Characterization and Design of 3D-Stacked Memory for Image Signal Processing on AR/VR Devices

Lita Yang, Changjung Kao, Sriseshan Srikanth, Daniel Morris, H. Ekin Sumbul, Tony F. Wu, Huichu Liu, Edith Beigné
Reality Labs, Meta, Inc. (Formerly known as Facebook, Inc.)
Sunnyvale, CA, USA
{yanglita,chriskao,seshan,dhmorris,ekinsumbul,tonyfwu,huichu,edith.beigne}@meta.com

ABSTRACT

Image Signal Processing (ISP) is an important component in augmented and virtual reality (AR/VR) applications. With the goal of running these applications on battery-powered edge devices, the ISP unit must satisfy rigorous power, performance, and form factor requirements. However, ISP workloads incur large memory footprints and intensive DRAM accesses that are prohibitively expensive for the stringent requirements of all-day wearable AR/VR products. Recent progress in 3D integration provides a promising solution for increasing memory capacities for iso-footprint, while achieving lower I/O power with shorter, vertical 3D interconnections. In this work, we explore and characterize two types of advanced 3D-stacked memories for ISP workloads: 3D-SRAM and 3D-DRAM. Our analysis demonstrates that by allocating additional 3D-stacked local memory to the ISP unit, we reduce expensive off-chip DRAM accesses by 57-92%, allowing us to deploy larger ISP workloads within power budgets not previously feasible with the 2D ISP baseline architecture. Comparing the two 3D-stacked memories, we observe that the use of 3D-DRAM reduces the total ISP power consumption by up to 53%, while 3D-SRAM achieves up to 32% power savings due to significant leakage contribution at increasing SRAM capacities. Finally, we propose a 3D-stacked hybrid memory ISP solution, combining both 3D-SRAM and 3D-DRAM, which can further improve the ISP power efficiency by an additional 9-16% on top of a 3D-DRAM-only memory architecture. To our knowledge, this is the first study to explore the benefits of advanced 3D-stacked memory for deploying ISP workloads on AR/VR devices.

CCS CONCEPTS

• **Hardware → Memory and dense storage**; • **Computer systems organization → Heterogeneous (hybrid) systems**.

KEYWORDS

3D integration, 3D-stacked memory, image signal processing, augmented reality, virtual reality

1 INTRODUCTION

Recent advancements in augmented and virtual reality (AR/VR) technologies are poised to become the next generation computing platforms, ushering in opportunities for new application domains, such as immersive education and presence, interactive entertainment and productivity, and personalized and contextual artificial intelligence (AI) [1]. Delivering a rich, context-aware, and accessible AR/VR user interface in real-time within the constraints of a socially acceptable AR/VR glasses form factor, however, require major innovations in key technical areas of power, performance, thermal, and silicon area [6]. To improve the quality of AR/VR images and video applications, optimizing the Image Signal Processing (ISP) subsystem within the AR/VR System-on-Chip (SoC) is vital towards delivering seamless all-day wearable smart glasses with high resolution and high-fidelity mono/stereo image and video capture. However, image signal processing on mobile devices often suffers from a severe "memory wall" bottleneck. Many ISP workloads have a deep and wide pipeline that require high memory bandwidth with relatively low computational density [4]. Increasing demand for high resolution and high-fidelity images and video quality also exacerbates the memory capacity problem on-device, requiring ever increasingly more memory to deploy larger ISP workloads while maintaining low power consumption for longer battery life [7].

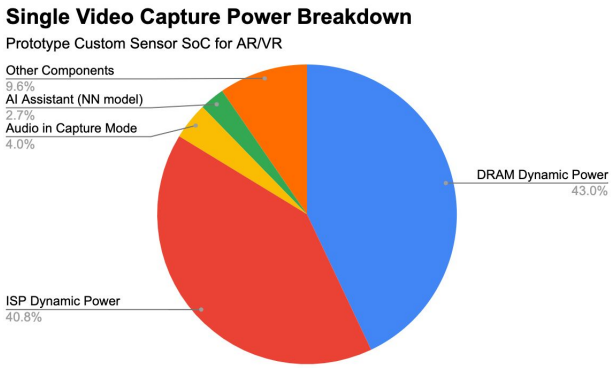

Figure 1: Power breakdown for single video capture on a prototype custom sensor SoC for AR/VR smartglasses (adapted from [1]). A significant portion of the total power goes to off-chip DRAM dynamic power (43%) and the ISP dynamic power (40.8%) consumption.

As shown in Figure 1, even with a custom SoC designed for low-power AR/VR smart glasses prototypes, the total power consumption is dominated by the DRAM dynamic power (43%) and ISP dynamic power (40.8%). Additionally, AR/VR SoC architectures must support not just one subsystem but multiple subsystems, including audio, machine learning (ML), computer vision (CV) and ISP, which all need to access the same Shared Memory (SMEM) and/or DRAM [6]. This makes expanding SMEM in the 2D direction extra challenging because we cannot increase the amount of 2D SRAM

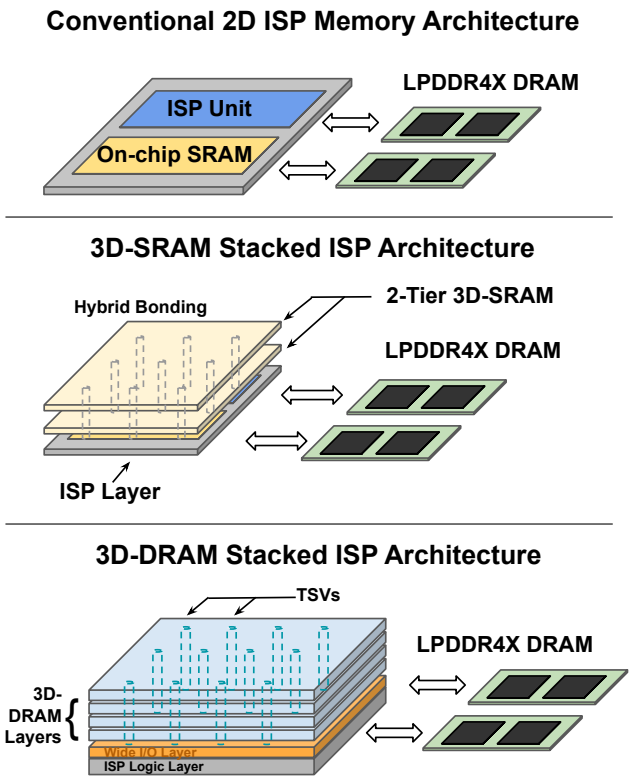

Figure 2: Memory technologies analyzed/compared for our proposed ISP architecture: (1) Conventional 2D ISP memory architecture with low-power DRAM (LPDDR4X), (2) ISP with 3D-stacked SRAM, and (3) ISP with 3D-DRAM. Conceptually the logic layer is illustrated as located on the bottom tier, but could be placed anywhere in the 3D stack.

due to footprint constraints, and we cannot allocate additional memory to the ISP subsystem due to sharing of the SoC's SMEM with other subsystems. The prohibitive cost of DRAM (due to power and bandwidth requirements) limits the throughput and image/video resolutions that can be deployed on-device and is the biggest challenge to delivering high resolution and high-fidelity images/videos on AR/VR devices. With limited I/O pins and area/power budget in AR/VR devices, it is difficult and near infeasible for traditional 2D architectures to satisfy the high throughput demands for qualitative image processing.

To address these challenges, we leverage recent advancements in 3D integration and packaging techniques. The ability to expand in the z-direction (3D vertical integration) enables additional memory stacked directly on top of the logic die per compute IP with lower latency, lower power, and higher bandwidth 3D connections. In this work, we investigate two types of 3D-stacked memory for ISP workloads: 3D-SRAM and 3D-DRAM, as illustrated in Figure 2. 3D-SRAM [9, 11] integrates additional SRAM on logic die through TSVs in the vertical direction, and multiple tiers of SRAM can be used to increase the SRAM capacity without incurring footprint overheads. 3D-DRAM [12] rearchitects the conventional DRAM memory die to optimize for much lower dynamic power consumption using a very large number of 3D connections, each running at relatively lower speeds than conventional DRAM I/Os but are ultra-low powered. To quantitatively analyze the benefits of 3D memory, we build a modeling tool to evaluate the power consumption under different ISP workloads. Our key findings are:

- 3D-stacked memory reduces total ISP power consumption by 37-53% using 3D-DRAM, while 3D-SRAM achieves 25-32% power savings. This is due to the increasing contribution of SRAM leakage to the total power when scaling to larger on-chip memory capacities allocated for 3D local memory.
- The use of 3D-stacked memory to increase local memory allocation reduces expensive LPDDR4X off-chip DRAM accesses by 57-92%, allowing us to deploy larger ISP workloads within power budgets not previously feasible with the 2D ISP baseline.
- Using a hybrid 3D-stacked memory architecture, combining both 3D-SRAM and 3D-DRAM, we can potentially further improve the 3D-DRAM-only system with an additional 9-16% improvement in power savings.

2 BACKGROUND AND MOTIVATION

2.1 Image Signal Processing (ISP)

The ISP subsystem lies in between the camera sensor and the CV/ML pipeline, which captures video from camera sensors and transforms the raw data to a post-processed visualized image the user can see without noise and artifacts. In AR/VR devices, the image taken by photosensors is in the format of Bayer image, which arranges RGB color filters on a 2D matrix. To convert the Bayer pattern to an RGB image and produce sufficient output image quality, complex ISP pipelines are needed. Traditional ISP pipelines contain a series of tasks that focus on different perspectives of the image, such as demosaicing, denoising, white balance, and pixel correction [2]. The pipeline stages of these heterogeneous tasks tend to have low computation density (operations/byte) but require massive parallelism at the pixel level. As a result, the ISP units are extremely memory-intensive due to the large amount of image data loading to/from memory. Additionally, a portion of this expensive off-chip DRAM access is irregular/random depending on the scene being captured, which results in significantly lower bandwidth utilization and increased DRAM power from activation and precharge overheads [8]. Since ISP workloads in AR/VR SoCs may not necessarily run in isolation, the available DRAM bandwidth, thermals and power are also highly contended in these systems.

Figure 3 summarizes the memory bandwidth and footprint requirements of our key ISP workloads running on our in-house custom ISP compute IP. First, we see that to process a FHD video, we require at least 2-4 GB/s of average bandwidth without having to compress or compromise image quality. Second, we observe the ISP workloads require large memory footprints, necessitating off-chip memory. Even for the smallest ISP workloads, we require >200 MB of uncompressed footprint and >100 MB of compressed footprint for FHD video processing. With limited on-chip SRAM (<10 MB) in conventional 2D ISP logic, this results in extensive and expensive

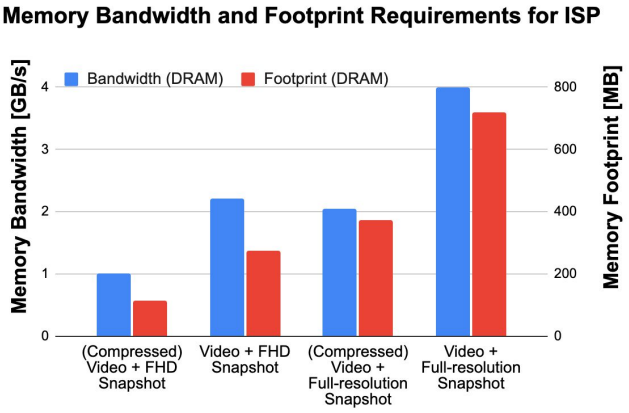

Figure 3: Memory bandwidth and footprint requirements of ISP workloads, measured with our in-house custom ISP. We explore four key ISP workloads with varying memory requirements.

off-chip DRAM accesses which burn a significant amount of total system power.

2.2 Background on 3D-Stacked Memory

Current 2D DRAM technologies are experiencing severe scaling challenges, resulting in an increasing performance gap between compute and memory. 3D-stacked memory is one of the recent technologies put forth to address this widening gap. By integrating one or multiple memory dies with the logic die in the vertical dimension, 3D-stacked memory enables several benefits not previously available in 2D, including high bandwidth and ultra-low power 3D connections and the ability to achieve the same or smaller footprints with larger memory capacities. The use of high-density 3D interconnects allows us to achieve very localized benefits at the IP and accelerator-level [11] compared with standard SoC memory disaggregation techniques and architectures commonly used in datacenter applications.

In this work, we investigate two types of advanced 3D-stacked memory, 3D-SRAM and 3D-DRAM, shown previously in Figure 2. We focus on hybrid bonding for 3D-SRAM [9, 11], which uses face-to-face (F2F) bond pads to stack 2D wafers and integrates additional SRAM on the logic die with Through-Silicon-Vias (TSVs), offering significant bandwidth and power benefits over state-of-the-art micro-bump approaches. To increase the SRAM capacity without increasing the x-y footprint, we can stack multiple tiers of SRAM in the z-direction as shown in Figure 2. A key benefit is that the short wirelengths and high-density connections reduce the interconnect lengths significantly, resulting in higher bandwidth connections, lower latency, and lower energy consumption when accessing 3D-stacked memory [9, 11]. 3D-DRAM [12] rearchitects the conventional DRAM memory die for lower power and higher bandwidth using a very large number of pins, each of which runs at relatively lower speed compared with conventional LPDDR4X DRAM I/Os but are ultra-low powered. The wide interface that is enabled by 3D F2F interconnects through vertical redistribution layer (RDL) routing allows for much lower dynamic and leakage power compared to conventional DRAM. As shown in Figure 2, with 3D stacking technology, it is possible to have many more I/O connections because they are not limited to the periphery of the device, and interconnect distances can be much shorter than the track lengths across the chip due to 3D-stacking between very thin dies.

2.3 Related Work

Recent advancements in 3D integration and stacking technologies have focused primarily in two categories: (1) datacenter applications such as AMD's 3D V-Cache [10] and (2) 3D stacking for image sensors [3]. Our work is uniquely positioned for 3D accelerator architectures for advanced 3D-stacking technologies targeted for AR/VR devices, not previously explored with ISP mobile applications. Prior work in mobile ISP has focused deployment primarily on mobile phones which are similarly constrained by requirements such as power, frame-rate and mobile phone thickness [2, 5, 7]. We borrow concepts from the mobile ISP community in terms of optimization and algorithmic advancements for AR/VR devices but have stricter form factor and power requirements, requiring the use of more advanced technologies to reduce or eliminate power-hungry off-chip DRAM accesses on AR/VR devices. Additionally, there is similar work on processing-in-memory (PIM) and other near-compute memory architectures for ISP for datacenter applications [4], however, we make the distinction that our approach optimizes 3D-stacked memories for low-power and footprint-constrained AR/VR applications and does not modify the ISP compute IP, but simply augments the IP with additional 3D-stacked memory "on-chip" with very advanced short 3D F2F bonding.

3 EVALUATION SETUP

3.1 ISP Workloads

We model four in-house ISP workloads that target key AR/VR use cases shown in Figure 3: FHD Video ISP + 2MP Snapshot (also known as Full HD or 1080p, which translates to roughly 1920 x 1080 pixels = 2MP) with/without compression, and FHD Video ISP + (Full-Resolution) 12MP Snapshot with/without compression. With the snapshot enabled, the user has the option to make a snapshot (either in FHD or Full-Resolution) anytime during the video streaming to be stored for further post-processing or for the user's image generation. This requires the ISP unit to buffer the full-resolution raw image to DRAM for each frame whenever the frame is snapshotted. As we can see from Figure 3, the four ISP workloads have varying DRAM footprint and bandwidth requirements, depending on the video/snapshot quality required and compression can be applied to reduce the DRAM footprint and bandwidth by trading off video quality.

Since it is traditionally challenging to quantify what is sufficient image quality or identify what will adequately cover all the corner cases the ISP can handle robustly [2], we chose these four ISP workloads to cover a range of image/video qualities and use cases which empirically produce high-quality images/video for our application. In an ideal scenario, we would like to deploy the largest ISP workload (FHD Video + Full-Resolution 12MP Snapshot), however, the power/footprint and bandwidth requirements make this

Table 1: Memory Modeling Specifications

Specification	LPDDR4X	3D-SRAM	3D-DRAM
Density $[MB/mm^2]$	10	4	8
Dynamic Energy $[pJ/B]$	65	2	7
Leakage Power $[\mu W/MB]$	3	350	17

Table 2: 3D Memory Configurations

Name	Memory Configuration	2D vs. 3D Footprint
2D ISP Baseline	Baseline	1x
2D-SRAM_16MB	Baseline ISP + 16 MB (2D) SRAM	1.4x
3D-SRAM_32MB	32 MB stacked on top of the ISP compute unit	1.9x
3D-SRAM_64MB	2 x 32 MB (2-tiers of 3D-SRAM) stacked on top of ISP compute	2.8x
3D-SRAM_128MB	4 x 32 MB (4-tiers of 3D-SRAM) stacked on top of ISP compute	4.6x
3D-DRAM_64MB	64 MB 3D-DRAM stacked on top of the ISP compute unit	1.8x
3D-DRAM_128MB	2 x 64 MB (2-tiers of 3D-DRAM) stacked on top of ISP compute	2.7x
3D-DRAM_256MB	4 x 64 MB (4-tiers of 3D-DRAM) stacked on top of ISP compute	4.4x

workload infeasible for current 2D AR/VR SoCs. Our current 2D baseline ISP can run the Compressed FHD Video + 2MP snapshot workload (the smallest ISP use case) and meets an acceptable power budget but suffers from lower video and snapshot quality. With this in mind, our modeling objective is two-fold in assessing the four key in-house ISP workloads: (1) quantify the benefits (power savings) achievable using 3D-stacked memories (3D-SRAM, 3D-DRAM) within the same footprint as the baseline 2D ISP IP, and (2) demonstrate the feasibility of deploying larger ISP workloads with a similar power budget to the baseline 2D ISP running what is currently feasible today (Compressed FHD video + 2MP snapshot).

3.2 3D Memory Configurations

Table 1 summarizes the specifications used for modeling our proposed 3D-stacked memories, in which we compare the storage density, dynamic (access) energy, and leakage power of LPDDR4X, 3D-SRAM, and 3D-DRAM. 3D-SRAM numbers in 7nm technology were obtained from measurement results from [9, 11] and LPDDR4X and 3D-DRAM numbers are specifications based off DRAM technology. Traditional LPDDR4X DRAM has high cell density and low leakage power but consumes high dynamic energy. 3D-SRAM has the lowest dynamic energy but has high leakage power and the lowest density. 3D-DRAM is a trade-off between the two other memory technologies, with the key features to note that compared to LPDDR4X, 3D-DRAM has 7-9x lower access energy with a balance of slightly higher leakage power while achieving similar memory density to LPDDR4X DRAM.

To maintain the same footprint with the original ISP unit, we adopt different configurations for the comparison of the 2D/3D architectures as shown in Table 2. Column 3 of Table 2 illustrates the 2D footprint needed for these configurations versus for the 3D-stacked memory configurations, which would be iso-footprint to the 2D ISP baseline. For the 2D architectures, we have both the baseline ISP and ISP + 16 MB of additional 2D-SRAM. We note the latter incurs a slight 1.4x footprint overhead, which is non-ideal for our AR/VR SoC footprint constraints. For 3D-SRAM, we configure 1-tier, 2-tiers, and 4-tiers of 3D-SRAM on top of the ISP logic die using TSV interconnections. Each 3D-SRAM tier consists of 32 MB SRAM. Similarly, for 3D-DRAM, we stack 1-tier, 2-tiers, and 4-tiers of 64 MB of 3D-DRAM on top of the ISP logic die for 64, 128, and 256 MB of total capacity. Note that when the workload footprint exceeds the capacity of the 3D-SRAM or 3D-DRAM, we assume the data will spill to conventional LPDDR4X DRAM in all cases.

4 RESULTS AND ANALYSIS

In this section, we outline our modeling results and architectural findings. We first analyze the power breakdown and savings using our proposed 3D-stacked memories versus the 2D ISP baseline. Then, we characterize the trade-off between 3D-SRAM and 3D-DRAM. Finally, we study the effectiveness of our memory allocation scheme by proposing a hybrid memory configuration combining both 3D-SRAM and 3D-DRAM.

4.1 Power Savings Using 3D-Stacked Memory

Figure 4 presents the modeling results from analyzing the power breakdown with our proposed 2D versus 3D memory configurations. Focusing on the reduction of LPDDR4X DRAM dynamic power, we observe that LPDDR4X DRAM dynamic power can be reduced by 32-53% across the four ISP workloads. Notably, for the FHD Video + 2MP Snapshot workloads (both compressed and uncompressed), we can reduce LPDDR4X DRAM access power to near negligible by using 3D-DRAM (128 MB for the compressed workload and 256 MB for the uncompressed version). However, for the FHD Video + 12MP Snapshot workload (compressed and uncompressed) we observe that LPDDR4X DRAM is still a significant portion of the total power since we are not able to completely fit the workloads in 3D-SRAM and 3D-DRAM. Thus, some data will still need to be buffered in LPDDR4X DRAM due to the limited 3D-stacked memory capacity for the larger ISP workloads.

Comparing with each workload's 2D ISP baseline, adding 3D-DRAM up to 256 MB greatly improves the power efficiency by reducing the amount of off-chip DRAM access traffic and consequently the DRAM dynamic power. We see across the four ISP workloads power reduction ranging from 37-53% power savings compared with their respective 2D baseline ISP implementation. On the other hand, if we expand the ISP unit with 32-128 MB 3D-SRAM, we observe limited power savings (up to 32%). Even though 3D-SRAM achieves smaller dynamic power consumption compared to 3D-DRAM, its leakage power begins to contribute a significant portion of the total power consumption as we scale up in SRAM memory capacity (ranging from 5% - 16% of the total power consumption).

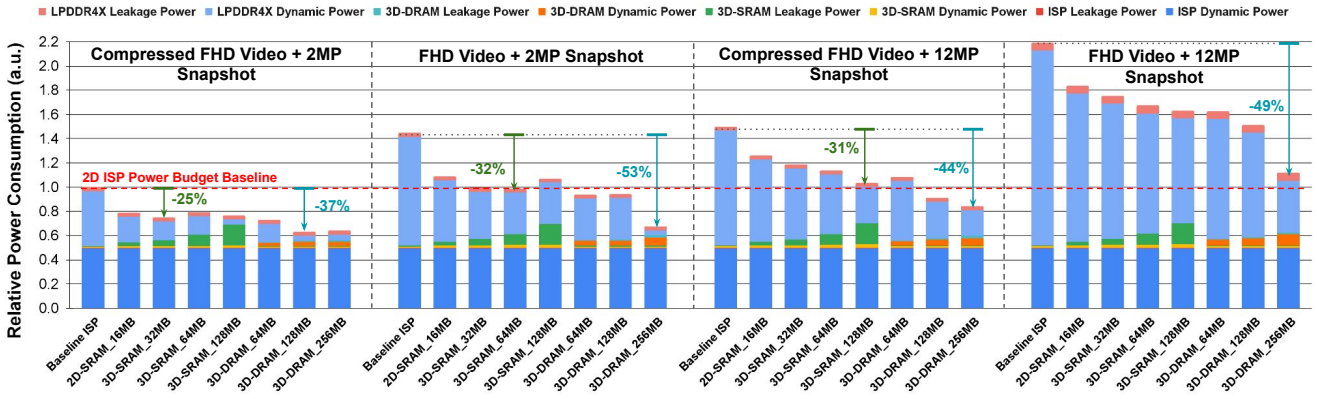

Figure 4: Power breakdown of the four key ISP workloads across the proposed 3D vs. 2D memory configuration settings, comparing 2D baseline ISP using LPDDR4X DRAM versus 2D-SRAM, 3D-SRAM, and 3D-DRAM. The improvement lines (colored in green for 3D-SRAM and colored in cyan for 3D-DRAM) indicate the largest power savings for 3D-SRAM and 3D-DRAM across the different configurations. The red 2D ISP power budget line illustrates which 3D configuration points meet the baseline 2D ISP power budget (e.g., 3D-DRAM_256MB meets the budget except for FHD Video + 12MP Snapshot).

When looking at the 2D ISP power budget (based on the Compressed FHD Video + 2MP Snapshot 2D ISP baseline), we note that in addition to reducing the power consumption of the ISP, we are now able to deploy larger ISP workloads (i.e., uncompressed FHD Video + 2MP Snapshot and Compressed FHD Video + 12MP Snapshot) within the tight power budgets allotted for ISP. While the uncompressed FHD Video + 12MP Snapshot workload is still over budget using 3D-DRAM_256MB, we note there may be opportunities to stack additional 3D-DRAM to more tiers to get closer to our goal of deploying the highest resolution workload.

4.2 3D-SRAM vs. 3D-DRAM Power Trade-off

A key observation from the previous section's analysis is that the smaller (compressed) ISP workloads which already have reduced LPDDR4X DRAM accesses achieve significant power savings by adding 1-tier of 3D-SRAM or 3D-DRAM, while larger ISP workloads will require more advanced stacking or multiple 3D tiers to achieve the required power budgets. To assess just the memory power comparison trade-off/sweet spot between 3D-SRAM and 3D-DRAM with increased capacity, we focus on the FHD Video + 2MP Snapshot use case. For a direct memory power comparison between the two 3D-stacked memories, we add the same memory capacity of each (3D-SRAM and 3D-DRAM) to the ISP unit and compare the total memory power consumption.

As shown in Figure 5, when the added memory capacity is <10-20 MB, both 3D-SRAM and 3D-DRAM can reduce the memory power by taking over the data traffic to LPDDR4X DRAM. However, when scaling beyond this trade-off point, 3D-DRAM demonstrates better memory power scalability while 3D-SRAM, due to the large leakage power introduced with increasing on-chip SRAM capacity, reaches diminishing returns beyond 20 MB. In total, we expect up to 59% of memory power savings for adding 3D-SRAM capacity and up to 92% of memory power savings for adding 3D-DRAM for the FHD Video + 2MP Snapshot workloads (compressed and uncompressed).

4.3 Case Study: 3D-SRAM and 3D-DRAM Hybrid Architecture

Given the trade-offs between 3D-SRAM and 3D-DRAM, we propose a case study for a hybrid memory hierarchy combining both 3D memory technologies to balance the dynamic and leakage power consumption. We propose two hybrid memory configurations: (1) Partition-1 with 25% 3D-SRAM and 75% 3D-DRAM; (2) Partition-2 with 50% 3D-SRAM and 50% 3D-DRAM. Note this is not meant to be a comprehensive analysis since we could configure and allocate many different percentages of memory and would be a full study and design space exploration by itself, but this case study demonstrates the potential benefits we could achieve combining both memory technologies.

Figure 6 illustrates a conceptual drawing of our hybrid memory architecture, where we plot the relative power savings with Partitions 1/2 over the All 3D-SRAM/3D-DRAM baselines across varying on-chip memory capacities. The hybrid memory architecture assumes we place data with high bandwidth density in 3D-SRAM until scaling on-chip memory capacities require 3D-DRAM to reduce leakage consumption. Figure 6 shows that Partition 1 presents better trade-offs with memory capacities over 16 MB while Partition 2 is optimal for small <16 MB requirements. The hybrid options provide an overall 20-49% improvement to the All 3D-SRAM baseline, while the improvements to All 3D-DRAM option start to converge with Partition 1 (only 9-16% improvements) as we increase on-chip memory capacity beyond 48 MB.

5 CONCLUSION

In this paper, we model the power efficiency of our in-house ISP workloads using advanced 3D-stacking memory technology. We characterize two types of 3D memories: 3D-SRAM and 3D-DRAM. We find that while 3D-SRAM provides significant reduction in dynamic/access power, its leakage can contribute a large portion of the total power consumption. On the other hand, 3D-DRAM

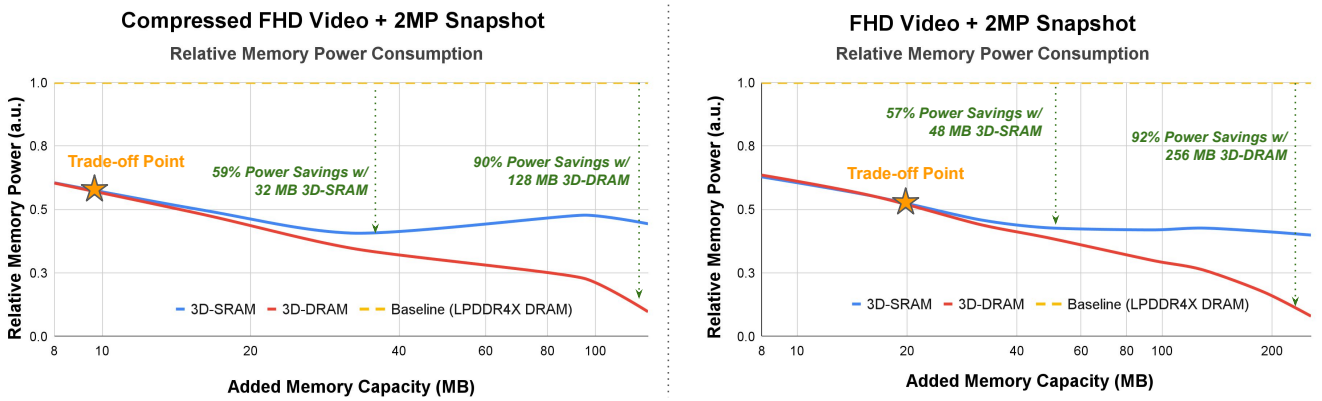

Figure 5: Comparison between 3D-SRAM and 3D-DRAM total memory power consumption for FHD Video + 2MP Snapshot ISP with and without compression. We note that the trade-off point for the Compressed FHD Video + 2MP Snapshot occurs at around 10 MB, while the uncompressed workload occurs higher at around 20 MB. Beyond these points (indicated by the star), 3D-SRAM consumes more memory power than 3D-DRAM at high memory capacity due to leakage, illustrating the better scalability of 3D-DRAM for workloads requiring more than 10-20 MB of memory capacity.

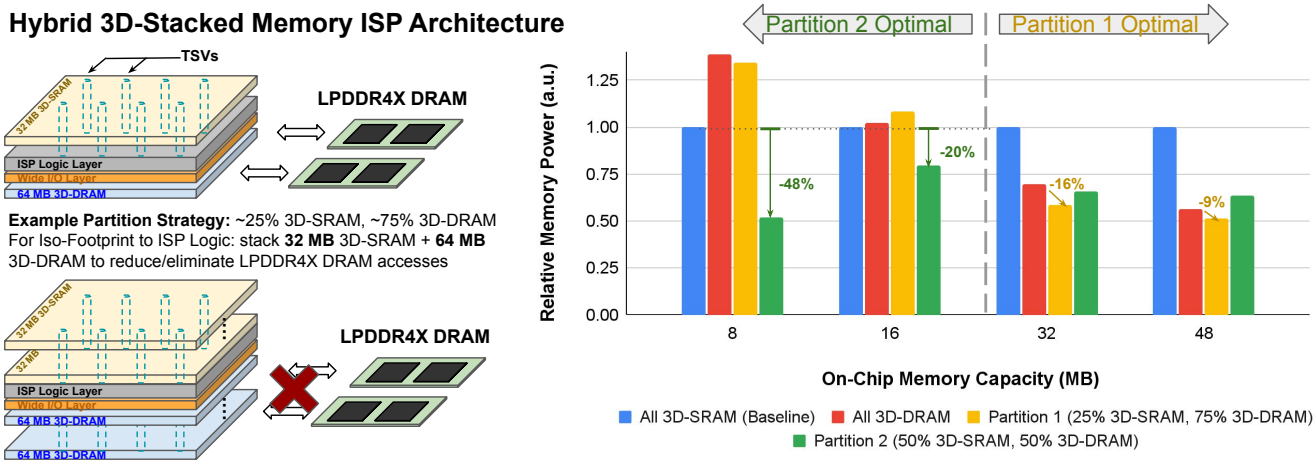

Figure 6: Conceptual drawing of a 3D-stacked hybrid memory configuration using a combination of 3D-SRAM and 3D-DRAM. Partition 1 is optimal for larger memory capacities (>16 MB) while Partition 2 is better for memory requirements below 16 MB.

demonstrates better trade-offs for leakage versus dynamic power and can consistently reduce the memory power with increasing capacity to 10s of MB. By combining both memory technologies, we can further improve the power efficiency by up to 16% compared with All 3D-DRAM-only systems, where we expect total system power savings of 53% improvement over the 2D ISP baseline. Future work will aim to expand on this hybrid architecture and optimal partitioning.

While this work focuses on 3D-stacked memory using 3D-SRAM and 3D-DRAM, the methodology could be extended to other types of 3D-stacked memory (e.g., RRAM) and 2.5D solutions (e.g., HBM), and used orthogonally/in-conjunction with other advanced memory solutions (e.g., PIM). One could integrate the 3D modeling parameters into a multi-level memory hierarchy with 2D/2.5D/3D architectures to enable even larger ISP workloads with higher memory capacity requirements and multiple tiers of compute and memory units. Additionally, the proposed 3D memory allocation strategy could be extended to support dynamic allocation (instead of static) during run-time to support multiple AR/VR uses cases and memory requirements.

ACKNOWLEDGMENTS

The authors would like to thank Jilan Lin and Elnaz Ansari who initiated this project during Jilan's internship at Meta and provided all the modeling analysis work. We would also like to thank Pietro Caragiulo who helped review the paper and provided feedback on 3D-DRAM. We would also like to specially thank the ISP architecture team, including Yiqian Min, Shashank Rao, and Lavanya

Subramanian, for providing ISP architectural data and traces and answering questions related to the ISP architecture.

REFERENCES

[1] Michael Abrash. 2021. Creating the Future: Augmented Reality, the next Human-Machine Interface. In *2021 IEEE International Electron Devices Meeting (IEDM)*. 1–11. https://doi.org/10.1109/IEDM19574.2021.9720526

[2] Mauricio Delbracio, Damien Kelly, Michael S. Brown, and Peyman Milanfar. 2021. Mobile Computational Photography: A Tour. *CoRR* abs/2102.09000 (2021). arXiv:2102.09000 https://arxiv.org/abs/2102.09000

[3] Jorge Gomez, Saavan Patel, Syed Shakib Sarwar, Ziyun Li, Raffaele Capoccia, Zhao Wang, Reid Pinkham, Andrew Berkovich, Tsung-Hsun Tsai, Barbara De Salvo, and Chiao Liu. 2022. Distributed On-Sensor Compute System for AR/VR Devices: A Semi-Analytical Simulation Framework for Power Estimation. arXiv:2203.07474 [cs.AR] https://arxiv.org/abs/2203.07474

[4] Peng Gu, Xinfeng Xie, Yufei Ding, Guoyang Chen, Weifeng Zhang, Dimin Niu, and Yuan Xie. 2020. iPIM: Programmable In-Memory Image Processing Accelerator Using Near-Bank Architecture. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. 804–817. https://doi.org/10.1109/ISCA45697.2020.00071

[5] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Trans. Graph.* 35, 6, Article 192 (dec 2016), 12 pages. https://doi.org/10.1145/2980179.2980254

[6] Karl Kaiser, Dinesh Patil, and Edith Beigne. 2023. A prototype 5nm custom sensor SoC for Augmented Reality/Virtual Reality targeting Smartglasses with embedded computer vision, audio, security and ML. In *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. 1–2. https://doi.org/10.23919/VLSITechnologyandCir57934.2023.10185381

[7] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. 2013. Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines. *SIGPLAN Not.* 48, 6 (jun 2013), 519–530. https://doi.org/10.1145/2499370.2462176

[8] Sriseshan Srikanth, Lavanya Subramanian, Sreenivas Subramoney, Thomas M. Conte, and Hong Wang. 2018. Tackling memory access latency through DRAM row management. In *Proceedings of the International Symposium on Memory Systems* (Alexandria, Virginia, USA) *(MEMSYS '18)*. Association for Computing Machinery, New York, NY, USA, 137–147. https://doi.org/10.1145/3240302.3240314

[9] Tony F. Wu, Huichu Liu, H. Ekin Sumbul, Lita Yang, Dipti Baheti, Jeremy Coriell, William Koven, Anu Krishnan, Mohit Mittal, Matheus Trevisan Moreira, Max Waugaman, Laurent Ye, and Edith Beigné. 2024. 11.2 A 3D integrated Prototype System-on-Chip for Augmented Reality Applications Using Face-to-Face Wafer Bonded 7nm Logic at <2μm Pitch with up to 40% Energy Reduction at Iso-Area Footprint. In *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 67. 210–212. https://doi.org/10.1109/ISSCC49657.2024.10454529

[10] John Wuu, Rahul Agarwal, Michael Ciraula, Carl Dietz, Brett Johnson, Dave Johnson, Russell Schreiber, Raja Swaminathan, Will Walker, and Samuel Naffziger. 2022. 3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 65. 428–429. https://doi.org/10.1109/ISSCC42614.2022.9731565

[11] Lita Yang, Robert M. Radway, Yu-Hsin Chen, Tony F. Wu, Huichu Liu, Elnaz Ansari, Vikas Chandra, Subhasish Mitra, and Edith Beigné. 2022. Three-Dimensional Stacked Neural Network Accelerator Architectures for AR/VR Applications. *IEEE Micro* 42, 6 (nov 2022), 116–124. https://doi.org/10.1109/MM.2022.3202254

[12] Tao Zhang, Cong Xu, Ke Chen, Guangyu Sun, and Yuan Xie. 2014. 3D-SWIFT: a high-performance 3D-stacked wide IO DRAM. In *Proceedings of the 24th Edition of the Great Lakes Symposium on VLSI* (Houston, Texas, USA) *(GLSVLSI '14)*. Association for Computing Machinery, New York, NY, USA, 51–56. https://doi.org/10.1145/2591513.2591529