Ultra-Scaled E-Tree-Based SRAM Design and Optimization With Interconnect Focus

Zhenlin Pei, *Graduate Student Member, IEEE*, Hsiao-Hsuan Liu, *Graduate Student Member, IEEE*, Mahta Mayahinia, *Graduate Student Member, IEEE*, Mehdi B. Tahoori, *Fellow, IEEE*, Francky Catthoor, *Fellow, IEEE*, Zsolt Tőkei, *Member, IEEE*, Dawit Burusie Abdi, *Member, IEEE*, James Myers, *Member, IEEE*, and Chenyun Pan, *Senior Member, IEEE*

***Abstract*— SRAM performance is highly dominated by interconnects as technology scales down because of the significant parasitic resistance and capacitance in the interconnect. This paper introduces a framework for the co-design of technology, interconnect, and cache memory with tag array overhead, to optimize the performance of cache memory using a variety of emerging interconnect technologies. In addition, we introduce an innovative E-Tree interconnect aimed at further decreasing the average interconnect length with the consideration of realistic workloads and benchmark against its traditional H-Tree counterparts in terms of various performance metrics, such as energy-delay-area product (EDAP) or energy-delay product (EDP) in the SRAM cache memory system. A comprehensive investigation of design space is conducted, employing realistic, deeply scaled subarray designs across a range of cutting-edge technology nodes. Furthermore, the case study examines various cache memory system design parameters to assess the true potential of emerging interconnect technologies in achieving optimal performance at the cache memory system.**

***Index Terms*— Ultra-scaled SRAM design, E-tree, workload, graphene, benchmarking, technology/memory co-design.**

I. Introduction

SRAM is a crucial component in modern processors, occupying a significant portion of the chip area, especially for artificial intelligence (AI)/deep learning algorithms that contain large amounts of stored weights, which are best stored in on-chip SRAM for optimal performance and energy efficiency [1], [2]. SRAM's compatibility with standard core processes of CMOS, higher density compared to flip-flops, and seamless embedding on the logic die, make it a prevalent choice in applications ranging from register files to caches [3]. The SRAM cache array comprises bitlines (BLs), wordlines (WLs), and H-Trees, covering a broad spectrum of interconnect lengths and widths across global, intermediate, and local levels, leading to a dominant impact on the performance in the cache system [4]. The development of emerging interconnect technologies, such as graphene, provides new opportunities as the potential replacement of traditional copper (Cu)-based interconnects to further reduce the delay and energy overheads [5], [6].

Previous research has explored emerging interconnect options for SRAM applications using both academic and industry-standard cell libraries [4], [7]. However, the benefits of emerging interconnects for the advanced technology nodes have not been studied based on intra-SRAM microarchitecture exploration. Given the complex relationship among transistors, devices, and interconnects, it is imperative to conduct a thorough investigation by comparing the advanced, realistic, industry-standard SRAM cell libraries and subarray design under microarchitecture exploration for the cache memory system.

In terms of the interconnect performance overhead, existing work has shown that the H-Tree interconnects primarily influence the energy and delay at the cache level due to the large interconnect parasitic resistance and capacitance [4], [7]. Although the H-tree exhibits minimal skew and demonstrates robust performance in the presence of variations with simple control logic [8], accessing the instance(s) that is (are) adjacent to the cache array root pin results in a delay that is commensurate with the delay of accessing the most distant instance in the SRAM cache array due to its symmetry structure. To enhance the SRAM cache memory system performance, it is crucial to re-design the tree interconnect architecture/technology while considering the distance between the data locations and the cache array root pin under various realistic workload scenarios.

To further reduce the delay overhead, repeater insertion is widely used to strategically place inverters or buffers along the timing path or interconnect [4], [7]. However, most existing work sets a fixed delay overhead target without careful optimization and consideration of the trade-off between energy and delay. The significant impact of lengthy tree interconnects on energy and delay makes it critical to redesign the

Manuscript received 21 May 2024; revised 5 July 2024; accepted 29 July 2024. This work was supported in part by the Interuniversity Microelectronics Centre (IMEC), in part by the Advanced Scientific Computing Research (ASCR) Program of U.S. Department of Energy (DOE) under Award DE-SC0022881, and in part by the National Science Foundation (NSF) under Grant CCF-2219753. This article was recommended by Associate Editor X. Fong. *(Corresponding author: Zhenlin Pei.)*

Zhenlin Pei and Chenyun Pan are with the Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX 76010 USA (e-mail: zhenlin.pei@mavs.uta.edu).

Hsiao-Hsuan Liu and Francky Catthoor are with imec, 3001 Leuven, Belgium, and also with the Department of Electrical Engineering, KU Leuven, 3000 Leuven, Belgium.

Mahta Mayahinia and Mehdi B. Tahoori are with Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany.

Zsolt Tőkei, Dawit Burusie Abdi, and James Myers are with imec, 3001 Leuven, Belgium.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2024.3438164.

Digital Object Identifier 10.1109/TCSI.2024.3438164

interconnects with optimized repeater size and spacing to obtain the true benefit of the device and interconnect for the SRAM cache-level performance. Other than repeater insertion, cache design parameters, such as subarray size/design, bank organizations, and array tree interconnect level are important or necessary features to optimize the performance, including area efficiency and energy consumption of a cache memory system. Optimizing these parameters requires trade-offs, and the selection of values depends on the specific requirements/constraints imposed by the target algorithms, or applications with industry-standard cell libraries/subarray, which are neglected by existing works [9], [10], [11], [12].

To reduce the delay and energy for accessing memory cells that are closer to the cache array root pin, this paper will present an E-Tree network design technology option with the objective of reducing the average length of interconnects based on the real workload. Then, we will examine the potential for center-pin technology to further decrease the average length of array tree interconnect, which can be complemented with 3D stacking [13]. The proposed E-Tree interconnect design opens up new avenues for optimizing SRAM cache systems with tag array overhead, allowing frequently accessed data to be relocated closer to the input pins at the array and bank levels. Because workload analysis is essential for optimizing VLSI designs to meet performance, power, and area constraints, we will include both synthetic and realistic assumptions of workload and will quantify the impacts of these assumptions on optimal SRAM cache memory design and determine the true benefits of E-Tree in comparison to traditional H-Tree counterparts. For center-pin technology and side-pin technology options, we also investigate different cache tag array organizations or placements, whose impact on the overall cache system performance is quantified.

In addition, a novel memory/technology co-design framework will be developed to effectively tackle cache-level challenges inherent in SRAM, including large resistance of back-end-of-line (BEOL) metal, and cache and cell size inefficiencies at multiple cutting-edge technology nodes. We adopt IMEC's standard high-density (HD) SRAM design in 14-, 10-, 5-, and 3-Å-compatible technology nodes [14]. Furthermore, dedicated cache subarrays are incorporated, whose structure consists of the sense amplifier (SA), column multiplexer, bitcell array, BLs, WLs, and write driver (WD). Moreover, we will investigate optimal repeater insertion and delay overhead to balance delay and energy for cache-level performance and examine the impact of various key/essential design parameters, including the subarray size/design, the number of vertical banks, and the tree metal level in the cache.

It should be noted that the primary focus of this paper is on the LLC. The main reason for this focus is that the delay and energy impact of tree interconnects are particularly significant for large LLCs. To reduce the average tree length and its associated latency and energy, we propose the E-Tree. However, for the L1 and L2 cache, the energy and latency are more significantly influenced by the SRAM subarray, which limits the benefits of the proposed E-tree design, particularly for the L1 cache.

The principal contributions of the study are presented below.

- A framework for co-designing, the cache memory system with tag array overhead, interconnect, and technology node in microarchitecture exploration is developed. The framework incorporates validated experimental subarray design under a variety of ultra-scaled advanced device technology nodes.
- We propose and quantify the benefit/constraint of E-Tree design technology with real workload memory allocation to improve the data average access delay and energy efficiency.
- The cache array E-/H-Tree is redesigned by placing repeaters strategically along the path in a range of delay overheads compared to optimal delay to obtain the true benefit of the device and trees for the SRAM cache system performance.
- We investigate the impact of cache design parameters, such as the subarray size/design, the number of vertical banks, and the array E-Tree interconnect level on cache performance.

II. Modeling Approaches

A. E-Tree Interconnect Technology Design

To improve energy consumption and delay associated with lengthy H-Tree interconnects at the cache array level, an E-Tree technology option is proposed by extending our preliminary work [15]. It should be noted that the most recent and important upgrade/update of the Cacti-based framework incorporates the overhead associated with the tag array within the cache memory system, including the tag array within the data array for center-pin technology, as depicted in Fig. 1 (b)∼(e). The impact of distinct tag array overhead on cache performance will be examined. To assess the advantages and limitations of the proposed E-Tree and center-pin technology, a comprehensive investigation and comparison of the cache system performance with tag array overhead under various degrees of non-uniform workloads will be conducted, including both realistic and synthetic workloads. Furthermore, the framework has been developed for the co-design of the interconnect and ultra-scaled cutting-edge technology node in the cache memory systems, with consideration of tag array overhead in microarchitecture exploration. This includes experimental subarray designs using various ultra-scaled cutting-edge technology nodes from IMEC, such as A3, A5, A10, and A14.

Fig. 1 (a) illustrates the cache in conventional H-Tree interconnect with side-pin technology, utilizing H-Tree interconnects across all hierarchy levels, including array, bank, and mat. Fig. 1 (b)∼(e) depicts our proposed cache memory system, which incorporates the H-Tree or E-Tree interconnects with center-pin technology for bank-level/array-level interconnections under the worst case, namely, horizontal or vertical direction timing path which inputs/outputs tag/data arrays. Horizontal wires entering each hierarchy are subdivided into interconnects in the vertical direction that are shared with every two adjacent mats or banks. The proposed CACTI-based framework for cache system signal flow includes a loop path through the logic cores, the tag array, and the data array as

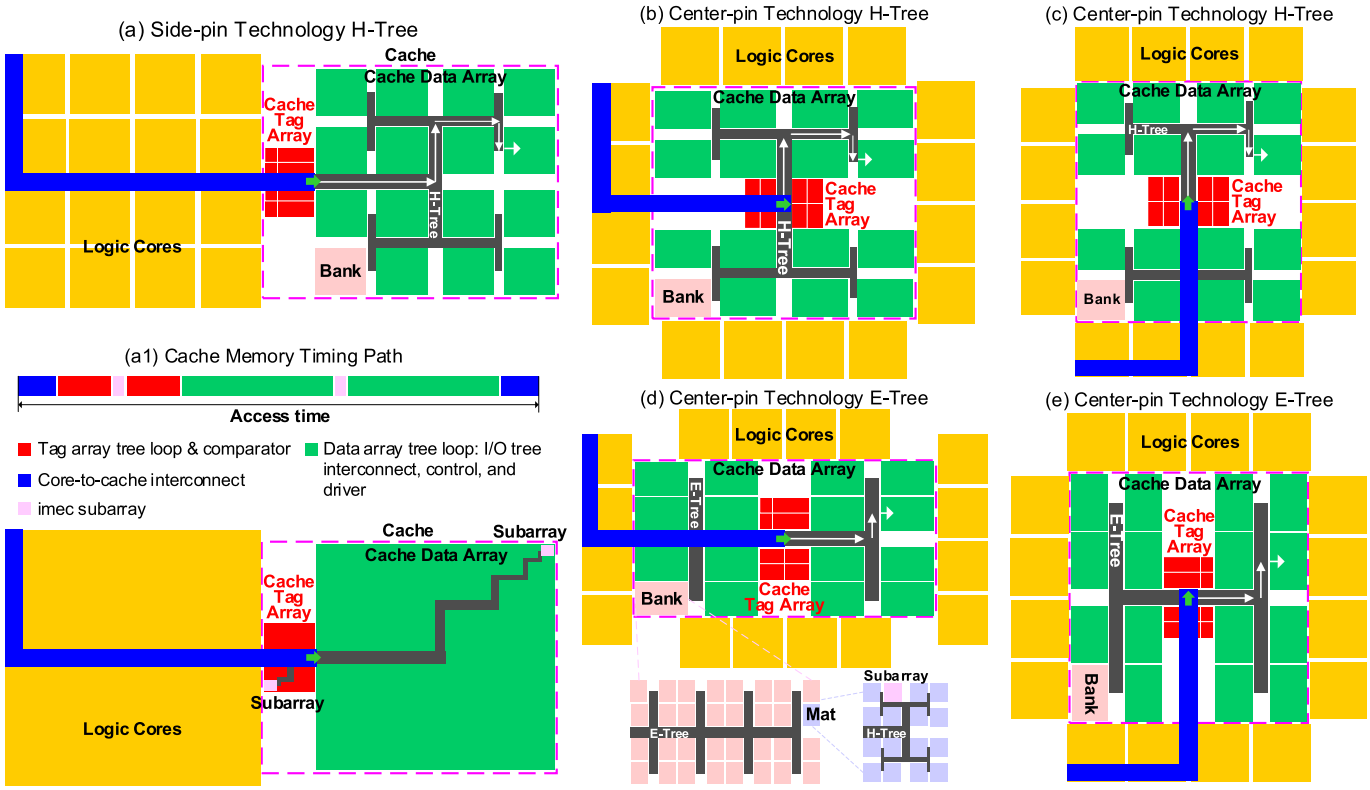


Fig. 1. Schematic of the SRAM cache memory system utilizing (a) conventional H-Tree interconnects with side-pin technology. Traditional H-Tree interconnects with the center-pin technology option under the worst case of (b) horizontal direction or (c) vertical direction timing path which inputs and outputs the data and tag arrays. The proposed E-Tree interconnects with the center-pin technology option under the worst case of (d) horizontal direction or (e) vertical direction timing path which inputs and outputs the data and tag arrays. (a1) shows the general timing path of the SRAM cache memory system through the logic cores (block in yellow), the cache tag array (block in red), and the cache data array (block in green), i.e., access time in the cache access mode of sequential. The critical timing path is based on the farthest subarrays in the tag and data arrays. The green arrows denote the locations of the root pins within the cache array. The width (1um) of the core-to-cache (core-cache) interconnect in traditional Cu (blue lines) is larger than that of the array-level interconnect (black or gray lines). The tag array tree interconnect design is identical to its counterpart in the data array within the cache. The tag array tree interconnect is not displayed here due to space limitations. The impact of distinct tag array organizations on the performance of the cache system is investigated. The cache area, represented by the pink dashed box, is equivalent to the area occupied by the logic cores.

depicted in Fig. 1 (a1), where tree interconnects in tag and data arrays dominate the total access time/delay.

The area and energy consumption of the SRAM cache memory system are also determined by logic cores, core-to-cache (core-cache) interconnect, and arrays. We assume that (1) the access delay to the tag array is based on the scenario of the worst-case shown in the schematic of Fig. 1, even if there are cores closer to the tag array, (2) the core-to-tag interconnects are on the top level with the widest width, so the number of buffer utilization is very low, leaving enough room for the tag array below the blue wires. The primary benefit of E-Tree interconnects lies in their ability to minimize interconnection lengths, especially when accessing data located nearer to the cache array root pin (as denoted by the green arrows) or bank inputs. Due to the asymmetric routing, additional logic circuits for timing control are involved, which are not considered in this study. We plan to delve deeper into this aspect and assess its architectural-level implications in our future research, including demultiplexers for the logic cores. The findings outlined in Section III will demonstrate the maximum potential advantages of the E-Tree interconnections.

B. Workload Modeling

The average length (in the space/timing) of the proposed E-Tree design/interconnect depends on the access probability which depends on the workload. We will consider two workload scenarios: synthetic and realistic workloads. The synthetic workload aims to capture the impact of non-uniform workload on E-Tree-based SRAM performance efficiently. It assumes that there exists a negative correlation between the probability of access and the distance from the cache array root pin to the subarray, namely quantified by a probability factor α, as shown in equation (1). If the probability factor α is close to 0, it indicates a uniform workload. As α increases, there is more access to data closer to the array root pin. Thus, sweeping α provides an efficient way to quantify the potential advantages of E-Tree. In equation (1), $P_{subarray\_i}$ and $L_{subarray\_i}$ are the access probability to subarray i and length of interconnect (in the space) to subarray i from the array root pin, respectively.

The real workload includes a list of benchmarks from SPEC CPU 2017 [16], which provides insights into how practical applications can benefit from our proposed E-Tree technology. As shown in equation (2), the access probability depends on the number of accesses to the banks based on real benchmarks,

TABLE I
NON-UNIFORM WORKLOAD FROM SPEC CPU 2017

Benchmark	Benchmark Program General Category
602.gcc	GNU C Language optimizing compiler
605.mcf	Combinatorial Optimization & Vehicle Scheduling
607.cactus	Physics: General/Numerical Relativity
623.xalan	XSLT processor for transforming XML to HTML
625.x264	Video compression
641.leela	AI: Go engine with Monte Carlo & selective tree search
649.fotonik	Computational Electromagnetics (CEM)
654.roms	Regional Ocean Modeling System

TABLE II
WORKLOAD SIMULATION ASSUMPTION PARAMETERS

Parameter	Value
CPU Model	X86 Out-of-Order CPU
CPU Clock Frequency	3 GHz
CPU Physical Address	64 bits
SRAM Cache Line Size	64 B
L1 Size/Associativity/Access(Cycle)/Bank	32kB/8/3/--
L2 Size/Associativity/Access(Cycle)/Bank	512kB/16/9/4
L3 Size/Associativity/Access(Cycle)/Bank	128MB/16/33/16

where $Access_{bank\_i}$ is the total access number to the bank i.

$$\text{Synthetic Workload:} \quad P_{subarray\_i} \propto \frac{1}{L^{\alpha}_{subarray\_i}} \tag{1}$$

$$\text{Realistic Workload:} \quad P_{subarray\_i} \propto \frac{Access_{bank\_i}}{\sum_{bank\_i}^{all} Access_{bank\_i}} \tag{2}$$

$$\sum_{i=1}^{all} P_{subarray\_i} = 1 \tag{3}$$

$$L_{average} = \sum_{i=1}^{all} \left(L_{subarray\_i} \cdot P_{subarray\_i} \right) \tag{4}$$

The proposed CACTI-based framework enables obtaining E-Tree wire distribution and access probability distribution. Realistic workloads of non-uniform are executed from SPEC CPU 2017, and a list of benchmarks is shown in Table I [16]. Note that the benchmark workload applications are based on a single core. L3 traces are collected, including the tick of the L3 access, address, status (hit or miss), type (read or write), and the data during the write operation. The traces for realistic workloads are produced through the gem5 simulator extended version, a widely accepted microarchitectural simulator [16], [17], [18], [19]. Realistic workload simulation assumption parameters are listed in Table II. We propose organizing the cache into multiple independent units, e.g., cache banks within the array. We assume that addresses are sequentially mapped to each bank to distribute access across multiple banks. It is crucial to distinguish between set associativity and multi-banking.

C. Enabling Methods

1) Subarray Models at Ultra-Scaled Technology Nodes: To efficiently and accurately analyze large cache module performance, several simulations were reported using Cadence Virtuoso/Spectre and Synopsys Hspice to extract latency and energy data for the standard HD subarray across various validated technology nodes, including 3-, 5-, 10-, and 14-Å-compatible technology, i.e., A3, A5, A10, and A14, respectively [14], [20], [21], [22]. Extensive electrical-level simulations have validated the accuracy of the model. The device models used in our experiments are sourced from the IMEC standard-cell library. For interconnect repeaters, we extract device parameters based on the classic delay equations of the repeater and interconnect, such as temperature-dependent leakage current, gate and drain capacitance, threshold voltage, and ON current, through simulations using Cadence Spectre [20].

The simulation is performed at room temperature, nominal V_{DD} of 0.7V, and typical-typical corner by Cadence Virtuoso/Spectre and Synopsys Hspice/QuickCap [20], [21], [22], [23]. The SRAM subarray circuit consists of timing/address control, column multiplexer, row decoder (RD), pre-charge circuitry, bitcell array, sense amplifiers (SAs), and write drivers (WDs). The size of bitcell array is 288 columns (WLs)×256 rows (BLs), namely 288C × 256R. The data on delay, energy, and area are achieved on the bitcell of worst-case, i.e., the farthest bitcell of SRAM from RD, SA, and WD. The SRAM operation varies based on the n-type- or p-type-pass-gate transistor, following mechanisms for BL charging and discharging [14], [24], [25].

2) Cache-Level Memory Models: We adopt and modify the open-source CACTI simulator, a widely recognized tool, to optimize the cache memory system [26], [27]. CACTI systematically adjusts SRAM cache memory organization parameters to achieve optimal metrics defined by the users, such as minimum energy-delay product (EDP) or EDAP. The cache access timing path involves E-/H-Tree input and output from both inside and outside the bank, including tag and data arrays, along with a timing path from the subarray described in Section II-A and Fig. 1. We assume the logic core area equals the total area of the cache, e.g., the width of cache times height of cache, as shown in Fig. 1. The original CACTI was validated through SPICE simulations and reliable data from classic commercial caches, such as L3 cache from Intel at the technology node of 65nm and SPARC L2 cache from Sun at the technology node of 90nm [26]. The validated cache model simulator allows users or researchers to explore various emerging technology/device/interconnect options and organization parameters in the early or initial stage of the SRAM cache memory system design accurately and efficiently.

To integrate the IMEC-designed subarray, key performance metrics in the original CACTI, such as area, energy, and delay, are incorporated based on true and reliable values extracted from the experimental simulations and data. The parameters such as the number of output sense amplifiers, columns and rows with each subarray, and column decoders adhere to values and guidelines provided by IMEC, influencing cache organization exploration. This study involves crucial trade-offs among interconnect parameters and the cache design, including optimal interconnect repeater insertion and delay overhead, subarray size, the number of vertical banks, as well as the comparison among a variety of realistic workloads from SPEC CPU 2017 benchmarks and subarray standard HD designs in various ultra-scaled technology

nodes, and all aim at optimizing SRAM at the cache-level performance metrics. By comparing a variety of interconnect parameters with Cu-based counterparts, we aim to provide valuable insights to SRAM cache system developers and material engineers, identifying the genuine benefits of showing potential graphene-based tree interconnects to implement energy-efficient computing memory systems.

3) Interconnect Modeling: We select four promising interconnect materials to evaluate their effects on the cache memory system performance based on the updated version of established interconnect modeling techniques [4], [7], [15], [28]. These options include (i) baseline Cu (copper), (ii) Graphene-capped Ru (Ruthenium) or Gra+Ru, (iii) Graphene-capped Cu (copper) or Gra+Cu, and (iv) thick graphene or Thick Gra [4], [5], [6], [7], [15], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. To model the energy or delay of inter-array and inter-subarray interconnects, including array E-/H-Tree interconnects and core-to-cache interconnects, we follow the original CACTI work to sweep repeater size and spacing to balance delay and energy [26]. Interconnect-level parameters are extracted using Cadence Spectre and Synopsys Raphael [20], [47]. The interconnect geometry is based on the interconnect design of IMEC at A14 and A10 nodes, including the interconnect pitch, width, spacing, and thickness.

In the case of a lengthy interconnect with repeater insertion, the single pole time constant circuit model for the interconnect, as illustrated in Fig. 2, can be expressed in equation (5) [48],

$$\tau = \frac{R_{o1}}{H}C_{pu} + 0.5R_{pu}C_{pu}l + R_{pu}HC_{g1} + \frac{1}{l}R_{o1}\left(C_{d1}+C_{g1}\right) \tag{5}$$

$$\frac{\partial \tau}{\partial l} = 0 \overset{solve\ l}{\Longrightarrow} l_{opt} = \sqrt{\frac{2R_{o1}\left(C_{d1}+C_{g1}\right)}{R_{pu}\cdot C_{pu}}} \tag{6}$$

$$\frac{\partial \tau}{\partial H} = 0 \overset{solve H}{\Longrightarrow} H_{opt} = \sqrt{\frac{R_{o1}\cdot C_{pu}}{R_{pu}\cdot C_{g1}}} \tag{7}$$

$$\begin{aligned} t_{c\_opt} &= R_oC_w + 0.5R_wC_w + R_wC_g + R_o\left(C_d+C_g\right) \\ &= R_{o1}C_{pu}\frac{l_{opt}}{H_{opt}} + \frac{1}{2}R_{pu}C_{pu}l_{opt}l_{opt} + R_{pu}l_{opt}C_{g1}H_{opt} \\ &\quad + R_{o1}\left(C_{d1}+C_{g1}\right) = R_{o1}\sqrt{2C_{g1}\left(C_{d1}+C_{g1}\right)} \\ &\quad + R_{o1}\left(C_{d1}+C_{g1}\right) + R_{o1}\sqrt{2C_{g1}\left(C_{d1}+C_{g1}\right)} \\ &\quad + R_{o1}\left(C_{d1}+C_{g1}\right) = 2R_{o1}\sqrt{2C_{g1}\left(C_{d1}+C_{g1}\right)} \\ &\quad + 2R_{o1}\left(C_{d1}+C_{g1}\right) = 2R_{o1}\sqrt{C_{d1}+C_{g1}} \\ &\quad \times\left(\sqrt{2C_{g1}}+\sqrt{C_{d1}+C_{g1}}\right) \end{aligned} \tag{8}$$

$$T_{pu} = \frac{0.693\cdot t_{c\_opt}}{l_{opt}} \tag{9}$$

where l and H are the repeater spacing and size, respectively. t_{c_opt} is the time constant of the interconnect segment, which is determined by the repeater parameters, with optimal repeater spacing l_{opt} and size H_{opt} normalized to the minimum value. The repeater parameters: R_{o1}, C_{d1}, and C_{g1} are the driving resistance, drain capacitance, and gate capacitance of the repeater utilizing minimum-sized FinFET with 1 fin. T_{pu} is the wire delay per unit length which is determined by resistance per unit length R_{pu}, capacitance per unit length C_{pu}, and repeater parameters. To minimize interconnect EDP, the repeater size and spacing are swept to identify the optimal repeater placement design, which allows for various delay overheads [48].

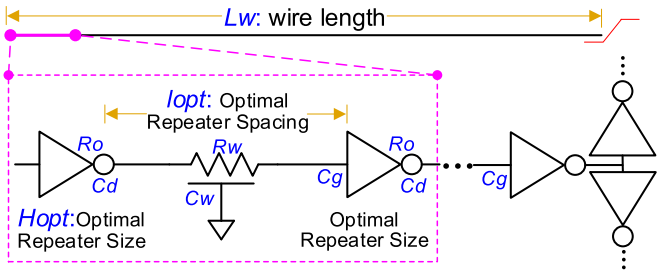


Fig. 2. RC-based circuit model of the interconnect with the optimal repeater size H_{opt} and spacing l_{opt}. R_o is the driving resistance of the optimal repeater, C_g and C_d are the gate and drain capacitance of the optimal repeater, respectively, R_w and C_w are the resistance and capacitance of the interconnect. The additional quantum resistance $R_{quantum}$ and contact resistance R_{con} are added to each side of the graphene-based wire.

III. Simulation Results

Building upon the approaches and modeling outlined in Section II, we investigate the performance of a cache memory system implemented with the standard HD subarray at various technology nodes, including A3, A5, A10, and A14, using E-Tree interconnects based on realistic workloads, whose benchmarks are listed in Table I. Optimal repeaters will be inserted with side-/center-pin technology options for four interconnect materials, i.e., baseline Cu, Gra+Ru, Gra+Cu, and Thick Gra. Three interconnect levels and different cache array organizations and sizes will be investigated in the case study, whose values are listed in Table III. Unless specified elsewhere, the default SRAM is direct mapped; the baseline tree interconnect level is 5; the repeater size and spacing are obtained based on a 30% delay overhead compared to the optimal repeater insertion in the default original CACTI configuration file. This tends to have smaller delays with energy-efficient computing [48]. The workload is uniformly distributed. The graphene's effective mean free path (MFP) is mainly influenced by various factors, including the roughness of its edge and the properties of the substrate material [4], [7], [15].

A. Workload-Dependent Performance Analysis for Cache With E-Tree Interconnects

To fully utilize the benefit of the proposed E-Tree interconnect, we will explore the impact of access probability and average tree length under synthetic and realistic workloads.

1) Synthetic Workload: Based on the workload modeling methodology and approach introduced in Section II-A and II-B, Fig. 3 illustrates the access probability and interconnect length to each bank from a cache that employs center-pin and side-pin technologies, assuming the access probability

TABLE III
CACTI FRAMEWORK AND SIMULATION PARAMETERS

Parameter	Value
Number of Banks	4,16,64
Cache Size (MB)	64~1024
Standard High-Density Subarray Size (kb)	16~64
High-Density Subarray Technology Nodes	A3, A5, A10, A14
Core-Cache Interconnect in Cu Aspect Ratio	0.1
Core-Cache Interconnect in Cu Width (μm)	1
Inter-Subarray Interconnect Pitch (nm) M3/5/6	28/28/80
Inter-Subarray Interconnect Width (nm) M3/5/6	16/16/40
Inter-Subarray Interconnect Height (nm) M3/5/6	49/28/80
Mean-Free-Path for 1μm Wide Graphene (nm)	460
Graphene Interconnect Contact Resistance (Ω·μm)	100

to each subarray is assumed inversely proportional to the length of tree interconnect, namely $\alpha = 1$. The probability of access to each bank is the cumulative probability of accessing subarrays within that bank. Notably, the banks near array input pins (indicated by green arrows) exhibit a larger access probability and shorter interconnect length.

For the SRAM cache memory system employing side-pin technology, Fig. 4 (a)(b) illustrate the number of interconnects and the probability of access at different interconnect lengths, respectively. The E-Tree features shorter interconnect average lengths compared to H-Tree. This is primarily because E-Tree utilizes shorter interconnects that access the subarray directly near the input pins across the bank and array hierarchy levels. Comparatively, the cache memory system with E-Tree design with center-pin technology, as demonstrated in Fig. 4 (c) and (d), has a smaller average interconnect tree length than its side-pin counterpart, thanks to the closer proximity to the input pins. The average length in the center-pin technology E-Tree is the shortest based on the specified workload assumption.

To quantify the proposed E-Tree interconnect advantage, we optimize the cache system performance by employing our co-design extended framework tailored for synthetic workloads outlined in Section II-B. Fig. 5 illustrates various metrics for both pin type options under various probability factors α for a 128MB cache. The SRAM cache memory system employing E-Tree design with center-pin technology option demonstrates superior performance compared to its side-pin counterparts. This is attributed to the larger length of the initial interconnect segment inherent in the side-pin technology configuration, resulting in a significant increase in delay overhead caused by a larger average interconnect length within the arrays.

In Fig. 5 (a)(b) and (c)(d), segmented bar charts illustrate the delay and energy of read and write per access for various probability factors α under side-/center-pin technologies employing traditional Cu and thick graphene interconnects, respectively. The primary factor contributing to the total delay/access time is the E-Tree interconnects within arrays, which are constrained by the narrow interconnect width within the cache array at the metal level of the intermediate. In general, tree interconnects using thick graphene provide a smaller delay compared to their Cu counterpart due to graphene's smaller Resistance Per Unit length (RPU) caused by its long MFP. The delay of core-cache interconnects in Cu remains relatively small as those interconnects are situated at the metal level of the global, featuring a larger interconnect width. But core-cache interconnects primarily dominate the overall energy consumption due to their long lengths. By varying assumptions regarding workload, both energy and delay exhibit a reduction in magnitude with an increase in the probability factor, as a consequence of the shorter average length (in the space/timing) of the E-Tree interconnect within the SRAM cache.

To consider area, energy, and delay comprehensively, Fig. 5 (e)(f) and (g)(h) depict the EDP of the SRAM cache memory system and cache-level energy-delay-area product (EDAP) against probability factors α in center-pin and side-pin technologies under Cu and thick graphene interconnects, respectively. The cache memory system using a center-pin technology with thick graphene interconnect offers the most optimal overall performance, due to its shorter interconnect length and smaller RPU. A larger probability factor is associated with a larger delay in the center-pin access E-Tree compared to the side-pin access counterpart. It is due to the increased overhead of the tag array in the center-pin access under the workload exhibiting higher non-uniformity, leading to a larger average length of the tree in the data array.

2) Realistic Workload: It is essential to quantify and validate the benefits of cache using the proposed E-Tree design based on realistic workloads.

To assess the influence of realistic workloads on the cache system performance, Fig. 6 illustrates the EDP, energy, and delay under center-pin and side-pin technologies for the baseline uniform workload and various non-uniform workloads from SPEC CPU 2017 benchmarks outlined in Section II-B. The L3 footprint of benchmark applications is different. The largest one is for the 649.fotonik workload whose size is smaller than our cache capacity of 128MB. The baseline workload assumes that the number of accesses to each bank is the same. A uniform workload, a specific type of synthetic workload with a probability factor of 0, serves as the baseline. The cache under non-uniform workloads provides more advantages compared to the uniform workload. This is because of the smaller average tree length caused by more frequent access to the data whose location is closer to the array root pin. Therefore, a more evenly distributed workload leads to a longer average interconnect tree length and a larger EDP. In a workload with a larger degree of non-uniformity, the delay of center-pin access is larger than the side-pin counterpart due to the larger overhead of the tag array.

The proposed E-Tree design trade-off is not straightforward, as the average length (in the space and timing) is contingent upon the access probability, which is influenced by various degrees of non-uniformity workloads, including synthetic and realistic workloads. It is anticipated that the overall interconnect overhead will be reduced due to the shorter average interconnect length, particularly in the context of non-uniform workloads. However, in the event of highly non-uniform workloads, the E-Tree with center-pin technology may encounter a larger delay in comparison to its side-pin counterpart, due to the increased tag array overhead, as depicted in Fig. 5 (a)(b) and Fig. 6 (a)(b). Consequently, the advantages and limitations of the proposed E-Tree technology must be meticulously

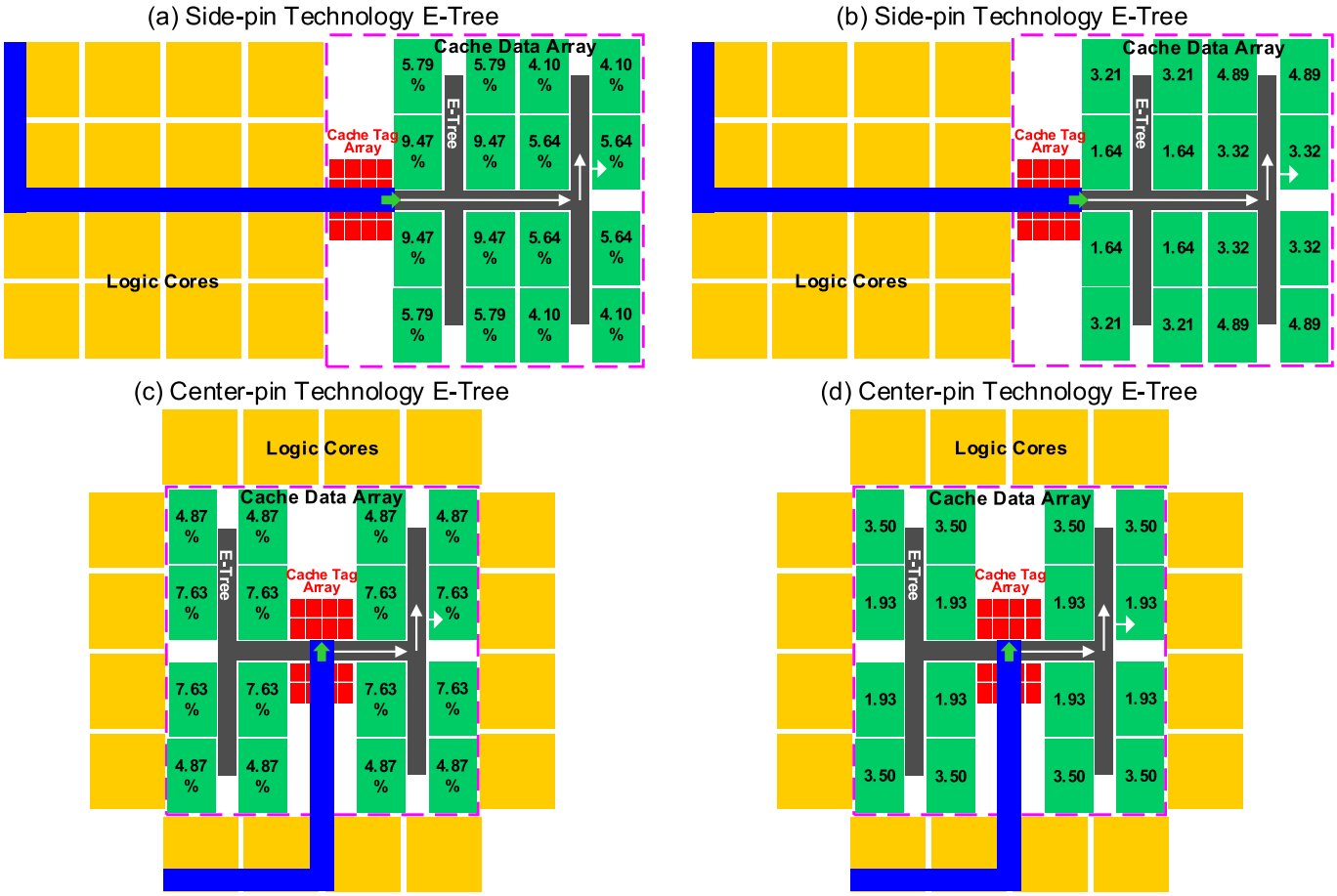


Fig. 3. (a)(c) Access probability of banks and (b)(d) interconnect length (unit: mm) to banks from the cache array root pin (green arrows) for a 128MB cache with 16 banks using the standard HD subarray of 288C × 256R implemented with E-Tree at A14 node. (a)(b) are for the side-pin technology option, and (c)(d) are for the center-pin technology option. Both options pertain to the E-Tree interconnect technology. Note that the probability factor α is 1. The probability of access and wire distributions of the cache memory system for E-Tree under the side-pin technology option and center-pin technology option are illustrated in Fig. 4.

examined in the context of diverse computing hardware and software scenarios.

> Insights from Results
> The cache utilizing the proposed E-Tree interconnect under non-uniform workloads provides more benefits to the cache compared to uniform workloads.

B. Scalability Analysis Using SRAM Subarray Under Various Technology Nodes

The technology scaling plays a pivotal role in enhancing the energy efficiency and work performance of the cache memory system. Here, we adopt the standard HD subarray design at various technology nodes down to the A3 node to investigate the scalability of the SRAM [14]. To continue scaling SRAM beyond 14-Å-compatible technology nodes, several options of devices have been investigated, including complementary field-effect transistor (CFET) and nanosheet (NS) [49].

Fig. 7 shows the cache performance under sequential or monolithic CFET with Fin-on-Fin or NS-on-NS SRAM designs in A5 and A3. Additionally, a six-transistor (6T) hybrid CFET SRAM can be designed with either an n-type or p-type pass-gate (PG) for A3 [14]. Fig. 7 shows the cache-level performance comparison using H-/E-Tree interconnects with traditional Cu, where the delay is mainly dependent on the Tree due to its large length and smaller wire width. The core-cache and tree interconnects dominate the overall write and read energy because of the large interconnect length and a large number of data bits. Overall, a smaller technology node helps to largely improve the energy and delay/access time due to a smaller interconnect length thanks to the smaller subarray area. To account for both delay and energy, Fig. 7 (e) and (f) show the EDP in relation to the tree interconnect type for subarray in various technology nodes under side-pin technology and center-pin technology, respectively. Under the consideration of cache area, the cache using SRAM subarray in small technology nodes provides a greater advantage in terms of the EDAP compared to counterparts with large technology nodes, as shown in Fig. 7 (g) and (h). In general, SRAM subarrays using small technology nodes are favored to reduce EDAP and EDP of the cache memory system due to the small area overhead. The reasons for the improvement of E-Tree with center-pin technology compared to side-pin

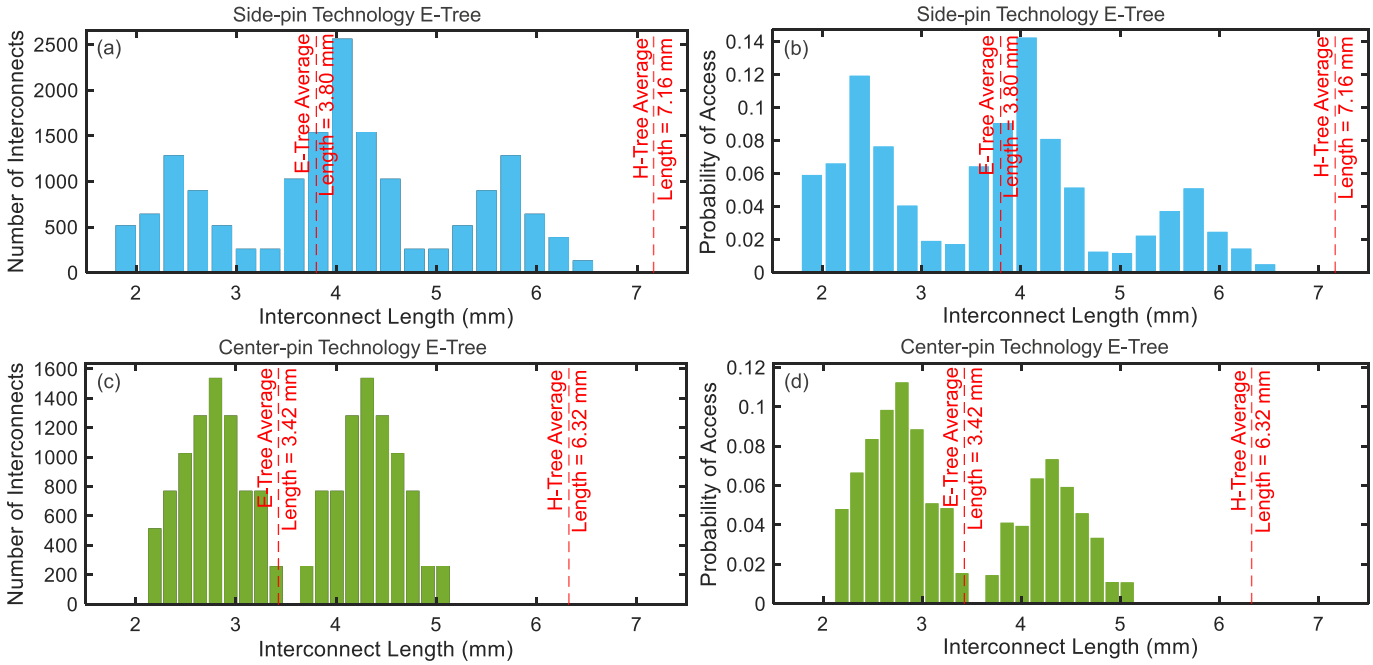


Fig. 4. (a)(c) Number of interconnects and (b)(d) probability of access vs. length of interconnect (unit: mm) for two E-Tree interconnect technologies in an SRAM cache memory system of 128MB cache with 16 banks using the standard HD subarray of 288C × 256R implemented with E-Tree at A14 node. (a)(b) are for the side-pin technology (bars in blue), and (c)(d) are for the center-pin technology (bars in green). Note that the probability factor α is 1. The probability of accessing and the interconnect length (unit: mm) to banks from the cache array root pin of the cache system for center-pin and side-pin technologies are depicted in Fig. 3.

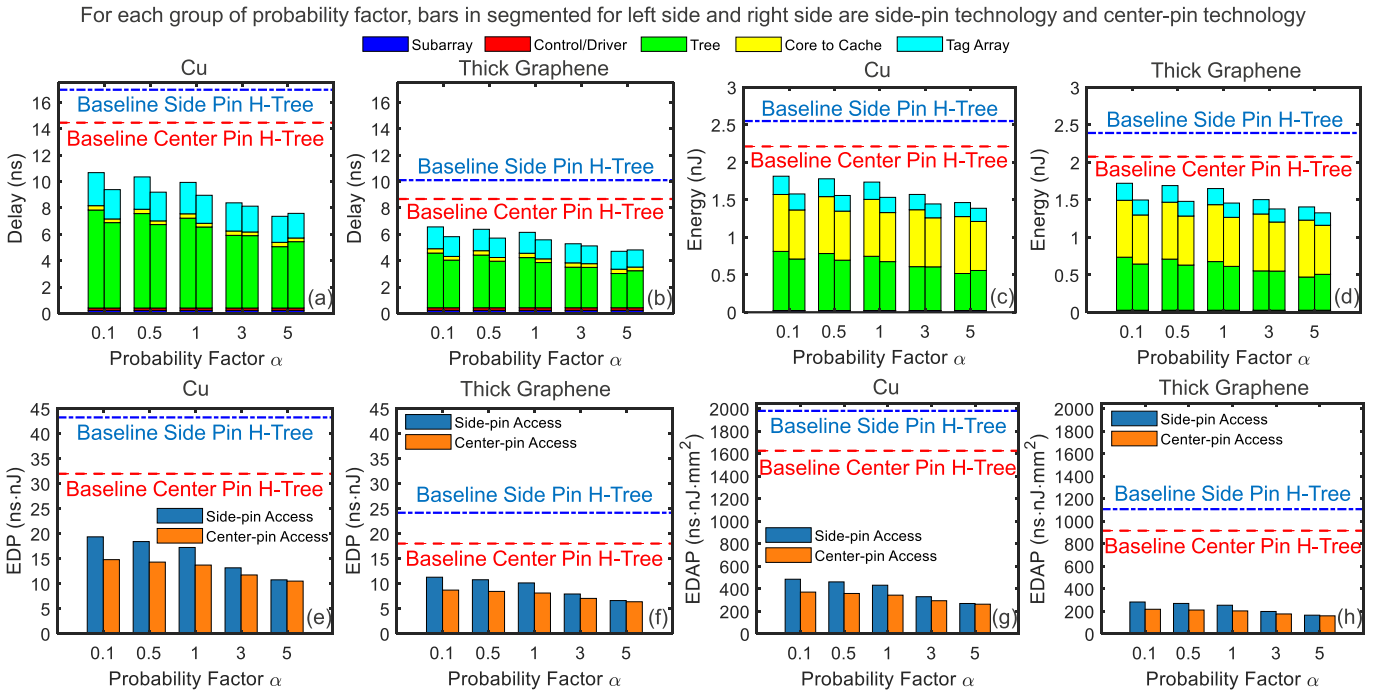


Fig. 5. (a) Delay, (c) read and write energy per access (power delay product) segmented bar chart, (e) energy-delay product (EDP) of the SRAM cache memory system, and (g) energy-delay-area product (EDAP) of cache memory system versus the proposed probability factor α for a 128MB cache with 16 banks using the standard HD subarray of 288C × 256R at A14 node using traditional Cu interconnects as the baseline for the interconnect material selection for the center-pin technology and side-pin technology, respectively. (b) Delay, (d) read and write energy, (f) EDP, and (h) EDAP of the system versus probability factor α for the cache array employing interconnects in thick graphene. In each group of probability factor α, the right side bar and left side bar in segmented represent the E-Tree under the center-pin technology and side-pin technology, respectively. The dashed lines show the performance of the cache under the baseline H-Tree interconnects with center-pin technology (dashed lines in red) and side-pin technology (dashed lines in blue), respectively. A probability factor of 0 means a uniform workload.

technology H-Tree have been described in the previous subsection III-A.1.

The proposed E-Tree network design technology introduces a novel approach to reducing delay and energy consumption

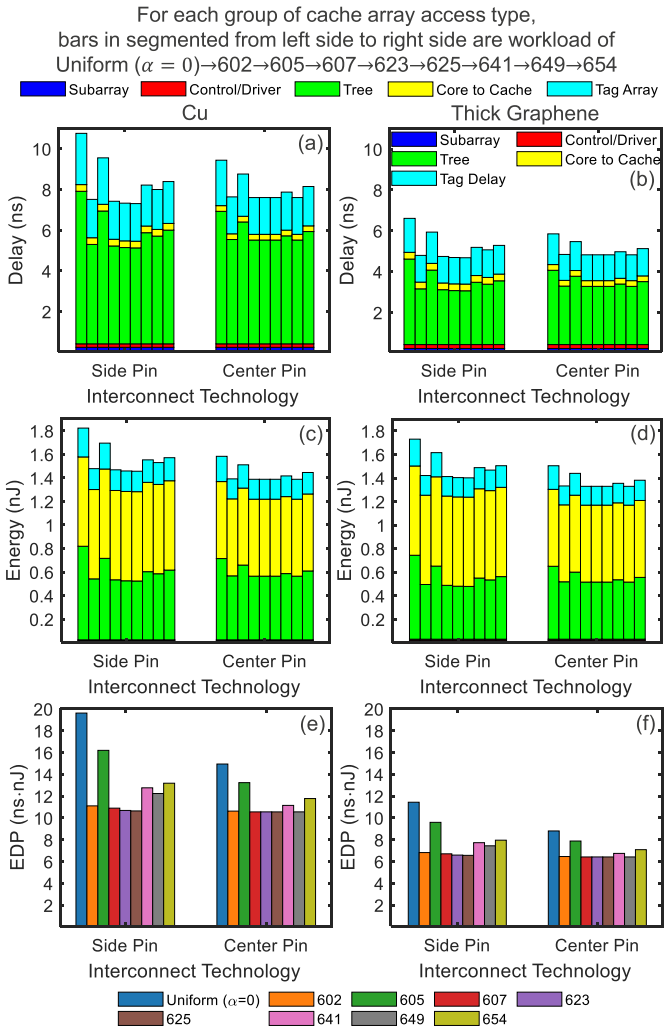


Fig. 6. (a) Delay, (c) read and write energy, and (e) EDP versus interconnect technology option at uniform and non-uniform workloads for a 128MB cache with 16 banks using the standard HD subarray of 288C × 256R implemented with E-Tree at A14 node using traditional Cu interconnects as baseline. (b) Delay, (d) read and write energy, and (f) EDP of SRAM cache memory versus interconnect technology at uniform/non-uniform workloads under thick graphene interconnect. A probability factor α of 0 means a uniform workload.

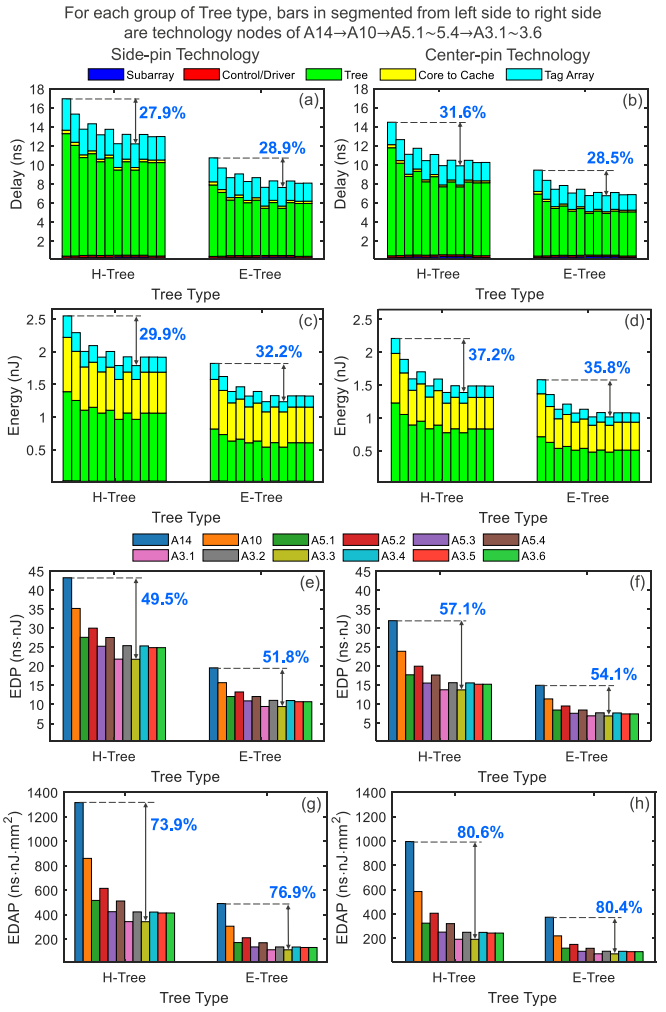


Fig. 7. (a) Delay, (c) read and write energy, (e) EDP, and (g) EDAP versus tree type under side-pin technology for a 128MB cache with 16 banks using the standard HD subarray with 288C × 256R at A14, A10, A5, and A3 technology nodes in Cu interconnect for the uniform workload. (b) Delay, (d) read and write energy (f) EDP, and (h) EDAP versus tree type for center-pin technology.

for memory cells situated in close proximity to the cache array root pin. The technology aims to minimize the average interconnect length, including the tag array overhead, in ultra-scaled cutting-edge technology nodes.

> Insights from Results
>
> To reduce cache-level overall EDAP and EDP while minimizing area overhead, utilizing SRAM subarrays with smaller technology nodes is critical.

C. Impact of Repeater Insertion Design on Cache Memory System Performance

As previously outlined in Section II-C.3, the inserted repeater size and spacing strongly affect the delay or energy per unit length of the array tree interconnect. To properly balance cache-level delay and energy, Fig. 8 shows the cache memory system performance for various delay overheads by relaxing repeater spacing and size. Four different interconnect materials are investigated and compared, i.e., baseline Cu, thick graphene, Gra+Cu, and Gra+Ru.

Under a large delay overhead target, fewer repeaters with smaller sizes will be inserted, which saves energy at the cost of a larger delay. Clear tradeoffs can be observed between delay and energy to minimize the SRAM cache memory system's overall EDP, as shown in Fig. 8 (e)~(h). The cache

TABLE IV

STANDARD HD SUBARRAY DESIGN SPECIFICATION AT A14 TECHNOLOGY NODE

Subarray Design	Number of Subarray Columns	Number of Subarray Rows	Subarray Area (mm^2)
0	288	256	1.08×10^{-3}
1	128	128	2.79×10^{-4}
2	128	256	5.05×10^{-4}
3	256	128	5.34×10^{-4}
4	256	256	9.66×10^{-4}
5	512	128	1.04×10^{-3}

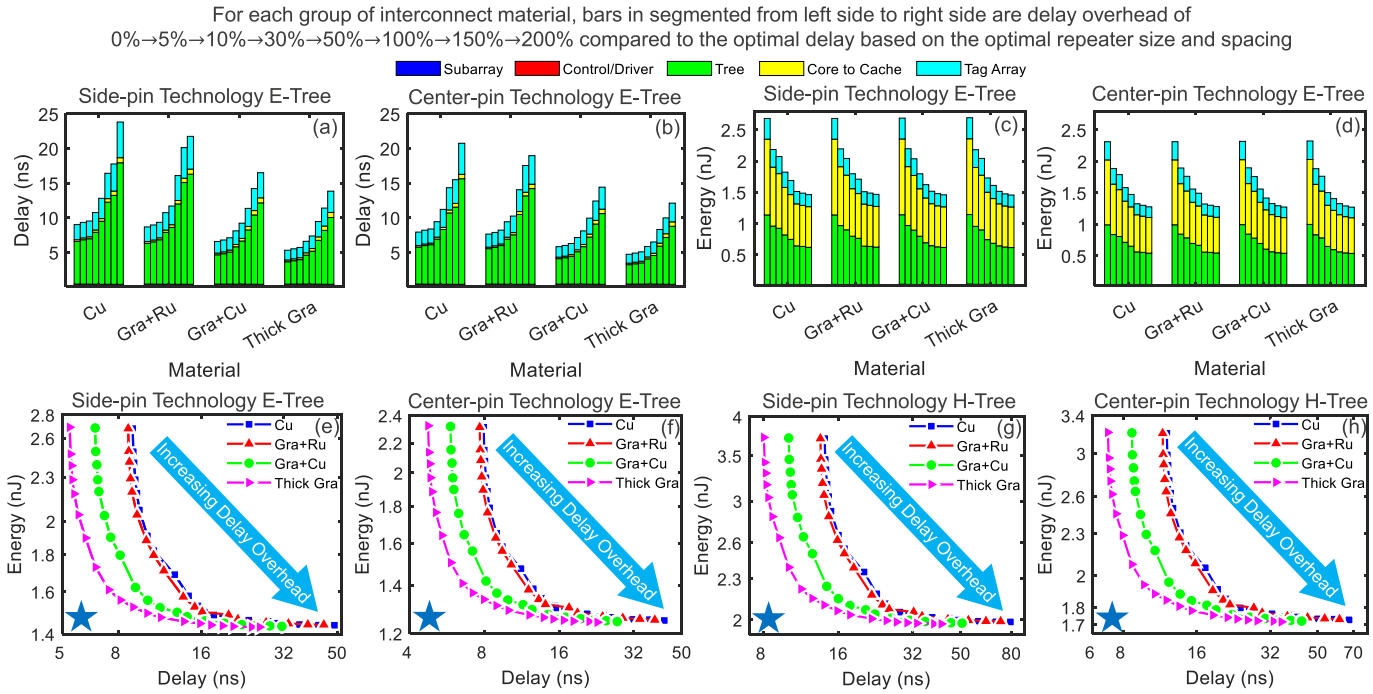


Fig. 8. (a) Delay, (c) read and write energy versus interconnect material, and (e) read and write energy versus delay at delay overhead for a 128MB cache with 16 banks using the standard HD subarray of 288C × 256R implemented with E-Tree interconnect and side-pin technology at A14 node under the uniform workload. (b) Delay, (d) read and write energy versus interconnect material, and (f) read and write energy versus delay under E-Tree interconnect and center-pin technology. (g) and (h) show read and write energy versus delay under the conventional H-Tree for the side-pin technology and center-pin technology, respectively. For each curve of interconnect material, the markers are the delay overhead from 0% to 500% compared to the optimal delay based on the optimal repeater size and spacing. The blue star indicates the favored corner.

using thick graphene outperforms other materials due to its good electrical conductivity. The energy per access is mainly dominated by the cache array tree and core-cache interconnect whose capacitance per unit length depends on interconnect geometry. The performance of the E-Tree-based cache is closer to the preferred corner compared to the H-Tree counterparts thanks to its shorter average interconnect length within arrays.

> Insights from Results
>
> Under a large delay overhead target, fewer repeaters with smaller sizes are inserted, saving energy at the cost of a larger delay. There are clear tradeoffs between delay and energy to minimize overall EDP.

D. Optimization of Cache Design Parameters

In this subsection, we will quantify and optimize several key or essential cache array-level design parameters, including subarray size/design, cache size, the number of vertical banks, and interconnect level for optimal cache-level performance.

1) Impact of Subarray Design/Size and Cache Size on Cache Performance: The subarray design and size are critical to balance the interconnects through local and global. Thus, we investigate delay and energy versus cache size under various subarray designs/sizes, as shown in Fig. 9.

The optimal subarray design and size exist to optimize the cache system EDP. This is because the delay/access time is predominantly influenced either by the subarray or array tree interconnect if the subarray is too large or too small. For example, based on the standard subarray design specification shown in Table IV, for a relatively small subarray, although the area of subarray Design 2 increases by 81.0% compared to Design 1, with half the number of subarrays and less overhead from the tree interconnects, the overall cache area decreases by 15.9% and 26.6% for cache array E-Tree under side-pin technology and center-pin technology, respectively, within the cache size of 128MB. In contrast, for a relatively large subarray, e.g., Design 5, a larger subarray leads to an overall increase in the array area due to the larger dominance of the subarray area.

2) Impact of Number of Vertical Banks and Interconnect Metal Level on Cache Memory System Performance: To optimize the cache organization, we further examine the cache array using different numbers of vertical banks under a fixed total number of banks of 64. The optimal number of vertical banks can be noted to minimize delay, energy, EDP, and EDAP due to the minimal length of the cache array tree or memory system core-to-cache interconnect caused by the cache organization. In addition, we investigate and compare the cache performance using tree interconnects at three different metal levels. From the results shown in Fig. 10, the arrays that use M5 for inter-subarray show a reduction of more than 20% in EDP. Furthermore, we investigate the tree interconnect at three metal levels, where the cache using M6 interconnects provides the lowest EDP due to the large delay benefit under a large width, where up to 54.3% and 49.9% reduction in EDAP and EDP compared to the unoptimized scenario, respectively.

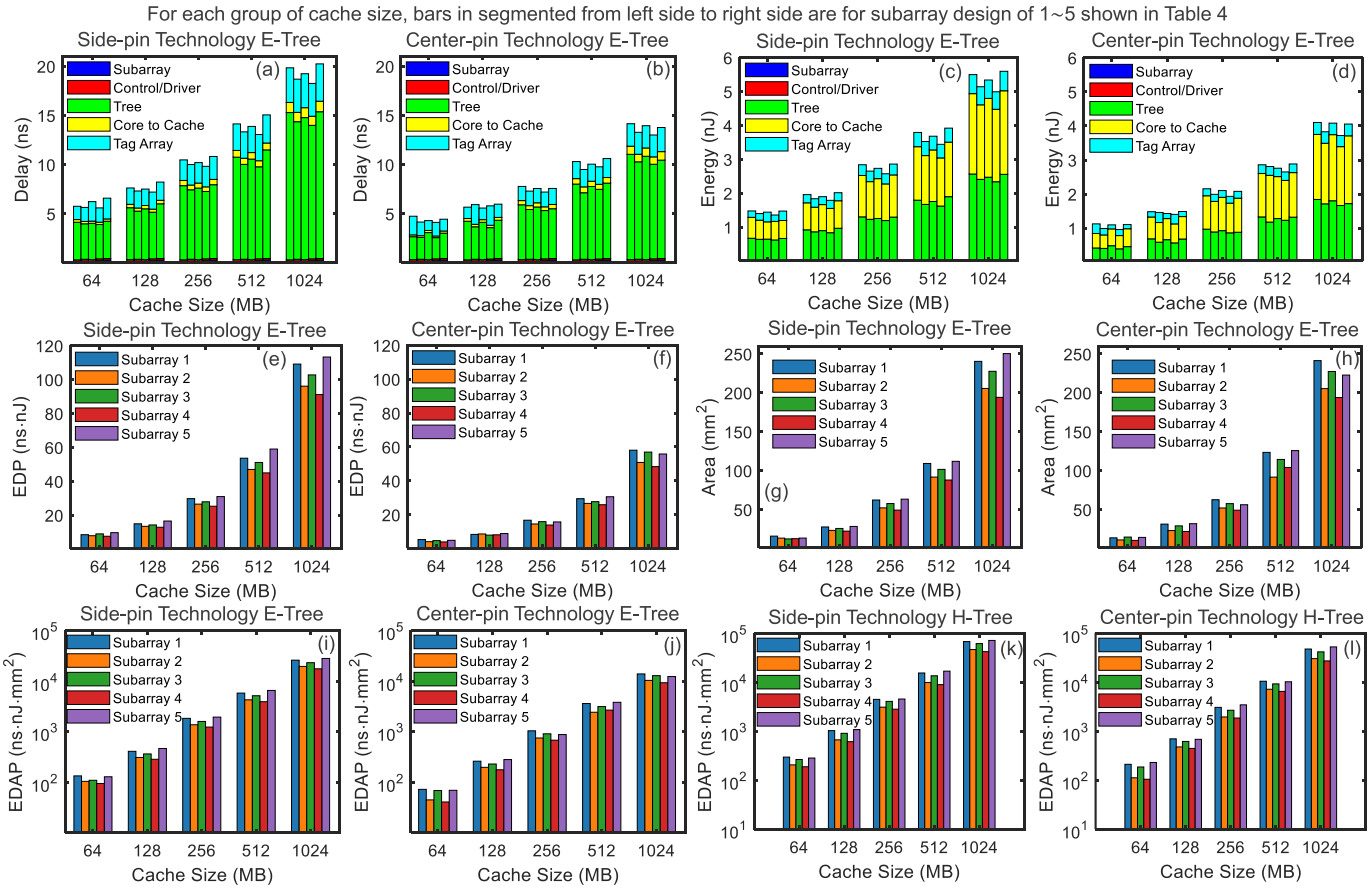


Fig. 9. (a) Delay, (c) read and write energy, (e) EDP, (g) area, and (i) EDAP versus SRAM cache size for various standard HD subarray designs with 4 banks at A14 node using thick graphene interconnect implemented with E-Tree and side-pin technology option under the uniform workload. (b) Delay, (d) read and write energy, (f) EDP, (h) area, and (j) EDAP versus cache size under E-Tree and center-pin technology option. EDAP versus cache size for various standard HD subarray designs with conventional H-Tree interconnects under (k) side-pin technology option and (l) center-pin technology option, respectively.

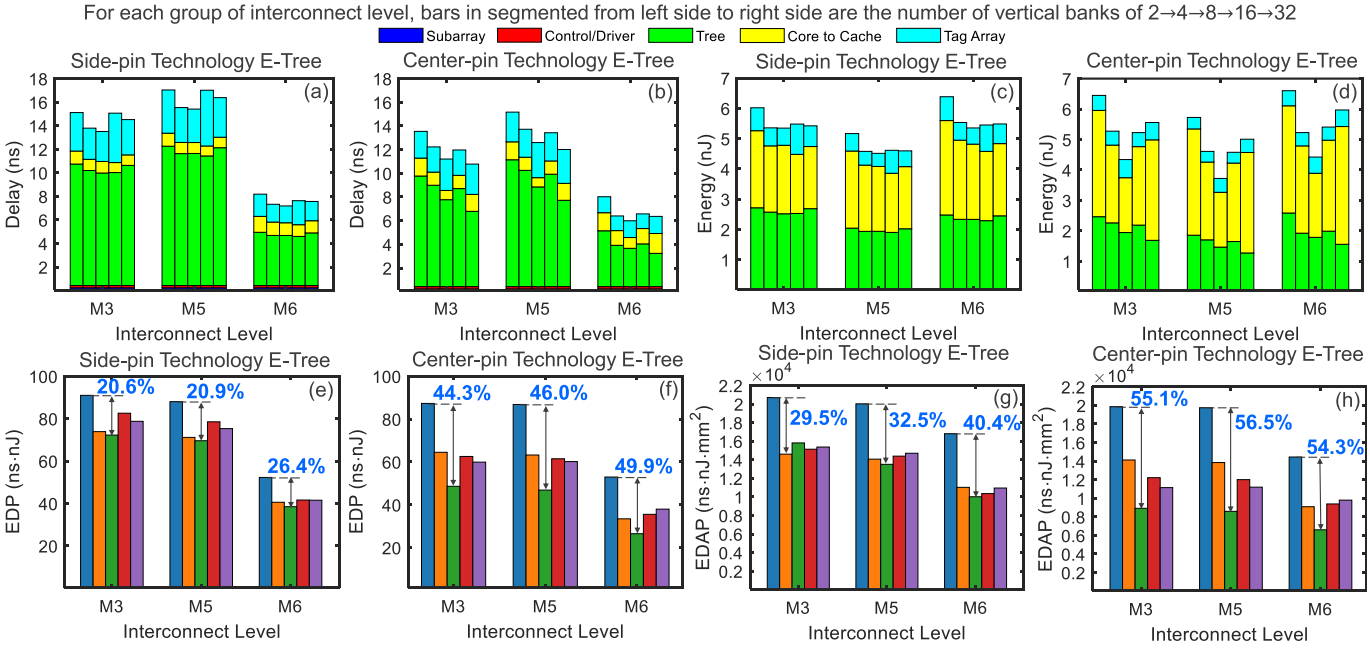


Fig. 10. (a) Delay, (c) read and write energy, (e) EDP, and (g) EDAP versus interconnect level under various numbers of vertical banks with a total number of 64 banks for a 1GB cache using the standard HD subarray of 288C × 256R at A14 node using thick graphene interconnect implemented with E-Tree and side-pin technology under the uniform workload. (b) Delay, (d) read and write energy, (f) EDP, and (h) EDAP versus interconnect level under various numbers of vertical banks for the center-pin technology option and E-Tree interconnect.

Insights from Results

Due to the minimal tree length, the use of M5 for inter-subarray interconnects under the optimal number of vertical banks reduces EDAP and EDP by 56.5% and 46.0%, respectively. The use of M6 interconnects in the cache results in the smallest EDAP/EDP and the large reduction due to the significant delay benefit when using a large width.

IV. Conclusion

A framework has been developed for the co-design of cache with tag array overhead, technology, and interconnect, aiming to optimize device and interconnect technologies at ultra-scaled cutting-edge technology nodes for maximizing SRAM cache memory system performance efficiently. The proposed E-Tree design/interconnect technology can further reduce the overall overhead from the interconnects due to their smaller average interconnect length, especially under non-uniform workloads. However, under highly non-uniform workloads, the center-pin technology may experience a larger delay than its side-pin counterpart, due to an increased tag array overhead. It is shown that the technology node exerts a significant influence on the SRAM cache memory system performance in terms of its overall EDAP. The cache using subarray at the A3 node offers the best performance, primarily attributed to its minimal area overhead, where up to 80.4% and 54.1% reduction in EDAP and EDP for a cache using E-Tree and center-pin technology can be noted. The SRAM cache using thick graphene interconnects demonstrates superior cache-level performance compared to the other material options. This is due to the long MFP and small RPU. In the context of optimal cache design, the optimal subarray design and size exist to maximize the cache memory system performance. Furthermore, the optimal number of vertical banks exists to optimize the work performance of the SRAM cache memory system. Up to 56.5% and 46.0% reduction in EDAP and EDP can be observed for 8 vertical banks compared to 2 vertical banks under 64 banks within the cache.

References

[1] C. Berry et al., "2.7 IBM z15: A 12-core 5.2 GHz microprocessor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 54–56, doi: 10.1109/ISSCC19947.2020.9063030.

[2] D. Wolpert et al., "Cores, cache, content, and characterization: IBM's second generation 14-nm product, z15," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 98–111, Jan. 2021, doi: 10.1109/JSSC.2020.3030062.

[3] L. Hamouche, "Design of SRAM for CMOS 32nm," M.S. thesis, Dept. Elect. Eng., INSA de Lyon, Villeurbanne, France, 2011.

[4] Z. Pei et al., "Graphene-based interconnect exploration for large SRAM caches for ultrascaled technology nodes," *IEEE Trans. Electron Devices*, vol. 70, no. 1, pp. 230–238, Jan. 2023, doi: 10.1109/TED.2022.3225512.

[5] J. Jiang, J. H. Chu, and K. Banerjee, "CMOS-compatible doped-multilayer-graphene interconnects for next-generation VLSI," in *IEDM Tech. Dig.*, Dec. 2018, p. 34, doi: 10.1109/IEDM.2018.8614535.

[6] A. Contino et al., "Circuit delay and power benchmark of graphene against Cu interconnects," presented at the IEEE Int. Interconnect Technol. Conf. (IITC), Brussels, Belgium, Jan. 2019.

[7] Z. Pei, F. Catthoor, Z. Tokei, and C. Pan, "Beyond-Cu intermediate-length interconnect exploration for SRAM application," *IEEE Trans. Nanotechnol.*, vol. 21, pp. 367–373, 2022, doi: 10.1109/TNANO.2022.3157952.

[8] A. B. Kahng, J. Lienig, I. L. Markov, and J. Hu, *VLSI Physical Design: From Graph Partitioning to Timing Closure*. New York, NY, USA: Springer, 2011, doi: 10.1007/978-90-481-9591-6.

[9] W. Gomes et al., "Ponte vecchio: A multi-tile 3D stacked processor for exascale computing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 65, Feb. 2022, pp. 42–44, doi: 10.1109/ISSCC42614.2022.9731673.

[10] X. Wang, C. Augustine, E. Nurvitadhi, R. Iyer, L. Zhao, and R. Das, "Cache compression with efficient in-SRAM data comparison," in *Proc. IEEE Int. Conf. Netw., Archit. Storage (NAS)*, Oct. 2021, pp. 1–8, doi: 10.1109/NAS51552.2021.9605440.

[11] S. Tayal et al., "Incorporating bottom-up approach into device/circuit co-design for SRAM-based cache memory applications," *IEEE Trans. Electron Devices*, vol. 69, no. 11, pp. 6127–6132, Nov. 2022, doi: 10.1109/TED.2022.3210070.

[12] R. Zhang, K. Yang, Z. Liu, T. Liu, W. Cai, and L. Milor, "A comprehensive framework for analysis of time-dependent performance-reliability degradation of SRAM cache memory," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 5, pp. 857–870, May 2021, doi: 10.1109/TVLSI.2021.3056674.

[13] L. Yang et al., "Three-dimensional stacked neural network accelerator architectures for AR/VR applications," *IEEE Micro*, vol. 42, no. 6, pp. 116–124, Nov. 2022, doi: 10.1109/MM.2022.3202254.

[14] H.-H. Liu et al., "CFET SRAM with double-sided interconnect design and DTCO benchmark," *IEEE Trans. Electron Devices*, vol. 70, no. 10, pp. 5099–5106, Oct. 2023, doi: 10.1109/TED.2023.3305322.

[15] Z. Pei et al., "Technology/memory co-design and co-optimization using E-Tree interconnect," in *Proc. Great Lakes Symp. VLSI*, Jun. 2023, pp. 159–162, doi: 10.1145/3583781.3590311.

[16] J. Bucek, K.-D. Lange, and J. V. Kistowski, "SPEC CPU2017: Next-generation compute benchmark," in *Proc. Companion ACM/SPEC Int. Conf. Perform. Eng.*, Apr. 2018, pp. 41–42, doi: 10.1145/3185768.3185771.

[17] S. M. Nair et al., "Workload-aware electromigration analysis in emerging spintronic memory arrays," *IEEE Trans. Device Mater. Rel.*, vol. 21, no. 2, pp. 258–266, Jun. 2021, doi: 10.1109/TDMR.2021.3074251.

[18] N. Binkert et al., "The gem5 simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, May 2011, doi: 10.1145/2024716.2024718.

[19] T. Marinelli, J. I. G. Pérez, C. Tenllado, M. Komalan, M. Gupta, and F. Catthoor, "Microarchitectural exploration of STT-MRAM last-level cache parameters for energy-efficient devices," *ACM Trans. Embedded Comput. Syst.*, vol. 21, no. 1, pp. 1–20, Jan. 2022, doi: 10.1145/3490391.

[20] *Spectre*, Cadence, San Jose, CA, USA, 2023.

[21] *HSPICE*, Synopsys, Mountain View, CA, USA, 2023.

[22] *Virtuoso*, Cadence, San Jose, CA, USA, 2023.

[23] *QuickCap*, Synopsys, Mountain View, CA, USA, 2023.

[24] H.-H. Liu et al., "CFET SRAM DTCO, interconnect guideline, and benchmark for CMOS scaling," *IEEE Trans. Electron Devices*, vol. 70, no. 3, pp. 883–890, Mar. 2023, doi: 10.1109/TED.2023.3235701.

[25] H.-H. Liu et al., "DTCO of sequential and monolithic CFET SRAM," in *Proc. DTCO Comput. Patterning II*, Apr. 2023, pp. 219–225, doi: 10.1117/12.2657524.

[26] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, "CACTI 5.1," HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2008-20, 2008.

[27] 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 Trans. Archit. Code Optim.*, vol. 14, no. 2, pp. 1–25, Jun. 2017, doi: 10.1145/3085572.

[28] Z. Pei et al., "Emerging interconnect exploration for SRAM application using nonconventional H-Tree and center-pin access," in *Proc. 24th Int. Symp. Quality Electron. Design (ISQED)*, Apr. 2023, p. 1, doi: 10.1109/ISQED57927.2023.10129316.

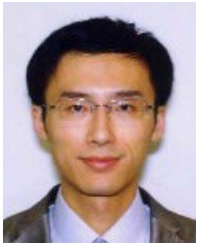
[29] K. I. Bolotin et al., "Ultrahigh electron mobility in suspended graphene," *Solid State Commun.*, vol. 146, nos. 9–10, pp. 351–355, Jun. 2008, doi: 10.1016/j.ssc.2008.02.024.

[30] S. Achra et al., "Characterization of interface interactions between graphene and ruthenium," presented at the IEEE Int. Interconnect Technol. Conf. (IITC), San Jose, CA, USA, Oct. 2020, doi: 10.1109/IITC47697.2020.9515595.

[31] W. S. Leong, H. Gong, and J. T. L. Thong, "Low-contact-resistance graphene devices with nickel-etched-graphene contacts," *ACS Nano*, vol. 8, no. 1, pp. 994–1001, Jan. 2014, doi: 10.1021/nn405834b.

[32] T. Nogami, "Overview of interconnect technology for 7nm node and beyond—New materials and technologies to extend Cu and to enable alternative conductors (invited)," in *Proc. Electron Devices Technol. Manuf. Conf. (EDTM)*, Mar. 2019, pp. 38–40, doi: 10.1109/EDTM.2019.8731225.

[33] S. Achra et al., "Metal induced charge transfer doping in graphene-ruthenium hybrid interconnects," *Carbon*, vol. 183, pp. 999–1011, Oct. 2021, doi: 10.1016/j.carbon.2021.07.070.
[34] S. Achra et al., "Graphene-ruthenium hybrid interconnects," presented at the IEEE Int. Interconnect Technol. Conf. (IITC), Brussels, Belgium, Jul. 2019.
[35] X. Zhang et al., "Ruthenium interconnect resistivity and reliability at 48 nm pitch," in *Proc. IEEE Int. Interconnect Technol. Conf. Adv. Metallization Conf. (IITC/AMC)*, May 2016, pp. 31–33, doi: 10.1109/IITC-AMC.2016.7507650.
[36] C. Pan and A. Naeemi, "A proposal for a novel hybrid interconnect technology for the end of roadmap," *IEEE Electron Device Lett.*, vol. 35, no. 2, pp. 250–252, Feb. 2014, doi: 10.1109/LED.2013.2291783.
[37] 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, Aug. 2019, doi: 10.1016/j.carbon.2019.04.101.
[38] S. Datta, *Quantum Transport: Atom To Transistor*. Cambridge, U.K.: Cambridge Univ. Press, 2005, doi: 10.1017/CBO9781139164313.
[39] T. Yu, E.-K. Lee, B. Briggs, B. Nagabhirava, and B. Yu, "Bilayer graphene/copper hybrid on-chip interconnect: A reliability study," *IEEE Trans. Nanotechnol.*, vol. 10, no. 4, pp. 710–714, Jul. 2011, doi: 10.1109/TNANO.2010.2071395.
[40] S. Sun and D. Jiao, "Multiphysics modeling and simulation of 3-D Cu—Graphene hybrid nanointerconnects," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 2, pp. 490–500, Feb. 2020, doi: 10.1109/TMTT.2019.2955123.
[41] J. Jiang, J. Kang, J. H. Chu, and K. Banerjee, "All-carbon interconnect scheme integrating graphene-wires and carbon-nanotube-vias," in *IEDM Tech. Dig.*, Dec. 2017, p. 14, doi: 10.1109/IEDM.2017.8268389.
[42] N. C. Wang, S. Sinha, B. Cline, C. D. English, G. Yeric, and E. Pop, "Replacing copper interconnects with graphene at a 7-nm node," in *Proc. IEEE Int. Interconnect Technol. Conf. (IITC)*, May 2017, pp. 1–3, doi: 10.1109/IITC-AMC.2017.7968949.
[43] J. Jiang et al., "Intercalation doped multilayer-graphene-nanoribbons for next-generation interconnects," *Nano Lett.*, vol. 17, no. 3, pp. 1482–1488, Mar. 2017, doi: 10.1021/acs.nanolett.6b04516.
[44] C. Pan, P. Raghavan, A. Ceyhan, F. Catthoor, Z. Tokei, and A. Naeemi, "Technology/circuit/system co-optimization and benchmarking for multilayer graphene interconnects at sub-10-nm technology node," *IEEE Trans. Electron Devices*, vol. 62, no. 5, pp. 1530–1536, May 2015, doi: 10.1109/TED.2015.2409875.
[45] A. Hazra and S. Basu, "Graphene nanoribbon as potential on-chip interconnect material—A review," *C*, vol. 4, no. 3, p. 49, 2018, doi: 10.3390/c4030049.
[46] G. K. Mekala, Y. Agrawal, and R. Chandel, "Modelling and performance analysis of dielectric inserted side contact multilayer graphene nanoribbon interconnects," *IET Circuits, Devices Syst.*, vol. 11, no. 3, pp. 232–240, May 2017, doi: 10.1049/iet-cds.2016.0376.
[47] *Raphael*, Synopsys, Mountain View, CA, USA, 2023.
[48] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, "CACTI 6.0: A tool to model large caches," HP Laboratories, Palo Alto, CA, USA, Tech. Rep. HPL-2009-85, 2009.
[49] B. Chehab et al., "Design-technology co-optimization of sequential and monolithic CFET as enabler of technology node beyond 2nm," in *Proc. Design-Process-Technol. Co-Optimization XV*, Apr. 2021, pp. 59–63, doi: 10.1117/12.2583395.



Zhenlin Pei (Graduate Student Member, IEEE) received the M.S. degree in electrical engineering from Columbia University, New York, NY, USA. He is currently pursuing the Ph.D. degree in electrical engineering with The University of Texas at Arlington, Arlington, TX, USA. He was a Senior Design Engineer with the IP Group Research and Development for Tapeout, Cadence Design Systems, Inc., for four years. His current research interests include modeling and optimization of energy-efficient computing systems, spanning from the device level to the system level, and utilizing emerging interconnect and device technologies at ultra-scaled advanced technology nodes.

Hsiao-Hsuan Liu (Graduate Student Member, IEEE) received the B.S. degree in optics and photonics from National Central University, Taoyuan, Taiwan, in 2017, and the M.S. degree from the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan, in 2019. She is currently pursuing the Ph.D. degree in electrical engineering with KU Leuven in collaboration with imec. Her current research interests include SRAM design and technology co-optimization based on nanosheet, forksheet, and complimentary FET technology.

Mahta Mayahinia (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from Shahid Beheshti University, Tehran, Iran, in 2015, and the M.Sc. degree in computer system architecture from the Sharif University of Technology, Tehran, in 2018. In 2020, she joined the Chair of Dependable Nano Computing (CDNC) of Professor Tahoori, KIT University, Karlsruhe, Germany. Her current research interests include VLSI design, computer architecture, computation in memory, and non-volatile memories.

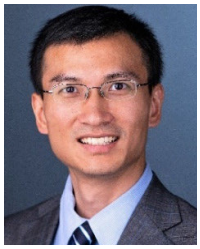
Mehdi B. Tahoori (Fellow, IEEE) received the B.S. degree in computer engineering from the Sharif University of Technology, Tehran, Iran, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2002 and 2003, respectively. He is currently a Full Professor with Karlsruhe Institute of Technology, Karlsruhe, Germany. He was a recipient of the National Science Foundation Early Faculty Development (CAREER) Award and European Research Council (ERC) Advanced Grant. He received a number of best paper awards at various conferences and journals, including ICCAD, FPL, ACM TODAES, and IEEE TVLSI.

Francky Catthoor (Fellow, IEEE) received the Ph.D. degree in EE from Katholieke Universiteit Leuven (KU Leuven), Belgium, in 1987. From 1987 to 2000, he headed several research domains in the area of synthesis techniques and architectural methodologies. Since 2000, he has been strongly involved in other activities at imec, Leuven, Belgium, including co-exploration of application, computer architecture and deep submicron technology aspects, biomedical systems and the IoT sensor nodes, and photo-voltaic modules combined with renewable energy systems, all at imec. He is currently an imec Senior Fellow. He is also a part-time Full Professor with the EE Department, KU Leuven. He has been an associate editor of several IEEE and ACM journals. He was elected a Fellow of the IEEE in 2005.

Zsolt Tőkei (Member, IEEE) received the M.S. degree in physics from University Kossuth, Debrecen, Hungary, in 1994, and the Ph.D. degree in physics and materials science, in 1997, in the framework of a co-directed thesis between Hungarian University Kossuth and French University Aix Marseille-III. He joined imec in 1999 and, since then, he has been holding various technical positions in the organization. First, as a Process Engineer and a Researcher of copper low-k interconnects, then headed the metal section. Later he became the Principal Scientist and the Program Director of Nano-Interconnects. In 1998, he started working at the Max-Planck Institute of Düsseldorf, Germany, as a Post-Doctoral Researcher. Joining imec, he continued working on a range of interconnect issues, including scaling, metallization, electrical characterization, module integration, reliability, and system aspects. He is an imec Fellow and the Program Director of Nano-Interconnects at imec.

Dawit Burusie Abdi (Member, IEEE) received the B.Sc. degree in electrical engineering from Jimma University, Jimma, Ethiopia, in 2009, the M.Tech. degree in microelectronics from Addis Ababa University, Addis Ababa, Ethiopia, in 2011, and the Ph.D. degree from Indian Institute of Technology Delhi (IITD), New Delhi, India, in 2017. He was an Assistant Professor with Addis Ababa Science and Technology University, Addis Ababa, from 2016 to 2020, and a Post-Doctoral Fellow with imec, Leuven, Belgium, linked with KU Leuven, Leuven, from October 2020 to June 2022. He is currently with imec, as a Memory Macro Researcher, where his research focuses on memory bit-cells to macro level DTCO and memory compilers.

James Myers (Member, IEEE) received the M.Eng. degree in electrical and electronic engineering from Imperial College London, London. He spent 15 years at Arm, leading research from low power circuits and systems, through printed electronics, to DTCO activities. He joined imec in 2022 to start a new U.K. site and lead the system technology co-optimization research program, with the aim to build upon established DTCO practices to overcome the numerous scaling challenges foreseen for future systems. He holds 60 patents, taped out 20 SoCs, presented at ISSCC and VLSI Symposium, and published in IEDM and Nature.



Chenyun Pan (Senior Member, IEEE) received the B.S. degree in microelectronics from Shanghai Jiao Tong University, Shanghai, China, in 2010, and the Ph.D. degree in ECE from Georgia Institute of Technology, in 2015. He is currently an Assistant Professor with the Department of Electrical Engineering, The University of Texas at Arlington. He has published over 70 peer-reviewed IEEE journal and conference papers. His research interests include device-, circuit-, and system-level modeling and optimization for energy-efficient Boolean and non-Boolean computing systems based on various emerging device and interconnect technologies. He was a recipient of two best paper awards in the IEEE International Symposium on Quality Electronic Design and the IEEE Conference on IC Design and Technology, Research Spotlight Award in the School of ECE at Georgia Tech, and the Early Career Research Program Award from U.S. Department of Energy.