Received 18 June 2024; revised 30 August 2024 and 18 September 2024; accepted 23 September 2024.
Date of publication 26 September 2024; date of current version 15 October 2024.

Digital Object Identifier 10.1109/JXCDC.2024.3468386

Toward Fine-Grained Partitioning of Low-Level SRAM Caches for Emerging 3D-IC Designs

SUDIPTA DAS[1,2], BHAWANA KUMARI[1,3], SIVA SATYENDRA SAHOO[1], YUKAI CHEN[1] (Member, IEEE), JAMES MYERS[1], DRAGOMIR MILOJEVIC[1,4], DWAIPAYAN BISWAS[1] (Member, IEEE), and JULIEN RYCKAERT[1]

[1]Imec, 3001 Leuven, Belgium
[2]Department of Electronics and Computer Science, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[3]Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium
[4]Bio, Electro And Mechanical Systems (BEAMS), Université Libre de Bruxelles, 1050 Bruxelles, Belgium
CORRESPONDING AUTHOR: B. KUMARI (bhawana.kumari@imec.be).

ABSTRACT Scaling on-chip memory capacity is one of the primary approaches to mitigate memory wall bottlenecks. Various 2.5-D/3-D integration schemes, leveraging novel partitioning, are being actively explored to improve system performance. However, fine-grained functional partitioning of memory macros is not widely reported. As static RAM (SRAM) scaling stagnates with emerging CMOS logic roadmap, we propose a partitioning of low-level (faster access) caches in 3-D using an array under CMOS (AuC) technology paradigm. Our study focuses on partitioning and optimization of SRAM bit-cells and peripheral circuits, enabling heterogeneous integration, achieving up to 12% higher operating frequency with 50% leakage power reduction in the memory macros. Applied on a 64-bit mobile system-on-chip (SoC) CPU core, we achieve up to 60% higher energy efficiency compared with 2-D baseline and 14% increase in operating frequency compared with standard memory-on-logic 3-D partitioning scheme.

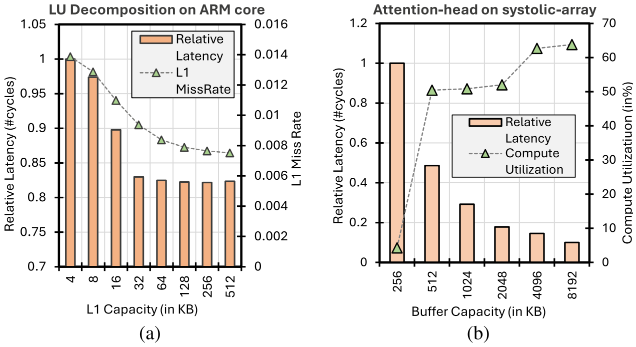
INDEX TERMS 3-D integrated circuit (3D-IC), ARM, array under CMOS (AuC), mobile computing, monolithic integration, partitioning, sequential integration, static RAM (SRAM) design, through-silicon-via (TSV)/bump pitch.

I. INTRODUCTION

The memory wall problem remains one of the major bottlenecks to system-level performance improvements in modern computing systems [1]. The effect of the mismatch in the scaling of compute performance and main memory speed is further exacerbated for AI workloads that rely on intensive data movement [2]. In recent times, the architectural efforts toward overcoming the memory wall have included processing-in/near-memory (PIM), memory disaggregation, etc. [3], [4]. While memory disaggregation targets the memory capacity issue and applies primarily to data centers, PIM architectures, targeting the memory bandwidth issue, are being actively explored across edge- and high-performance computing. In addition, increasing the low-level cache capacity remains one of the primary methods for improving performance across different workloads and processors.

For instance, Fig. 1 shows the performance improvement due to the scaling of first-level cache capacity, based on architecture-level simulations.[1] Fig. 1(a) shows the reduction in the latency due to increased L1 cache capacity in a low-power ARM core, while executing a memory-bound workload: LU decomposition. Similarly, Fig. 1(b) shows the latency with increasing buffer capacity in a systolic-array-based accelerator with 32 × 32 multiply-and-accumulate units, for the processing of an attention head and feedforward layers of a transformer. For both the cases, increasing the low-level cache capacity leads to better workload performance. Reduced cache miss rate and improved compute utilization lead to improved performance for the mobile core and the accelerator, respectively. With the increasing

[1] Gem5 and SCALESimv2 were used for the simulations while assuming single-cycle access latency for the L1 caches and buffers.



(a) (b)

FIGURE 1. Workload performance gains with low-level cache capacity scaling using architecture-level simulations. (a) Latency and L1 miss rate in a commercial mobile core. (b) Latency and overall compute utilization in a systolic-array-based accelerator core.

complexity of edge devices, driven by latency and communication concerns, scaling the low-level cache capacity can enable better quality of service. For instance, the latest mobile cores from Apple, the A17 Pro, include up to 256 kB of L1 cache, a 4× increase from the 64 kB in Apple A11 from 2017. However, in addition to integrating larger cache capacities at core frequency, edge computing poses additional challenges in terms of footprint (form factor-driven) power, energy (battery-operated), and temperature (reliability).

Consequently, various novel integration schemes have been explored to improve the performance, power, area, cost, and thermal (PPACT) of edge computing systems. Specifically, vertical integration methods such as monolithic and sequential 3-D integrated circuits (3D-ICs) are being actively explored. From a technology perspective, novel through-silicon-via (TSV) and 3D-IC bonding approaches are being used for enabling high-capacity, fast, local memory. Similarly, partitioning schemes such as memory-on-logic (MoL), logic-on-memory (LoM), and logic-on-logic (LoL) are being actively explored for system-level design optimization in 3D-ICs. The partitioning schemes focus on the separation of the memory and logic components onto different stacks (for monolithic 3-D) or tiers (for sequential 3-D).

However, in most related works, the exploration with respect to partitioning does not extend to the memory macros. As a result, using homogeneous memory macros does not fully exploit the benefits of 3-D integration. To this end, we present a comparison of 3D-IC design methods, focusing primarily on the partitioning of low-level caches.

Our novel contributions include: wide

1) We propose a novel partitioning scheme of memory macros in array under CMOS (AuC) technology that enables heterogeneous 3-D integration. Specifically, it involves placing all the peripheral circuits in the logic tier alongside other standard cells while the other tier comprises only static RAM (SRAM) bit-cells. With the proposed schemes, we report up to 12% improved performance using homogeneous AuC technology macros compared with 2-D SRAM macros. Furthermore, using heterogeneous AuC macros, we report up to 50% lower leakage power.
2) We present a comparative study of different integration schemes for a commercial mobile core. Specifically, we use the proposed memory macro schemes for the L1 cache of the mobile core to compare the power, performance, and area metrics for 2-D and various 3-D integration schemes.
3) The proposed 3-D partitioning scheme allows SRAM bit-cells and periphery logic to be optimized separately in the back-end-of-line (BEOL). Furthermore, it also enables the SRAM periphery logic to be placed near system-level core logic in the same advanced node. The study shows this memory-level optimization results in up to 60% higher energy efficiency compared with 2-D and at least 14% higher operating frequency compared with standard 3-D MoL partitioning in the commercial mobile core.

The rest of this article is organized as follows. Section II provides a brief overview of the background and related works for emerging integration methods and memory macro design. The vertical staking methodologies used in the current article are presented in Sections III and IV, respectively. In Section V, the results from the experimental evaluation are presented, followed by a discussion on the scope of related future research in Section VI.

II. BACKGROUND AND RELATED WORKS

A. MEMORY MACRO DESIGN

The low-level caches, tightly coupled with the core, comprise architectural elements designed to optimize the performance by reducing the miss rate. Depending on the capacity, and other architecture-level specifications such as data/instruction cache, cache-line size, and associativity, the memory component of the cache is physically realized with multiple memory macros. Technology-specific metrics are also considered during the design of the appropriate macros, usually implemented as single/multiple subarrays (SAs). The SRAM SA is designed with bit-cell array and their corresponding peripheries. The peripheries can be categorized as row peripheries (word-line drivers, row decoders) and column peripheries (sense amplifiers, write drivers, etc.). These SAs can be combined to form a macro of required memory capacity. The major factors contributing to the performance and power of the macro are word-line and bitline metal resistances and capacitances which keep deteriorating with scaling. SRAM bit-cell area is reaching its limits in advanced CMOS technology nodes due to restricted scaling of poly-pitch (PP) and metal pitch (MP) [5].

New device architectures such as forksheet [6], nanosheet [7], and CFET [8] have been able to improve the SRAM density at scaled technology nodes. However, cache performance deteriorates with scaling as metal resistance increases [9], [10]. In modern mobile and high-performance processors, larger caches (such as L2 and above) are built from smaller memory macros to reduce delays within each macro. In a 2-D layout, this leads to long wires for routing all the macros, which increases interconnect delays and worsens the memory wall problem. To tackle this issue, several 3-D integration technologies were proposed showing that 3-D integration enables overall wirelength reduction by shortening the connections between logic and memory [11], [12], [13], [14].

B. EMERGING INTEGRATION

Various 3-D integration approaches have been recently proposed to address device and memory scaling challenges in modern electronics. 3D-IC technologies such as micro-bumping, hybrid bonding, and sequential 3-D have gained popularity in recent times. In micro-bumping 3D-ICs, two dies are vertically stacked using a dense array of micro-bumps in a face-to-face (F2F) configuration, ensuring high yield and reliability.

A 3-D die stacking technique using μ-bump technology was introduced by Intel [15]. This method also allows for heterogeneous 3-D die stacking, providing significant flexibility in technology selection and intellectual property (IP) configurations. Hybrid bonding technology enables 3-D integration using F2F bond pads to stack two predesigned 2-D wafers through the BEOL layers. Kim et al. [16] proposed a direct Cu–Cu thermo-compression wafer-level bonding and stacking process for 3-D stacked IC bonding. Since bond pads are smaller than TSV, hybrid bonding 3D-ICs offer high-density vertical integration. A hybrid bonding high-level cache-on-logic partitioned system-on-chip (SoC) improves performance while maintaining footprint scalability [17]. However, due to the large 3-D interconnect pitch and capacitance, hybrid bonding is not always suitable for partitioning timing-critical low-level caches such as the L1 cache of processor cores.

Sequential integration is an emerging technology that integrates device layers sequentially in the vertical direction, using nano intertier via (ITV) for fine-grained integration. Vandooren et al. [18] demonstrated sequentially stacked FinFETs with high alignment accuracy showing a footprint reduction of up to 50%. AuC is a unique partitioning scheme proposed by Salahuddin et al. [19] where memory and logic are partitioned vertically to achieve system-level performance and cost benefits.

How different tiers are partitioned plays a major role in optimizing the overall PPACT aspects of the chip. Partitioning schemes such as MoL/LoM and LoL have shown significant potential for future high-density designs, facilitated by aggressive TSV and bump scaling [20]. Among these, MoL/LoM appears to offer the most feasible partitioning, considering existing designs, providing gains from 3-D integration without compromising the system's thermal properties that much [21]. This is a critical consideration for emerging 3D-ICs, making MoL a widely accepted scheme across various designs. Therefore, it is essential to explore fine-grained, optimized partitioning strategies for splitting memory and logic elements in 3-D implementations.

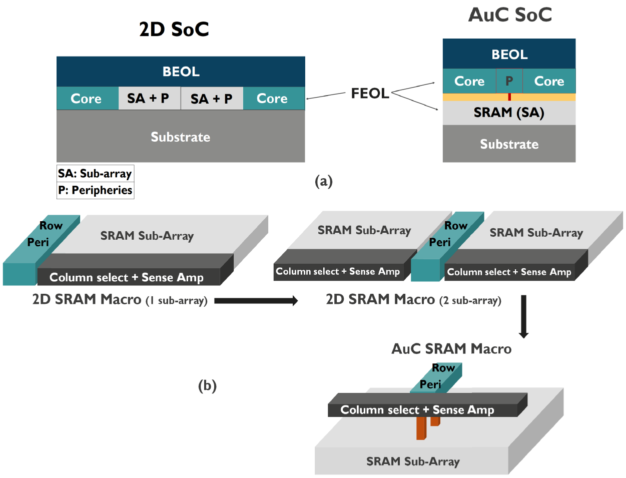


FIGURE 2. **(a) Schematic illustration of the conventional 2-D IC and the novel AuC integration scheme. (b) SRAM macro SA arrangements in case of 2-D and AuC.**

C. SUMMARY

Sequential 3-D technology appears to be the most promising candidate for logic and high-speed cache partitioning due to its low parasitic ITVs. This article evaluates the potential of memory on logic partitioning schemes (with hybrid bonding and sequential integration methods) through design co-optimization (DTCO)/system technology co-optimization (STCO). The partitioning scheme proposed in previous literature [19], where some of the periphery circuits are placed with SRAM bit-cell tier, restricts the decoupling of the array and periphery circuits efficiently. The proposed memory-element and logic partitioning in this article, which decouples the SRAM bit-cell arrays and all the peripheries, allows the optimization in BEOL of the SRAM array tier independently. The metal aspect ratio of word-lines and bitlines is increased by two times of that of 2-D to reduce the parasitics. This partitioning scheme also gives the freedom of heterogeneous integration, where the periphery circuits correspond to nanosheet technology (same as that of logic standard cells) while SRAM bit-cells correspond to FinFET technology.

III. MEMORY MACRO DESIGN

The partitioning of CMOS Logic and SRAM arrays in 3-D integration scheme is illustrated in Fig. 2, highlighting the placement of SRAM SA and control peripheries. The SRAM memory macros are placed in the bottom tier, while the periphery circuits sit in the top tier along with the other logic circuits of the core/processor as shown in Fig. 2. The schematic differentiation between the conventional 2-D and AuC SoC is described in Fig. 2(a). In case of 2-D IC, the logic core and SRAM SA along with its control peripheries (SA + P) are placed side by side on the same plane. Whereas

TABLE 1. **SRAM memory configurations for ARM.**

Config	Macro Instance	Subarray Instance	Memory capacity (Kb)
2D (1S)	128x38; 128x50; 128x60; 160x118; 256x32; 512x12; 512x32; 2048x39; 4096x22	64x76; 128x50; 64x120; 80x236; 128x64; 256x24; 256x64; 256x312; 256x352	0.6; 0.8; 0.9; 2.3; 1; 0.75; 2; 9.75; 11
2D (2S) and AuC	128x38; 128x50; 128x60; 160x118; 256x32; 512x12; 512x32; 2048x39; 4096x22	64x38; 128x25; 64x60; 80x118; 128x32; 256x12; 256x32; 256x156; 256x176	0.6; 0.8; 0.9; 2.3; 1; 0.75; 2; 9.75; 11

in AuC, core logic and SRAM periphery (P) are on the top tier, while the SRAM memory SAs are in the bottom tier. Supervias are used to propagate signals between SRAM SA and their peripheries (shown by red via connections).

Fig. 2(b) shows the design considerations of 2-D and AuC SRAM macros. The 2-D SRAM macros are designed in two ways, i.e., macros consisting of one SA with row peripheries at the edge and macros with two SAs with shared row peripheries in the middle. The memory macro in AuC is arranged such that the two SA s are merged to a single SA in the bottom tier and the shared row periphery is pulled up to be placed in the top tier along with column peripheries and other core logic. The supervia connections are made between word-line drivers (top-tier) and word-lines (bottom-tier) and sense amplifiers (top-tier) to column select (bottom-tier) as shown in the figure.

Table 1 provides the information on different SA configurations used in 2-D 1S, 2-D 2S, and AuC macros. Iso-capacity memory macros are considered in all the scenarios for a fair comparison. One macro instance corresponds to two SA configurations (i.e., 1S and 2S) (e.g., 128 × 38 macro instance corresponds to 64 × 76 in 1S and 64 × 38 in 2S configurations, respectively, as shown in Table 1. For the purposes of our analysis, we have assumed the capacities of the 64 × 38 and 256 × 176 arrays to be 1 and 10 kb, respectively. Although the actual capacities of these arrays are approximately 0.6 and 11 kB, we have rounded these values to simplify the analysis and to provide a consistent basis for comparison.

This architecture presents a unique advantage: the independent optimization of SRAM (bottom tier) and core logic (top tier) transistors and BEOL processes. By decoupling the SRAM array in the bottom tier from that of the logic tier, we unlock the flexibility to optimize each component separately. Notably, the degradation of SRAM performance in scaled technology nodes, attributed to word-line and bitline resistance, necessitates innovative approaches. In this study, we leverage this flexibility to optimize the BEOL of the SRAM array, aiming to enhance its performance.

IV. PHYSICAL DESIGN

A. PDK, 2-D, AuC, AND 3-D INTEGRATION

The physical implementation involves a 2-D process design kit (PDK) based on an A10 nanosheet technology. Furthermore, we consider an N3 FinFET device [22] for the bit-cells in the heterogeneous integration case of AuC. We use a five-track standard cell library characterized at 0.7 V and

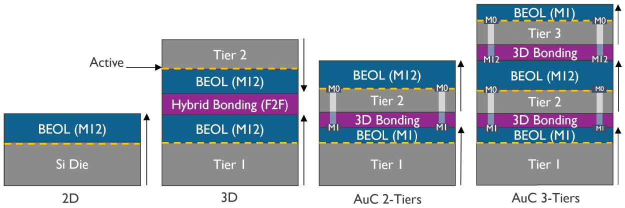


FIGURE 3. **Technology integration methodology.**

TABLE 2. **Technology assumptions.**

Config	Tech.	Pitch	MIV/Bump	Block Metal Layer	Memory Metal Layer	Process
2D	A10	-	-	M12	M5	Monolithic
3D (MoL)	A10	1.12 µm	M12	M12	M5	Sequential
AuC	A10, N3	0.15 µm	M2	M12	M3	Monolithic / Sequential

25 °C. The front-side BEOL stack includes 13 routing metal layers (M0–M12). We generate timing and geometry views of the memories for 2-D and 3-D implementation using an in-house memory compiler, simulating the complete SRAM SA operation for different operating modes. The same memory compiler has been used to generate AuC memories but with consideration of proper resistance (R), capacitance (C), and physical shape as discussed in Section III. For the AuC block-level integration, we assume that the device growth will be monolithic for each plane separated by an insulator and thin silicon layer as shown in Fig. 3 with a pitch of 0.15 μm enabled by the use of ITV. The same assumption is true for even sequential integration with the same advanced pitch following the trend of TSV scaling [23], [24]. For 3-D MoL integration, we assumed F2F and wafer-to-wafer hybrid bonding (W2W-HB) technology. Three-dimensional pitch of 1.12 μm is considered to provide sufficient 3-D interconnect density for MoL partitioning. The technology assumptions for AuC and 3-D are summarized in Table 2.

B. PHYSICAL IMPLEMENTATIONS AND ARCHITECTURE

We have used a low-power ARM core for the implementation and normalized the extracted data to prevent revealing proprietary information for the commercial processor. An almost equal split between memory blocks and logic modules in A10 technology makes this IP an ideal choice for such explorations. On top of that, it ensures the feasibility of AuC like memory optimization in current industrial systems even without any specific system-level modifications to extract gains.

Five different implementations have been selected for this exploration (a) 2-D baseline, (b) 3-D MoL, (c) AuC two-tier, (d) AuC three-tier, and (e) AuC heterogeneous. Two-dimensional and AuC implementations are done using traditional 2-D place and route (PnR) implementation flow from Cadence using Genus for synthesis and Innovus for PnR. AuC memories have been integrated after macro-level optimization as discussed in Section III where block-level

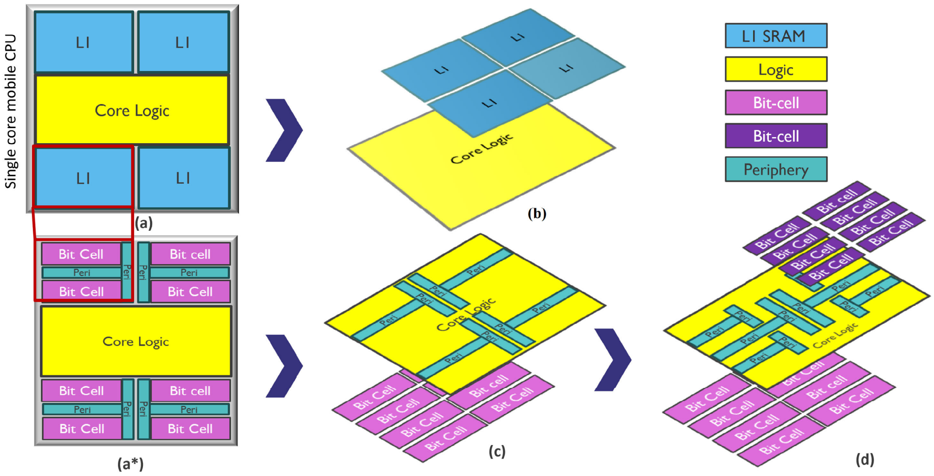


FIGURE 4. **Partitioning scenarios (a) 2-D, (a*) transparent 2-D, (b) 3-D F2F-HB MoL, (c) AuC two-tier, and (d) AuC three-tier.**

logic standard cells are allowed to sit on top of the bit-cells since they are in different plane in reality. Memory peripheral standard cells are sharing the same plane since in reality both are implemented by only logic cells. For 3-D MoL implementation, we have used integrity 3D-IC with state-of-the-art concurrent 3-D flow. This allows the implementation tool to optimize both the dies concurrently during PnR.

In our study, we use buried power rail (BPR) and back-side power delivery network (BS-PDN) with three metal layers. For AuC, we implement a traditional 2-D power grid across logic plane and consider BS-PDN for bit-cell plane. Furthermore, for 3-D MoL stacking, we again implement a standard 2-D power grid across all the dies, excluding power exchange between the dies (3-D power structures are not included). The omission of IR-drop analysis is due to its scope falling beyond the focus of this article. However, it is important to note that the choice of BPR and BS-PDN significantly impacts the thermal properties of the stack, hence placing the logic die on top (near to heat-sink) is mostly preferable.

C. PARTITIONING AND FLOORPLAN CONSIDERATIONS

We investigated two partitioning scenarios for the AuC: two-tier and three-tier, as depicted in Fig. 4. In addition, we included a transparent version of the AuC in Fig. 4(a*), although it does not represent an actual implementation. This transparency aids readers in understanding the internal structure of the memory, including how periphery and bit-cells are partitioned. In the two-tier case, only block logic is allowed to reside on top of the bit-cells. In the three-tier scenario, both periphery and block logic can be placed either above or below the bit-cells. Specifically, some bit-cells are positioned in the top tier, while others are in the bottom tier. However, the memory peripheries of the bottom tier cannot overlap with the bottom tier bit-cells, and the same restriction applies to the top tier memory periphery, as illustrated in Fig. 5.

We have examined only one partitioning scheme, denoted as the MoL as our 3-D reference scenario. This partitioning choice aligns with related research methodologies in the field of 3-D stacked caches [17], [25] and allows a direct comparison with AuC. Our primary focus remains on the AuC, a fine-grained 3-D partitioning approach. All the implementations (2-D, 3-D, and AuC) are developed with approximately equal design utilization of 60%. To maintain consistency across design explorations, we deducted specific empty silicon area from both the 3-D and AuC two-tier floorplans [as shown in Fig. 5(b) and (c)]. The two-tier footprint of 3-D and AuC is comparable, representing approximately 50% of the 2-D footprint. Notably, the AuC three-tier configuration achieved an additional 35% footprint reduction compared with the two-tier setup due to more staking options. Unfortunately, further optimization in terms of area and utilization was constrained by the memory peripheral shape and the physical size of the ARM core in the considered technology node. The SRAM peripheral area, being only a small fraction (9%) of the total logic die, minimally impacts the overall utilization and congestion of the top die. The AuC integration method confines routing blockages to metal 1, allowing the top die to use all the lower metal layers from M0, thus reducing congestion and avoiding disruption to the top logic die routing.

V. EXPERIMENTS AND RESULTS

A. MACRO DESIGN

Using circuit simulations, read access delay and leakage power of AuC macro design is evaluated and compared with the conventional 2-D SRAM macro counterpart. All the simulations for 2-D and homogeneous AuC SRAM macros are implemented in an advanced A10 nanosheet technology node. The heterogeneous AuC SRAM macros are implemented using a custom mix of FinFET (bit-cells) and nanosheet (peripheries) technology in A10 and N3 nodes, respectively. Circuit simulations for both 2-D and AuC

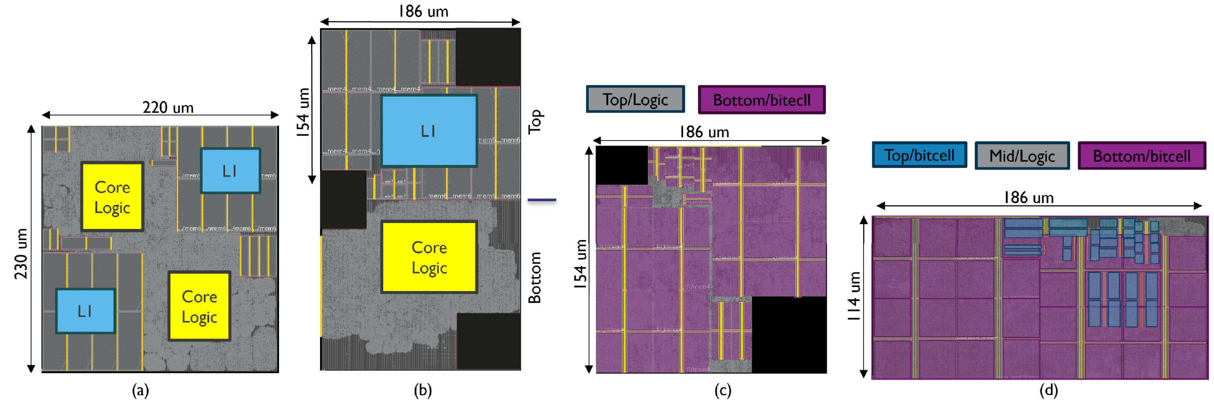


FIGURE 5. Floorplans (a) 2-D, (b) 3-D MoL, (c) AuC two-tier, and (d) AuC three-tier.

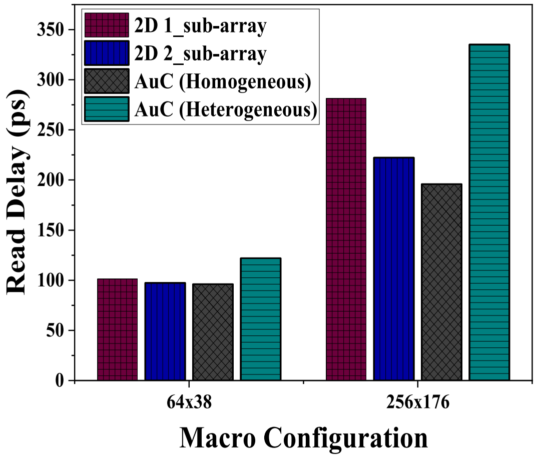


FIGURE 6. Read delay comparison of 2-D and AuC technologies for different SA configurations.

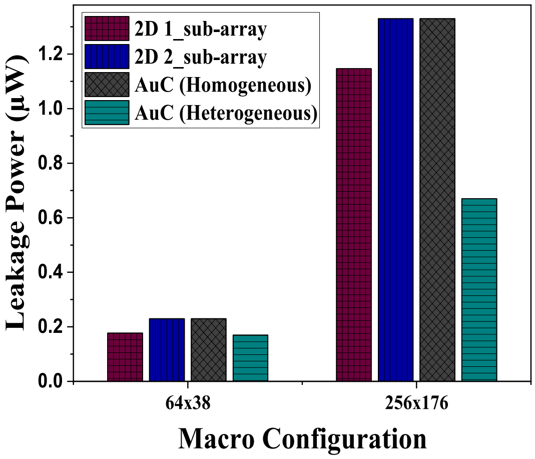
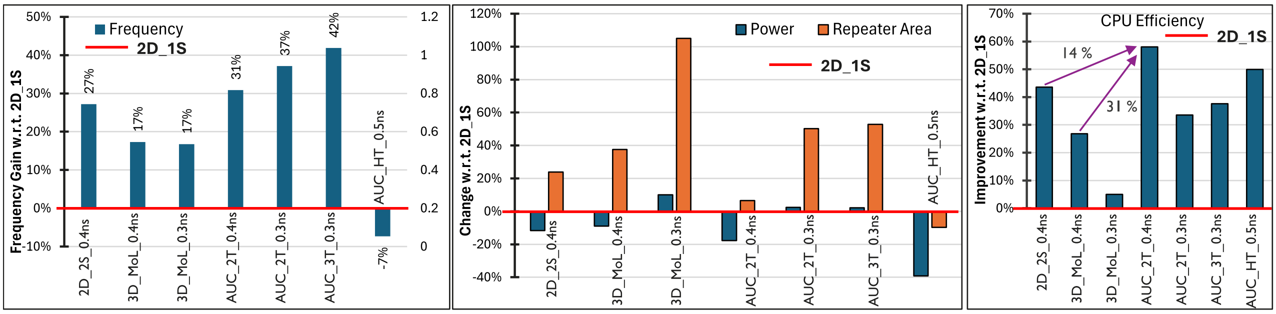


FIGURE 7. Leakage power comparison of 2-D and AuC technologies for different SA configurations.

SRAM macros are performed in Cadence Spectre circuit simulator [26] at a typical corner using a device compact model for nanosheet and FinFET [7], [27]. Two-dimensional macro configuration with two SAs is considered as the baseline 2-D in this study for a fair comparison with AuC macros. For SA performance analysis, we explore two distinct SRAM SA configurations 64 × 38 (#WL is 64 and #BL is 38) and 256 × 176 (#WL is 256 and #BL is 176) denoting the macro capacity of 1 and 10 kb, respectively. In this study, we modify the aspect ratios of WL and BL within the AuC technology framework, by 2× and compare their performance against the 2-D baseline.

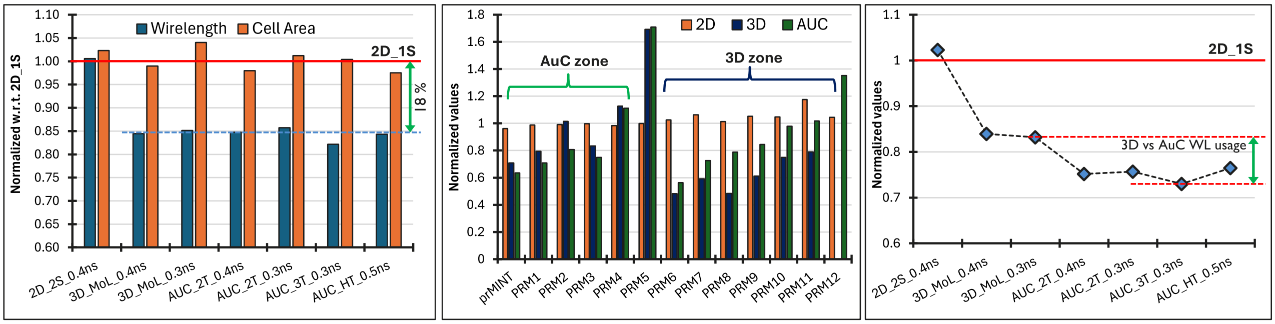
The read access delay comparison between 2-D macros and AuC for different SA configurations is shown in Fig. 6. Three-dimensional macro with two SA configuration performs faster when compared with 2-D macro with one SA by ∼4% and ∼20% in case of 64 × 38 and 256 × 176 SA, respectively. The performance gain in two SA configuration is obtained due to shorter word-line length (reduced *RC* delay) connecting the row periphery to the worst case bit-cell (corner bit-cell). When compared with the 2-D baseline, homogeneous AuC technology is slightly faster in case of 1-kB macro capacity (64 × 38 SA), whereas for larger memory capacity of 10 kB with SA configuration of 256 × 176, homogeneous AuC shows ∼12% performance improvement because it contributes to lower word-line and bitline resistances as a result of increased AR of metal lines by 2×. The heterogeneous AuC counterpart slows down by around 25% in case of 64 × 38 SA configuration, while in case of 256 × 176 SA, it shows a large degradation in performance by ∼50%. In heterogeneous AuC technology, the SRAM bit-cell arrays and its peripheries correspond to FinFET and nanosheet technologies, respectively. FinFET enabled SRAM bit-cells majorly contribute toward the performance degradation in heterogeneous AuC when compared with 2-D where both the peripheries and the bit-cells are enabled by relatively faster nanosheet transistors.

Fig. 7 shows the comparison of leakage power between 2-D and AuC macros for different SA configurations. Heterogeneous AuC technology dissipates less leakage power when



(a) (b) (c)

FIGURE 8. **Performance analysis (a) frequency, (b) power versus repeater area, and (c) CPU efficiency.**



(a) (b) (c)

FIGURE 9. **Wirelength statistics (a) total wirelength versus cell area, (b) metal distribution, and (c) memory to flop wirelength.**

compared with 2-D baseline in both 64 × 38 and 256 × 176 SA configurations by ∼26% and ∼50%, respectively. Since the SRAM bit-cells in heterogeneous AuC correspond to FinFET technology (less leaky), they contribute to a significant low leakage power than 2-D and homogeneous AuC counterparts.

B. SYSTEM-LEVEL INTEGRATION

Multiple configurations of 3-D and AuC integration have been assessed with respect to 2-D single and double SA (Section III) SRAM memory instances in block level using an ARM core with several target frequencies. The implementations represent best in the class in each category which came after rigorous optimization of timing constrains and floorplan iterations. Block-level performance in terms of achieved frequency and power consumption is depicted in Fig. 8. A PnR run is considered valid if the worst negative slack (WNS) is negative and it is absolute value is less than 10% of the target period. Furthermore, the count of failing paths (FPs) should be lower than 1000. This design methodology ensures realistic area and power estimates. For our current work, the delay and power estimation have been performed for the standard cell library settings mentioned in Section IV-A. The power numbers reported are based on the activity annotation from the *Dhrystone* workload. On the X-axis, we use the notation I_C_F, where I indicates the integration option, C denotes the configuration used, and F represents the PnR target period expressed in nanoseconds. The Y-axis data are normalized with respect to the 2-D single SA, which serves as the baseline at 0% level. In addition, we present the single-core efficiency [Fig. 8(c)], which is a function of the power and delay product

$$\text{CPU efficiency} = \frac{1}{\text{Total Power} \times \text{Eff. Period}}. \tag{1}$$

Our analysis shows that all the implementations using AuC memories can operate at a higher frequency than 2-D baseline and further exhibit a frequency gain of at least 14% compared with the 3-D implementation. This is because the faster memory access along with area and wirelength savings significantly reduce the critical path delay at the block level. The frequency gain is not that visible only in AuC heterogeneous case because of the much slower SRAM bit-cells in FinFET technology. But that is also a very minor (only 7%) penalty in delay, thanks to the area and resulting wirelength savings in AuC integration. This reduction in frequency can further be complimented by 40% power reduction and 10% less repeater area in Fig. 8(b). AuC heterogeneous can bring down the power consumption as low as −40%. This is possible because of the heterogeneous technology itself. Compared with advanced note technology, SRAM blocks with older node (FinFET) bit-cells consume significantly less energy, albeit at a slower speed. Moreover, the other AuC cases also highlight power optimization up to 10% more than 3-D for both memory and block-level compact optimization. If we look into the matrix of power and delay product, namely, CPU efficiency in Fig. 8(c), we find AuC two-tier and heterogeneous integration is highly efficient showing lower per unit power consumption for a given target frequency.

Further investigating wirelength statistics in Fig. 9, it is unsurprising that both 3-D MoL and AuC use approximately 18% less routing resources due to vertical integration capability. This reduction in the wirelength brings the core logic closer, contributing to delay reduction. Notably, in our benchmark design with only 500 K cells, this wirelength gain (∼18%) will be even more significant for larger designs.

One interesting observation arises when examining the metal distribution and memory-flop wirelength, as depicted in Fig. 9. The use of AuC demonstrates greater efficiency in saving lower metal layers compared with 3-D MoL. This advantage stems from the reduced utilization of metal layers (up to M3) within the SRAM memory block itself (as discussed in Section III). Such an advantage is not feasible in 2-D or even 3-D designs that rely on traditional memory blocks. Notably, highly resistive lower metal layers are costlier to fabricate. However, AuC presents significant potential for savings in this regard. The abrupt increase in M5 [as indicated in Fig. 9(b)] cannot be avoided in 3-D and AuC designs due to the placement of block-level I/Os on this metal layer. Unlike 2-D designs, where blockages within the memory block restrict the use of M5, the PnR tool tends to favor M5 (highest Mx layer in the used technology) in 3-D and AuC layouts, especially when minimal blockages exist at this layer. Finally, another noteworthy observation suggests that AuC outperforms 3-D MoL in terms of metal savings [Fig. 9(c)]. This advantage results from optimizations in both block area and memory metal layers. As design spaces grow larger, this phenomenon is expected to be even more pronounced.

VI. CONCLUSION

This study introduces a novel partitioning scheme for low-level SRAM caches using AuC technology, demonstrating enhancements in performance, overall CPU efficiency, and leakage power reduction. Preliminary investigations on a single-core mobile CPU highlight the significant impact of this next-generation advanced integration technique on small memory banks, with potential amplification at the system level. These findings suggest substantial gains in terms of power, performance, area, and cost benefits for larger SoCs with more integrated functionalities. This also underscores the promise of AuC technology for broader applications and emphasizes the importance of fine-grained functional partitioning in 3D-IC designs.

REFERENCES

[1] W. A. Wulf and S. A. McKee, “Hitting the memory wall: Implications of the obvious,” *ACM Comput. Arch. News*, vol. 23, no. 1, pp. 20–24, Mar. 1995.

[2] A. Gholami et al., “AI and memory wall,” 2024, *arXiv:2403.14123*.

[3] A. Gebregiorgis et al., “A survey on memory-centric computer architectures,” *ACM J. Emerg. Technol. Comput. Syst.*, vol. 18, no. 4, pp. 1–50, Oct. 2022.

[4] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch, “Disaggregated memory for expansion and sharing in blade servers,” *SIGARCH Comput. Archit. News*, vol. 37, no. 3, pp. 267–278, 1145.

[5] S. B. Samavedam et al., “Future logic scaling: Towards atomic channels and deconstructed chips,” in *IEDM Tech. Dig.*, Dec. 2020, pp. 1–10.

[6] P. Weckx et al., “Novel forksheet device architecture as ultimate logic scaling device towards 2nm,” in *IEDM Tech. Dig.*, Dec. 2019, pp. 36.5.1–36.5.4.

[7] D. Jang et al., “Device exploration of NanoSheet transistors for sub-7-nm technology node,” *IEEE Trans. Electron Devices*, vol. 64, no. 6, pp. 2707–2713, Jun. 2017.

[8] M. K. Gupta et al., “The complementary FET (CFET) 6T-SRAM,” *IEEE Trans. Electron Devices*, vol. 68, no. 12, pp. 6106–6111, Dec. 2021.

[9] S. M. Salahuddin et al., “SRAM with buried power distribution to improve write margin and performance in advanced technology nodes,” *IEEE Electron Device Lett.*, vol. 40, no. 8, pp. 1261–1264, Aug. 2019.

[10] R. Mathur et al., “Buried bitline for sub-5nm SRAM design,” in *IEDM Tech. Dig.*, Dec. 2020, pp. 20.2.1–20.2.4.

[11] Md. A. Baig et al., “3-D monolithic stacking of complementary-FET on CMOS for next generation compute-in-memory SRAM,” *IEEE J. Electron Devices Soc.*, vol. 11, pp. 107–113, 2023.

[12] L. Zhu et al., “High-performance logic-on-memory monolithic 3-D IC designs for arm cortex-A processors,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 6, pp. 1152–1163, Jun. 2021.

[13] S. Srinivasa et al., “ROBIN: Monolithic-3D SRAM for enhanced robustness with in-memory computation support,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 7, pp. 2533–2545, Jul. 2019.

[14] Y. Yu and N. K. Jha, “A monolithic 3D hybrid architecture for energy-efficient computation,” *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 4, no. 4, pp. 533–547, Oct. 2018.

[15] D. B. Ingerly et al., “Foveros: 3D integration and the use of face-to-face chip stacking for logic devices,” in *IEDM Tech. Dig.*, Dec. 2019, pp. 9.6.1–19.6.4.

[16] S. E. Kim and S. Kim, “Wafer level Cu–Cu direct bonding for 3D integration,” *Microelectron. Eng.*, vol. 137, pp. 158–163, Apr. 2015.

[17] R. Chen et al., “3D-optimized SRAM macro design and application to memory-on-logic 3D-IC at advanced nodes,” in *IEDM Tech. Dig.*, Dec. 2020, pp. 15.2.1–15.2.4.

[18] A. Vandooren et al., “First demonstration of 3D stacked finfets at a 45 nm fin pitch and 110 nm gate pitch technology on 300 mm wafers,” in *IEDM Tech. Dig.*, Dec. 2018, pp. 7.1.1–7.1.4.

[19] S. M. Salahuddin et al., “Thermal stress-aware CMOS–SRAM partitioning in sequential 3-D technology,” *IEEE Trans. Electron Devices*, vol. 67, no. 11, pp. 4631–4635, Nov. 2020.

[20] Y. Xie, “Processor architecture design using 3D integration technology,” in *Proc. 23rd Int. Conf. VLSI Design*, vol. (10), Jan. 2010, pp. 446–451.

[21] B. Black et al., “Die stacking (3D) microarchitecture,” in *Proc. 39th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Dec. 2006, pp. 469–479.

[22] S. Yang et al., “PPA and scaling potential of backside power options in N2 and A14 nanosheet technology,” in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2023, pp. 1–2.

[23] E. Beyne, “A view on the 3D technology landscape: Design and technology options for 3D systems-on-chip,” *IMEC*, Oct. 2021. Accessed: May 27, 2024. [Online]. Available: https://www.imec-int.com/en/articles/view-3d-technology-landscape

[24] E. Beyne, D. Milojevic, G. Van der Plas, and G. Beyer, “3D SoC integration, beyond 2.5D chiplets,” in *IEDM Tech. Dig.*, Dec. 2021, pp. 3.6.1–3.6.4.

[25] A. Agnesina et al., “Power, performance, area, and cost analysis of face-to-face-bonded 3-D ICs,” *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 13, no. 3, pp. 300–314, Mar. 2023.

[26] *Spectre-Cadence® Spectre® Circuit Simulator*.

[27] M. K. Gupta et al., “A comprehensive study of nanosheet and forksheet SRAM for beyond N5 node,” *IEEE Trans. Electron Devices*, vol. 68, no. 8, pp. 3819–3825, Aug. 2021.