scenarios are possible. It is up to the designer to make a choice which macros should be moved to another die, i.e. he needs to make a partitioning decision. Because of the functional partitioning, Memory die (wafer) now can use different FEOL/BEOL configuration offer more opportunities for wafer cost optimization. Further, logic to memory interconnect could be shortened and delay/latency improved. Obvious disadvantage of this approach is that it is very likely that more 3D connections will be generated during partitioning, requiring tighter 3D structure pitches with very little, or ideally no area overheads.

Which of the above scenarios is better for 3D system integration can not be decided upfront. Potential benefits will strongly depend on the design itself: which level of cache is partitioned, cache capacity and configuration, data path width, logic-to-memory interconnect etc.

3. PARTITIONING MEMORY HIERARCHY

To analyze the impact of different partitioning scenarios on the benefits of the 3D system integration we will use OpenPiton SoC generation framework and Ariane core. The SoC instance contains 4 tiles. Each tile is composed of a single Ariane core (64-bit, low-power, 6-stage, RISC-V in-order issue, out-of-order execute); L1, L15 & L2 cache memory sub-systems; a Network-on-Chip (NoC) router for inter-tile communication. For the SoC place &

route we used imec iN5 PDK, without Buried Power Rail and we assume the 3D pitch of 1μm. SoC architecture, 2D floorplan, different SoC sub-systems and memory-to-logic connectivity are shown on Fig. 6.

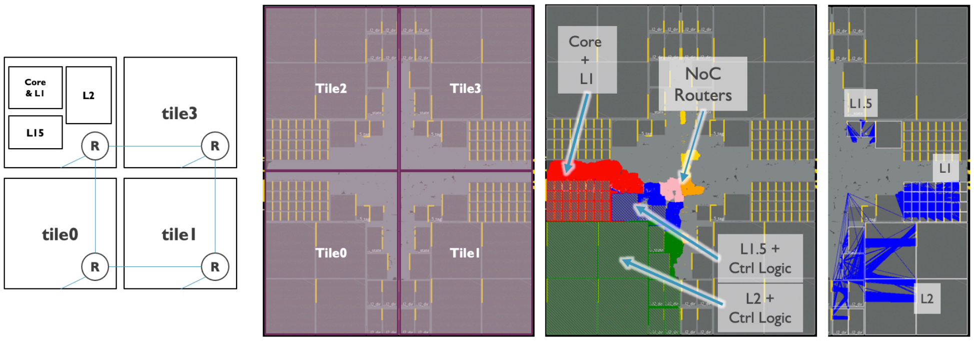
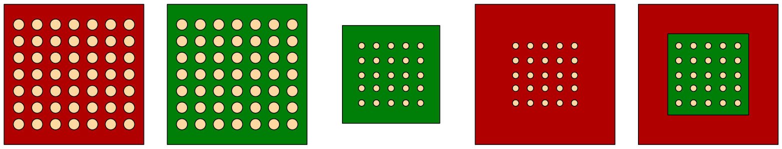


Figure 6: OpenPiton-Ariane core SoC, 2D implementation, different sub-systems and memory-logic connectivity

Different partitioning schemes defined in the previous section are applied to the SoC above to establish four following partitioning options. Option A. Partitioning of the complete L2 sub-system (Scheme1) – all L2 memory macros and the control logic are removed from the Logic die and moved to the Memory die. Option B. Partitioning of L2 data macros (Scheme2) – only the data (i.e. biggest) macros of L2 are moved to the memory die, resulting in 4 macros per tile. Option C. Partitioning of *all* L2 macros – that is data macros (like the previous option), but also all tag & directory macros (10 macros per tile). Option D. Partitioning of *all* SoC macros – all macros of L2, L15 and L1 cache sub-systems (total of 45 macros per tile).

Looking at the 2D area of different SoC sub-systems, and assuming no system re-design (i.e. cache memory capacity modification) certain partitioning options yield fairly different Memory & Logic die sizes, that need to be handled during P&R. Proposed solution (depicted on Fig. 7) requires multiple P&R iterations and will introduce some fan-out routing overhead to compensate for the difference in the die sizes. First we run top (Memory) die assuming that the whole core area is covered with Cu pads. Automated 3D pin assignment will connect all macro pins to 3D structures, so that the connection length is minimized (a.). Then we proceed with the first run of the bottom (Logic) die with predefined 3D net x,y locations (from the Memory die run); here we assume exactly the same die size of the bottom die as for the top die (b.). If the logic die Design Utilization (DU) is too low, we adjust the floorplan area, and therefore the area allocated for the 3D structures to reach the desired DU. Note that depending on the new value of the DU and the number of the 3D nets we might adjust the 3D structure size/pitch in the second run of the bottom die (c.). Once the desired DU is reached for the bottom die, it will define the appropriate Cu-Pad array and we can now proceed with the 2nd (final) top die run. With the appropriate bump area of the top die we can run the final P&R run for the bottom die (e.).



a. Memory die 1st run

b. Logic die 1st run

c. Logic die 2nd run

d. Memory die 2nd run with final 3D location

e. Logic die with final 3D location

Figure 7: Methodology to deal with unequal Top/Bottom die sizes at place & route level

Different partitioning options have been implemented in 3D, using Die-by-Die flow[6] (Fig. 8). For the Option A. we observe low 3D pin count, concentrated in the top die (control logic), while in the Option B. 3D pins are concentrated around data macros pins. For Option C. we see more pin clusters due to much denser inter-die connectivity. Small logic die forces concentrated placement of 3D pins for Option D.

Post route design data base has been used to extract statistics shown in table Fig. 9. Following observations can be made. First, we see that for all implementation options, top and bottom die have matched DU to minimize

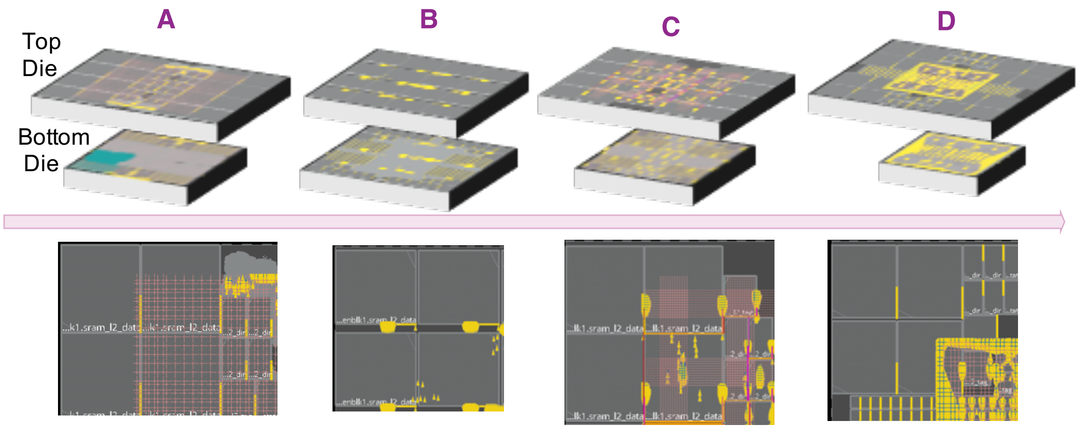


Figure 8: 3D Place & Route of different partitioning options

intra-die connectivity, i.e. total design wirelength (A.). Then for the partitioning Scheme1, the controller logic in the memory die requires significant amount of routing resources (B.) limiting the benefits of the 3D system integration since we observe only 7% of total wirelength savings compared to 2D. Further, more memory is moved out of the logic die, more gains we get (C.), but these gains come at the expense of more 3D connections that are order of the magnitude higher for different options of the Scheme2 compared to Scheme1 (D.). Finally, unequal die sizes cause limited fanout routing in the memory die, acceptable in this case, having in mind the benefits observed in the Logic die (E.).

Scenario	Design Utilisation (1st/2nd die)	Total system wirelength			Relative to 2D	3D Nets
		Bottom die [um]	Top die [um]	Total [um]		
A	A. 90/68	B. 3843697	1556164	5399861	C. -7%	D. 920
B	88/69	4715483	E. 193942	4909425	-17%	3028
C	89/75	4281665	311167	4592832	-25%	5872
D	87/74	3678292	499202	4177494	-38%	19264

Figure 9: Total system wirelength comparison for different SoC partitioning scenarios

4. POWERING THE STACK

3D-Stack is powered from the package and through the bottom die substrate using μTSVs & nTSVs as illustrated on Fig. 10. With μTSVs, that are larger but less resistive structures (5μm/1Ω), the Back Side (BS) is connected to one of the upper metal layers (M1). On the other hand nTSVs, much smaller but more resistive structures (0.09μm/10Ω), are used to connect the BS to Power Tap Cells (PTC) through Buried Power Rail (BPR).[4] Bottom die Power Delivery Network (PDN) could be pushed to the BS (BS-PDN[7]) offloading front-side BEOL routing resources, which is good for advanced CMOS, where high pin densities increase routing congestion. To quantify differences of these power delivery options we need a 3D aware IR-drop tool. In this work we have adapted the existing 2D multi-die rail analysis tool (Voltus, Cadence) to support 3D IR-drop. From 3D P&R, physical design data and parasitics of each die are passed to power simulation. Power interfaces, 3D power structures mapping and package spice model are provided trough custom made scripts. Proposed flow has been used to demonstrate the comparison between BS-PDN & FS-PDN (Fig. 10 b.) assuming the same amount of Vdd/Vss pins (shown bottom die only). We observe that while the IR-drop target of <10% voltage supply has been met for both BS-PDN & FS-PDN, BS-PDN exhibits 20% lower maximum & 63% less lower average IR-drop. Further we have compared two different 2D PDN options (Fig. 10 c.): I) PDN1 – using very short metal segments to connect Vdd, Vss stripes with 3D structures; II) PDN2 – use of full stripes on the last metal layer to connect Vdd, Vss to 3D structures. Two solutions are expected to trade-off BEOL usage with IR-drop.

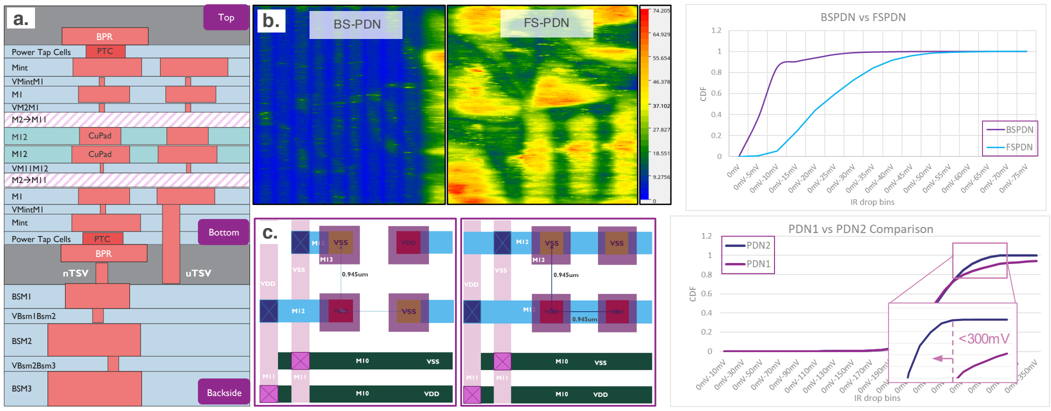


Figure 10: Power delivery to the stack and 3D aware IR-drop assessment examples

5. CONCLUSION

To allow simultaneous execution of more threads per IC package, high-performance SoCs are constantly increasing the number of cores implemented on a single die, with deeper memory hierarchies & with more memory capacity per cache layer. When implemented in advanced CMOS, such SoCs face memory wall, particularly prominent in intermediate caches (L2, L3). Solving memory wall requires co-optimization of memory tech, memory macro design & Logic-to-Memory interconnect. In this paper we investigate Logic-to-Memory interconnect optimization of intermediate caches using high density 3D structures (hybrid CuCu bonding with 1μm pitch) & advanced CMOS (imec iN5). Different 3D memory partitioning scenarios have been analyzed after 3D P&R of a 4-core MPSoC. Our experiments show that moving the full memory cache sub-system with all memory macros and the controller logic don't necessarily bring the highest benefits. Generally, moving macros only yield better results and more macros are moved, better the gains are (up to 40% of the total wirelength savings when compared to 2D). Benefits are due not only to shortened Logic-to-Memory wires, but also due to much shorter 2D nets in the logic die. Further advantage of this partitioning scheme is in the potential wafer cost optimization enabled by functional partitioning. These benefits come at the cost of increased number of 3D connections. Our estimations confirm that future pitch target of 500nm for the Cu Pads will be sufficient for the partitioning of intermediate caches in advanced nodes. Finally, we have demonstrated design enablement of IR-drop assessment of a 3D-stack using μTSVs or nTSVs with BPR and FS-&BS-PDN. The combination of the BS-PDN and nTSVs showed good IR-drop, making them good candidate for 3D power delivery in advanced nodes.

REFERENCES

[1] Perumkunnil, M. and al., "System exploration and technology demonstration of 3D Wafer-to-Wafer integrated STT-MRAM based caches for advanced Mobile SoCs," in [*IEEE International Electron Devices Meeting (IEDM)*], (202-).

[2] Aingaran, K. and al., "M7: Oracle's Next-Generation Sparc Processor," *IEEE Micro* **35**(2), 36–45 (2015).

[3] Jourdain, A. and al., "Extreme Wafer Thinning and nano-TSV processing for 3D Heterogeneous Integration," *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)* , 42–48 (2020).

[4] Ryckaert, J. and al., "Enabling sub-5nm cmos technology scaling thinner and taller!," in [*2019 IEEE International Electron Devices Meeting (IEDM)*], 29.4.1–29.4.4 (2019).

[5] Peng, L. and al., "Advances in sicn-sicn bonding with high accuracy wafer-to-wafer (w2w) stacking technology," in [*2018 IEEE International Interconnect Technology Conference (IITC)*], 179–181 (2018).

[6] Sisto, Giuliano and al., "Design enablement of fine pitch Face-to-Face 3D system integration using Die-by-Die Place & Route," in [*IEEE International 3D Systems Integration Conference (3DIC)*], (Oct. 2019).

[7] Hossen, M. O., Chava, B., Van der Plas, G., Beyne, E., and Bakir, M. S., "Power delivery network (pdn) modeling for backside-pdn configurations with buried power rails and μ tsvs," *IEEE Transactions on Electron Devices* **67**(1), 11–17 (2020).