Technology/Memory Co-Design and Co-Optimization Using E-Tree Interconnect

ABSTRACT

As technology scales down, interconnects become to dominate the delay and energy of VLSI systems. For on-chip SRAM, a major portion of delay and energy is contributed by the H-Tree interconnects. In this paper, we propose an E-Tree interconnect technology option to minimize the H-Tree delay and energy overheads based on an efficient interconnect technology/memory co-design framework for nonuniform workloads. It integrates a realistic cell library to enable a large design space exploration for emerging interconnect technology based on various workloads assumptions. Three emerging interconnect materials are studied and benchmarked against their traditional Copper counterparts for optimal SRAM performance, such as the energy-delay product (EDP) and energy-delay-area product (EDAP). Various array- and interconnect-level design parameters are co-designed for optimal performance.

CCS CONCEPTS

• Hardware→Very large scale integration design; • Hardware→ Integrated circuits→ Interconnect →Metallic interconnect; • Hardware→Integrated circuits→Semiconductor memory→Static memory.

KEYWORDS

Interconnect, E-Tree design, technology/memory co-optimization, SRAM, workload, center-pin access, emerging interconnect material.

1 INTRODUCTION

SRAM is one of the major components in on-chip VLSI systems due to its high density, great compatibility with industry-standard CMOS processes, good cost-efficiency, and lower leakage energy compared to DRAM [1]. One major limitation of the on-chip SRAM is its large delay and energy overheads associated with interconnects, including both local interconnects, i.e. bitline/wordline, and intermediate/global interconnects, i.e. H-Tree interconnects [2]. The large performance overhead is mainly caused by the large resistivity of traditional Copper interconnects that are suffered from the increasing size effect and impact of barrier thickness [3-6].

To minimize the interconnect delay and energy overheads, large research efforts have been performed to address interconnect challenges. Some previous work has architected large-scale SRAM with 3D integration technology [7-9]. However, one of the main challenges is the cost aspect due to the requirements of fabrication, such as thermal management and reliability. Another challenge is the issue of manufacturing complexity and yields because 3D interconnect SRAM technology requires highly precise alignment and bonding of multiple layers of memory, which can be challenging and expensive to achieve at scale.

On the material side, emerging interconnect technologies and processes, such as Graphene, Cobalt, and Ruthenium [10-13], have been proposed. Some existing work has investigated beyond-Cu interconnects for the SRAM application [14], showing that the cache-level delay and energy are mainly dominated by H-Tree interconnects. Therefore, alternative H-Tree interconnect design options are critical to further improve the overall cache-level performance, which will be the focus of this paper.

The traditional H-Tree interconnect provides minimal interconnect skew and good robustness against variations due to the symmetry of the H-Tree. In addition, H-Tree is easy to balance by construction with simple control logic [15]. However, due to the symmetry, accessing the cell that is right beside the root pin will have the same delay as accessing the farthest cell in the SRAM array. To improve the SRAM performance, it is important to redesign the interconnect technology and take into account the distance between the root pin and the location of the data.

In this paper, we propose an E-Tree interconnect technology to reduce the average interconnect length. The cell closer to the root pin will achieve a faster access time and lower energy dissipation due to the shorter interconnect. The proposed E-Tree design brings new opportunities to system-level optimization, where frequently used data can be moved closer to the input pin. We will investigate different workload assumptions and quantify their impacts on the optimal cache design and performance metrics.

In addition, we will study a center-pin access option to further reduce the interconnect length. The corresponding logic core placement will be taken into account for accessing the cache array. This work will use an experimentally verified sub-5nm technology library to investigate the true advantages of advanced interconnect materials at ultra-scaled technology nodes [16]. Based on the device technology, a cache subarray is designed, whose organization is composed of address control, row decoder, column multiplexer, write driver, sense amplifier, and array cell matrix. Last but not least, we will investigate key tradeoffs among various emerging interconnect design parameters, including interconnect geometry design, such as width and aspect ratio, and cache size.

The main contributions of the work are highlighted below.

1. We propose an E-Tree interconnect design to minimize the interconnect delay and energy overheads for the SRAM array.
2. We analyze the impact of different workload assumptions on the optimal cache design and performance metrics.
3. We investigate different access pin options, including side-pin and center-pin technologies, to co-optimize with emerging interconnect technologies.
4. Four interconnect material options are benchmarked to understand the true advantages of graphene-based interconnects on cache-level performance.

2 MODELING APPROACHES

In this section, we will illustrate the proposed E-Tree interconnect technology as well as the modeling for interconnect materials, repeater insertion, and array-level performance modeling approach.

2.1 E-Tree Interconnect Technology Design

Because the cache-level performance suffers from the delay and energy associated with the long H-Tree interconnects, we propose E-Tree technology options to reduce the average interconnect

length and further improve the array-level performance. Figure 1(a) shows the traditional cache array with side-pin access, where all three levels of hierarchies, including array, bank, and mat, use H-Tree interconnects. Figure 1(b) shows the proposed cache array using E-Tree interconnect for array- and bank-level interconnects. The horizontal interconnects coming into each hierarchy will split into vertical interconnects that are shared by every two columns of banks or mats. The main advantage of using E-Tree interconnects is to reduce the length of the interconnects when accessing the data that are physically located closer to the root pin (red arrow) or bank/mat inputs. Note that this type of asymmetric routing requires extra timing control logic circuits, which have not been included in this work. We will perform a more detailed design as well as study the architectural-level impact in our future work. The results presented in Section 3 will showcase the upper bound of the potential benefits of the proposed E-Tree interconnect network.

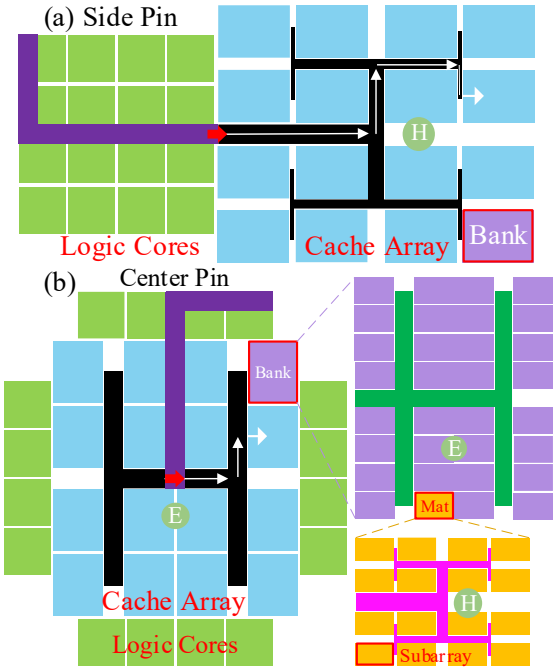

Figure 1: Schematic of cache using (a) traditional H-Tree interconnect with side-pin access and (b) proposed E-Tree interconnect with center-pin access. For the center-pin access, the access interconnect (purple line) is shown for the worst-case scenario, where the farthest logic core is connected to the root pin. The arrow in red indicates the root pin location.

For simplicity, the workload assumption for the proposed E-Tree is that the probability of access to each subarray is negatively correlated to the distance between the root pin and subarray. To quantify the workload, we assume the following access pattern:

$$P_{subarray\_i} \propto \frac{1}{L_{subarray\_i}^{\alpha}} \quad (1)$$

$$\sum_{i=1}^{all} P_{subarray\_i} = 1 \quad (2)$$

$$L_{average} = \sum_{i=1}^{all} \left(L_{subarray\_i} \cdot P_{subarray\_i} \right) \quad (3)$$

where α is the cache access probability factor, $P_{subarray\_i}$ and $L_{subarray\_i}$ are the access probability and interconnect length from the pin to the subarray i, respectively, $L_{average}$ is the average E-Tree length based on access probability for nonuniform workloads. For the center-pin technology shown in Figure 1(b), the logic cores are distributed around the cache array. For both side- and center-pin access options, we assume that the core area is equal to the total subarray area. The worst-case scenario is considered to calculate the logic cores-to-cache interconnect length, meaning that the interconnects connect from the corner of the logic cores to the root pin of the cache array, shown as the purple lines in Figure 1.

2.2 Interconnect Modeling

For the inter-array interconnects, such as logic-to-cache access interconnects and H-Tree/E-Tree interconnects, the delay of interconnect with repeater insertion based on the optimal repeater spacing and size is modeled from the existing work based on the original CACTI work [2, 14, 17], and the schematic of the circuit model for interconnects is shown in Figure 2. Device-level parameters, such as the drain/gate capacitance and output resistance, are extracted based on realistic device technology in Synopsys HSPICE and Cadence Spectre simulations [16, 18, 19].

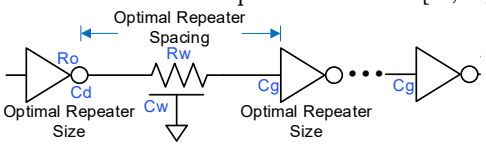

Figure 2: Circuit model of interconnects based on optimal repeater insertion. C_g and C_d are gate and drain capacitance of the device, respectively, R_o is the output resistance of the repeater, C_w and R_w are interconnect capacitance and resistance, respectively, and additional quantum and contact resistance $R_{quantum}$ and R_{con} are added on each side of the graphene interconnect.

2.3 Cache Array and Subarray Modeling

CACTI, one well-known and open-source simulator, is adopted to optimize the SRAM cache [17, 20]. CACTI sweeps the cache and array organization parameters to get optimal parameters for the target metric defined by the user. The original CACTI model has been already verified by SPICE simulation and data reported on the commercial caches from the Intel L3 cache at 65nm and Sun SPARC L2 cache at 90nm [20]. By the validated cache simulator, various configurations of interconnect and organization parameters can be explored efficiently with good accuracy at the early stage of design.

To incorporate the realistic subarray, key performance metrics, including energy, area, and delay for various components from the original CACTI model are updated based on the actual data obtained from the results of realistic experiments and simulations by Synopsys HSPICE and Cadence Spectre [18, 19]. In addition, we have developed a high-level SRAM subarray model based on equations to enable efficient and accurate analysis of the energy dissipation and latency for the large cache using various interconnect material options. Extensive electrical-level simulations have been performed to validate the accuracy of the compact model.

The cache capacity varies from multiple Mbits to even larger than 1 Gbits for the last-level caches (LLCs) based on the latest published high-performance processors [21-23]. It motivates us to investigate a large range of SRAM systems and how they interact with the different interconnect technology designs and materials.

2.4 Interconnect Materials

Four promising options of interconnect materials are adopted to quantify the impacts of materials on the performance of cache array-level based on the existing modeling work, including (1) Cu as the baseline, (2) graphene-capped Ruthenium, (3) graphene-capped Copper (Cu), and (4) thick graphene [2, 10, 13, 14, 24-32].

For the baseline, the Copper (Cu) interconnect, whose resistance model follows the existing work in the grain boundary reflectivity

value of 0.5 and the side wall specularity value of 0.5 which calibrated based on experimental results [27, 31]. For the general graphene-based interconnect material option, the current flowing through one single-layer graphene can be obtained from the Landauer formula [33], which is a function of the graphene effective mean free path (MFP). The MFP value depends on several factors, such as the edge roughness of graphene and the property of the substrate material. The MFP value has been fitted based on the data of mobility extracted from experimental results by the semiclassical equation [2, 14, 32, 34].

For graphene-capped Ru interconnect, the resistance per unit length is obtained based on experimental data under different thicknesses [25]. For graphene-capped Copper (Cu), the electrons scatter inside Cu less frequently, and 3× of the grain size value is used to capture this effect based on the published experimental work [28, 29]. The interconnect capacitance per unit length is extracted for various wire geometry by Synopsys Raphael [35].

3 SIMULATION AND EXPLORATION RESULTS

In this section, we will perform the interconnect/cache co-design based on different workload assumptions for the proposed E-Tree interconnect network with side- and center-pin technologies. Four interconnect materials introduced in Section 2 (i.e., Copper, graphene-capped Ruthenium, graphene-capped Copper, and thick graphene) will be investigated and benchmarked. Unless specified elsewhere, the SRAM cache level, material, and interconnect design parameters and their default value used in the modeling and simulation are listed in Table 1.

Table 1: Parameters Used in the Modeling and Simulation

Parameter	Value
Cache Size (MB)	128
Number of Banks	16
Associativity	1
Number of Subarray Rows	128
Number of Subarray Columns	128
Core-to-cache Cu Interconnect Width (μm)	1
Core-to-cache Cu Interconnect Aspect Ratio	0.1
Intra-subarray Interconnect Width (nm)	11
Inter-subarray Interconnect Width (nm)	28
Intra-subarray Interconnect Aspect Ratio	4
Inter-subarray Interconnect Aspect Ratio	1
Graphene Mean-Free-Path at W = 1μm (nm)	460
Graphene Contact Resistance (Ω·μm)	100

3.1 Impact of E-Tree on Wire Distribution and Access Probability

To better analyze the cache performance, we first investigate the impact of the E-Tree network on the interconnect length and access probability for each bank, mat, and subarray. Figure 3 shows the interconnect length and access probability to each bank for the cache using side- and center-pin access. The access probability to a bank is the sum of the probability of access to subarrays in this bank. One can observe that the bank close to the input pin (red arrow) has a shorter interconnect length and higher access probability.

For the cache using side-pin access, Figure 4 (a) and (b) show the probability of access and the number of interconnects for each subarray under different lengths of interconnect, respectively. The average interconnect length of the E-Tree is smaller than the H-Tree counterpart because there are short interconnects that directly access the subarray that is close to the input pin at three levels of hierarchies, including mat, bank, and array. Compared to the average interconnect length from the cache using the E-tree with side-pin access, the one using the center-pin access shown in Figure 4 (c) and (d) is shorter due to the closer distance to the pin.

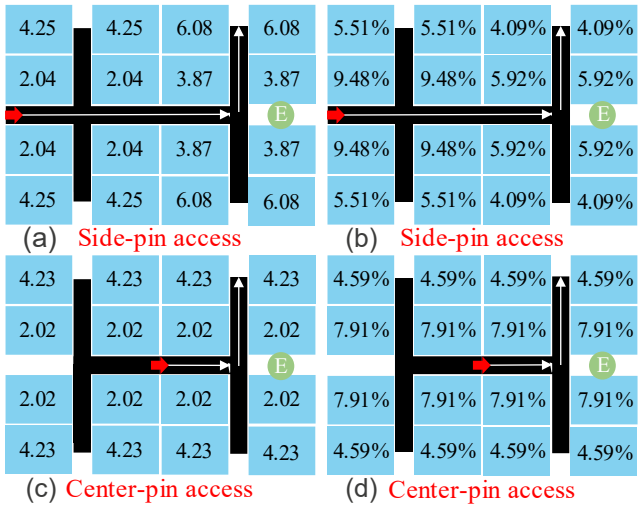

Figure 3: (a)(c) Interconnect length in (μm) from the root pin to the bank and (b)(d) access probability from different E-Tree technology options under the cache size of 128MB. The side-pin access is for (a)(b) and the center-pin access is for (c)(d). The arrows in red indicate the root pin locations.

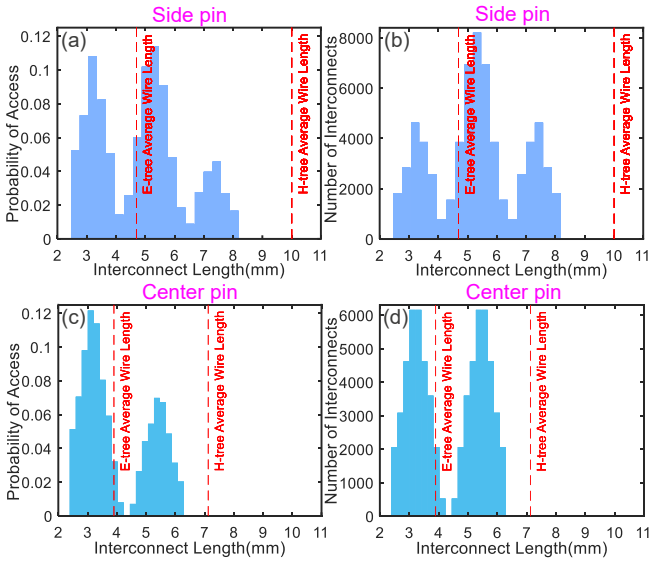

Figure 4: (a)(c) Probability of access and (b)(d) the number of interconnects versus interconnect length from different E-Tree technologies under the cache size of 128MB. The side-pin access is for (a)(b) and the center-pin access is for (c)(d).

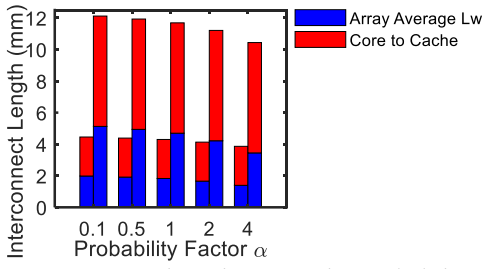

Figure 5: Interconnect length versus the probability factor α for E-Tree with side-pin access. For each probability factor α, the left and right bars are for cache sizes of 16MB and 128MB, respectively. H-Tree length is 5.997mm under 16MB.

3.2 Impact of Workload for E-Tree Interconnect on Cache Performance

Based on the average interconnect length obtained in the previous subsection, we perform the cache-level performance optimization using the co-design framework for nonuniform workloads described in Section 2. Under different workload assumptions, Figure 5 shows the interconnect length versus cache access probability factor α for the cache using side-pin access under the cache size of 16MB and 128MB. The interconnect length of the E-Tree decreases with probability factor α because of the higher access probability to data that are closer to the input pin, as shown in Figure 5.

Figure 6 (a) and (b) show the breakdown bar charts of delay and energy for different probability factors α under side-pin access. The overall delay is mainly dominated by the array E-Tree interconnects due to the smaller interconnect width at the intermediate metal level in the array. The delay for the core-to-cache interconnects is relatively small because these interconnects locate at the global metal level with a large interconnect width. However, the overall energy is dominated by the core-to-cache interconnects due to their longer lengths. Note that the energy is shown with the log scale due to the large energy difference for different energy components. For different workload assumptions, both delay and energy decrease with the increase of the probability factor because of the decreasing average E-Tree length at the array level as shown in Figure 5. To take delay, energy, and area into account, Figure 6 (c) and (d) show the energy-delay product (EDP) and energy-delay-area product (EDAP) versus the probability factors α for different cache sizes under side-pin access. The interconnect material is thick graphene.

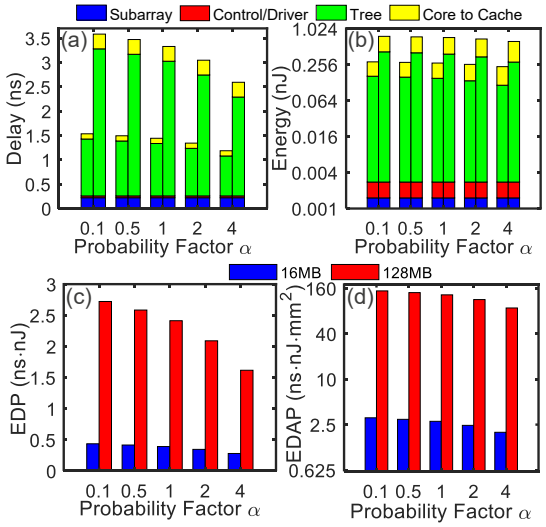

Figure 6: (a) Delay, (b) energy, (c) EDP, and (d) EDAP versus probability factor α for E-Tree with side-pin access in thick graphene. For each probability factor α, the left and right bars are for the cache size of 16MB and 128MB, respectively. The delay of cache using H-Tree is 2.47ns under 16MB.

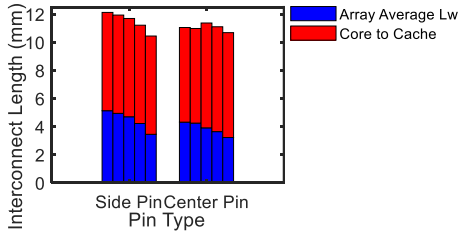

Figure 7: Interconnect length versus the E-Tree interconnect technology option under the cache size of 128MB. For side-pin and center-pin access, the bars from left to right are probability factor α of 0.1, 0.5, 1, 2, and 4.

3.3 Impact of Interconnect Access Pin Types on Cache Performance

To quantify the potential benefits of the proposed center-pin technology, Figure 7 and Figure 8 show various metrics versus two pin types. The cache using the E-Tree with center-pin access outperforms the side-pin counterparts because the first critical segment length in the array for the side-pin access is large, leading to a significant average interconnect length and delay overhead. The cache using E-tree with center-pin access outperforms the side-pin access counterpart due to a smaller interconnect length, as shown in Figure 7. To take delay, energy, and area into account, Figure 8 (c) and (d) show the EDP and EDAP versus the interconnect technology option for different cache access probability factors α.

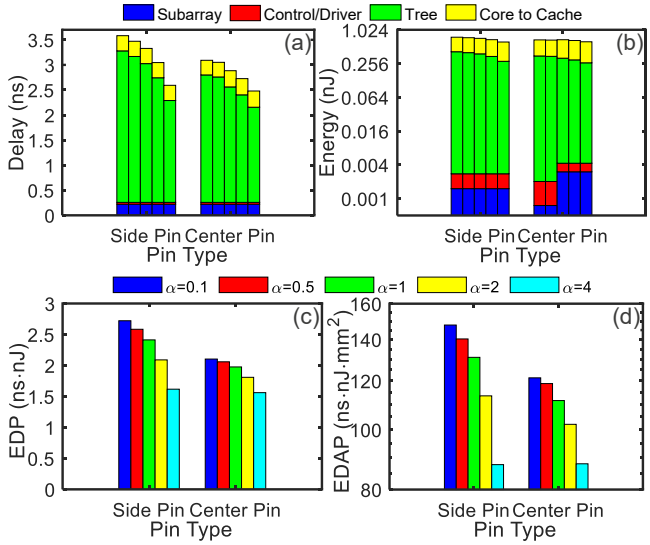

Figure 8: (a) Delay, (b) energy, (c) EDP, and (d) EDAP versus E-Tree interconnect technology option in ideal thick graphene under the cache size of 128MB. For side-pin and center-pin access, the bars from left to right are probability factor α of 0.1, 0.5, 1, 2, and 4.

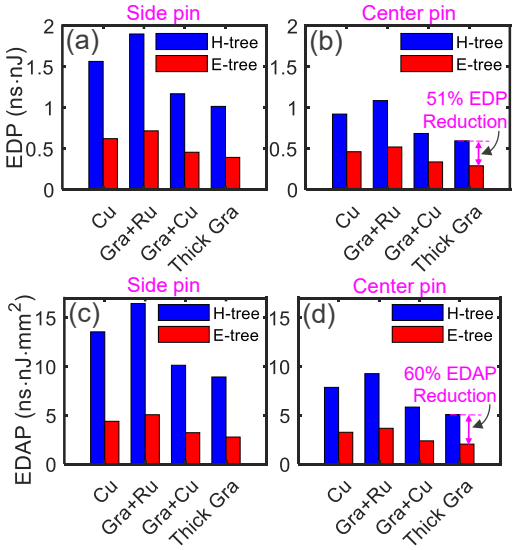

Figure 9: (a)(b) EDP and (c)(d) EDAP versus the interconnect material for H-Tree and E-Tree for interconnect technology options in optimal interconnect width and aspect ratio under the cache size of 128MB. The side-pin access is for (a) (c) and the center-pin access is for (b)(d).

3.4 Comparisons of H-Tree and E-Tree Networks Using Various Interconnect Material

To benchmark different interconnect technology options, Figure 9 shows optimal EDP and EDAP versus interconnect material for the cache using traditional H-Tree and proposed E-Tree design. In general, cache using thick graphene E-Tree with center-pin access outperforms its side-pin-based counterparts in terms of EDP and EDAP due to the relatively large advantage in interconnect resistance. The cache using graphene interconnect E-Tree with center-pin access provides the best performance, where up to 51% and 60% reduction in EDP and EDAP can be observed compared to the traditional H-Tree counterparts, as shown in Figure 9(b) and (d).

3.5 Impact of Interconnect Geometry on Cache Performance

Finally, we investigate the impact of various interconnect geometries on the cache performance, including aspect ratio and width. Figure 10 (a) and (b) show the breakdown bar charts of delay and energy for different inter-subarray interconnect aspect ratios under different access pin options with E-Tree. In general, delay decreases with the inter-subarray interconnect aspect ratio due to small resistance per unit length. Meanwhile, the energy of the side-pin- and center-pin-based cache increases with the inter-subarray interconnect aspect ratio due to the large capacitance. In short, the cache performance is either (i) limited by the delay if the inter-subarray aspect ratio is small or (ii) limited by the energy if the inter-subarray aspect ratio is large. Therefore, an optimal inter-subarray aspect ratio exists to minimize the overall EDP and EDAP, as shown in Figure 10 (c) and (d).

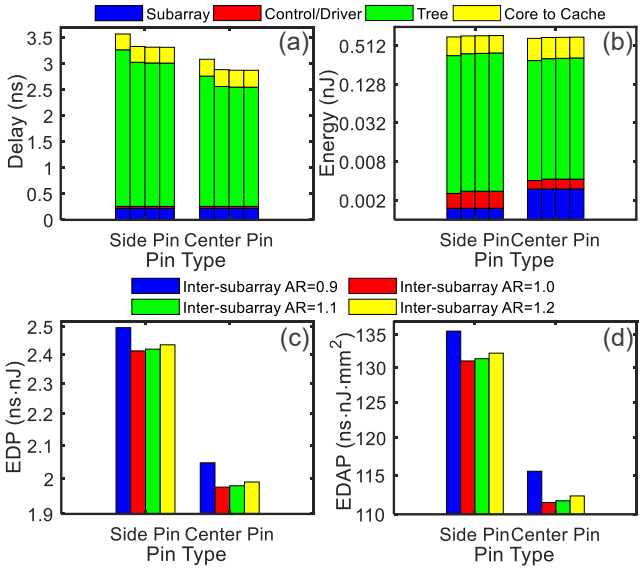

Figure 10: (a) Delay, (b) energy, (c) EDP, and (d) EDAP versus the E-Tree technology option for different inter-subarray interconnect aspect ratios in ideal thick graphene under the cache size of 128MB. For each pin type, bars from left to right are the inter-subarray interconnect aspect ratio of 0.9, 1.0, 1.1, and 1.2.

To quantify the interconnect width impact on different array-level performances, we use a width scaling factor (WSF) to multiply the standard interconnect width. Figure 11 (a)(b) and (c)(d) show the comparisons of breakdown bar charts of delay and energy for different inter-subarray interconnect widths under different interconnect technology options, respectively. In general, the delay decreases with the increase of the inter-subarray E-Tree interconnect width thanks to the smaller resistance per unit length. The energy increases due to the large interconnect length caused by the area overhead, as shown in Figure 12. Compared to the cache using E-Tree, the length of the cache using H-Tree is more sensitive to the interconnect width due to its longer length and larger area overhead. The cache using the center-pin technology outperforms its side-pin E-Tree counterpart due to a similar reason in the previous subsection 3.3.

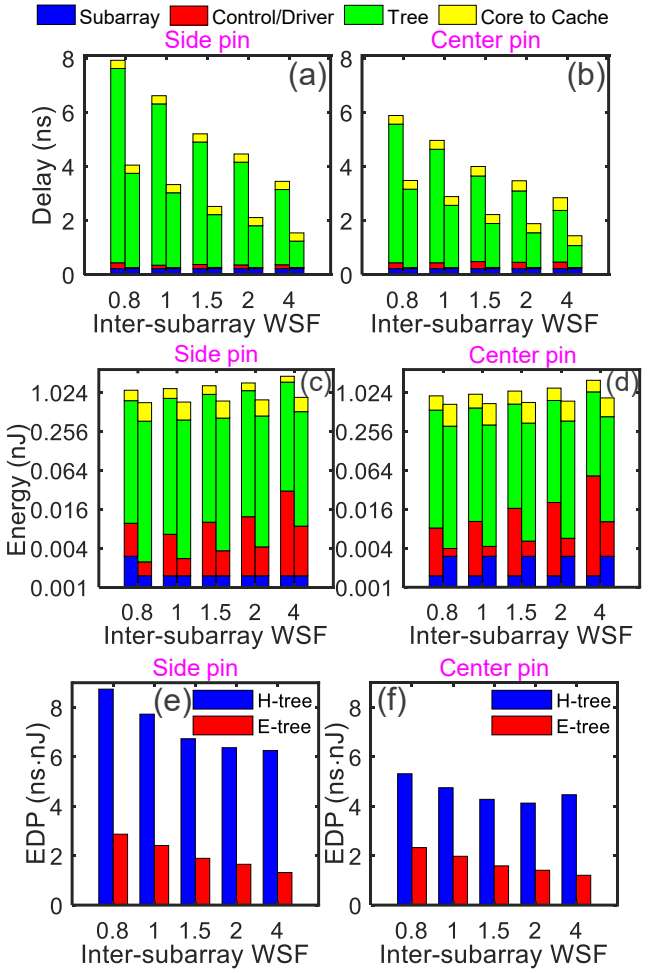

Figure 11: (a)(b) Delay, (c)(d) energy, and (e)(f) EDP versus the inter-subarray interconnect width scaling factor (WSF) for H-Tree and E-Tree using thick graphene under the cache size of 128MB. The side-pin access is for (a), (c), and (e) and the center-pin access is for (b), (d), and (f). For each interconnect WSF, the left and right bars are for the H-Tree and E-Tree, respectively.

To take both delay and energy into account, Figure 11 (e) and (f) show the comparison of EDP versus inter-subarray interconnect width for different interconnect technology options. In general, the EDP decreases with the increase of width due to the smaller delay caused by small resistance per unit length. To take the area into account, Figure 13 (a)(b) and (c)(d) show the area and EDAP versus the inter-subarray interconnect width with the same configurations, respectively. The optimal inter-subarray interconnect width exists to minimize the EDAP performance of cache using E-Tree with center-pin and side-pin access because its longer interconnect length induces a larger area overhead at a large interconnect width.

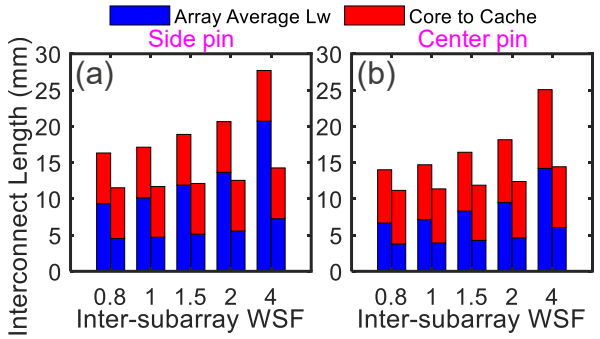

Figure 12: Interconnect length versus the inter-subarray interconnect WSF for (a) side-pin and (b) center-pin access using ideal thick graphene under the cache size of 128MB. For each interconnect WSF, the left and right bars are for the cache using H-Tree and E-Tree, respectively.

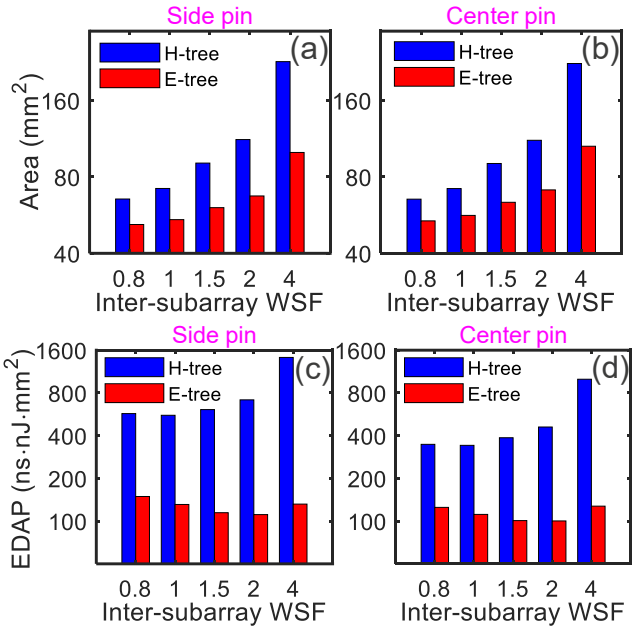

Figure 13: (a)(b) Area and (c)(d) EDAP versus the inter-subarray interconnect WSF using ideal thick graphene under the cache size of 128MB. The side-pin access is for (a) and (c) and the center-pin access is for (b) and (d). For each interconnect WSF, the left and right bars are for the cache using H-Tree and E-Tree, respectively.

4 CONCLUSION

In this paper, we propose a novel E-Tree interconnect technology option to substantially reduce the average length of the interconnect, leading to a smaller overhead in access delay and energy. Two access strategies are investigated, including side-pin and center-pin access, for different workload assumptions. In addition, three novel interconnect materials are benchmarked against their traditional Cu H-Tree interconnect counterpart. The SRAM cache system using E-Tree with thick graphene interconnect and center-pin access provides the best performance, where up to 51% and 60% reduction in EDP and EDAP can be observed compared to the thick graphene counterparts using H-Tree interconnects.

REFERENCES

[1] M. K. Gupta *et al.*, "A comprehensive study of nanosheet and forksheet SRAM for beyond N5 node," *IEEE Transactions on Electron Devices,* vol. 68, no. 8, pp. 3819-3825, 2021.

[2] Z. Pei *et al.*, "Graphene-Based Interconnect Exploration for Large SRAM Caches for Ultrascaled Technology Nodes," *IEEE Transactions on Electron Devices,* vol. 70, no. 1, pp. 230-238, 2022.

[3] R. Brain, "Interconnect scaling: Challenges and opportunities," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 9.3. 1-9.3. 4: IEEE.

[4] G. Bonilla, N. Lanzillo, C.-K. Hu, C. Penny, and A. Kumar, "Interconnect scaling challenges, and opportunities to enable system-level performance beyond 30 nm pitch," in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 20.4. 1-20.4. 4: IEEE.

[5] D. Prasad, A. Ceyhan, C. Pan, and A. Naeemi, "Adapting interconnect technology to multigate transistors for optimum performance," *IEEE Transactions on Electron Devices,* vol. 62, no. 12, pp. 3938-3944, 2015.

[6] K. Cho *et al.*, "SRAM write-and performance-assist cells for reducing interconnect resistance effects increased with technology scaling," *IEEE Journal of Solid-State Circuits,* vol. 57, no. 4, pp. 1039-1048, 2022.

[7] J. Kong, Y.-H. Gong, and S. W. Chung, "Architecting large-scale SRAM arrays with monolithic 3D integration," in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2017, pp. 1-6: IEEE.

[8] R. Chen *et al.*, "3D-optimized SRAM macro design and application to memory-on-logic 3D-IC at advanced nodes," in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 15.2. 1-15.2. 4: IEEE.

[9] S. Srinivasa *et al.*, "A monolithic-3D SRAM design with enhanced robustness and in-memory computation support," in *Proceedings of the International Symposium on Low Power Electronics and Design*, 2018, pp. 1-6.

[10] S. Achra *et al.*, "Metal induced charge transfer doping in graphene-ruthenium hybrid interconnects," *Carbon,* vol. 183, pp. 999-1011, 2021.

[11] S. Dutta *et al.*, "Highly scaled ruthenium interconnects," *IEEE Electron Device Letters,* vol. 38, no. 7, pp. 949-951, 2017.

[12] O. V. Pedreira *et al.*, "Reliability study on cobalt and ruthenium as alternative metals for advanced interconnects," in *2017 IEEE International Reliability Physics Symposium (IRPS)*, 2017, pp. 6B-2.1-6B-2.8: IEEE.

[13] T. Nogami, "Overview of interconnect technology for 7nm node and beyond-New materials and technologies to extend Cu and to enable alternative conductors," in *2019 Electron Devices Technology and Manufacturing Conference (EDTM)*, 2019, pp. 38-40: IEEE.

[14] Z. Pei, F. Catthoor, Z. Tokei, and C. Pan, "Beyond-Cu Intermediate-Length Interconnect Exploration for SRAM Application," *IEEE Transactions on Nanotechnology*, 2022.

[15] A. B. Kahng, J. Lienig, I. L. Markov, and J. Hu, *VLSI Physical Design: From Graph Partitioning to Timing Closure*. Springer Publishing Company, Incorporated, 2011.

[16] S. Y. Sherazi *et al.*, "Standard-cell design architecture options below 5nm node: The ultimate scaling of FinFET and Nanosheet," in *Design-Process-Technology Co-optimization for Manufacturability XIII*, 2019, vol. 10962, p. 1096202: SPIE.

[17] R. Balasubramonian, A. B. Kahng, N. Muralimanohar, A. Shafiee, and V. Srinivas, "CACTI 7: New tools for interconnect exploration in innovative off-chip memories," *ACM Transactions on Architecture and Code Optimization (TACO),* vol. 14, no. 2, pp. 1-25, 2017.

[18] HSPICE, *Synopsys, Mountain View, CA, USA*, 2022.

[19] Spectre, "Cadence," *San Jose, CA, USA*, 2022.

[20] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, "CACTI 5.1," Technical Report HPL-2008-20, HP Labs2008.

[21] W. Gomes *et al.*, "Ponte Vecchio: A Multi-Tile 3D Stacked Processor for Exascale Computing," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, 2022, vol. 65, pp. 42-44: IEEE.

[22] R. M. Rao *et al.*, "POWER10™: A 16-Core SMT8 Server Processor With 2TB/s Off-Chip Bandwidth in 7nm Technology," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, 2022, vol. 65, pp. 48-50: IEEE.

[23] A. Nayak *et al.*, "A 5nm 3.4 GHz Tri-Gear ARMv9 CPU Subsystem in a Fully Integrated 5G Flagship Mobile SoC," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, 2022, vol. 65, pp. 50-52: IEEE.

[24] S. Achra *et al.*, "Characterization of Interface Interactions between Graphene and Ruthenium," presented at the IEEE International Interconnect Technology Conference (IITC), San Jose, California, USA, 2020.

[25] S. Achra *et al.*, "Graphene-Ruthenium hybrid interconnects," presented at the IEEE International Interconnect Technology Conference (IITC), Brussels, Belgium, 2019.

[26] X. Zhang *et al.*, "Ruthenium interconnect resistivity and reliability at 48 nm pitch," in *2016 IEEE International Interconnect Technology Conference/Advanced Metallization Conference (IITC/AMC)*, 2016, pp. 31-33: IEEE.

[27] C. Pan and A. Naeemi, "A Proposal for a Novel Hybrid Interconnect Technology for the End of Roadmap," *Electron Device Letters, IEEE,* vol. 35, no. 2, pp. 250-252, 2014.

[28] H. C. Lee *et al.*, "Toward near-bulk resistivity of Cu for next-generation nano-interconnects: Graphene-coated Cu," *Carbon,* vol. 149, pp. 656-663, 2019.

[29] T. Yu, E.-K. Lee, B. Briggs, B. Nagabhirava, and B. Yu, "Bilayer graphene/copper hybrid on-chip interconnect: A reliability study," *IEEE transactions on nanotechnology,* vol. 10, no. 4, pp. 710-714, 2010.

[30] W. S. Leong, H. Gong, and J. T. Thong, "Low-contact-resistance graphene devices with nickel-etched-graphene contacts," *ACS nano,* vol. 8, no. 1, pp. 994-1001, 2014.

[31] I. Ciofi *et al.*, "Impact of wire geometry on interconnect RC and circuit delay," *IEEE Transactions on Electron Devices,* vol. 63, no. 6, pp. 2488-2496, 2016.

[32] A. Contino *et al.*, "Circuit Delay and Power Benchmark of Graphene against Cu Interconnects," presented at the IEEE International Interconnect Technology Conference (IITC), Brussels, Belgium, 2019.

[33] S. Datta, *Quantum transport: atom to transistor*. Cambridge University Press, 2005.

[34] K. I. Bolotin *et al.*, "Ultrahigh electron mobility in suspended graphene," *Solid State Communications,* vol. 146, no. 9, pp. 351-355, 2008.

[35] Raphael, *Synopsys, Mountain View, CA, USA*, 2022.