Three-Dimensional Stacked Neural Network Accelerator Architectures for AR/VR Applications

Lita Yang, Robert M. Radway, Yu-Hsin Chen, Tony F. Wu, Huichu Liu, Elnaz Ansari, and Vikas Chandra, *Reality Labs, Meta, Sunnyvale, CA, 94089, USA*

Subhasish Mitra, *Stanford University, Stanford, CA, 94305, USA*

Edith Beigné, *Reality Labs, Meta, Sunnyvale, CA, 94089, USA*

Three-dimensional integration offers architectural and performance benefits for scaling augmented/virtual reality (AR/VR) models on highly resource-constrained edge devices. Two-dimensional off-chip memory interfaces are too prohibitively energy intensive and bandwidth (BW) limited for AR/VR devices. To solve this, we propose using advanced 3-D stacking technology for high-density vertical integration to local memory and compute, increasing memory capacity within the same footprint at iso-BW with improvements in energy and latency. We evaluate 3-D architectures for a prototype AR/VR accelerator to demonstrate up to 3.9× latency reduction and 1.6× lower energy compared to a 2-D configuration within a smaller/similar footprint. Additionally, we show the feasibility of deploying higher resolution AR/VR models by stacking multiple tiers of memory, providing a pathway to break the footprint constraints of 2-D architectures. The use of high-density 3-D interconnects allows us to demonstrate localized benefits at the accelerator-level compared with standard system-on-chip memory disaggregation techniques/architectures.

The success of deep learning algorithms has led to breakthroughs in using neural networks (NNs) for state-of-the-art performance in augmented/virtual reality (AR/VR) applications. Emerging artificial intelligence (AI) and machine learning (ML)-enabled AR/VR tasks include object detection, image segmentation, eye and hand tracking, and depth estimation.[1] Deploying these NNs onto edge devices, such as AR glasses and wearables, would enable a new paradigm of next-generation human interaction and computing.

As more AR/VR workloads become NN-heavy, AI/ML accelerators become the dominant energy consumer in full-scale systems on chips (SoCs).[2] The small form factors of these devices, however, impose stringent demands on footprint, memory capacity, energy efficiency for long(er) battery life and real-time processing latency to enable a seamless and enjoyable user experience.[1] Even with custom silicon chips highly optimized for NN acceleration, over 50% of the total power goes to memory accesses and strict power budgets make off-chip dynamic random-access memory (DRAM) accesses too prohibitively expensive for AR/VR devices.[1]

Additionally, given the tight footprint constraints, 2-D solutions are unable to continue scaling in the X-Y directions to enable deployment of large AR/VR NN models onto edge devices due to power, memory capacity, and BW constraints.[3] Three-dimensional integration approaches have been proposed in recent years to cope with device scaling challenges, including micro-bumping, hybrid bonding, and monolithic 3-D integrated circuits (ICs).[4] In particular, hybrid bonding enables high-density, fine-pitch 3-D interconnect integration by using face-to-face (F2F) bond pads to stack 2-D wafers. This enables higher density vertical integration, allowing for high BWs with low-energy and low-latency access to local memory and computing.

In this article, we explore the design tradeoffs of 3-D stacked NN accelerator architectures for AR/VR

0272-1732 © 2022 IEEE
Digital Object Identifier 10.1109/MM.2022.3202254
Date of current version 28 October 2022.

workloads and highlight the improvements (energy, latency) enabled by 3-D hybrid bonding within a smaller or similar footprint to our 2-D baseline with no off-chip DRAM accesses. Our key findings and objectives are as follows.

1) *Evaluate energy and latency improvements with 3-D stacked NN accelerator architectures:* Using 3-D integration to increase activation memory from 1 MB to 2–8 MB improves latency by 1.2–1.9× and energy by 1.1–1.4× for 31% smaller footprint. For a slight footprint increase (38% overhead), using 3-D stacking to increase activation memory from 1 MB to 4–16 MB and the multiply-accumulate (MAC) array from 1K to 2K results in energy savings of 1.3–1.6× and latency reduction of 2.4–3.9× for depth estimation, denoising, and super-resolution AR/VR NN models.
2) *Enable deployment of larger AR/VR workloads in edge devices:* The ability to stack up to 16 MB of activation memory within a similar footprint to our 2-D baseline enables deployment of larger super-resolution models (1,024 × 1,024 resolution) not previously feasible with a 2-D form factor.
3) *Explore the 2-D to 3-D design space with a preliminary modeling framework:* The results show greater benefits for a 3-D architecture design over a 2-D folded into 3-D. The use of 3-D hybrid bonding demonstrates localized benefits at the accelerator-level compared with standard SoC memory disaggregation techniques/architectures.

MOTIVATION

Challenges With AR/VR Workloads

The key features of deploying AR/VR NNs onto edge devices include 1) large activation memory footprints due to high input/output resolution requirements, and 2) the need to support a wide set of convolution operations and different types of computer-vision-based AR/VR models on a single accelerator for cost reasons. Because of this, our evaluated prototype ML accelerator is designed with a scalable compute array to meet frame rate (FPS) requirements and support versatility to deploy various AR/VR workloads. Given the large activation memory requirements, we model an accelerator that is scaled-up from Sumbul et al.[5] with an architecture similar to Shao et al.[6] using a distributed on-chip buffer interspersed with compute units.

Even with a highly optimized accelerator targeting AR/VR workloads, memory accesses due to large activation memory requirements remain a challenge for meeting energy and latency requirements. Since weights are fixed during inference while activations differ for each input sample, techniques such as quantization, pruning, and retraining are often applied to weights only and activations are less likely to be compressed before runtime leading to large activation memory footprints in edge devices. Figure 1 illustrates a sweep of AR/VR workloads with high heterogeneity across weight and activation memory requirements, and how a few workloads (i.e., HRNet, Agg, denoising, and super-resolution) are more activation heavy than weight heavy.

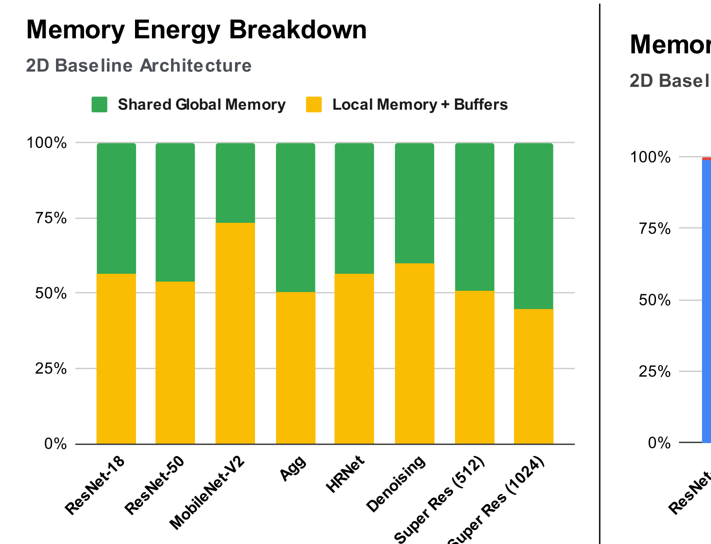
THE USE OF HIGH-DENSITY 3-D INTERCONNECTS ALLOWS US TO DEMONSTRATE LOCALIZED BENEFITS AT THE ACCELERATOR-LEVEL COMPARED WITH STANDARD SYSTEM-ON-CHIP MEMORY DISAGGREGATION TECHNIQUES/ ARCHITECTURES.

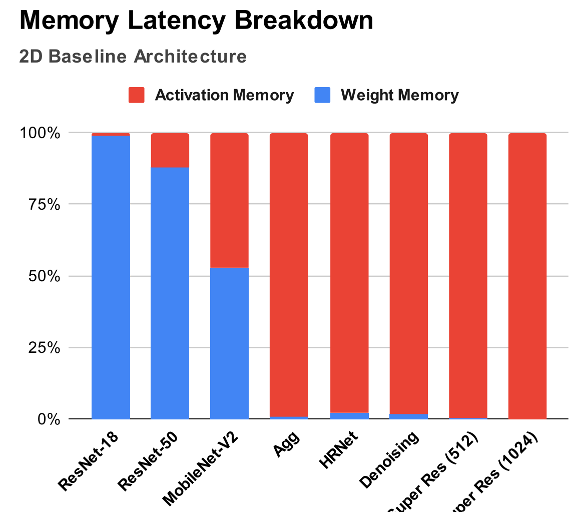
In particular, super-resolution becomes prohibitively expensive as we scale up the resolution requirement in Figure 2(a). With 4 MB of on-chip activation memory, the external BW requirements grow super linearly with increasing resolution. With 1,024 × 1,024 resolution at 90 FPS, the BW requirement exceeds that of typical DRAM/LPDDR4 peak BW of 16 GB/s, making it infeasible to support larger super-resolution cases. This is particularly challenging since classification accuracy, image fidelity, and overall AR/VR user experiences improve with higher resolutions, however, these networks typically need tens to hundreds of MB of memory, necessitating off-chip spilling to external memory.

Figure 2(b) demonstrates that if we were to increase the internal static random-access memory (SRAM) to 8–16 MB, this would alleviate the amount of off-chip spilling and reduce the BW requirement to acceptable ranges. The challenge with increasing internal SRAM capacity in 2-D is the highly stringent footprint constraints of AR/VR use cases, especially for AR glasses or wearables. A large central SRAM in 2-D would not only be prohibitively area expensive but would cause wiring density congestion and increase energy overhead between the compute cores and SRAM.

Three-Dimensional Integration for Increasing Activation Memory Capacity and Scalability

To solve the footprint, BW, and scalability issues of increasing on-chip activation memory, we propose using





AR/VR Workload/Task	NN Model
Object Classification	ResNet-18, ResNet-50 [7]
Object Detection	MobileNet-V2 [8]
Denoising	UNet + Feature-Align [9]
Depth Estimation	High-Res Net (HRNet) [10]
Depth Estimation	3D Aggregation (Agg) [11]
Super Resolution (512)	MFSR [12] @ (512×512) resolution
Super Resolution (1024)	MFSR [12] @ (1024×1024) resolution

FIGURE 1. Memory energy and latency breakdowns for AR/VR workloads, from literature, running on a 2-D ML accelerator baseline architecture (1K MAC array, 1-MB activation memory). Classic object classification and detection models are not as activation-heavy compared to depth estimation models (Agg, HRnet), denoising, and super-resolution. From the breakdown, we see a significant portion of the total memory energy (∼50%) is due to memory spillage to the shared global memory for these workloads.

External BW Requirement (GBps) vs. Resolution
Super Resolution, 4 MB Activation Memory, 90 FPS
BW (GBps)
200.00
150.00
100.00
50.00
0.00
1.00
28.98
165.87
512x512
1024x1024
2048x2048
Resolution
(a)

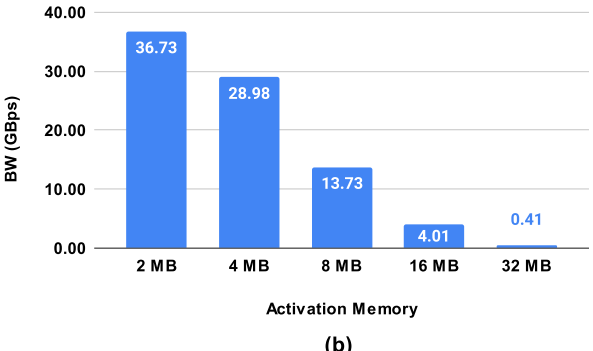


FIGURE 2. (a) External BW requirements for super-resolution at 4-MB activation memory, scaling resolution from 512×512 to 2,048×2,048; and (b) reduction of BW requirements as we increase local activation memory capacity for super-resolution at 1,024×1,024 image resolution.

Total Capacity*	2D Read/Write (R/W) Energy^	3D 2-Tier Read/Write (R/W) Energy**	3D 4-Tier Read/Write (R/W) Energy**
8 MB	8 MB (2D)	2 x 4 MB (3D/2-Tier)	4 x 2 MB (3D/4-Tier)
	1x	0.75x - 0.89x	1.13x - 1.15x
16 MB	16 MB (2D)	2 x 8 MB (3D/2-Tier)	4 x 4 MB (3D/4-Tier)
	1x	0.79x - 0.90x	1.01x - 1.08x
32 MB	32 MB (2D)	2 x 16 MB (3D/2-Tier)	4 x 8 MB (3D/4-Tier)
	1x	0.72x - 0.85x	0.87x - 0.99x

* 1 MB bank size, 8 ports
** Read/Write (R/W) energy @ 0.7V Vdd (7nm), ranges indicate Avg to Max energy
^ 2D R/W Energy Ratios (compared to 8 MB): 16 MB (2D) ~1.1-1.3x; 32 MB (2D) ~1.3-1.8x

FIGURE 3. Read/write energy comparison for the same memory capacity (8–32 MB) split across 3-D tiers (2–4 tier partitions).[3] Two-Tier 3-D access energy is lower than the 2-D equivalent due to shorter wirelengths but starts to increase or become iso-energy to the 2-D configuration at 4 tiers due to TSV energy overhead.

3-D integration to increase the internal SRAM capacity but within a similar X-Y footprint. Not only does 3-D stacking mitigate the footprint and BW restrictions, it can reduce both the latency and energy consumption with much shorter wirelengths and high-density connections.

To illustrate this, we performed floorplanning experiments for a large multibank SRAM design. Energy numbers were extracted from place-and-route experiments using a 7-nm technology process design kit. Figure 3 shows for the same memory capacities compared with 2-D, we can achieve similar or lower energy accesses due to shorter wirelengths going to 2- or 4-tier stacking using F2F bonding with hybrid bumps. When considering the energy overhead costs of through-silicon vias (TSVs), we start to see diminishing returns in stacking more than 4 tiers when looking at the example of 4-tier 3-D stacking for 8 MB of capacity. This, however, becomes closer in energy to the 2-D configuration as we scale up in capacity to 16–32 MB and we expect TSV technology/overhead to improve over time.

For the purposes of our analysis, we explore the energy versus footprint (memory capacity) design trade-offs between the 2-D baseline, 3-D 2-tier, and 3-D 4-tier configurations but constrain the design space to footprints close to the 2-D baseline (scaling from 1–16 MB). While we also expect improvements in latency and BW with 3-D stacking, we pessimistically assume similar latency and BW to the 2-D baseline SRAM accesses for a conservative estimate for our analysis.

METHODOLOGY

ML Accelerator Analytical Model for 3-D Stacking Simulation

We use an in-house ML accelerator simulator which provides estimation on a wide range of hardware metrics, such as execution time (latency), energy, input/output traffic, and resource utilization of our evaluated prototype ML accelerator running AR/VR models. This analytical model-based simulator takes an ML model (i.e., from PyTorch), extracts the operators, and calculates the expected performance metrics based on TOPS and energy tables based on the accelerator architecture. While not cycle-accurate, the simulator has been internally verified with cycle-accurate simulations and is used as a tool to provide early-stage performance estimations to guide architectural design space trends and decisions for different ML models.

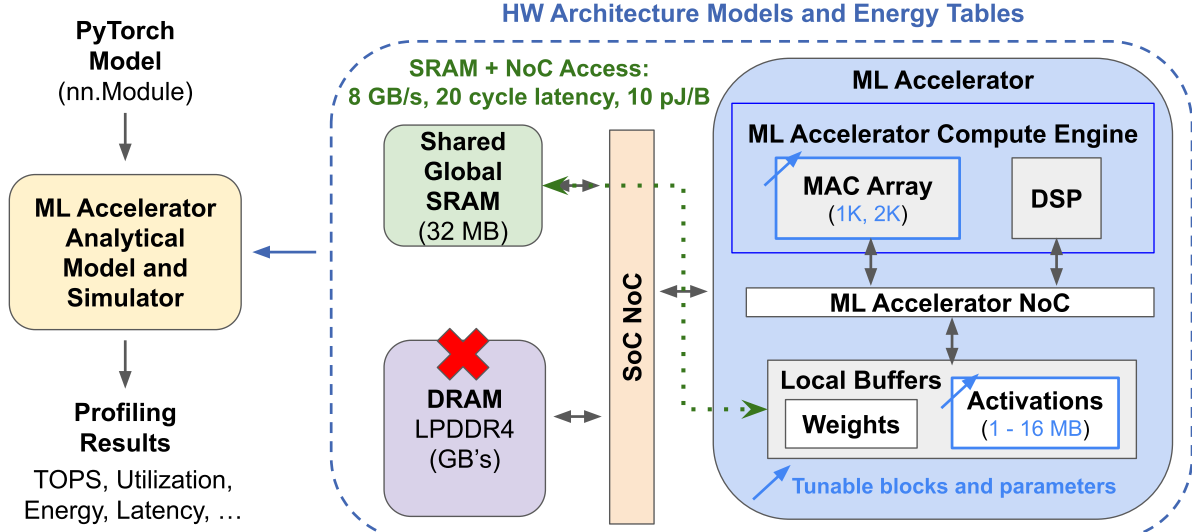


FIGURE 4. ML accelerator simulator and model architecture setup based on a scaled-up version of the Sumbul et al.'s work[5] with distributed buffers. The local SRAM/buffers are scaled up to 16 MB for activations and the MAC array is scaled from 1K to 2K. Since DRAM accesses are prohibitively expensive, we assume all memory accesses external to the accelerator come from the shared global SRAM via the SoC NoC.

Three-Dimensional Stacked Configurations

We configure the simulator to take in larger activation memories and account for energy overheads going to expanded 3-D memories. Given DRAM accesses are prohibitively expensive, we constrain our workload choices to NNs that fit within the 32 MB shared SRAM accessed via the SoC network-on-chip (NoC). Figure 4 illustrates the block diagram of our setup.

With the goal of maintaining a similar footprint to our 2-D baseline, we swept the design space parameters in Figure 5 with two key configurations: 1) use 3-D stacking to achieve a smaller footprint and increase the activation memory to 2–8 MB; and 2) incur a small footprint overhead (38%) to increase the MAC array to 2K and stack from 4–16 MB of activation memory. This provides a representative exploration to compare the design tradeoffs between energy, latency, and footprint for different 3-D configuration options compared with our 2-D baseline. Note we only consider the footprint of the ML accelerator but include the energy of both local and shared memory accesses. Our goal is to cut memory spilling by increasing the local memory to reduce shared memory accesses external to the accelerator.

RESULTS AND FINDINGS

The following sections outline the results from running our ML simulator for AR/VR workloads suffering from high activation memory requirements: depth estimation (HRNet, Agg), denoising, and super-resolution.

Depth Estimation

Depth estimation enables 3-D experiences such as augmented calling for AR glasses.[3] Accurate per-pixel depth prediction is critical for a natural and immersive calling experience. The depth estimation pipeline consists of two key NN models: High-Resolution Net (HRNet)[10] and 3-D aggregation (Agg).[11]

Figure 6 illustrates the energy and latency savings for HRNet and Agg across the 3-D configurations compared with the 2-D baseline. Focusing on the shared memory energy and memory latency, we see that increasing the local activation memory to 4 MB for HRNet and 8 MB for Agg reduces the shared memory accesses and consequently reduces the memory energy and latency to negligible levels. Our simulator estimates a total savings of ~1.35× in energy and 1.4–2.6× latency savings for HRNet, and a total savings of ~1.45× in energy and 1.6–3× latency savings for Agg.

Denoising

Denoising is another important workload for AR/VR applications since image restoration is needed for a seamless and enjoyable user experience and denoising also helps improve detection rates given limited sensor resolution.[9] Increasing from 1 to 2 MB significantly reduces shared memory accesses in Figure 7

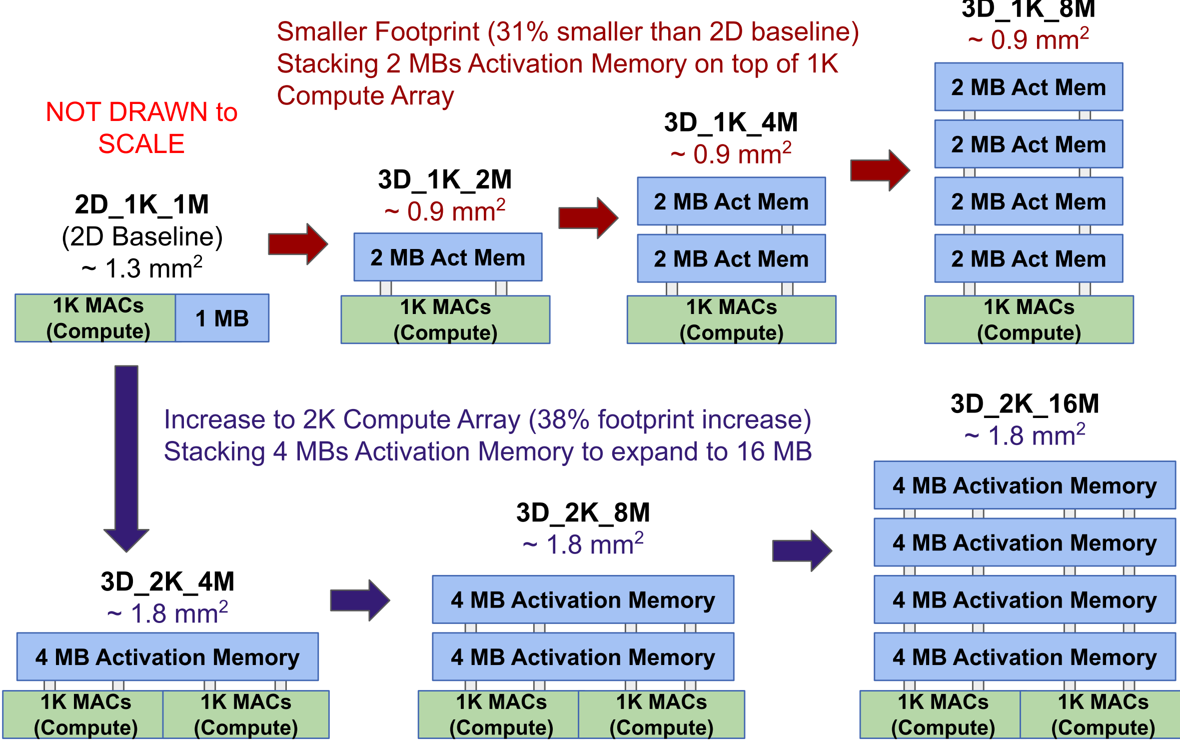


FIGURE 5. Three-dimensional configurations comparing to our 2-D baseline for 1) a smaller footprint, stacking 2 MB of activation memory in the *Z*-direction up to 8 MB; and 2) increase from 1K to 2K MACs (38% footprint increase) to stack 4 MB of activation memory up to a total of 16 MB in the *Z*-direction.

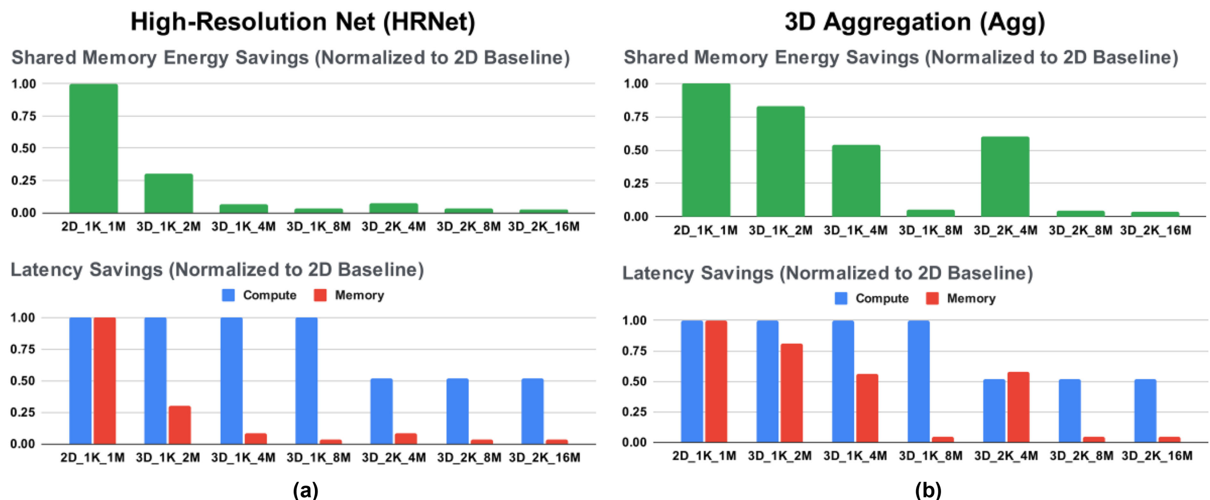
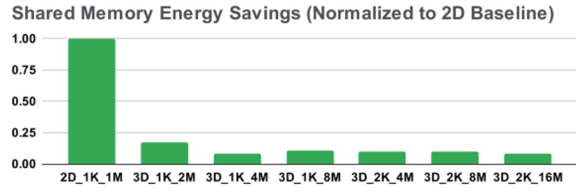


FIGURE 6. Shared memory energy and compute/memory latency savings for depth estimation NN models. (a) HRNet and (b) Agg normalized to our 2-D baseline. Reduction of spilling to shared memory to negligible levels occurs at 4 MB of activation memory for HRNet and 8 MB of activation memory for Agg.

Denoising



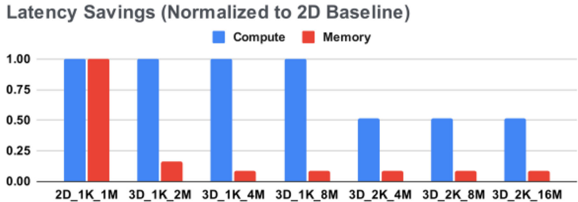


FIGURE 7. Shared memory energy and compute/memory latency savings for denoising normalized to our 2-D baseline. Reduction of spilling to shared memory to negligible occurs around 2 MB of activation memory and reaches diminishing returns scaling to larger capacities.

due to the relatively small size of the model and we see diminishing returns increasing the activation memory beyond 2 MB. Our simulator estimates ~1.35× total energy savings and 1.4–2.8× total latency savings going from 2-D to 3-D. If we are area-limited and cannot stack beyond 2 MB, this is a good candidate workload that benefits from 3-D integration and results in a smaller footprint than the 2-D baseline.

Super Resolution

Super-resolution in an AR/VR system is used to upscale low-resolution images to improve image quality.[12] As shown in Figure 2, the ability to deploy super-resolution models at higher resolutions is limited by the on-chip memory capacities and subsequently off-chip BW requirements for meeting FPS requirements.

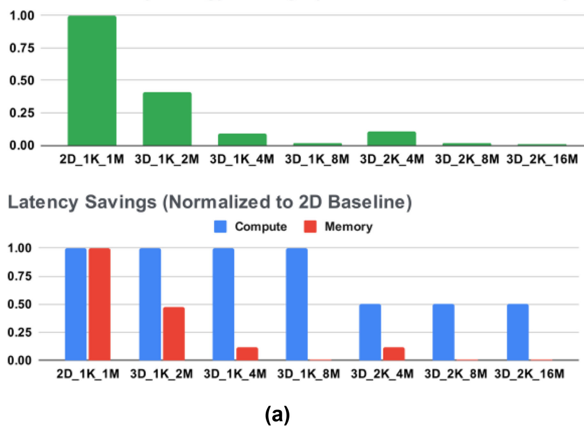
Figure 8 illustrates energy and latency improvements for super-resolution at (512 × 512) and (1,024 × 1,024) resolutions. The sweet spot for reducing shared memory accesses occurs at 8 MB for (512 × 512) and 16 MB for (1,024 × 1,024). This aligns with our observation from Figure 2 that going to 16 MB of activation memory reduces the external memory BW requirements to acceptable ranges. Our simulator estimates ~1.45× total energy savings and 1.9–3.7× latency reduction for super-resolution at (512 × 512), and 1.3–1.6× energy and 1.6–3.8× latency reduction for super-resolution at (1,024 × 1,024).

We note that super-resolution at (1,024 × 1,024) resolution would not have been previously deployable in this footprint in 2-D, illustrating how 3-D integration enables deployment of a larger NN model without incurring significant footprint overhead, while simultaneously improving both energy and latency of the accelerator.

CONCLUSION

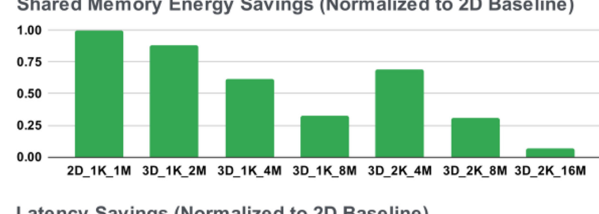
Table 1 summarizes our results from our proposed 3-D configurations and the expected performance (energy, latency) improvements and calculated footprints in 2-

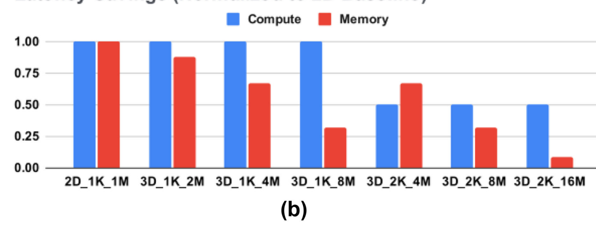
Super Resolution (512 x 512)



(a)

Super Resolution (1024 x 1024)





(b)

FIGURE 8. Shared memory energy and compute/memory latency savings for super-resolution at (512 × 512) and (1,024 × 1,024) resolutions normalized to our 2-D baseline. Optimal reduction of shared memory spilling occurs at 8 MB for (512 × 512) and 16 MB for (1,024 × 1,024).

TABLE 1. Summary table.

	2-D_1K_1M (Baseline)	3-D_1K_2M	3-D_1K_4M	3-D_1K_8M	3-D_2K_4M	3-D_2K_8M	3-D_2K_16M
# of MACs	1K	1K	1K	1K	2K	2K	2K
Activation Memory	1 MB	2 MB	4 MB (2 x 2 MB)[a]	8 MB (4 x 2 MB)[a]	4 MB	8 MB (2 x 4 MB)[a]	16 MB (4 x 4 MB)[a]
2-D Footprint [b]	~1.3 mm^2	~1.7 mm^2	~2.5 mm^2	~4.1 mm^2	~3.4 mm^2	~5.0 mm^2	~8.2 mm^2
3-D Footprint (Dom. Block)	N / A	~0.9 mm^2 (Compute)	~0.9 mm^2 (Compute)	~0.9 mm^2 (Compute)	~1.8 mm^2 (Compute)	~1.8 mm^2 (Compute)	~1.8 mm^2 (Compute)
Footprint (*F*) Overhead	1×	0.69× (-31%)	0.69× (-31%)	0.69× (-31%)	1.38× (38%)	1.38× (38%)	1.38× (38%)
Latency (*L*) Speedup	1×	Avg: 1.2× Max: 1.4×	Avg: 1.4× Max: 1.7×	Avg: 1.6× Max: 1.9×	Avg: 2.4× Max: 3.1×	Avg: 2.9× Max: 3.7×	Avg: 3.2× Max: 3.9×
Energy (*E*) Improvement	1×	Avg: 1.1× Max: 1.2×	Avg: 1.2× Max: 1.4×	Avg: 1.3× Max: 1.4×	Avg: 1.3× Max: 1.5×	Avg: 1.4× Max: 1.5×	Avg: 1.5× Max: 1.6×
FoM 1 ~($E \times F$)[c]	1×	Avg: 1.7× Max: 1.8×	Avg: 1.8× Max: 2.0×	Avg: 1.9× Max: 2.1×	Avg: 1.0× Max: 1.1×	Avg: 1.0× Max: 1.1×	Avg: 1.1× Max: 1.2×
FoM 2 ~($E \times L \times F$)[d]	1×	Avg: 2.0× Max: 2.5×	Avg: 2.5× Max: 3.4×	Avg: 3.1× Max: 3.9×	Avg: 2.3× Max: 3.3×	Avg: 3.0× Max: 4.1×	Avg: 3.4× Max: 4.5×

[a]($N \times K$ MB) represents *N* tiers of stacking *K* MBs of activation memory as illustrated in Figures 3 and 5.
[b]Estimated footprint if a 2-D configuration with the same number of MACs and activation memory sizes were used; all footprint estimates are compared with the 2-D baseline footprint of 1.3 mm^2.
[c]Figure of Merit (FoM) 1 ~($E \times F$) = (Energy Improvement/Footprint Overhead).
[d]Figure of Merit (FoM) 2 ~($E \times L \times F$) = (Energy Improvement × Latency Speedup/Footprint Overhead).

D and 3-D. Using a Figure of Merit (FoM) for energy-footprint product (FoM 1), we note the optimal 3-D configuration occurs at 3-D_1K_8M (smallest footprint for 8-MB capacity). If latency is critical and we can afford the additional 38% footprint overhead, using FoM 2 for the energy-latency-footprint product results in the optimal configuration at 3-D_2K_16M.

In summary, our proposed 3-D configurations for increasing activation memory from 1 MB to 2–8 MB improves latency by 1.2–1.9× and energy by 1.1–1.4× for a smaller footprint (31% smaller), and for a slight footprint increase (38% overhead), increasing activation memory from 1 MB to 4–16 MB and number of MACs from 1K to 2K results in an energy reduction of 1.3–1.6× and latency savings of 2.4–3.9× for depth estimation, denoising, and super-resolution AR/VR models. Additionally, we show that super-resolution at (1,024 × 1,024) resolution (not previously deployable in 2-D) can fit in a similar footprint with improvements in both energy and latency.

For future work, we plan to expand the memory capacities in both the *X-Y* and *Z*-directions to see the point at which there are diminishing returns (TSV/footprint overhead). Additionally, we plan to integrate DRAM access estimates to build a memory hierarchy to analyze deploying larger AR/VR models.

ACKNOWLEDGMENTS

The authors would like to thank Daniel Morris, Ekin Sumbul, Avishek Biswas, and William Koven of the Silicon Research Team at Reality Labs for their inputs on this article. The authors would like to thank Rakesh Ranjan, Meng Li, and Hyoukjun Kwon for support on ML accelerator modeling. The work of Robert M. Radway and Subhasish Mitra was supported by the DARPA 3DSoC Program and the Stanford SystemX Alliance.

REFERENCES

1. M. Abrash, "Creating the future: Augmented reality, the next human-machine interface," in *Proc. IEEE Int. Electron Devices Meeting*, 2021, pp. 1.2.1–1.2.11.
2. S. Rabii et al., "Computational and technology directions for augmented reality systems," in *Proc. IEEE Symp. VLSI Circuits, Plenary*, 2019.
3. E. Beigné, "3D Stacking opportunities for augmented reality hardware systems," in *Proc. IEEE DATE*, 2022.
4. J. Kim, L. Zhu, H. M. Torun, M. Swaminathan, and S. K. Lim, "Micro-bumping, hybrid bonding, or monolithic? A PPA study for heterogeneous 3D IC options," in *Proc. 58th ACM/IEEE Des. Autom. Conf.*, 2021, pp. 1189–1194.

5. H. E. Sumbul 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," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2022, pp. 1–8.
6. Y. Shao et al., "Simba: Scaling deep-learning inference with multi-chip-module-based architecture," in *Proc. MICRO*, 2019.
7. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2016.
8. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
9. L. D. Young et al., "Feature-Align network with knowledge distillation for efficient denoising," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops*, 2022, pp. 709–718.
10. J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
11. Z. Li et al., "Temporally consistent online depth estimation in dynamic scenes," 2021, *arXiv:2111.09337*.
12. G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Deep burst super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9205–9214.

LITA YANG is a research scientist at Reality Labs, Meta, Sunnyvale, CA, 94089, USA. Her research interests include hardware-software codesign for efficient AI/ML accelerators and energy-efficient edge computing devices. Yang received a Ph.D. degree in electrical engineering from Stanford University, Stanford, CA. Contact her at yanglita@fb.com.

ROBERT M. RADWAY is currently working toward a Ph.D. degree at Stanford University, Stanford, CA, 94089, USA. His research interests include multichip systems for application scale-up, 3-D monolithic/stacked/assembled ICs, and edge AI inference and training using emerging nonvolatile memories. Contact him at radway@stanford.edu.

YU-HSIN CHEN is a research scientist at Meta, Sunnyvale, CA, 94089, USA. His research focuses on hardware/software codesign to enable efficient on-device AI for AR/VR systems. Chen received a Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA. Contact him at yhchen@fb.com.

TONY F. WU is a research scientist at Reality Labs, Meta, Sunnyvale, CA, 94089, USA. His research interests include energy-efficient edge computing devices. Wu received a Ph.D. degree in electrical engineering from Stanford University, Stanford, CA. Contact him at tonyfwu@fb.com.

HUICHU LIU is a research scientist at Reality Labs, Meta, Sunnyvale, CA, 94089, USA. Her research interests include new memory design, technology-hardware codesign for energy-efficient AI/ML accelerators. Liu received a Ph.D. degree in electrical engineering from Pennsylvania State University, State College, PA, USA. Contact her at huichu@fb.com.

ELNAZ ANSARI is a research scientist at Reality Labs, Meta, Sunnyvale, CA, 94089, USA. Her research interests include hardware-software codesign for energy-efficient edge computing systems. Ansari received a Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA. Contact her at elnazans@fb.com.

VIKAS CHANDRA is the director of AI at Reality Labs, Meta, Sunnyvale, CA, 94089, USA, responsible for developing computer vision, machine perception, and speech/NLU technologies for AR glasses. Chandra received a Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA. Contact him at vchandra@fb.com.

SUBHASISH MITRA is a professor of electrical engineering and computer science at Stanford University Stanford, CA, 94305, USA. His research ranges across robust computing, nano systems, electronic design automation, and neurosciences. Mitra is a Fellow of IEEE and ACM. Contact him at subh@stanford.edu.

EDITH BEIGNÉ is the research director of AR/VR Silicon at Reality Labs, Meta, Sunnyvale, CA, 94089, USA. Her research interests include low-power digital and mixed-signal circuits, and design with emerging technologies applied to AR applications. Beigné was the Technical Chair of ISSCC 2022. Contact her at edith.beigne@fb.com.