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

Zhenlin Pei†
Department of Electrical Engineering, The University of Texas at Arlington
zhenlin.pei@mavs.uta.edu

Mahta Mayahinia
Department of Computer Science, Karlsruhe Institute of Technology
mahta.mayahinia@kit.edu

Hsiao-Hsuan Liu
IMEC and Department of Electrical Engineering, Katholieke Universiteit Leuven
Samantha.Liu@imec.be

Mehdi Tahoori
Department of Computer Science, Karlsruhe Institute of Technology
mehdi.tahoori@kit.edu

Francky Catthoor
IMEC and Department of Electrical Engineering, Katholieke Universiteit Leuven
Francky.Catthoor@imec.be

Zsolt Tokei
IMEC
Zsolt.Tokei@imec.be

Chenyun Pan
Department of Electrical Engineering, The University of Texas at Arlington
chenyun.pan@uta.edu

ABSTRACT

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 to minimize the H-Tree delay and energy overheads based on an efficient interconnect technology/memory co-design framework for nonuniform workloads. Various array- and interconnect-level design parameters are co-designed for optimal performance using three emerging interconnect materials with a realistic cell library.

CCS CONCEPTS

• Hardware→Integrated circuits→Interconnect; • Hardware→ Integrated circuits→Semiconductor memory→Static memory.

KEYWORDS

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

ACM Reference Format:

Zhenlin Pei, Mahta Mayahinia, Hsiao-Hsuan Liu, Mehdi Tahoori, Francky Catthoor, Zsolt Tokei, and Chenyun Pan. 2023. Technology/Memory Co-Design and Co-Optimization Using E-Tree Interconnect. In *Proceedings of the Great Lakes Symposium on VLSI 2023 (GLSVLSI'23), June 5-7, 2023, Knoxville, TN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3583781.3590311

1 INTRODUCTION

SRAM is one of the major components in on-chip VLSI systems [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 wire delay and energy overheads, large research efforts have been performed to address interconnect challenges, such as 3D integration technology [7-9]. On the material side, some existing work has investigated beyond-Cu interconnects for the SRAM application [10], showing that the cache-level delay and energy are mainly dominated by H-Tree interconnects. The traditional H-Tree provides minimal 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 [11]. 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 wire length. The cell closer to the root pin will achieve a smaller delay and lower energy dissipation due to the shorter wire. 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 optimal cache performance. In addition, we will study center-pin access to further reduce the wire length. The corresponding logic cores 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 [12]. 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.

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 performance metrics.
3. We investigate different access pin options, including side-pin and center-pin technologies, to co-optimize with emerging interconnect technologies.

GLSVLSI '23, June 5–7, 2023, Knoxville, TN, USA

ACM 979-8-4007-0125-2/23/06...
https://doi.org/10.1145/3583781.3590311

4. Four interconnect material options are benchmarked to understand the true advantages of graphene-based interconnects on cache-level performance.

2 MODELING APPROACHES

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 wire 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 show the upper bound of the potential benefits of the proposed E-Tree interconnect network.

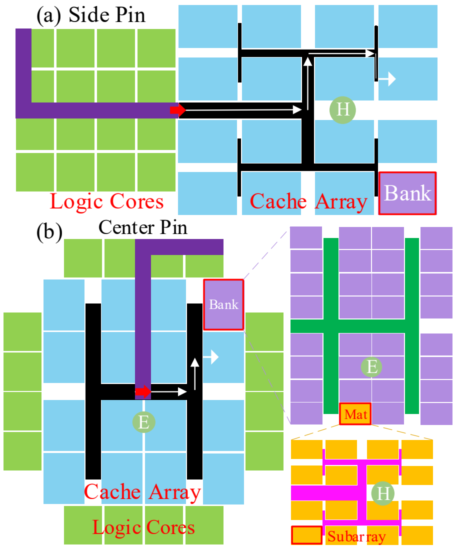

Figure 1: Schematic of cache using (a) traditional H-Tree with side-pin access and (b) proposed E-Tree with center-pin access. The arrows in red indicate the root pin locations.

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 the subarray:

$$P_{subarray\_i} \propto \frac{1}{L^{\alpha}_{subarray\_i}} \tag{1}$$

$$\sum_{i=1}^{all} P_{subarray\_i} = 1 \tag{2}$$

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

where α is the cache access probability factor, $P_{subarray\_i}$ and $L_{subarray\_i}$ are the access probability and wire length from the root 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-pin 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 wire length, meaning that the interconnects connect from the corner of the logic cores to the root pin of the cache array, as the purple lines shown in Figure 1.

2.2 Cache Array and Subarray Modeling

CACTI, one well-known and open-source simulator, is adopted to optimize the SRAM cache [13]. CACTI sweeps the cache organization parameters to get optimal parameters for the target defined by the user. 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. 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 materials. Extensive electrical-level simulations have been performed to validate the accuracy of the compact model.

2.3 Interconnect Materials and Modelings

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, 14-20]. For the inter-array interconnects, such as logic cores-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, 10, 13]. Device-level and interconnect parameters are extracted using Synopsys HSPICE and RAPHAEL.

3 SIMULATION RESULTS

In this section, we will perform the interconnect/cache co-design based on different workload assumptions for the proposed E-Tree interconnect with side-pin and center-pin technologies. Four material options 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, interconnect, and material design parameters and their default value used in the 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
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

To better analyze the cache performance, we first investigate the impact of the E-Tree network on the wire length and access probability for each bank, mat, and subarray. Figure 2 shows the wire length and access probability to each bank for the cache using side-pin and center-pin access. The access probability to a bank is the sum of the access probability to subarrays in this bank. One can observe that the bank close to the input pin (red arrow) has a shorter wire length and higher access probability.

For the cache using side-pin access, Figure 3 (a) and (b) show the probability of access and the number of interconnects to each subarray under different lengths of interconnects, respectively. The average wire 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 wire length from the cache using the E-tree with side-pin access, the one using the center-pin access shown in Figure 3 (c) and (d) is shorter due to the closer distance to the pin.

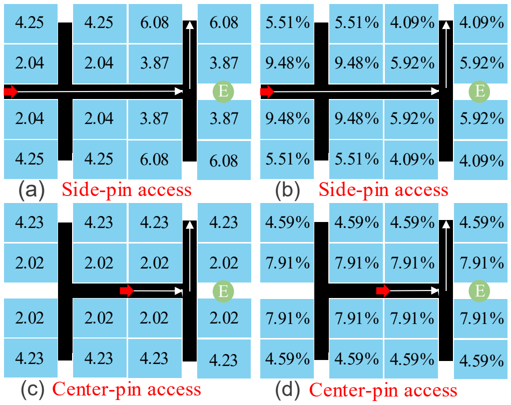

Figure 2: (a)(c) Wire length in (mm) from the root pin to the bank and (b)(d) access probability. The side-pin access is for (a)(b) and the center-pin access is for (c)(d). The arrows in red indicate the root pin locations.

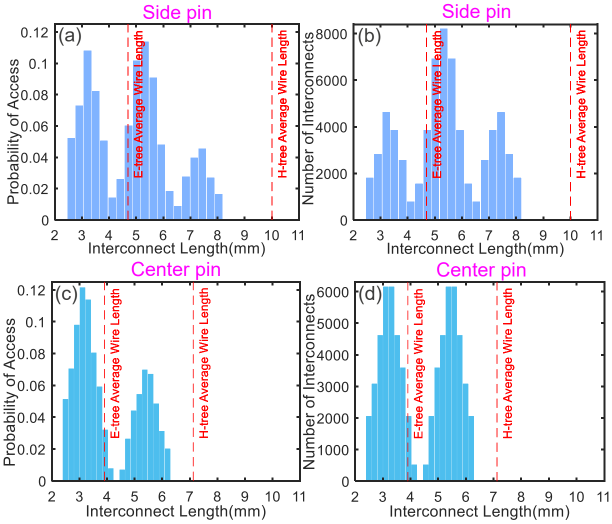

Figure 3: (a)(c) Probability of access and (b)(d) the number of interconnects versus wire length from different E-Tree technologies under the cache size of 128MB. The side-pin access is for (a)(b) and the center-pin access is for (c)(d).

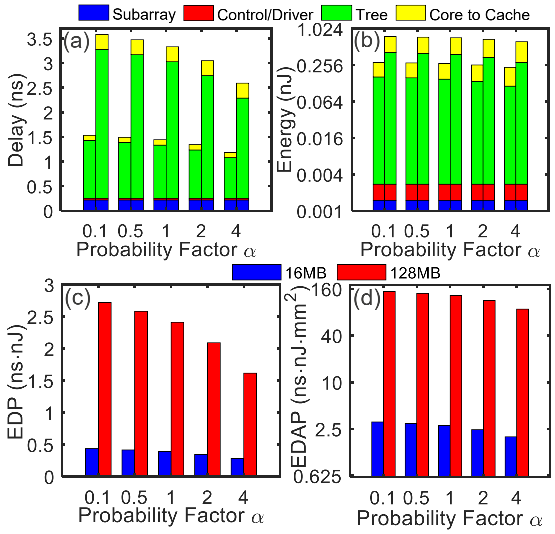

Figure 4: (a) Delay, (b) energy, (c) EDP, and (d) EDAP versus probability factor α for E-Tree with side-pin access in thick graphene. For each probability factor α, the left and right bars are for the cache size of 16MB and 128MB, respectively. The delay of cache using H-Tree is 2.47ns under 16MB.

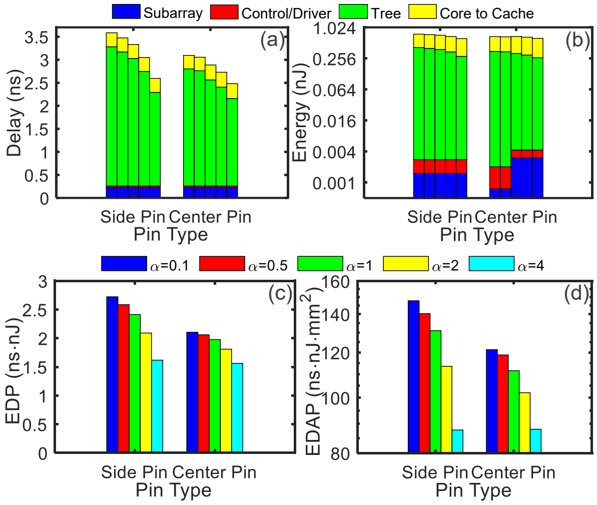

Figure 5: (a) Delay, (b) energy, (c) EDP, and (d) EDAP versus E-Tree interconnect technology option in thick graphene. For side-pin and center-pin access, the bars from left to right are probability factor α of 0.1, 0.5, 1, 2, and 4.

3.2 Impact of Workload on Cache Performance

Based on the average wire 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. Figure 4 (a) and (b) show the breakdown bar charts of delay and energy for different probability factors α under side-pin access with thick graphene. 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 relatively 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. To take delay, energy, and area into account, Figure 4 (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.

3.3 Impact of Interconnect Access Pin Types

To quantify the potential benefits of the proposed center-pin technology, Figure 5 shows various metrics versus two pin types under the cache size of 128MB. The cache using the E-Tree with center-pin access outperforms the side-pin counterparts because the first critical interconnect segment length in the array for the side-pin access is large, leading to a significant average wire length and delay overhead. To take delay, energy, and area into account, Figure 5 (c) and (d) show the EDP/EDAP versus the interconnect technology option for different cache access probability factors α.

3.4 Comparisons of H-Tree and E-Tree Using Various Interconnect Materials

To benchmark different interconnect technology options, Figure 6 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 58% and 67% reduction in EDP and EDAP can be observed compared to the traditional H-Tree counterparts, as shown in Figure 6 (b)(d).

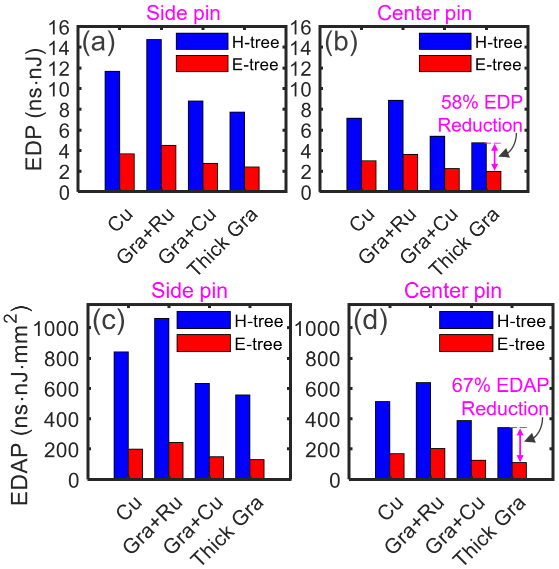

Figure 6: (a)(b) EDP and (c)(d) EDAP versus the interconnect material for H-Tree and E-Tree with different interconnect technology options in optimal interconnect width and aspect ratio under the cache size of 128MB. The side-pin access is for (a)(c) and the center-pin access is for (b)(d).

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 counterparts. The SRAM cache system using E-Tree with center-pin access and thick graphene interconnect provides the best performance, where up to 58% and 67% reduction in EDP and EDAP can be observed compared to the thick graphene counterparts in the H-Tree network.

ACKNOWLEDGEMENTS

This work was partially funded by IMEC, the Department of Energy (DoE) under Award DE-SC0022881, and in part by the National Science Foundation (NSF) under Grant CCF-2219753.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] C. Pan and A. Naeemi, "A paradigm shift in local interconnect technology design in the era of nanoscale multigate and gate-all-around devices," *Electron Device Letters, IEEE,* vol. 36, no. 3, pp. 274-276, 2015.