2023 Design, Automation & Test in Europe Conference (DATE 2023)

Ultra-Dense 3D Physical Design Unlocks New Architectural Design Points with Large Benefits

Tathagata Srimani[1*], Robert M. Radway[1*], Jinwoo Kim[2*], Kartik Prabhu[1], Dennis Rich[1], Carlo Gilardi[1], Priyanka Raina[1], Max Shulaker[3], Sung Kyu Lim[2] and Subhasish Mitra[1]

[1] Stanford University, [2] Georgia Institute of Technology; [3] Massachusetts Institute of Technology; * Equal Contribution

Email: tsrimani@stanford.edu; radway@stanford.edu; jinwookim@gatech.edu

***Abstract*— This paper focuses on iso-on-chip-memory-capacity and iso-footprint Energy-Delay-Product (EDP) benefits of ultra-dense 3D, e.g., monolithic 3D (M3D), computing systems vs. corresponding 2D designs. Simply folding existing 2D designs into corresponding M3D physical designs yields limited EDP benefits (~1.4×). New M3D architectural design points that exploit M3D physical design are crucial for large M3D EDP benefits. We perform comprehensive architectural exploration and detailed M3D physical design using foundry M3D process design kit and standard cell library for front-end-of-line (FEOL) Si CMOS logic, on-chip back-end-of-line (BEOL) memory, and a single layer of on-chip BEOL FETs. We find new M3D AI/ML accelerator architectural design points that have iso-footprint, iso-on-chip-memory-capacity EDP benefits ranging from 5.3× to 11.5× vs. corresponding 2D designs (containing only FEOL Si CMOS and on-chip BEOL memory). We also present an analytical framework to derive architectural insights into these benefits, showing that our principles extend to many architectural design points across various device technologies.**

Keywords*— *Monolithic 3D Integration, 3D Architectures. 3D Physical Design, Back-End-of-Line (BEOL) Technologies

I. Introduction

21st-century abundant-data applications, e.g., Artificial Intelligence/Machine Learning (AI/ML), place unprecedented demands on computing systems. Simultaneously, traditional 2D lithographic scaling is getting increasingly difficult – the *miniaturization wall*. New integration techniques are needed to overcome these challenges and enable large energy and speed benefits. Ultra-dense 3D, e.g., *monolithic 3D (M3D)*, integration [1] is one such approach. By tightly integrating many tiers of logic and memory in 3D with ultra-dense 3D inter-layer vias (*ILVs*, i.e., the same metal vias used in today's back-end-of-line – *BEOL* – processes for metal routing), M3D systems can deliver large benefits in computing speed and energy efficiency.

Several prior studies have focused on the system-level benefits of M3D. Some (e.g., [1-2]) have shown how futuristic M3D systems leveraging multiple interleaved tiers of compute and memory can achieve large improvements in system-level energy-delay product (*EDP*) vs. baseline 2D systems. However, these benefits are obtained mostly by bringing large amounts of off-chip memory on-chip [1-2] – i.e., the M3D chips have much larger amounts of on-chip memory vs. the corresponding 2D chips (*not* iso-on-chip-memory-capacity). Other studies (e.g., [3-4]) have explored M3D EDP benefits vs. 2D chips obtained **solely** by leveraging 3D physical design for a given architecture design point. These studies focused on folding existing 2D designs into M3D (i.e., iso-on-chip-memory-capacity, iso-architecture design point) with optimized place and route across 3D tiers (resulting in ~50% reduced footprint in M3D designs along with ~20% reduced wirelength and buffer sizes). Such M3D folding approaches offer limited EDP benefits (e.g., ~1.1-1.4× EDP benefits [3-4]). In this paper, we show large M3D EDP benefits, in the range of 5.3×-11.5×, despite iso-on-chip-memory-capacity and iso-footprint comparisons vs. 2D designs. We achieve such large benefits by deriving new M3D architectural design points that are enabled by M3D physical design (Fig. 1). Our M3D physical design uses a foundry M3D technology and its associated design infrastructure.

The key highlights of this paper are:

(1) Instead of using predictive models, we demonstrate large EDP benefits of M3D using a foundry M3D technology with foundry M3D process design kit (*PDK*) and foundry M3D standard cell library calibrated using foundry hardware data. The foundry M3D technology integrates Carbon Nanotube FETs (*CNFETs*) for BEOL transistors and Resistive RAM (*RRAM*) for BEOL memory cells on top of silicon CMOS [5].

(2) Using foundry M3D technology, we present a case study for iso-on-chip-memory-capacity and iso-footprint quantification of M3D benefits vs. 2D. Beyond folding existing 2D designs into M3D (and relying solely on physical design), we derive new M3D architectural design points enabled by M3D physical design.

(3) Despite our iso-footprint and iso-on-chip-memory-capacity constraints, and the relaxed characteristics of BEOL technologies (newly) implemented in foundry M3D technology, our physical design case study demonstrates 5.7× to 7.5× EDP benefits of using M3D over AI/ML neural network workloads. These benefits come without 3D thermal concerns as peak power density increases by just 1%. We emphasize that these M3D benefits are derived from new

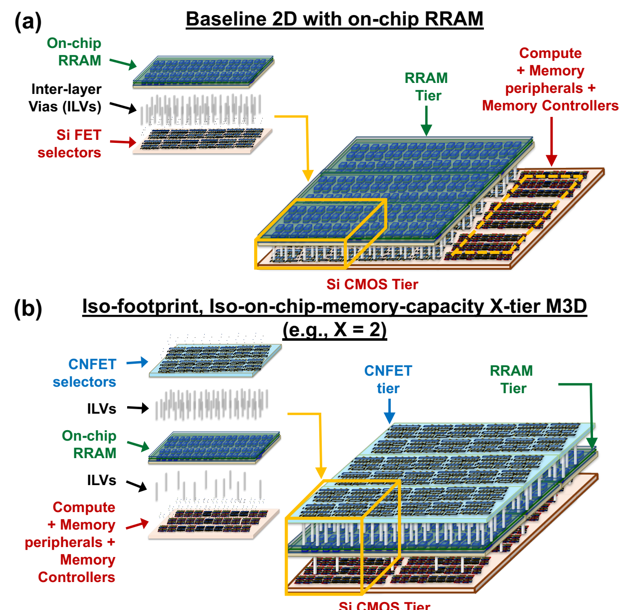
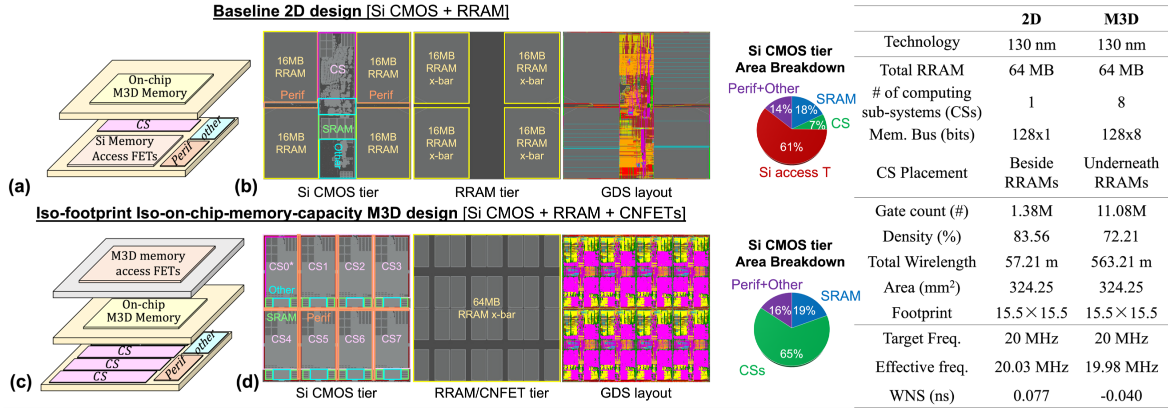


Fig. 1: (a) Baseline 2D with Si CMOS and on-chip RRAM. Si CMOS supports compute, memory access selectors, memory peripherals/controllers. (b) Iso-footprint, iso-on-chip-memory-capacity M3D (Si CMOS + RRAM + CNFETs) with CNFETs memory selectors.

978-3-9819263-7-8/DATE23/© 2023 EDAA
Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 13:03:34 UTC from IEEE Xplore. Restrictions apply.



	2D	M3D
Technology	130 nm	130 nm
Total RRAM	64 MB	64 MB
# of computing sub-systems (CSs)	1	8
Mem. Bus (bits)	128x1	128x8
CS Placement	Beside RRAMs	Underneath RRAMs
Gate count (#)	1.38M	11.08M
Density (%)	83.56	72.21
Total Wirelength	57.21 m	563.21 m
Area (mm²)	324.25	324.25
Footprint	15.5×15.5	15.5×15.5
Target Freq.	20 MHz	20 MHz
Effective freq.	20.03 MHz	19.98 MHz
WNS (ns)	0.077	-0.040

Fig. 2: AI accelerator SoC designs in baseline 2D: (a) 3D view chip and (b) GDS layout, and in iso-footprint, iso-on-chip-memory-capacity M3D: (c) 3D view chip and (d) GDS layout. CNFET-based RRAM access transistors frees Si CMOS space for 8× parallel computing sub-systems (CS) in M3D.

architectural design points enabled by M3D physical design and not from bringing off-chip memory on-chip or from changing the on-chip (BEOL) memory technology.

(4) We derive an explainable analytical framework which shows M3D design EDP benefits range from 5.3× to 11.5× across several AI/ML accelerator dataflows. Our framework is accurate, within 10% of our physical design results.

(5) Using our analytical M3D EDP framework, we find: (a) Greater than 5.3× benefits are possible with up to 1.6× wider BEOL memory selectors (b) BEOL ILV pitch cannot increase more than 1.3× before reducing M3D EDP benefits, (c) Additional incremental M3D layers quickly increases EDP benefits; these benefits plateau based on application parallelizability and thermal limits. This extends our M3D architectural design space understanding beyond a specific foundry technology and PDK.

This paper is organized as follows: Section II presents our M3D case study. Section III presents our analytical framework and quantifies M3D EDP benefits for several AI/ML accelerators, dataflows, and workloads, evaluating the impact of BEOL memory selector width, BEOL ILV pitch, and additional M3D layers. Section IV concludes this paper.

II. Case-Study: Iso-Footprint, Iso-On-Chip-Memory-Capacity M3D IC vs. 2D IC using Foundry M3D Technology

Ultra-dense M3D based on sequential fabrication of 3D layers [5-6] requires upper layers of logic and memory to be *BEOL-compatible*, i.e., fabricated at temperatures less than 400°C. Existing fine-pitch (e.g., < 100nm) ILVs – traditionally used in BEOL metal interconnects – are used for vertical connectivity between 3D layers. Low-temperature fabrication of several logic and memory technologies [7-8] (e.g., CoolCube low-temperature silicon FETs, CNFETs, 2D FETs, oxide semiconducting FETs, Resistive RAM, Magnetoresistive RAM, Ferroelectric FETs) naturally enables such M3D chips. For this case study, we focus on M3D integration of a single layer of CNFET and a single layer of RRAM on top silicon CMOS transistors because: (a) These technologies have already been established within a commercial foundry [5]. (b) M3D VLSI design infrastructure (foundry M3D PDK with foundry M3D standard cell library) has been developed [5]. (c) The M3D design infrastructure was made available to us. We extend our M3D design principles beyond these specific M3D technologies through our analytical framework in Sec. III.

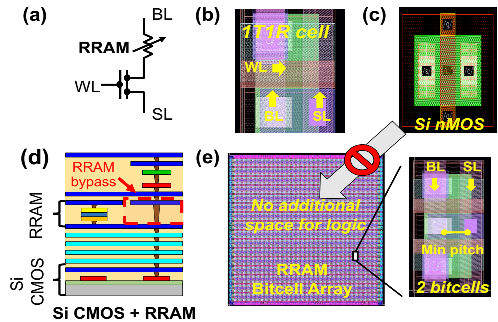


Fig. 3: RRAM and RRAM access transistor with labeled source line (SL), bit line (BL) and word line (WL): (a) schematic (b) layout (c) Layout of min width Si nMOS FET. (d) Vertical stack of RRAM over Si CMOS in PDK. (e) Layout of RRAM array at maximum bit-cell density. RRAM access transistors, no additional Si CMOS circuits (e.g., any transistors, or any larger circuits) can be placed below the RRAM array.

Our baseline 2D SoC design is based on an open-source AI/ML accelerator [9-10]. We refine this implementation, utilizing only RRAM for on-chip AI/ML model weight storage and leveraging optimizations reported by a recent state-of-the-art RRAM-on-Si-CMOS AI/ML accelerator SoC demonstrated in hardware (Fig. 2a, [10]). We chose this baseline architectural design point because: (a) We desire an iso-on-chip-memory-capacity comparison, so we use an accelerator design with ***only*** on-chip memory (i.e., our baseline starts with the benefits of eliminating expensive off-chip memory accesses) (b) The optimized accelerator architecture uses ~1/20th the SRAM buffers with 1/3rd the buffer access energy vs. a comparable architecture with off-chip memory (e.g., DRAM) leveraging on-chip RRAM's high bandwidth in reading AI/ML model weights (c) Since the foundry provides RRAM, a dense BEOL-compatible memory technology [11], an RRAM-based design provides an apples-to-apples comparison while reducing total area devoted to memory (i.e., using a 2D memory like SRAM would require additional area). RRAM non-volatility also helps eliminate idle energy (high for large SRAM-based accelerators) between sporadic AI/ML tasks common in edge applications.

Fig. 2a-b shows the baseline 2D SoC design. While [10] implements just 2 MB of RRAM per chip, we choose a RRAM capacity of 64 MB to fit large AI/ML workloads (e.g., ResNet-152, model size ~60M parameters). The accelerator computing sub-system (*'CS'* in Fig. 2b) is a 16×16 systolic array of processing elements (*PEs*), using a weight stationary dataflow (inputs streamed into the systolic array, weights remain stationary in the *PE*s), which has high utilization on AI/ML workloads such as convolutional neural networks [10]. For the baseline 2D design (Fig. 2a-b), RRAM devices are

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 13:03:34 UTC from IEEE Xplore. Restrictions apply.

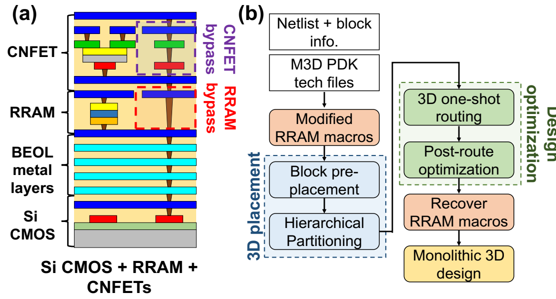


Fig. 4: (a) Vertical stack-up of foundry M3D PDK integrating CNFETs, RRAM and Si CMOS at a 130 nm node. (b) M3D RTL-to-GDS flow.

integrated vertically above RRAM access transistors implemented with Si nMOS or pMOS FETs (Fig. 3a-d). The Si CMOS tier underneath the RRAM cell array (the RRAM cells and RRAM access transistors) is therefore fully occupied (Fig. 3e), with no space available to place additional Si CMOS circuits below the RRAM cell array. Our computing sub-system (CS) is therefore placed adjacent to the RRAM cell arrays (Fig. 2b).

Our M3D implementation uses the foundry supplied CNFET layer above RRAM (Fig. 4a) to implement the RRAM access transistors. The Si CMOS tier underneath the RRAM cell array thus becomes available to place additional Si CMOS circuits. The BEOL metallization layers below the RRAM (highlighted in light blue, Figs. 3d, 4a) can then be used to route these additional Si CMOS circuits. Memory peripherals (sense amplifiers, controllers, etc.) are left in the Si CMOS tier creating blockages that must be placed and routed around. We must change the M3D architecture to utilize this additional Si CMOS space to the best advantage while staying within our iso-footprint and iso-on-chip-memory-capacity constraints. For this case study, we place 8× compute sub-systems operating in parallel, (each identical to the 2D configuration), with our RRAM capacity partitioned into 8× the banks to provide 8× total bandwidth.

The 130 nm technology node foundry M3D PDK (Fig. 4a) [5] is used for physical design of both the baseline 2D and iso-footprint, iso-on-chip-memory-capacity M3D designs. We restrict our baseline 2D implementation during synthesis to use only RRAM macros and Si CMOS devices; during place-and-route, a floorplan placement blockage in the CNFET layer ensures CNFET standard cells cannot be used (however all routing layers can be used – Fig. 3a). While the reported SoC on which we base our 2D design was originally optimized for the 40nm technology node [10], we relax our physical design target frequency (to 20MHz) to account for the difference in RRAM access between nodes, preserving the 40nm node architectural optimization benefits.

Our M3D RTL-to-GDS flow (Fig. 4b) uses standard Si EDA tools with custom place and route scripts (based on

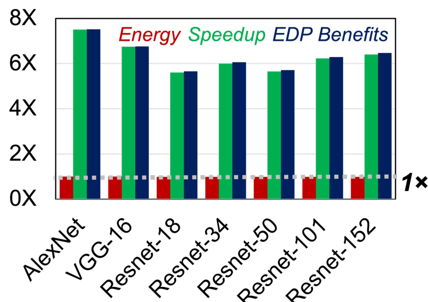


Fig. 5: Comparison of iso-footprint, iso-on-chip-memory-capacity M3D (Fig. 2c-d) vs. 2D (Fig. 2a-b) for different AI/ML model inference.

TABLE I: BENEFITS OF ISO-FOOTPRINT, ISO-ON-CHIP-MEMORY-CAPACITY M3D ACCELERATOR FOR ALL RESNET-18 LAYERS

Layer	Speedup	Energy	EDP benefits
CONV1+POOL	3.14×	1.0×	2.93×
L1.0 CONV1	3.72×	1.0×	3.73×
L1.0 CONV2	3.72×	0.99×	3.73×
L1.1 CONV1	3.72×	0.99×	3.73×
L1.1 CONV2	3.72×	0.99×	3.73×
L2.0DS	2.57×	1.0×	2.57×
L2.0 CONV1	6.0×	0.99×	7.37×
L2.0 CONV2	7.36×	0.99×	7.37×
L2.1 CONV1	7.36×	0.99×	7.37×
L2.1 CONV2	7.36×	0.99×	7.37×
L3.0 DS	2.52×	1.0×	2.51×
L3.0 CONV1	6.84×	0.99×	6.85×
L3.0 CONV2	7.67×	0.99×	7.68×
L3.1 CONV1	7.67×	0.99×	7.68×
L3.1 CONV2	7.67×	0.99×	7.68×
L4.0 DS	3.5×	1.0×	3.5×
L4.0 CONV1	7.37×	0.99×	7.4×
L4.0 CONV2	7.83×	0.99×	7.85×
L4.1 CONV1	7.83×	0.99×	7.85×
L4.1 CONV2	7.83×	0.99×	7.85×
Total	5.64×	0.99×	5.66×

techniques reported in [4]) and is compatible with state-of-the-art technology nodes. The M3D flow starts from a synthesized netlist using RRAM macro modules. RRAM macro blockages are modified from their 2D baseline to account for the partial blockage (i.e., in the CNFET and RRAM layers) of the RRAM cell arrays, and full blockage for the peripherals. Other macros (e.g., SRAM buffers) are pre-placed during floorplanning using their corresponding layers (e.g., Si CMOS). With the generated floorplan, a custom monolithic 3D place and route flow (Fig. 4b) generates a M3D physical design solution. Post-route optimization is performed to meet power and timing constraints. System-level EDP is assessed after physical design using the achieved frequency, post-physical-design power, and architectural simulations to determine the AI/ML workload cycle count for each design. Synopsys DC compiler was used for the RTL netlist synthesis with P&R performed using modified Cadence Innovus scripts. Power analysis is performed using Cadence Tempus with default activation factors. 2D and 3D designs are given identical target frequencies throughout. Fig. 2 summarizes the detailed post route comparisons of our two designs. From this case study we make several observations:

Observation 1: For a variety of AI/ML models (e.g., AlexNet, VGG, ResNet), our iso-footprint and iso-on-chip-memory-capacity M3D design achieves 5.7× to 7.5× speedup at 0.99× energy, resulting in 5.7× to 7.5× EDP benefits (Fig. 5). As an example, we show detailed layer by layer benefits for one model (ResNet-18) in Table I.

Observation 2: For our implemented M3D design, the power dissipated in the upper layers (CNFET and RRAM) is <1% of the total chip power (power-hungry memory peripherals/controllers are still located in Si CMOS). Thus, our M3D peak power density increases by just 1% vs. the 2D design, and our M3D implementation doesn't require additional thermal management ([1] reports similar findings).

Observation 3: Our 2D baseline already contains a dense, BEOL-compatible memory (RRAM). Using a non-BEOL-compatible memory (e.g., a Si CMOS SRAM that is 2× less dense) in our 2D baseline would result in a larger 2D baseline area, enabling additional computing subsystems in our M3D design (e.g., going from 8× *CSs* to 16× *CSs* increasing EDP benefits from 5.7× to 6.8×); our M3D benefits are therefore conservative vs. traditional non-BEOL-compatible memories.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 13:03:34 UTC from IEEE Xplore. Restrictions apply.

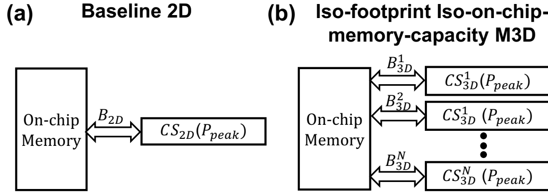


Fig. 6: (a) Baseline 2D chip with a computing sub-system (CS_{2D}) (b) Iso-footprint, iso-on-chip-memory-capacity M3D chip with N parallel computing sub-systems (CS_{3D}^1, …, CS_{3D}^N).

III. Analytical Framework

Our baseline 2D chip (Fig. 6a – implemented using Si CMOS in the bottom tier and on-chip RRAM in the second tier) has a total area A_{2D} and consists of a computing sub-system (or *'CS'* - which can contain many parallel processing elements or other circuits for computation) with area $A_{C,2D}$, on-chip memory area $A_{M,2D}$, and area of system buses and I/O $A_{bus,2D}$. Memory area has two components – memory cells – $A_{M,2D}^{cells}$ (e.g., the area occupied by RRAM bitcell array, Fig. 3, comprising of RRAM cells and RRAM access transistors), and memory peripheral circuits/controllers – $A_{M,2D}^{perif}$ (e.g., sense amplifiers, controllers, implemented in the Si CMOS tier). We define $\gamma_{2D}^{cells} = A_{M,2D}^{cell}/A_{C,2D}$ (the area ratio of memory cells and the *CS*), and $\gamma_{2D}^{perif} = A_{M,2D}^{perif}/A_{C,2D}$ (the area ratio of memory peripherals and the *CS*). The computing sub-system in our baseline 2D chip can perform at most P_{peak} operations per cycle (i.e., all the processing elements are active at full utilization) assuming a maximum memory bandwidth of B_{2D} bits per cycle for P_{peak}.

A. Speedup: We analyze execution time on the baseline 2D chip of a workload requiring F_0 compute operations on D_0 bits of data (available inside on-chip memory). Based on [12], the minimum time (in cycles) required to execute this workload in our baseline 2D chip (Fig. 6a) will be the maximum of data transfer and computation times (assuming that the workload is either limited by compute time or the memory access time):

$$T_{C,2D} = \max\left(\frac{D_0}{B_{2D}}, \frac{F_0}{P_{peak}}\right) \quad (1)$$

For an iso-footprint, iso-on-chip-memory-capacity M3D chip with an upper 3D tier of memory access transistors (e.g., CNFETs) and the freed space in the Si CMOS tier used to place N parallel *CS* in the M3D chip (Fig. 6b), then

$$N = \left\lfloor 1 + \frac{A_{M,2D}^{cells}}{A_{C,2D}} \right\rfloor = \left\lfloor 1 + \gamma_{2D}^{cells} \right\rfloor \quad (2)$$

Note that, while we perform this analysis at the granularity of a computing sub-system, the same analysis can be performed at a finer granularity e.g., at the level of processing elements constituting the computing sub-system. Given a total M3D memory bandwidth of B_{3D}, equally partitioned among these N parallel *CS*, memory bandwidth for each parallel *CS* (labeled as $B_{3D}^1, B_{3D}^2, \dots, B_{3D}^N$ in Fig. 6b) will be B_{3D}/N. We further assume that the F_0 compute operations in the given workload can be partitioned into at most $N^{\#}$ parallel partitions. Therefore, while executing this workload, at most N_{max} (where, $N_{max} = min(N^{\#}, N)$) parallel *CS*s in M3D can be utilized in parallel. Extending (1), execution time of this workload in M3D (in cycles) will be:

$$T_{C,3D} = max\left(\frac{D_0 N}{B_{3D}}, \frac{F_0}{N_{max} P_{peak}}\right) \quad (4)$$

So, speedup for M3D in this particular case will be:

$$speedup_{M3Dv2D} = \frac{T_{C,2D}}{T_{C,3D}} \quad (5)$$

B. Energy: Let the average memory read/write energy be α_{2D} J/bit (baseline 2D) and α_{3D} J/bit (M3D). For D_0 bits of memory access (including both reads and writes) in the given workload, memory access energy (in J) in the baseline 2D will be: $E_{M,2D} = \alpha_{2D} D_0$, and in M3D will be $E_{M,3D} = \alpha_{3D} D_0$. Now, we consider compute energy per operation of the *CS* in the baseline 2D chip is $E_{C,2D}$ J/op and for each of the N parallel *CS*s in the M3D chip is $E_{C,3D}$ J/op. Since the *CS* in the baseline 2D chip and the parallel *CS*s in the M3D chip is implemented in Si CMOS, $E_{C,3D} = E_{C,2D} = E_C$. Similarly, the idle energy of the *CS* in the baseline 2D chip and each of the N parallel *CS*s in the M3D chip is assumed as E_C^{idle} J/cycle. Additionally, we assume that the idle energy for the memory in baseline 2D and M3D will be $E_{M,2D}^{idle}$ J/cycle, and $E_{M,3D}^{idle}$ J/cycle respectively. Total energy spent executing the given workload for baseline 2D and M3D will be:

$$E_{2D} = \alpha_{2D} D_0 + E_{M,2D}^{idle}\left(T_{C,2D} - \frac{D_0}{B_{2D}}\right) + E_C^{idle}\left(T_{C,2D} - \frac{F_0}{P_{peak}}\right) + E_C F_0 \quad (6)$$

$$E_{3D} = \alpha_{3D} D_0 + E_{M,3D}^{idle}\left(T_{C,3D} - \frac{D_0 N}{B_{3D}}\right) + (N - N_{max}) E_C^{idle} T_{C3D} + N E_C^{idle}\left(T_{C,3D} - \frac{F_0}{N_{max} P_{peak}}\right) + E_C F_0 \quad (7)$$

C. EDP Benefits: Using (5), (6), (7), M3D EDP benefits are:

$$EDP_{M3Dv2D} = speedup_{M3Dv2D} \frac{E_{2D}}{E_{3D}} \quad (8)$$

With this analytical model we make several observations on the architecture changes that lead to M3D benefits:

<u>*Observation 4:*</u> We evaluate the EDP benefits of iso-footprint, iso-on-chip-memory-capacity M3D ICs vs. corresponding 2D designs for a variety of architectural design points (Table II) based on AI accelerator architectures with different organization of the *CS* (i.e., with different tiling of processing elements inside the *CS*, different memory hierarchies). For these design points, we estimate that M3D EDP benefits can range from 5.3×-11.5× over corresponding iso-footprint, iso-on-chip-memory-capacity baseline 2D designs (Fig. 7). Critically, for all these design points, our calculated EDP benefits are within 10% of the results predicted from a state-of-the-art architectural simulator for

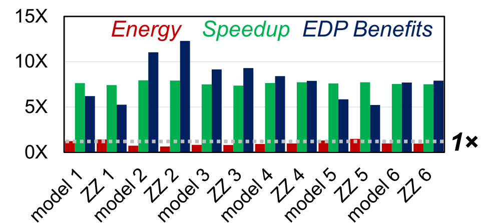


Fig. 7: Energy & delay benefits for different accelerator architectures for AlexNet (listed in Table II) inference. Arch. 1-5 are variants of popular AI accelerators [14-18] and Arch. 6 is a version of the accelerator we designed in Sec. II. We evaluate all architectures using ZigZag (labeled as ZZ) [13] and our analytical model. All architectures are normalized with the same total number of processing elements (PEs) in the *CS* and same on-chip RRAM capacity. Our results are within 10% of ZigZag predicted speedup, energy and EDP benefits.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 13:03:34 UTC from IEEE Xplore. Restrictions apply.

TABLE II: ACCELERATOR ARCHITECTURES EVALUATED

Arch #	PE spatial[a] (K, C, OX, OY)	Reg/PE or PE group (B)	SRAM local (KB)	SRAM global (MB)	On-chip RRAM
1	16,16,2,2	W: 1 O: 2	W: 64 I: 64 O: 256	2	256 MB
2	8,8,4,4	W: 1 O: 2	W: 32	2	256 MB
3	32,32,-,-	W: 128 O: 1K	-	2	256 MB
4	32,2,4,4	W: 1 O: 2	W: 64 I: 32	2	256 MB
5	32,-,8,4	W: 1 O: 4	W: 1 I: 1	2	256 MB
6	32,32	W&I: 2.2 O:1	I:32 O:32	0.5	256 MB

[a]K&C = output & input channels, OX&OY = output width and heights

AI/ML SoCs (Fig. 7, [13]) as well as with our physical design case study results (Sec. II).

Observation 5: For compute-bound workloads (i.e., where compute operations ≫ memory accesses), additional parallel *CS*s in M3D when given sufficient per-*CS* bandwidth (e.g., nearly the same or better than baseline 2D) results in better EDP (Fig. 8a). As an example, for a workload requiring 16 operations for every bit of memory access, our M3D design point will have 2.1× better EDP with a 2× increase in CSs with no change in bandwidth. For memory bound workloads (i.e., where compute operations ≪ memory accesses), increasing bandwidth to each CS (by placing more memory peripherals/controllers in Si CMOS tier vs. placing additional parallel CSs in M3D) will have better EDP benefits. As an example, for a memory bound workload which requires 16 bits of memory access for every compute operation, our M3D design point will see a 2.1× better EDP even with 2× fewer CSs but with 2× better bandwidth per CS.

Observation 6: Baseline 2D designs with a larger on-chip memory capacity will see more area freed up in the Si CMOS tier by moving memory access FETs in M3D. This results in higher compute parallelism and larger EDP benefits. As an example, for ResNet-18, M3D EDP benefits go up from 1× to 6.8× when RRAM capacity in baseline 2D designs is increased from 12 MB to 128 MB (Fig. 9).

D. Case 1 – M3D memory access FET drive strength: We consider relaxing the width of M3D BEOL memory access FETs (e.g., if the CNFETs in our case study achieve lower on-current vs. ideal as they are newly introduced technologies). With relaxed M3D FET widths, memory bitcell sizes in M3D will be proportionally larger. As we want to maintain our iso-on-chip-memory-capacity constraint, we must increase both the M3D and 2D footprints considering the larger area of a relaxed memory cell array in M3D (Fig. 10a). We can then re-optimize the 2D baseline with additional parallel *CS*s to match the M3D footprint (Fig. 10b). For a M3D FET width relaxation of δ, the total area for memory cells in M3D will be $A_{M,3D}^{cells} = \delta A_{M,2D}^{cells}$. Now if, $A_{M,3D}^{cells} < A_{2D}$ (the starting baseline 2D footprint), there is no increase required for the M3D chip footprint. However, if $A_{M,3D}^{cells} >$

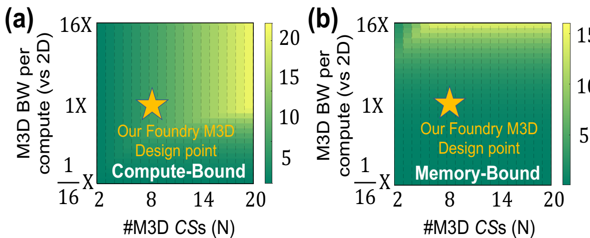


Fig. 8: M3D EDP benefits (vs. baseline 2D) varying bandwidth and number of parallel *CS*s.

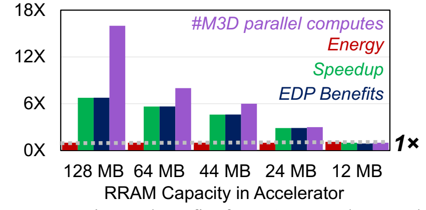


Fig. 9: RRAM capacity vs. benefits for M3D Accelerator shown in Sec. II, the DNN compute remain unchanged for ResNet-18 (~12M parameters, to fit in RRAM capacities from 12 MB to 128 MB).

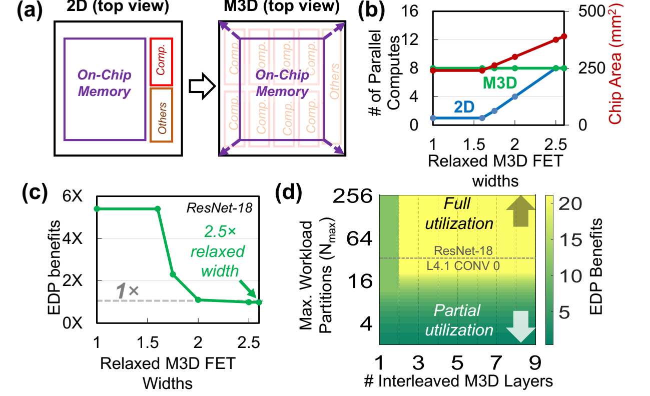


Fig. 10: (a) Moving memory access transistors to the upper M3D FETs relaxes memory area constraint in M3D. (b) Number of parallel *CS*s that be placed in M3D and in a commensurately large 2D baseline with relaxed M3D FET widths. (c) M3D EDP benefits remain even with 2.5× relaxed M3D FET widths. (d) EDP benefits with multiple interleaved 3D compute & memory tiers in M3D vs maximum parallel workload partitions (N_{max} - defined in sub-section A, Sec. III). For (d) workload has similar compute operations and memory accesses as ResNet-18.

A_{2D}, a commensurately larger 2D baseline can host up to N_{2D}^{new} parallel *CS*s in the Si CMOS tier with:

$$N_{2D}^{new} = \left\lfloor \frac{max\left(\delta A_{M,2D}^{cells} - A_{2D}, A_{C,2D}\right)}{A_{C,2D}} \right\rfloor \quad (9)$$

If the given workload can be maximally partitioned into $N^{\#}$ parallel partitions, $N_{max,2D}^{new} = min\left(N^{\#}, N_{2D}^{new}\right)$ parallel *CS*s can be fully utilized for this workload in the new 2D baseline. Using (5) and (9), the M3D speedup becomes:

$$speedup_{M3Dv2D}^{new} = \frac{T_{C,2D}^{new}}{T_{C,3D}} = \frac{max\left(\frac{D_0 N_{2D}^{new}}{B_{2D}}, \frac{F_0}{N_{max,2D}^{new} P_{peak}}\right)}{max\left(\frac{D_0 N}{B_{3D}}, \frac{F_0}{N_{max} P_{peak}}\right)} \quad (10)$$

Similarly, energy cost in the new baseline 2D will be:

$$E_{2D}^{new} = \alpha_{2D} D_0 + E_{M,2D}^{idle}\left(T_{C,2D}^{new} - \frac{D_0 N_{2D}^{new}}{B_{2D}}\right) + \left(N_{2D}^{new} - N_{max,2D}^{new}\right) E_C^{idle} T_{C,2D}^{new} + N_{2D}^{new} E_C^{idle}\left(T_{C,2D}^{new} - \frac{F_0}{N_{max,2D}^{new} P_{peak}}\right) + E_C F_0 \quad (11)$$

Thus, M3D EDP benefits for this case will be (Fig. 10c):

$$EDP_{M3Dv2D}^{new} = speedup_{M3Dv2D}^{new} \frac{E_{2D}^{new}}{E_{3D}} \quad (12)$$

Observation 7: Our M3D architecture has no loss of EDP benefits with up to 1.6× relaxed M3D memory access FET widths in the upper M3D tier (Fig. 10c – small benefits are retained even up to 2.5× relaxed M3D FET widths). Thus, our M3D architectures tolerate up to 40% lower M3D FET drive strength while still providing 5.7× EDP benefits.

E. Case 2 – M3D via pitch: M3D via density limits memory cell size, as every memory cell (e.g., RRAM) needs to be connected to the upper M3D layer of memory access FETs (e.g., CNFETs). If M3D vertical via pitch is β, and memory

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 13:03:34 UTC from IEEE Xplore. Restrictions apply.

cell area is via-pitch limited, then total memory cell area, $A_{M,3D}^{cells} = mk\beta^2$, where k is the memory capacity (in bits) and m is the number of 3D vertical vias needed per memory cell. EDP benefits for $A_{M,3D}^{cells} > A_{2D}$ can be evaluated following the same method shown above (in *Case 1*, i.e., EDP benefits are unchanged when $A_{M,3D}^{cells} < A_{2D}$).

Observation 8: While minor increases in M3D via pitch (e.g., up to 1.3×) do not change our M3D architecture EDP benefits, transitioning from fine-pitch M3D ILVs to coarse-pitch M3D vias (i.e., increasing up to 1.6× or more) result in limited to no benefit over 2D designs. Thus, ultra-dense M3D vias are key to these M3D architectural benefits.

F. Case 3 – Multiple interleaving M3D compute & memory tiers: By adding additional layers of compute and memory devices to a M3D chip we can increase both the memory capacity, memory bandwidth, and number of parallel *CS*s. In this vast arbitrary X-tier M3D design space, we consider a particular case with interleaved tiers of compute and memory in the M3D chip – i.e., some Y tiers of compute and Y tiers of memory (in our physical design case study Y = 1). Assuming no benefits from increased memory capacity, we extend our formulation for M3D EDP benefits by including additional M3D parallel *CS*s. If each memory tier has its own memory peripherals/controllers and I/O circuits, then we have $N = Y\lfloor 1 + \gamma_{2D}^{cells} + \gamma_{2D}^{perif} \rfloor$ total parallel *CS*s in M3D.

Observation 9: Adding just one pair of M3D compute and memory layers increase M3D EDP benefits from 5.7× to 6.9× for ResNet-18. However, these EDP benefits quickly plateau to 7.1× with additional layers as total parallel *CS*s in the M3D design exceeds maximum possible parallelizable workload partitions (Fig. 10d). For highly parallelizable workloads, this EDP benefit plateau can be substantially higher (e.g., ResNet-18 - layer L4.1 CONV 0 approaches 23×).

With multiple layers of stacked memory and compute, thermal management becomes necessary [19]. Let the resistance to ambient (e.g., determined by on-chip heat sink) as R_0 K/W. Then, if each j-th interleaving tier of compute and memory adds a thermal resistance of R_j K/W and a total power of P_j W, (including both compute power - $P_{C,j}$W and memory power - $P_{M,j}$W, i.e., $P_j = P_{C,j} + P_{M,j}$). Total temperature rise in this system can be modeled as:

$$Temp_{rise} = \sum_{i=1}^{Y}((\sum_{j=1}^{i} R_j) + R_0) \times P_i \qquad (17)$$

Observation 10: Given a maximum allowed temperature rise (typically ~60 K [20]), (17) quickly limits the maximum number of interleaved M3D tiers, which must be taken in consideration while calculating EDP for such systems.

IV. Conclusion

This paper establishes the following key points about ultra-dense monolithic 3D integration: (1) Even under a conservative iso-footprint and iso-on-chip-memory-capacity constraints, monolithic 3D integration of logic and memory provides large EDP benefits vs. comparable 2D designs with today's foundry M3D technologies. (2) These benefits are with newly implemented M3D technologies and will grow with further performance optimization (e.g., full CMOS on upper layers) and more features (e.g., additional layers). (3) The key to these results is co-optimizing architecture with physical design; such co-design may prove important for other emerging technologies beyond M3D. (4) Our analytical framework takes this analysis beyond a specific foundry technology deriving the unique characteristics of M3D physical design and corresponding architectural design points which should apply for many other M3D technologies. (5) While not specifically addressed in this paper, such large benefits suggest exploration of other ultra-dense 3D manufacturing techniques and any unique physical-design-co-optimized architectures they might offer.

Acknowledgment

This research was supported in part by DARPA 3DSoC, Stanford SystemX Alliance, Stanford Precourt Institute for Energy, U.S. Department of Energy, Ceremorphic, Samsung and ACCESS – AI Chip Center for Emerging Smart Systems, sponsored by InnoHK funding, Hong Kong SAR.

References

[1] M. M. S. Aly, T. F. Wu, A. Bartolo, Y. H. Malviya, W. Hwang, G. Hills, et al. "The N3XT approach to energy-efficient abundant-data computing." *Proceedings of the IEEE, 2018*, 107(1), pages- 19-48.

[2] R.M. Radway, K. Sethi, W-C. Chen, J. Kwon, S. Liu, et al. "The Future of Hardware Technologies for Computing: N3XT 3D MOSAIC, Illusion Scaleup, Co-Design." *IEEE International Electron Devices Meeting*, 2021, pp. 25-4.

[3] H. Park, K. Chang, B. W. Ku, J. Kim, E. Lee, D. Kim, A. Chaudhuri et al. "RTL-to-GDS tool flow and design-for-test solutions for monolithic 3D ICs." *ACM/IEEE Design Automation Conference*, 2019, pp. 1-4.

[4] J. Kim, G. Murali, P. Vanna-Iampikul, E. Lee, D. Kim, et al. "RTL-to-GDS design tools for monolithic 3D ICs." In *2020 IEEE/ACM International Conference On Computer Aided Design*, 2020, pp. 1-8.

[5] T. Srimani, G. Hills, M. Bishop, C. Lau, P. Kanhaiya, R. Ho, et al. "Heterogeneous integration of BEOL logic and memory in a commercial foundry: Multi-tier complementary carbon nanotube logic and resistive RAM at a 130 nm node." *IEEE Symp. VLSI Tech.*, 2020, pp. 1-2.

[6] M. Vinet, P. Batude, C. Fenouillet-Beranger, F. Clermidy, L. Brunet, O. Rozeau, J. M. Hartmannn et al. "Monolithic 3D integration: A powerful alternative to classical 2D scaling." *SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, 2014, pp. 1-3.

[7] C. D. Llorente, C. Le Royer, P. Batude, C. Fenouillet-Beranger, S. Martinie, C-MV Lu, F. Allain et al. "New insights on SOI Tunnel FETs with low-temperature process flow for CoolCube™ integration." *Solid-State Electronics* 144 (2018), pp. 78-85.

[8] S. Salahuddin, K. Ni, and S. Datta. "The era of hyper-scaling in electronics." *Nature Electronics*, 2018, 1(8), pp. 442-450.

[9] DNN-accelerator: https://code.stanford.edu/kprabhu7/dnn-accelerator

[10] M. Giordano, K. Prabhu, K. Koul, R.M. Radway, A. Gural, R. Doshi, Z.F. Khan et al. "Chimera: A 0.92 tops, 2.2 tops/w edge ai accelerator with 2 mbyte on-chip foundry resistive ram for efficient training and inference." *IEEE Symposium on VLSI Circuits*, *2021*, pp. 1-2.

[11] E.R. Hsieh, X. Zheng, B. Q. Le, Y. C. Shih, R. M. Radway, M. Nelson, S. Mitra, and S. Wong. "Four-bits-per-memory one-transistor-and-eight-resistive-random-access-memory (1T8R) array." *IEEE Electron Device Letters*, 2021, 42(3), pp. 335-338.

[12] M. Hill, and V.J. Reddi. "Gables: A roofline model for mobile socs." *IEEE International Symposium on High Performance Computer Architecture*, 2019, pp. 317-330.

[13] L. Mei, P. Houshmand, V. Jain, S. Giraldo, and M. Verhelst. "ZigZag: Enlarging joint architecture-mapping design space exploration for DNN accelerators." *IEEE Transactions on Computers,* 2021, 70(8), pp: 1160-1174.

[14] H. E. Sumbul, T. F. Wu, Y. Li, S. S. Sarwar, W. Koven, E. Murphy-Trotzky, et al. "System-Level Design and Integration of a Prototype AR/VR Hardware Featuring a Custom Low-Power DNN Accelerator Chip in 7nm Technology for Codec Avatars." *IEEE Custom Integrated Circuits Conference*, 2022, pp. 1-8.

[15] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, et al. "In-datacenter performance analysis of a tensor processing unit." *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1-12.

[16] A. Yazdanbakhsh, K. Seshadri, B. Akin, J. Laudon, and R.Narayanaswami. "An evaluation of edge tpu accelerators for convolutional neural networks." *arXiv preprint* (2021), arXiv:2102.10423.

[17] H. Liao, J. Tu, J. Xia, H. Liu, X. Zhou, H. Yuan, and Y. Hu. "Ascend: a scalable and unified architecture for ubiquitous deep neural network computing: Industry track paper." *IEEE International Symposium on High-Performance Computer Architecture*, 2021, pp. 789-801.

[18] E. Talpes, D. Das Sarma, G. Venkataramanan, P. Bannon, B. McGee, B. Floering, et al. "Compute solution for tesla's full self-driving computer." *IEEE Micro,* 2020, 40(2), pp: 25-35.

[19] P. Shukla, A. K. Coskun, V. F. Pavlidis, and E. Salman. "An overview of thermal challenges and opportunities for monolithic 3D ICs." *Great Lakes Symposium on VLSI,* 2019, pp. 439-444.

[20] W. Steller, F. Windrich, D. Bremner, S. Robertson, R. Mroßko, J. Keller, et al. "Microfluidic Interposer for High Performance Fluidic Chip Cooling." *7th IEEE Electronic System-Integration Technology Conference*, 2018, pp. 1-8

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 13:03:34 UTC from IEEE Xplore. Restrictions apply.