PPAT for Multi-Die 3D Accelerator Design in Advanced Nanosheet and CFET Technology

D. Milojevic [1,2], L.G. Rocha [1], J. Stephenson [3], L. Stefanidis [1,2], M. Stassar [1,2,4], M. Naeim [4],
M. Verhelst [5], D. Biswas[1], J. Myers[1]
[1] imec, Leuven, Belgium; [2] Université libre de Bruxelles, Belgium; [3] University of Newcastle, UK;
[4] Cadence Design System, San Jose, USA; [5] KU Leuven, Belgium

***Abstract*—In this paper, we demonstrate that Multi-Plane Systolic Arrays (MP-SAs) can execute workloads in up to 2× less compute cycles than conventional Single-Plane SAs (SP-SAs), given the same compute capacity. However, increasing the number of processing planes in a 2D-IC implementation introduces interconnect overheads. Deploying MP-SAs in multi-die 3D-ICs using advanced CMOS technology reduces interconnect by up to 30%, but results in only a 5% reduction in power consumption, mostly dominated by logic. Nevertheless, for a workload that shows 30% less compute cycles, MP-SAs in 3D-ICs can achieve 17% higher energy efficiency when compared to their SP-SAs counterparts implemented as 2D-IC. Finally, despite increased power density, we demonstrate that thermal challenges can be effectively managed with appropriate cooling strategies, even in 4-Dies 3D-IC configurations.**

***Index Terms*—3D systolic arrays, Nanosheet, CFET, Logic-on-Logic 3D-IC, hybrid bonding, slit TSV.**

I. Introduction, related work & contributions

Systolic Arrays (SAs) have been the architecture of choice for commercial Deep Neural Network (DNN) accelerators [1] due to their power efficiency & high performance when executing General Matrix Multiply (GEMM) operations. Multi-Plane SAs (MP-SAs) extend this approach by distributing computation & dataflow across the third dimension [2], addressing latency & mapping inefficiencies inherent in Single-Plane SAs (SP-SAs). MP-SAs are well-suited for 3D-IC integration & Logic-on-Logic (LoL) stacking [9]. While previous work focused on the architecture and workload mapping [8] or studied performance, power, area and thermal (PPAT) architectural trade-offs [7], they focused primarily on post-synthesis netlist information to derive physical design implications. To the best of our knowledge, no multi-die layout-level study has been performed for SAs. This work claims the following contributions: 1) We investigate distributed Output Stationary (dOS) dataflow and analyse execution speedup for a given GEMM workload; 2) We conduct PPA analysis on physically implemented 2D- & multi-die 3D-IC with N2 Nano-Sheet (N2-NS) & A7 CFET nodes, using an actual GEMM workload; 3) Conduct a thermally-aware power simulation of the MP-SA, both for 2D & multi-die 3D-IC up to 4-Dies.

II. System architecture & workload analysis

Traditional SP-SAs implement a mesh of Processing Elements (PEs), containing Multiply-Accumulate (MAC) units & local memory [1]. In SP-SAs, PEs are interconnected locally to 4 nearest neighbours, and thus require minimal PE2PE routing, Fig. 1 a). MP-SAs extended SP-SA connectivity, with multiple compute planes producing partial sums (PSUMs), forwarded from top- to bottom-most compute plane through an arbitrary number of mid-planes. Fig. 1 b) features MP-SAs with 4 PE compute planes. Each plane has dedicated activation & weight buffers, allowing simultaneous read operations. Weights & activations enter each plane through the first PE row & column, respectively. MP-SAs can provide workload performance uplift w.r.t. SP-SAs at iso-PE count, due to data computation rearrangement, improve parallelism & utilisation with reduced latencies for larger arrays. We investigate dOS dataflow in MP-SAs [2], organising the architecture as $P_x \times P_y \times P_z$ with $N \times K \times M$, where N & M are spatially unrolled on $P_x \times P_y$, and K is unrolled across P_z. Minimising $P_x \times P_y$ array dimensions & increasing P_z improves utilisation up to 2× at iso-compute due to less idle PEs from non-ideal GEMM partitioning. Workload performance assessment using GEMM operations are reported on Fig. 2. All results are normalised w.r.t. SP-SA with 128×128×1 PEs. Bars represent speed-up due to cycle count reduction. Red line represents array utilisation, light blue lines represent baseline utilisation. MP-SA shows speedup from 1.1× to 2× in a), c) & d) at iso-compute due to better mapping. Performance has near-linear scaling w.r.t compute.

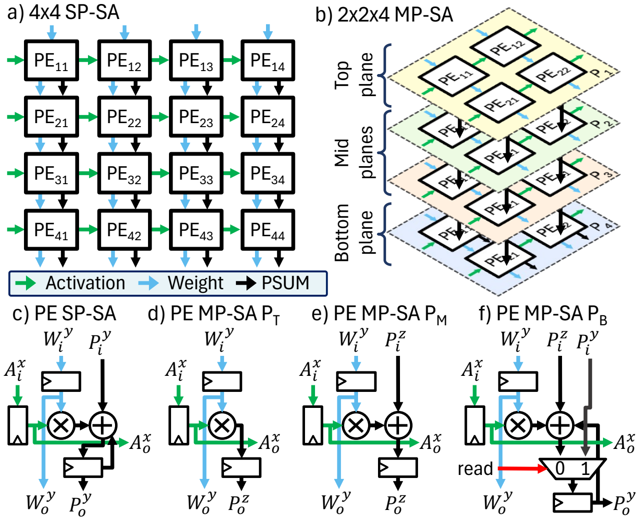


Fig. 1. a) SP-SA & b) MP-SAs and c), d), e), f) their PE architectures.

	Workload	N	K	M	Workload	N	K	M	
Vision	GEMM-1	3136	512	128	GEMM-3	144	768	3072	Transformer
	GEMM-2	676	1024	256	GEMM-4	32	4096	1024	

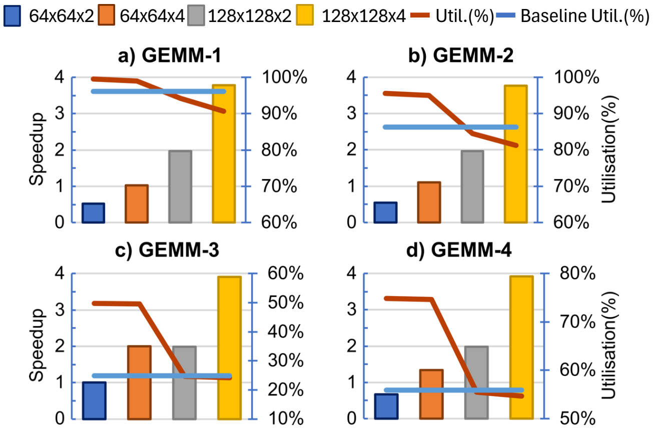


Fig. 2. Speed-up & utilisation for SP-SA & MP-SA for 4 workloads.

III. 2D- & 3D-IC Physical Design Considerations

For all SA configurations we assume PEs with 16-bit integer input words for both activation & weights, and 32-bit PSUMs. Note that in the SP-SA all PEs are identical, while in MP-SAs, PEs are specialised for top, mid and bottom planes (see Fig. 1). Following iso-compute SA configurations (~16384 PEs) are considered: SP-SA with 128×128×1PEs, MP-SAs with 64×64×4 & 74×74×3 PEs. Physical implementation of SAs has been performed using two PDKs: a) N2-NS: 4 sheets, 48nm gate & 22nm metal pitch, 6T cells [3] & b) A7-CFET: 42nm CPP, 19nm MP & 7T cells [4]. Due to little routing demand of the SAs, BEOL metal stack has been limited to 10 metal layers in all cases. Further BEOL optimisation is possible, especially for multi-die 3D-ICs. SP- & MP-SAs have been implemented in N2 & A7 to establish 2D-IC baseline. MP-SAs have been implemented with 2-, 3-, & 4-Dies 3D-IC, N2 PDK only. For 2-Dies 3D-IC we use Face-to-Face (F2F) stacking, with Copper-to-Copper (CuCu) hybrid bonding & Cu pad pitch of 0.48μm [5] corresponding to 6× last metal layer pitch. Note that in F2F configuration the top die is flipped, facing down the bottom die and the package. 3D-ICs with 3- & 4-Dies assumed Back-to-Face (B2F) integration, with all dies facing up. Connections between front & back side of each die has been made using slit TSVs [6] with ~1CPP width, 1 standard cell height (in plane dimensions of 40×110nm), 100μm depth (die thickness) & pitch of 0.24μm (including KOZ). To enable reasonable Place & Route (PnR) run-time for large SAs (128×128×1 PEs results in ~12MGates), a bottom-up hierarchical design flow is adopted. First, different PEs are implemented to generate geometrical and timing abstracts from their respective layouts. Top-level design is then constructed using these PE abstracts, with floor-plan being generated semi-automatically. Once routed, the final layout is obtained by replacing abstracted PEs with their standard-cell implementations. Due to bottom-up strategy, PE I/O placement is critical to minimising top-level routing. For 2D-ICs, all PE input pins are placed on one side, and all output pins on the opposite side, following the dataflow of the SA. For 3D-ICs, in-plane connections retain their locations like in 2D-IC. However, PSUM input & outputs are placed within the PE core area, to reduce routing overhead between PE pins and 3D interconnect structures. For SP-SA mapped on 2D-IC the floorplan is straightforward: PEs are manually placed in abutted fashion to minimise PE2PE routing. For MP-SAs we have a choice of abutting the SA planes, so that all PEs of a single plane are placed just like in a SP-SA, see Fig. 3 a). However, such approach is not good for large SAs since Plane-to-Plane (PL2PL) connectivity (red connections) will have to span the width (or height) of the complete plane (distance $D = N_{x/y} \times PE_{W/H}$) for each activation (or weight) bit. For MP-SAs a more effective solution would be to stagger PEs from different planes, and place them in abutted way. This way PL2PL routing will be minimised at the expense of some PE2PE routing for the PEs in the same plane, that need to hoover over the PE of another plane, seen as routing blockage. The length of these PE2PE connections correspond to the size of the PE ($D = PE_{W/H}$), shown in blue in the Fig. 3 b). 3D integration assumes LoL, PE-on-PE stacking with two options, where either one or two SA planes are mapped per 3D-IC tier (see Fig. 4). In 2-Dies 3D-ICs, the staggered PE floorplan is used to minimise PL2PL routing. Some PE2PE connections are optimised, marked green in Fig. 4 a), while others remain as in 2D-IC (marked red). When one SA plane is mapped per 3D-IC tier, Fig. 4 b), all PEs can be abutted as in SP-SAs, resulting in minimal PE2PE connectivity, while PL2PL connectivity is exclusively 3D. These floorplan choices are key towards interconnect optimisation & design PPA.

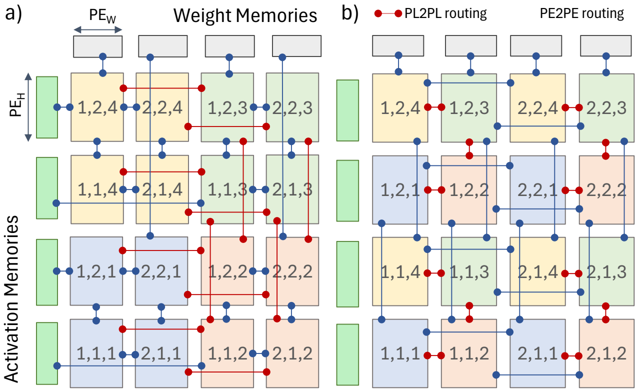


Fig. 3. Floorplan options for MP-SA, 2D-IC: a) naive – abutted PEs of the same plane & abutted planes, or b) staggered PEs from different planes.

IV. PPA Analysis with Thermally-Aware Power

Baseline 2D-IC implementation comparing N2-NS vs. A7-CFET of SP-SA with 128×128×1 PEs & MP-SA with 64×64×4 PEs is shown on Fig. 5. Note that reported power is statistical (20% toggle rate at inputs & flip-flops). Technology scaling promises are kept with ~45% less area, ~30% less power & ~20% less total system wire-length (WL). What

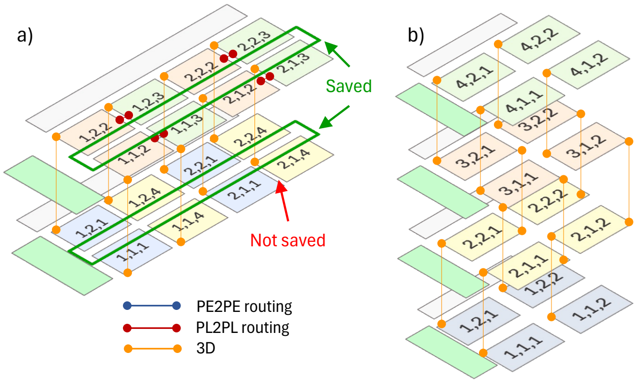


Fig. 4. Mapping of 2 planes a) & 1 plane b) per 3D-IC tier.

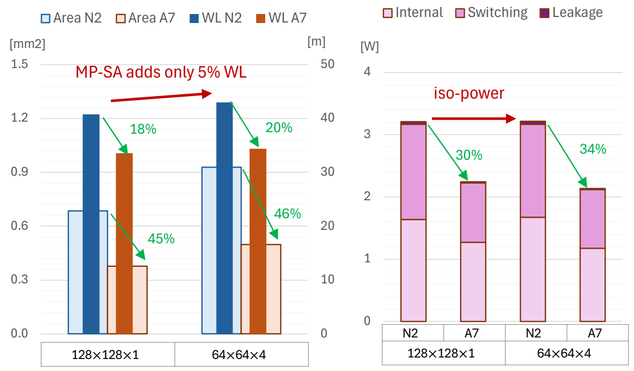


Fig. 5. Baseline 2D-IC comparison of Single- & Multi-plane SA at iso-compute: area, total system WL (left Fig.) & power breakdown (right Fig.).

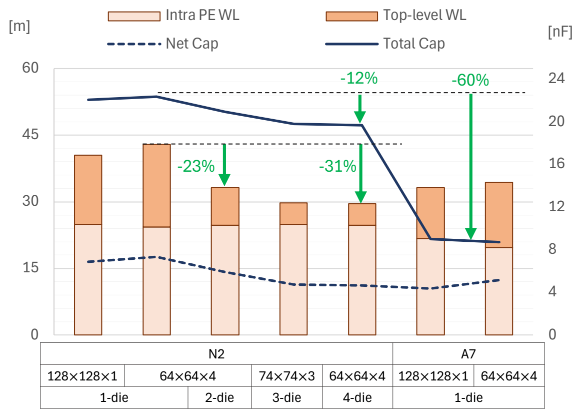


Fig. 6. Intra-PE & Top-Level WL (left axis), net & total (net+pin) capacitance for baseline 2D-IC & MP-SA mapped to multi-die 3D-IC.

comes as a surprise is MP-SA overhead compared to SP-SA: while there is 36% area more area due to per plane PE configuration (Fig. 1), there is only extra 5% total system WL. Adopted optimised placement with staggered PEs of MP-SAs in 2D-IC works well. Interconnect demand of baseline 2D-IC (N2 & A7) is compared to 2-, 3- & 4-Dies 3D-IC implementations (N2 only). In Fig. 6 we report Intra-PE & Top-level WL (left axis) and net vs. total (net+pin) capacitance (right axis). MP-SA with 64×64×4 PEs mapped on 2-Dies 3D-IC achieve 23% reduction of total system WL, while 4-Dies 3D-IC tops with 31% less interconnect than the equivalent 2D-IC. MP-SA with 74×74×3 implemented in 3-Dies 3D-IC has similar interconnect demand as 64×64×4. Looking at the capacitance values (right-axis) of the 4-Dies implementation compared to 2D-IC, it follows the WL reduction with 32% less total net cap. This however contributes to only 12% less total capacitance. Workload power simulation was conducted using a GEMM operation with $N \times K \times M$=128×128×128. It is important to note that this workload is not optimal for a $64 \times 64 \times 4$ MP-SA, as the matrix size exceeds the spatial dimensions of the SA. For activation inputs, toggling was assumed to be random, while for weights, two extreme scenarios were considered: a) Best Case (BC), where only one bit toggles between two successive line loads; b) Worst Case (WC), where 15 out of 16 bits toggle between two successive line loads. Power breakdown (left axis) and energy consumption (right axis) are reported in Fig. 7, comparing results from both baseline statistical and workload-annotated power simulations. For a SP-SA mapped onto a 2D-IC, statistical power estimation with 20% activity overestimates total power by 6× compared to the BC workload, with the overestimation being particularly pronounced in the switching power component. Additionally, in the WC scenario, power consumption doubles relative to BC, highlighting a strong dependency on the actual data being processed within the SA. While MP-SAs implemented on 3D-ICs don't necessarily reduce total power consumption compared to SP-SAs on 2D-ICs, they achieve 17% higher energy efficiency due to ∼30% less compute cycles, even for workloads that do not inherently favour MP-SAs. Note that 3D-IC integrations adds extra +5% energy savings on top of savings due to MP-SA. Power & thermal analysis of SP- & MP-SAs implemented on 2D- & 3D-ICs was conducted using post-routed layouts & a complete package cross-section definition. To optimise simulation time, power & thermal maps were tessellated using a 1.5μm-edge squared mesh to estimate extreme stack temperatures. Thermal simulations were performed using the N2 PDK (V_{dd} = 0.9V at 25, 75, & 115°C), within an iterative, thermally-aware

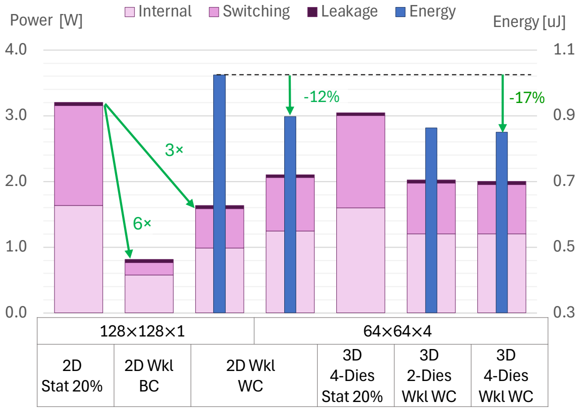
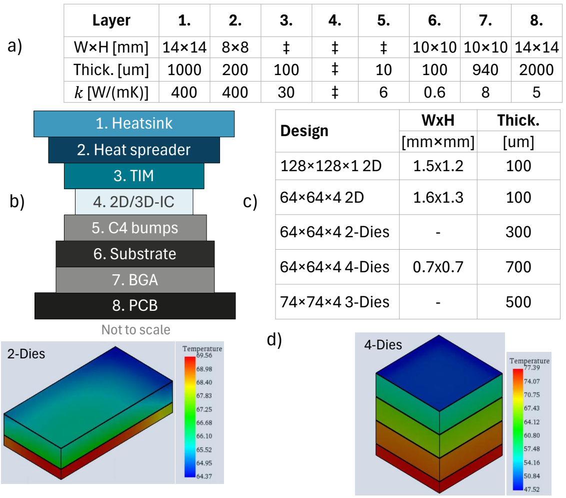


Fig. 7. Power breakdown & energy for statistical & workload simulations.

statistical power simulation flow. Initially, power simulations assumed a 25°C starting point, generating a thermal profile used to refine subsequent power simulations. This iterative process continued until achieving thermal convergence (minimal temperature difference between two successive iterations) or reaching a predefined limit (∼10 iterations), beyond which thermal runaway was expected. While the thermally-aware power simulation remains statistical, input and flip-flop activity factors were tuned to reflect the worst-case workload scenario. Fig. 8 illustrates: a) Layer dimensions, thicknesses, & assumed thermal conductivity (k); package cross-section is shown in b); design-dependent data is summarised in c). Fig. 9 illustrates power breakdown & thermal limits (T_{min}, T_{max}) for various SA configurations, multi-die integration approaches, & cooling assumptions. In 2D-IC implementations, both SP- & MP-SAs exhibit comparable and thermally acceptable profiles, even with a relatively low Heat Transfer Coefficient (HTC) of 0.015 W/cm^{2}K. However, a MP-SA in a 2-Dies F2F 3D-IC requires an improved HTC of 0.03 W/cm^{2}K to maintain manageable thermal conditions. In a 3-Dies stack, maintaining the same cooling conditions as a 2D-IC setup results in thermal runaway. Increasing the HTC to match that of a 2-Dies 3D-IC improves thermal performance, though temperatures remain higher than in the 2-Dies case. Lowering power density, for example, through Dynamic Voltage and Frequency Scaling (DVFS), could make even low-end cooling solutions viable for a 3-Dies configuration. However, at constant activity levels, a 4-Dies stack demands high-end cooling with an HTC of 0.05 W/cm^{2}K to ensure thermal stability.

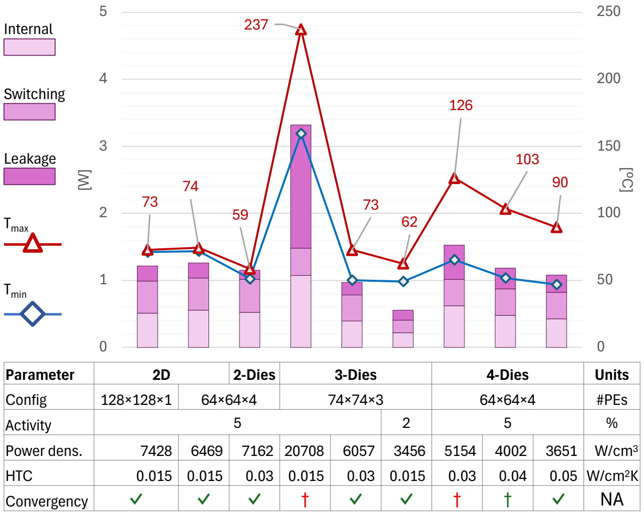

a)

Layer	1.	2.	3.	4.	5.	6.	7.	8.
W×H [mm]	14×14	8×8	‡	‡	‡	10×10	10×10	14×14
Thick. [um]	1000	200	100	‡	10	100	940	2000
k [W/(mK)]	400	400	30	‡	6	0.6	8	5

c)

Design	WxH [mm×mm]	Thick. [um]
128×128×1 2D	1.5x1.2	100
64×64×4 2D	1.6x1.3	100
64×64×4 2-Dies	-	300
64×64×4 4-Dies	0.7x0.7	700
74×74×4 3-Dies	-	500

Fig. 8. Package dimensions a), cross-section b) & design dependent data c) assumed for thermally-aware power simulation d).



Parameter	2D		2-Dies	3-Dies			4-Dies			Units
Config	128×128×1	64×64×4		74×74×3			64×64×4			#PEs
Activity	5					2	5			%
Power dens.	7428	6469	7162	20708	6057	3456	5154	4002	3651	W/cm³
HTC	0.015	0.015	0.03	0.015	0.03	0.015	0.03	0.04	0.05	W/cm²K
Convergency	✓	✓	✓	†	✓	✓	†	†	✓	NA

Fig. 9. Thermally aware power & thermal limits for various SA configurations

V. Conclusion

For small input words and a limited number of compute planes, the power consumption in MP-SAs is primarily driven by compute logic rather than interconnect, even in large SAs. The power savings offered by 3D integration are modest, reaching only up to 5% despite achieving a >10% total capacitance reduction. Additionally, 4-Dies 3D-IC configurations provide only marginal improvements over 2- & 3-Dies implementations. However, due to improved workload-architecture adaptation, MP-SAs require fewer compute cycles at iso-compute, leading to energy savings of up to 17% compared to SP-SAs in 2D-ICs when executing actual workloads. Our findings highlight a strong dependence of power consumption on workload characteristics & data movements within the SA. The outlook may shift with PEs that incorporate greater functionality, additional planes, and larger input words. Lastly, we demonstrate that multi-die stacks are thermally viable, with T_{max} remaining manageable through appropriate cooling strategies. We emphasise the importance of workload annotation and thermally aware power simulations to accurately assess power and thermal behaviour. In our future work we will explore the impact of memory sub-system & data movements to & from the SAs on system performance.

References

[1] A. C. Yüzügüler et al., “Scale-out Systolic Arrays”, ACM Trans. Archit. Code Optim. 20, 2, Article 27, 2023

[2] J. M. Joseph et al., “Architecture, Dataflow and Physical Design Implications of 3D-ICs for DNN-Accelerators”, ISQED, 2021, pp. 60-66.

[3] A. Farokhnejad et al., “N2 Nanosheet Pathfinding-PDK Including Back-Side PDN”, ESSERC, 2024, pp.17-20

[4] H.Kukner et al., “2-4 — Double-Row CFET: DTCO for Area Efficient A7 Technology Node”, IEDM, 2024

[5] B. Zhang et al.,“Scaling Cu/SiCN Wafer-to-Wafer Hybrid Bonding down to 400nm interconnect pitch”, ECTC,2024

[6] P. Zhao et al., “Backside Power Delivery With Relaxed Overlay for Backside Patterning Using Extreme Wafer Thinning and Molybdenum-Filled Slit Nano Through Silicon Vias”, IEEE Transactions on Electron Devices, vol. 71, no. 12, 2024

[7] J. M. Joseph et al., “Architecture, Dataflow and Physical Design Implications of 3D-ICs for DNN-Accelerators,” 22nd International Symposium on Quality Electronic Design (ISQED), CA, USA, 2021, pp. 60-66.

[8] Y. Wang, Y. Wang, H. Li, C. Shi and X. Li, “Systolic Cube: A Spatial 3D CNN Accelerator Architecture for Low Power Video Analysis,” 2019 56th ACM/IEEE Design Automation Conference (DAC), Las Vegas, NV, USA, 2019, pp. 1-6.

[9] Kung, H. T. et al., “Systolic Building Block for Logic-on-Logic 3D-IC Implementations of Convolutional Neural Networks,” 2019 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5.