Graphene-Based Interconnect Exploration for Large SRAM Caches for Ultrascaled Technology Nodes

Zhenlin Pei, Graduate Student Member, IEEE, Mahta Mayahinia, Graduate Student Member, IEEE, Hsiao-Hsuan Liu, Graduate Student Member, IEEE, Mehdi Tahoori, Fellow, IEEE, Francky Catthoor, Fellow, IEEE, Zsolt Tokei, Member, IEEE, and Chenyun Pan, Senior Member, IEEE

Abstract—Graphene-based interconnects are considered promising replacements for traditional copper (Cu) interconnect due to their great electric properties. In this article, an interconnect-memory co-design framework is developed to efficiently optimize various graphene-based interconnect technologies. Four interconnect materials and heterogeneous design schemes are benchmarked against their traditional Cu counterparts to optimize large cache-level SRAM performance in terms of delay and energy per access, energy-delay product (EDP), and energy-delay-area product (EDAP). A large design space exploration is performed based on realistic subarray design and device technology. Various interconnect- and array-level design parameters are studied to quantify the true potential of graphene-based wires for optimal memory performance.

Index Terms—Benchmarking, design/technology co-optimization, graphene, heterogeneous interconnect, SRAM.

I. Introduction

THE technology research becomes more "interconnect-centric" at sub-10-nm nodes due to the large resis-tance of traditional copper (Cu) interconnect caused by the ever-increasing size effect and impact of barrier thickness on wire and via resistance [1], [2], [3], [4]. To address the challenges of the interconnect, enormous research efforts have been performed, including those proposing new interconnect materials, such as ruthenium (Ru), cobalt (Co), and graphene [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. The barrierless Ru is one potential option due to its advantages: 1) competitive resistance to Cu in advanced nodes and 2) higher electromigration (EM) reliability with higher lifetimes compared to Co and Cu counterparts [5], [7], [8], [11], [12], [13], [14]. Co with barrier has higher EM reliability than Cu [11], [12], [13], [14]. However, challenges in the process integration of Co or Ru exist at a linewidth of 20 nm, including optimized chemical mechanical planarization (CMP) and scaled barriers/liners [6]. Graphene is considered a potential alternative to traditional interconnect material to enhance power-efficient systems due to its good electric properties, including excellent current conductivity and large mean free path (MFP) [15], [16]. In addition, it provides a small capacitance due to its thin geometry and quantum capacitance, which results in decreasing dynamic power dissipation. However, graphene interconnects have several challenges, including 1) large contact resistance, leading to limited usage for very short interconnects at local levels and 2) the quality control of graphene during the fabrication. In addition, difficulties exist in the fabrication of graphene. For example, some developed graphene films may not be purely single crystals. The uncontrolled fabrication manufactures multilayer graphene that turns into graphite, whose conductivity is lower due to the interlayer electron hopping [17]. These fabrication challenges sometimes lead to a limited number of available graphene layers, increasing its resistance and limiting the usage of graphene for thick interconnects at global levels compared to their Cu counterpart. As a result, graphene is more suitable for the intermediate-length interconnect, which is the focus of this research.

To evaluate the potential benefit of graphene interconnects, we choose a large SRAM cache as our benchmarking circuit because it is one of the major components in all digital processors, and the SRAM has good compatibility with industry CMOS processes, high density, great cost-efficiency, and lower leakage compared to the DRAM [18]. In addition, the interconnect of an SRAM array consists of wordlines (WLs),

Manuscript received 25 August 2022; revised 21 October 2022 and 25 November 2022; accepted 26 November 2022. Date of publication 6 December 2022; date of current version 3 January 2023. This work was supported in part by IMEC and the Department of Energy (DoE) under Award DE-SC0022881 and in part by the National Science Foundation (NSF) under Grant CCF-2219753. The review of this article was arranged by Editor P. Thadesar. (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).

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

Hsiao-Hsuan Liu and Francky Catthoor are with IMEC, 3001 Leuven, Belgium, and also with the Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven (KU Leuven), 3000 Leuven, Belgium.

Zsolt Tokei is with IMEC, 3001 Leuven, Belgium.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2022.3225512.

Digital Object Identifier 10.1109/TED.2022.3225512

bitlines (BLs), and H-trees, which span a large range of lengths and widths across different interconnect levels, which makes it an excellent study for benchmarking intermediate-length interconnects using emerging technologies. Based on the latest reported high-performance microprocessors [19], [20], [21], the cache capacity spans from the last-level cache with a size up to or even above 1 Gbits to smaller L2 and L3 cache sizes of a few Mbits. This motivates us to investigate a wide range of SRAM macros and how they interact with the design and different interconnect materials.

For traditional Cu interconnects, the effective *RC* delay increases due to its small effective MFP in ultrascaled dimensions, which degrades the writability and performance of bit-cells located far from the WL driver and write driver [4]. Existing work has investigated beyond-Cu intermediate-length interconnection for the SRAM application based on the ASU Predictive Technology Model (PTM) with the updated CACTI framework [22], [23], [24]. However, predictive models are known to have limited accuracy due to their extrapolation for key device-level parameters. The PTM cannot accurately capture the realistic device characteristics from the fabrication processes and complex physical behaviors, especially for devices at technology nodes beyond 10 nm. Because of the close interaction between the device and interconnect, it is critical to perform a rigorous simulation based on the realistic industry-standard cell library to investigate the true performance benefits of advanced interconnect materials at deeply scaled technology nodes.

In this article, we will use a technology library that has been experimentally verified. In addition, we will develop a dedicated cache subarray, whose structure consists of WLs, BLs, flip-flop, column mux, write driver, sense amplifier, and array matrix. The performance of the high-density subarray is modeled based on the realistic data extracted from experiments. Compared to the analytical subarray model from CACTI, the adopted subarray will provide more precise and meaningful tradeoffs among technology parameters and structural design, allowing an efficient and accurate design/technology exploration at the cache level.

Although the impact of the graphene-based interconnects on cache-level performance has been investigated [22], an ideal assumption of an unlimited number of graphene layers is made, which is not realistic considering the actual fabrication process. In this work, we will quantify the impact of the number of available graphene layers on the cache-level performance. In addition, to fully utilize the advan-tage of graphene, we propose heterogeneous interconnect design schemes, and different interconnect materials are used at ***intrasubarray*** level (e.g., BLs and WLs) and ***intersubarray*** level (e.g., H-trees). Key tradeoffs among a variety of heterogeneous interconnect parameters are investigated, including different material options, geometry design, such as aspect ratio and width, and cache size. The main contributions of the work are highlighted in the following.

1) An efficient interconnect/cache co-design framework is developed by incorporating realistic device technology and subarray design for ultrascaled technology nodes to realize co-exploration among the interconnect, technology, SRAM circuit, and cache array.
2) Four promising graphene-based interconnect options are benchmarked against Cu counterparts, and a large design space is explored to maximize the potential benefits of graphene for cache-level performance.
3) We propose heterogeneous interconnect design schemes to fully utilize graphene and quantify the impact of the number of graphene layers on the cache-level performance.
4) Valuable design insights are provided to cache designers and interconnect technologists to mutually acknowledge the process and material requirements and to design more appropriate interconnect materials for SRAM systems.

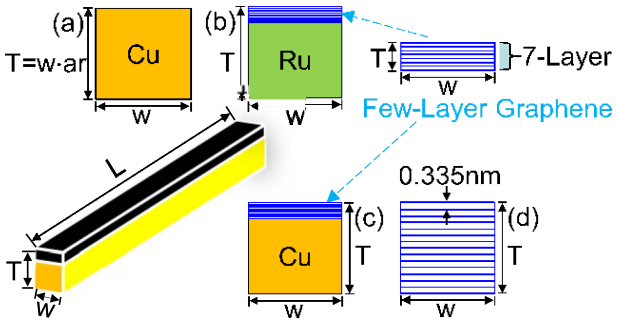

Fig. 1. (a) Cu for the baseline, (b) graphene-capped Ru, (c) graphene-capped Cu, and (d) thick graphene.

II. Modeling Approaches

A. Interconnect Modeling

Based on the existing modeling work, four promising types of interconnect materials are adopted to quantify their impacts on the cache array-level performance, including 1) Cu for the baseline; 2) graphene-capped Ru; 3) graphene-capped Cu; and 4) thick graphene [5], [7], [8], [14], [22], [25], [26], [27], [28], [29]. The sketch for the four interconnect materials is shown in Fig. 1. The total graphene thickness is the product of the gap distance of 0.335 nm and the number of graphene layers.

For the baseline Cu interconnect, its resistivity model follows the existing work with a side wall specularity of 0.5 and a grain boundary reflectivity of 0.5 that are calibrated based on the experimental data [26], [30]. For general graphene-based interconnects, the current flowing through a single-layer graphene is obtained by the Landauer formula [31], which is a function of the effective MFP of graphene. The MFP depends on several factors, including the graphene edge roughness and substrate material property. In the previous work [22], the MFP has been fitted based on the mobility extracted based on experimental data by the following semiclassical equation [32], [33]:

$$\sigma = en\mu = \frac{2e^2}{h}\sqrt{\pi n}\cdot \mathrm{MFP} \tag{1}$$

where n is the carrier density, σ is the conductivity, and μ is the mobility.

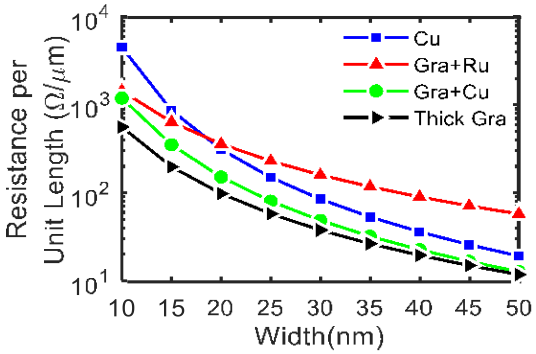

Fig. 2. Resistance per unit length versus width for four interconnect materials with an aspect ratio of 1.

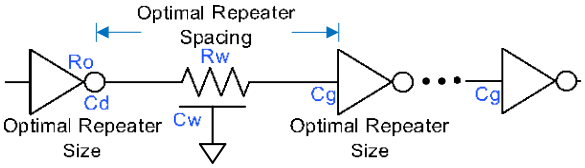

Fig. 3. Circuit model for interconnects with optimal repeater insertion. R_o is the repeater output resistance, C_d and C_g are the device drain and gate capacitance, respectively, R_w and C_w are the interconnect resistance and capacitance, respectively, and additional contact and quantum resistance R_{con} and $R_{quantum}$ are added on each side of the interconnect for graphene.

The MFP is set as 460 nm for a 1-μm-wide graphene based on the existing work [5], which scales down with the decrease of width due to the edge scattering. Based on the reported mobility from [33], the carrier density can be back-calculated as 4.85×10^{10} cm^{-2}. Assuming side contacts are used, the graphene contact resistance is given by $R_{con}/(W N_{layer})$, where W is the interconnect width, N_{layer} is the number of graphene layers, and R_{con} is the contact resistance of 100 $\cdot \mu$m [29]. The graphene quality and its MFP are affected by the fabrica-tion. A large MFP reduces the wire resistance, which helps to improve the access time. However, the MFP has less impact on energy because energy depends on the wire capacitance, which is determined by the interconnect geometry.

For graphene-capped Ru, the resistance per unit length is extracted based on experimental data for different thicknesses [8]. For graphene-capped Cu, the electrons scatter less frequently inside Cu, and 3× of the grain size is adopted to capture such an effect based on the existing experimental work [27], [28]. To compare different interconnect materials under a given aspect ratio of 1, the resistance per unit length is shown in Fig. 2, where thick graphene provides the best resistance at a small width due to its large MFP. The capacitance per unit length of the interconnect is extracted by Synopsys Raphael for various interconnect geometries [34]. In the H-tree, the interconnect delay model with repeater insertion under the optimal repeater size and spacing follows the previous work based on original models from CACTI [22], [23], and the circuit schematic is shown in Fig. 3. Device parameters are extracted based on practical device library at sub-5-nm technology node [35].

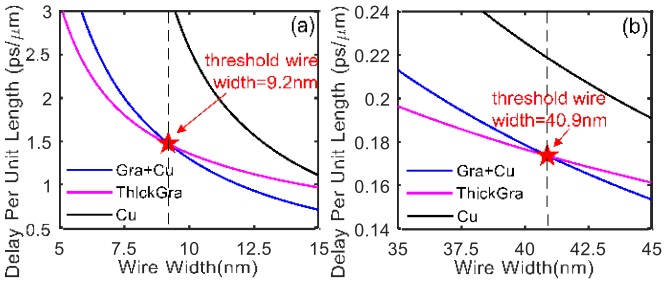

Fig. 4. Delay per unit length under the optimal repeater size and spacing versus interconnect width for graphene-capped Cu and thick graphene under the aspect ratio of 1 with a maximum number of graphene layers of (a) 10 and (b) 100. Note that the actual aspect ratio of graphene depends on the number of graphene layers. The thickness of graphene is (a) 3.35 and (b) 33.5 nm.

B. Heterogeneous Interconnect Design Under the Impact of Limited Number of Graphene Layers

The resistance per unit length shown in Fig. 2 assumes that an unlimited number of graphene layers can be achieved during the fabrication process. In reality, the graphene performance highly depends on the available number of graphene layers, defectivity levels, and the resistance per unit length of graphene may be larger than the graphene-capped Cu or even the traditional Cu counterparts if the number of graphene layers is limited. For example, in the intrasubarray level interconnect, the aspect ratio is usually large to reduce the resistance per unit length of the interconnect. Therefore, Cu interconnects can be made much thicker than graphene interconnects, and graphene cannot be competitive to Cu if only a few layers are available. To fully utilize the potential of graphene interconnects, we propose heterogeneous interconnects using a combination of graphene-capped Cu and thick graphene for inter- or intrasubarray levels depending on the width and aspect ratio.

Based on the interconnect repeater insertion model described in the previous subsection, Fig. 4 shows the delay per unit length versus the interconnect width under the optimal repeater insertion with a maximum number of graphene layers of 10 and 100. For the thick graphene with ten layers, a threshold wire width of 9.2 nm can be observed, below which thick graphene provides a better delay per unit length compared to its graphene-capped Cu counterpart. This is because the severe size effect of Cu increases the resistance per unit length substantially compared to the thick graphene counterpart. If the maximum number of graphene layers increases to 100, as shown in Fig. 4(b), the critical width increases to 40.9 nm, meaning that a wider range of interconnects can take advantage of the low resistance of thick graphene to minimize the delay. Here, we propose a heterogeneous interconnect design scheme to choose the best material based on the wire width at intra- and intersubarray levels. If the width is below (or above) the threshold, thick graphene (or graphene-capped Cu) will be used for that level.

In Fig. 4, only two different maximum numbers of graphene layers, 10 and 100, are investigated under a fixed aspect ratio of 1, and in Fig. 5, we sweep the aspect ratio and

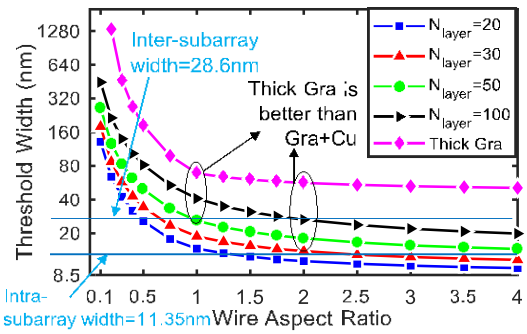

Fig. 5. Threshold width versus aspect ratio of interconnects with optimal repeater insertion at the maximum number of graphene layers.

the corresponding threshold widths are extracted for different numbers of graphene layers. For a given number of graphene layers, the threshold width keeps decreasing as the aspect ratio increases because Cu interconnects benefit from the large cross section area. From Fig. 5, designers can obtain valuable information regarding how to choose the best material based on the interconnect geometry and the number of available graphene layers. For example, two dashed lines in Fig. 5 show the intra- and intersubarray interconnect widths. For intrasubarray interconnects with an aspect ratio of 2, in order for graphene to outperform graphene-capped Cu in terms of delay per unit length, more than 30 graphene layers are needed; for intersubarray interconnects with an aspect ratio of 1, the minimum number of graphene layers to outperform graphene-capped Cu counterpart increases to 50. Note that the thickness of thick graphene is determined by the number of available layers instead of the aspect ratio, as shown in Fig. 4.

It can be observed that the required number of graphene layers will highly depend on the width as well as the optimal aspect ratio during the design of the cache. Here, only delay per unit length is considered as the target metric. During the optimization of the cache array-level performance in Section III, the energy will also be considered by minimizing the overall EDP. Depending on the tradeoff in delay and energy, the optimal aspect ratio of interconnects on different levels will be obtained, which determines the best material choices, i.e., heterogeneous interconnects. In the next section, we will investigate heterogeneous design schemes, where the intra- and intersubarray level interconnects use different combinations of the wire materials, to maximize cache-level performance.

C. Subarray Models

To enable a fast, accurate, and flexible analysis of latency and energy dissipation of large cache modules, we have developed a high-level equation-based model for the SRAM-based subarray. The accuracy of this model has been verified based on extensive electrical-level simulations. The device models are adopted and calibrated from imec experiments and standard-cell library [36]. Device parameters, including gate capacitance, drain capacitance, ON current, supply voltage, temperature-dependent leakage current, and threshold voltage,

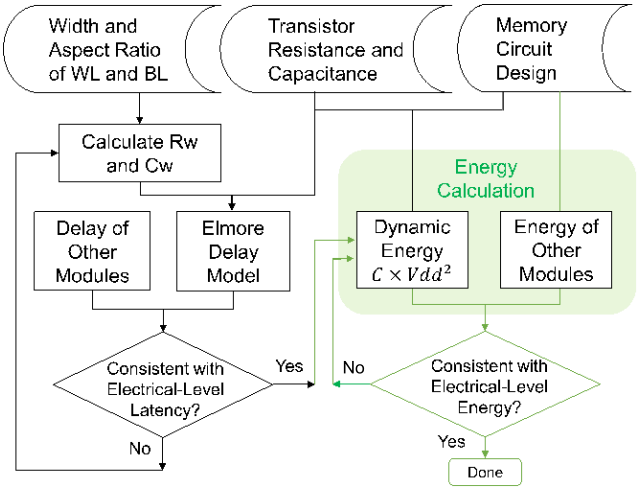

Fig. 6. Design methodology for the high-level modeling of the latency and energy of the SRAM subarray.

are extracted by Cadence Spectre and Synopsys HSPICE simulations [37], [38]. The overall methodology of the proposed high-level modeling for the energy and delay is shown in Fig. 6.

1) Modeling of the Subarray-Level Latency: The first step toward such high-level modeling is approximating the subarray-level latency. Based on CACTI and NVsim [23], [39], we model read and write latencies as follows:

$$T_{\text{read}} = t_{\text{decode}} + t_{\text{WL}} + t_{\text{BL}} \tag{2}$$

$$T_{\text{write}} = t_{\text{decode}} + t_{\text{WL}} + t_{\text{BL}} + t_{\text{bit-flip}} \tag{3}$$

where t_{decode}, t_{WL}, t_{BL}, and $t_{\text{bit-flip}}$ are the decoder latency, charging latency for the WL and BL, and bit-flip latency, respectively. The calculation of the decoder latency is fully aligned with mature high-level memory simulators [39]. On the other hand, bit-flip latency (in the write latency equation) is required for the content of the SRAM to be switched, which can be derived from a one-time electrical-level simulation. In high-level modeling, besides accuracy, speed and flexibility are also essential. In this regard, our latency approximation only covers the dominant latency terms within the subarray, and its accuracy has been verified based on the electrical-level simulations on the memory array and periphery.

Based on the induced *RC* load on the WL and BL, the required time for charging the WL and BL, i.e., t_{WL} and t_{BL}, can be determined. The resistive load of both the WL and BL originated from the interconnect parasitic resistance. The capacitance of the lines originated from the interconnect parasitic capacitance as well as the capacitive load from the access transistors. For the WL and BL, the gate and drain of the access transistor have contributed to the capacitive load, which is characterized based on electrical-level simulations. In this work, we have modeled the BL and WL as *RC* networks. Hence, the latency term corresponding to WL and BL charge latency can be calculated through the Elmore delay equation [40]. For calculating the interconnect parasitic resistance and capacitance, we have reused the equation from

the existing work [39]. Our proposed high-level modeling has been calibrated based on the subarray-level latency through electrical simulations. To have an accurate latency modeling and to calibrate the high-level model, we have swept the interconnect feature size in the acceptable range corresponding to the technology node [30].

2) Modeling of the Subarray-Level Energy: After calibrating the *RC* parasitic of the WL and BL, we proceed with the energy approximation. For the write operation, dynamic energy stored in the interconnect parasitic capacitor of the WL and BL, the bit-flipping, write drivers enabler, decoder, the timing control (write mode), as well as the nonnegligible leakage energy are the main terms that contribute to the total write energy dissipation. Like the dynamic energy stored in the interconnect parasitic capacitor, enabling the write driver also requires a capacitor to be charged. Therefore, these energy terms can be calculated as $(C \times V^2)$. In a similar way, the energy of the timing control module can be calculated by performing the integral on its dynamic power during its activation time. For the decoder energy as well as its latency, we have reused the models developed in the existing work [39].

For the subarray-level read operation, besides dynamic energy stored in the interconnect parasitic capacitor of the WL and BL, the decoder, the timing control (read mode), the leakage energy, sense amplifier, column multiplexer enabler, and output latches are the other energy contributors. To enable the multiplexer, a capacitor should be charged, and its corresponding energy term is $(C \times V^2)$. For the sense amplifier and output latches, performing the integral on their dynamic power during their activation time results in energy dissipation.

The energy consumption of the memory can also be calculated by electrical-level simulation of the full memory array and periphery. However, this approach is too slow and effortful, particularly for higher level design exploration. Therefore, developing a flexible high-level model for the energy terms can be quite helpful. For instance, for the bit-flip energy, besides the write drivers, we have considered only one SRAM cell, and the entire row and column have been represented as the corresponding *RC* load. The energy of the sense amplifier can be measured by considering the entire column, while only the equivalent *RC* load of the row is involved in the simulation.

As shown in Fig. 6, the terms of the main energy contributions are determined accurately through electrical-level simulations. Once the high-level model of the energy closely converges with the energy through the electrical simulation, *fitting parameters* (A_{read} and A_{write}) can be applied to the high-level model of the energy. Based on CACTI and NVsim [23], [39], the following equations show the model for the write and read energy:

$$E_{\text{Write}} = A_{\text{write}} \times E_{\text{decoder}} + E_{\text{timingCtrl(write)}} + E_{\text{WL}} + E_{\text{BL}} + E_{\text{bit-flip}} + E_{\text{WriteDriverEN}} + E_{\text{leakage}} \quad (4)$$

$$E_{\text{Read}} = A_{\text{read}} \times E_{\text{decoder}} + E_{\text{timingCtrl(read)}} + E_{\text{WL}} + E_{\text{BL}} + E_{\text{SA}} + E_{\text{colMux}} + E_{\text{outLatch}}) + E_{\text{leakage}} \quad (5)$$

where A_{write} and A_{read} are the fitting parameters for write and read energy, respectively. E_{decoder}, $E_{\text{timingCtrl(write)}}$, $E_{\text{timingCtrl(read)}}$, E_{WL}, E_{BL}, $E_{\text{bit-flip}}$, $E_{\text{WriteDriverEN}}$, E_{SA}, E_{colMux}, E_{outLatch}, and E_{leakage} are the decoder energy, energy for the write/read timing control, WL, BL, bit-flip, write driver enable, sense amplifier, column Mux, output latch, and leakage energy, respectively. Please note that developing the high-level equation-based model, e.g., (2)–(5), enables a faster and more scalable approach compared to time-consuming electrical-level simulations. Also, the calibration of the aforementioned equations based on the electrical-level simulations by utilizing the methodology presented in Fig. 6 ensures sufficient accuracy of the high-level model of the SRAM subarray. Therefore, such a high-level approximative model fits well to our interconnect/cache co-design framework.

D. Cache-Level Memory Models

An open-source and well-known simulator, CACTI, is adopted to optimize the SRAM array [23], [41]. CACTI sweeps the cache organization parameters to obtain optimal parameters for target metrics users defined. The array access critical path contains input and output H-tree from outside and inside of the bank and timing path from the subarray that is designed by imec researchers [36]. The original CACTI has been already validated by SPICE simulation and reported data from the commercial caches for Intel 65-nm L3 cache and Sun SPARC 90-nm L2 cache [41]. With a validated cache simulator, various wire configurations and design parameters can be efficiently explored at the early design stage with good accuracy.

To integrate the subarray designed by imec researchers, key performance metrics, such as delay, energy, and area for various components in the original CACTI, will be updated based on the actual values extracted from the realistic experimental data and simulation. In addition, the number of rows and columns, column decoders, MUXs, and output sense amplifiers follow the values provided by imec, which will affect the exploration of the cache organization [36]. In this work, critical tradeoffs among interconnect parameters are performed, including the interconnect width, aspect ratio, and the number of available graphene layers, to optimize cache-level SRAM performance. Generic guidelines to material technologists and system designers will be provided based on the comparison among different interconnect parameters and Cu-based counterparts to identify the true benefits of promising graphene-based interconnects and realize energy-efficient memory systems.

III. Simulation Results

Based on the modeling approaches in Section II, the performance analysis is performed at the cache level. We will investigate five interconnect materials (i.e., Cu, graphene-capped Ru, graphene-capped Cu, thick graphene, and heterogeneous interconnects, where Cu, graphene-capped Cu, and thick graphene are adopted for intra- and intersubarray levels, respectively) and four cache sizes (i.e., 0.5, 2, 16, and 128 MB) in the case study. The material, interconnect, and array-level design parameters used in the modeling and simulation are listed in Table I.

TABLE I
PARAMETERS USED IN THE MODELING AND SIMULATION

Parameter	Value
Cache Size (MB)	0.5/2/16/128
Number of Banks	4/16
Associativity	2
Number of Subarray Rows	256
Number of Subarray Columns	128
Intra-subarray Interconnect Width (nm)	11.35
Inter-subarray Interconnect Width (nm)	28.57
Intra-subarray Interconnect Aspect Ratio	1/2/3/4
Inter-subarray Interconnect Aspect Ratio	1
Cu Side Wall Specularity, p	0.5
Cu Grain Boundary Reflectivity Co-efficient, R	0.5
Graphene Mean-Free-Path at W = 1μm (nm)	460
Graphene Contact Resistance (Ω·μm)	100
Single Graphene Layer Thickness (nm)	0.335
Number of Graphene Layers of Graphene-capped Ru	7
Number of Graphene Layers of Graphene-capped Cu	7

A. Impact of Interconnect Geometry on Cache Performance

One key benefit of graphene-based interconnect is its large MFP, which potentially lowers the resistance. Because the resistance highly depends on the geometry, the impact of wire aspect ratio and width is investigated for intra- and intersubarray wires to maximize the SRAM array-level performance.

Under different intrasubarray interconnect aspect ratio assumptions, the breakdown bar charts of access time and energy for different interconnect widths are shown in Fig. 7. The interconnect width is scaled with a width scaling factor, which is applied to multiply the standard interconnect width to quantify the width impact on different array-level performances. The scaling factor of 1 corresponds to an intrasubarray interconnect width of 11.35 nm. Here, the default cache size is 16 MB with four banks using the subarray with 256 rows and 128 columns, and the associativity is two under the intersubarray interconnect aspect ratio of 1 and width scaling factor of 1. The delay and energy per access of the cache contain subarray and input and output H-tree of outside and inside the bank. Fig. 7 shows that the output H-tree dominates the overall energy due to a large number of data bits and interconnect length.

In general, a larger interconnect aspect ratio helps to improve the delay due to a larger cross section area and smaller interconnect resistance, but it increases the energy due to the larger line-to-line capacitance. In Fig. 7(a), when the intrasubarray interconnect width is small, the delay improves with the increase of the width due to the reduced interconnect resistance per unit length. However, as the width becomes large, both delay and energy increase because of the area overhead, which increases the total interconnect length. In short, the array-level performance is either 1) limited by the access time if the intrasubarray interconnect width is too small or large or 2) limited by the energy if the width is too large. Note that for the energy per access in Fig. 7(b), the y-axis is shown in a log scale due to the large span in different energy components.

Fig. 7(c) and (d) shows the delay and energy for a larger cache size of 128 MB. Compared to Fig. 7(a) and (b), the overall trend is similar, except for the fact that the delay and

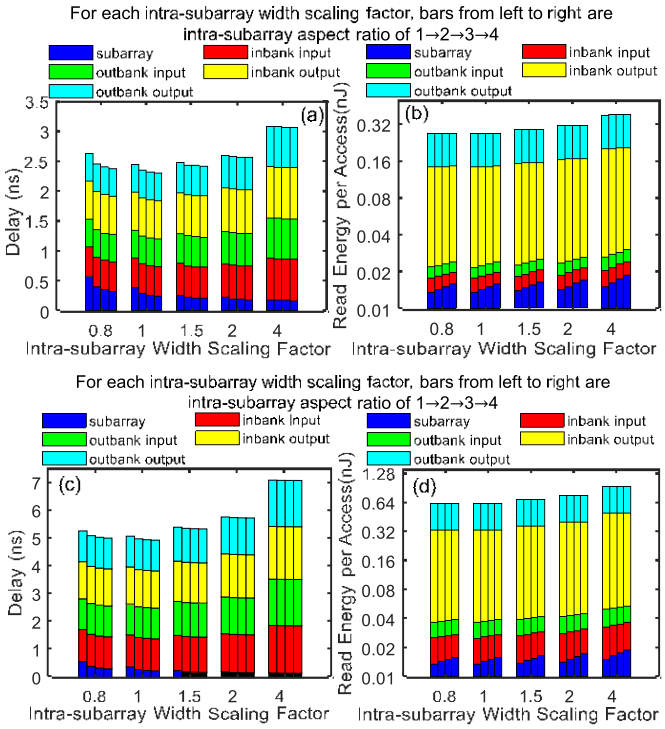

Fig. 7. (a) Access time and (b) read energy (power delay product) per access breakdown bar chart versus intrasubarray width scaling factor for a variety of aspect ratios using thick graphene interconnects. Here, the cache size is 16 MB with associativity of two with four banks using a subarray with 256 rows and 128 columns under the intersubarray interconnect aspect ratio and width scaling factor of 1. (c) and (d) Access time and read energy (power delay product) per access for a 128-MB cache with the same configurations as (a) and (b).

energy contributions from the H-tree inside the bank increase due to the increase in the bank size. To take both delay and energy into account, the EDP versus the intrasubarray interconnect width scaling factor for different aspect ratios is shown in Fig. 8(a) and (c), where an optimal width exists to minimize the EDP. Under the consideration of array area, optimal intrasubarray width and aspect ratio of interconnects exist to minimize the EDAP, as shown in Fig. 8(b) and (d). Overall, interconnects with a large aspect ratio at the nominal width are preferred to minimize the cache-level EDP and EDAP.

B. Impact of Number of Graphene Layers on Cache-Level Performance

As described in Section II-B, the number of available graphene layers strongly affects the resistance per unit length. To quantify the true advantage of graphene for cache-level performance, Fig. 9 shows the cache-level EDP and EDAP for various graphene-based interconnect options under different assumptions in the maximum number of graphene layers. Here, heterogeneous interconnect design schemes (cyan, black, and pink curves) are investigated, namely, using interconnect materials combinations for intra- and intersubarray wires.

In Fig. 9, when the number of graphene layers is large, using graphene for both intra- and intersubarray interconnects provides the best performance in terms of EDP and EDAP due

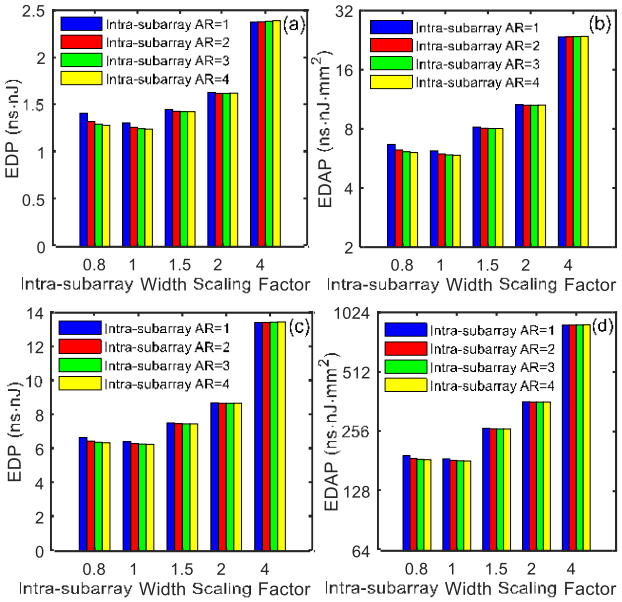

Fig. 8. (a) EDP and (b) EDAP versus intrasubarray graphene interconnect width scaling factor for a variety of aspect ratios. Here, the cache size is 16 MB with associativity of two with four banks using a subarray with 256 rows and 128 columns under the intersubarray interconnect aspect ratio and width scaling factor of 1. (c) and (d) EDP and EDAP for a 128-MB cache with the same configurations as (a) and (b).

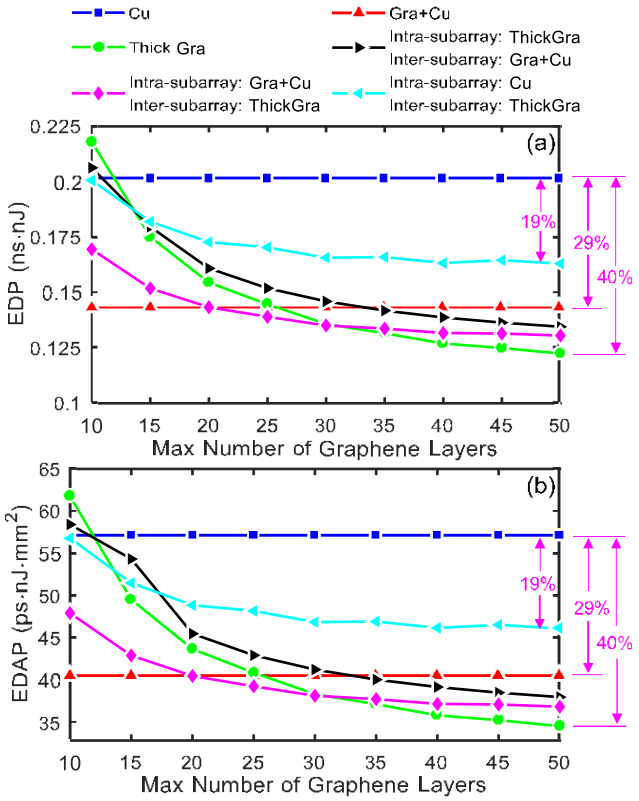

Fig. 9. (a) EDP and (b) EDAP comparison of cache using Cu, graphene-capped Cu, thick graphene, and heterogeneous interconnect design schemes versus the maximum number of graphene layers for a cache size of 0.5 MB for associativity of two with four banks using the subarray with 256 rows and 128 columns under the optimal intra- and intersubarray interconnect aspect ratio and width.

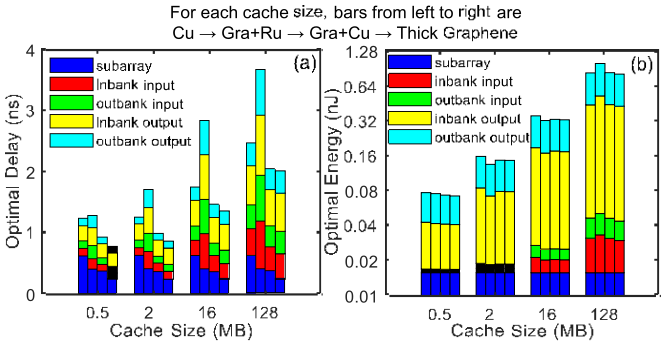

Fig. 10. Optimal (a) access time and (b) read energy (power delay product) per access breakdown bar chart versus cache size under associativity of two with four banks using subarray with 256 rows and 128 columns under the optimal intra- and intersubarray interconnect aspect ratio and width.

to the small delay per unit length, as shown in Figs. 4 and 5. However, the performance of graphene-based cache keeps decreasing as the number of graphene layers decreases due to the increasing resistance, especially for the cache using thick graphene for all interconnects. For the heterogeneous interconnect design scheme of using thick graphene only for intersubarray interconnects (pink curve), the SRAM can provide the best performance when there are ⍰25 layers of graphene layers. This is because the aspect ratio of intrasubarray interconnects is much larger than the intersubarray interconnects due to the narrow BL and WL width. The cache can overcome the limitation of the available number of graphene layers by using graphene-capped Cu for intrasubarray interconnects, while at the same time, taking advantage of the low resistance and capacitance of thick graphene for the intersubarray wires.

In conclusion, the optimal material choice highly depends on the number of available graphene layers. When the number of graphene layers is below 20, graphene-capped Cu is preferred for both intra- and intersubarray interconnects; heterogeneous interconnects using thick graphene only for intersubarray interconnects provide the best EDP when the available number of graphene layers is between 20 and 31, as shown in the pink curve. When the number of graphene layers is above 31, thick graphene can be used for all intermediate layers and up to 40% EDP reduction can be observed compared to the traditional Cu counterpart when 50 graphene layers are available.

C. Impact of Cache Size on Cache Performance

To quantify the impact of cache size on the array-level performance, Fig. 10 shows the optimal delay and energy for four materials under the optimal intra-/intersubarray interconnect aspect ratio and width. In general, the delay and energy increase with the increase in cache size due to the long wires, especially for the H-tree. The subarray energy is insensitive to the cache size due to the similar number of active subarrays. Using thick graphene provides the best delay and energy due to the small resistance and its large MFP and thin geometry. The delay of graphene-capped Ru is large because of its large resistance per unit length, leading to a large EDP,

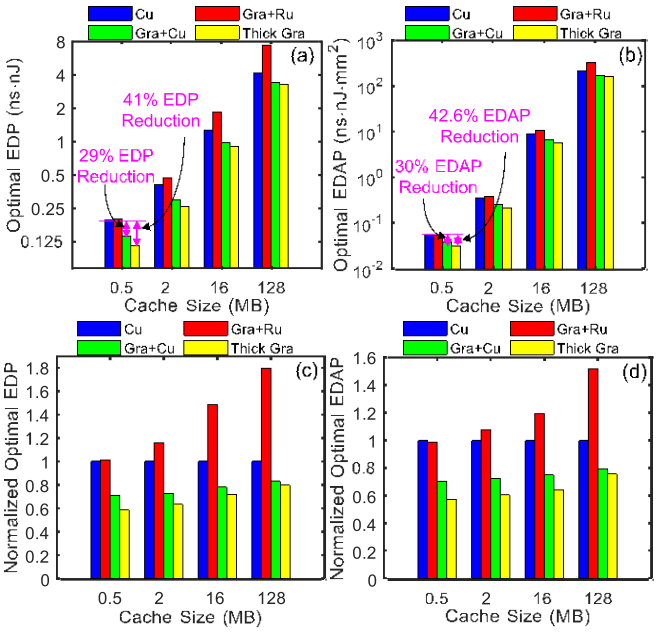

Fig. 11. Optimal (a) EDP and (b) EDAP and normalized (c) EDP and (d) EDAP versus cache size for four material options for associativity of two with four banks using subarray with 256 rows and 128 columns under the optimal intra- and intersubarray interconnect aspect ratio and width.

TABLE II
EDP REDUCTION COMPARED TO BASELINE CU

Cache Size	Gra+Ru	Gra+Cu	Thick Graphene
0.5MB	+2%	-29%	-41%
2MB	+16%	-27%	-37%
16MB	+49%	-22%	-28%
128MB	+79%	-17%	-20%

TABLE III
EDAP REDUCTION COMPARED TO BASELINE CU

Cache Size	Gra+Ru	Gra+Cu	Thick Graphene
0.5MB	-2%	-30%	-43%
2MB	+8%	-28%	-39%
16MB	+20%	-25%	-36%
128MB	+52%	-20%	-24%

as shown in Fig. 11(a). Graphene-capped Cu is the second-best choice, and up to 29% and 30% reduction in EDP and EDAP, respectively, can be observed compared to the Cu counterpart. The reduction of optimal EDP and EDAP is up to 41% and 42.6% for thick graphene compared to Cu counterparts.

To better visualize the relative performance of different materials for different cache sizes, Fig. 11(c) and (d) shows the normalized EDP and EDAP compared to the baseline Cu interconnect. One can observe that graphene-based interconnects provide a larger improvement at a smaller cache size. This is because the delay and energy are more dominated by the subarray, and graphene can provide more advan-tages for the intrasubarray-level interconnects compared to the intersubarray-level H-tree interconnects. The reduction of EDP and EDAP in percentage for graphene-based interconnect compared to the baseline Cu is shown in Tables II and III.

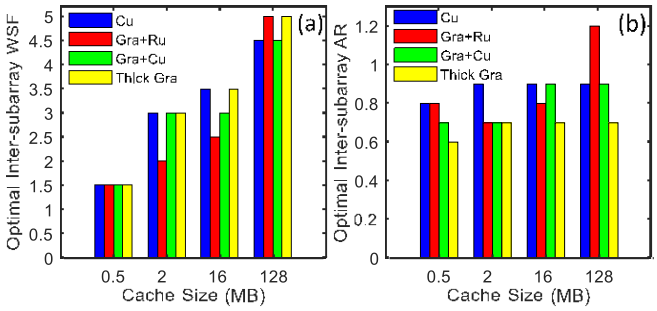

Fig. 12. Optimal intersubarray interconnect (a) width scaling factor and (b) aspect ratio versus cache size with associativity of two with four banks using subarray with 256 rows and 128 columns under the optimal intra- and intersubarray interconnect aspect ratio and width.

To visualize the optimal interconnect design parameters, Fig. 12 shows the optimal intersubarray interconnect width scaling factor and aspect ratio under given cache sizes for different materials. In general, a larger cache prefers to use a wide interconnect to reduce the delay overhead from long H-tree interconnects, while a small cache prefers a narrow width to reduce the area overhead. From Fig. 12(b), the optimal aspect ratio slightly increases with the cache size to reduce the interconnect resistance and properly balance the interconnect delay and energy dissipation. The large MFP and small resistance of graphene wires prefer to use a smaller aspect ratio compared to other materials, which can save energy dissipation.

IV. CONCLUSION

An interconnect-subarray-cache co-design framework is developed to efficiently optimize interconnect technologies to maximize cache-level performance. The available number of graphene layers has a large impact on the cache performance in terms of overall EDP. Under a limited number of graphene layers, using heterogeneous interconnects, where different materials are used for intra- and intersubarray levels, can provide the best performance in terms of EDP and EDAP. The cache-level performance of SRAM using thick graphene interconnects is the best among the four material options, and up to 41% and 42.6% reductions in EDP and EDAP, respectively, can be observed compared to Cu counterparts. Furthermore, a large cache prefers to use wide intersubarray interconnects with a large aspect ratio to maximize the cache-level performance.

REFERENCES

[1] R. Brain, "Interconnect scaling: Challenges and opportunities," in *IEDM Tech. Dig.*, Dec. 2016, pp. 9.3.1–9.3.4, doi: 10.1109/IEDM.2016.7838381.

[2] G. Bonilla, N. Lanzillo, C.-K. Hu, C. J. Penny, and A. Kumar, "Interconnect scaling challenges, and opportunities to enable system-level performance beyond 30 nm pitch," in *IEDM Tech. Dig.*, Dec. 2020, p. 20, doi: 10.1109/IEDM13553.2020.9372093.

[3] D. Prasad, A. Ceyhan, C. Pan, and A. Naeemi, "Adapting interconnect technology to multigate transistors for optimum performance," *IEEE Trans. Electron Devices*, vol. 62, no. 12, pp. 3938–3944, Dec. 2015, doi: 10.1109/TED.2015.2487888.

[4] K. Cho et al., "SRAM write- and performance-assist cells for reducing interconnect resistance effects increased with technology scaling," *IEEE J. Solid-State Circuits*, vol. 57, no. 4, pp. 1039–1048, Apr. 2022, doi: 10.1109/JSSC.2021.3138785.

[5] S. Achra et al., "Characterization of interface interactions between graphene and ruthenium," in *Proc. IEEE Int. Interconnect Technol. Conf. (IITC)*, Oct. 2020, pp. 133–135, doi: 10.1109/IITC47697.2020.9515595.

[6] 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.

[7] 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.

[8] S. Achra et al., "Graphene-ruthenium hybrid interconnects," presented at the IEEE Int. Interconnect Technol. Conf. (IITC), Brussels, Belgium, Jan. 2019.

[9] V. Huang, D. Shim, H. Simka, and A. Naeemi, "From interconnect materials and processes to chip level performance: Modeling and design for conventional and exploratory concepts," in *IEDM Tech. Dig.*, Dec. 2020, p. 32, doi: 10.1109/IEDM13553.2020.9371945.

[10] S. Dutta et al., "Highly scaled ruthenium interconnects," *IEEE Electron Device Lett.*, vol. 38, no. 7, pp. 949–951, Jul. 2017.

[11] O. V. Pedreira et al., "Reliability study on cobalt and ruthenium as alternative metals for advanced interconnects," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2017, pp. 6B-2.1–6B-2.8, doi: 10.1109/IRPS.2017.7936340.

[12] M. H. van der Veen et al., "Damascene benchmark of Ru, Co and Cu in scaled dimensions," in *Proc. IEEE Int. Interconnect Technol. Conf. (IITC)*, Jun. 2018, pp. 172–174.

[13] F. Griggio et al., "Reliability of dual-damascene local interconnects featuring cobalt on 10 nm logic technology," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2018, pp. 6E. 3-1–6E. 3-5, doi: 10.1109/IRPS.2018.8353641.

[14] 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.

[15] S. Hu, J. Chen, N. Yang, and B. Li, "Thermal transport in graphene with defect and doping: Phonon modes analysis," *Carbon*, vol. 116, pp. 139–144, May 2017, doi: 10.1016/j.carbon.2017.01.089.

[16] 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.

[17] A. Hazra and S. Basu, "Graphene nanoribbon as potential on-chip interconnect material—A review," *C*, vol. 4, no. 3, p. 49, Aug. 2018, doi: 10.3390/c4030049.

[18] M. K. Gupta et al., "A comprehensive study of nanosheet and forksheet SRAM for beyond N5 node," *IEEE Trans. Electron Devices*, vol. 68, no. 8, pp. 3819–3825, Aug. 2021, doi: 10.1109/TED.2021.3088392.

[19] 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*, Feb. 2022, pp. 42–44, doi: 10.1109/ISSCC42614.2022.9731673.

[20] R. M. Rao et al., "POWER10: A 16-core SMT8 server processor with 2TB/s off-chip bandwidth in 7nm technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2022, pp. 48–50, doi: 10.1109/ISSCC42614.2022.9731594.

[21] A. Nayak et al., "A 5nm 3.4GHz tri-gear ARMv9 CPU subsystem in a fully integrated 5G flagship mobile SoC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2022, pp. 50–52, doi: 10.1109/ISSCC42614.2022.9731604.

[22] 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.

[23] 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, Jul. 2017, doi: 10.1145/3085572.

[24] (2012). *Predictive Technology Model (PTM)*. [Online]. Available: https://ptm.asu.edu

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] I. Ciofi et al., "Impact of wire geometry on interconnect RC and circuit delay," *IEEE Trans. Electron Devices*, vol. 63, no. 6, pp. 2488–2496, Jun. 2016, doi: 10.1109/TED.2016.2554561.

[31] S. Datta, *Quantum Transport: Atom to Transistor*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

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

[33] 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.

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

[35] S. Y. Sherazi et al., "Standard-cell design architecture options below 5nm node: The ultimate scaling of FinFET and Nanosheet," *Proc. SPIE*, vol. 10962, Mar. 2019, Art. no. 1096202.

[36] H.-H. Liu et al., "Extended methodology to determine SRAM write margin in resistance-dominated technology node," *IEEE Trans. Electron Devices*, vol. 69, no. 6, pp. 3113–3117, Jun. 2022, doi: 10.1109/TED.2022.3165738.

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

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

[39] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012, doi: 10.1109/TCAD.2012.2185930.

[40] S. S. Sapatnekar, "RC interconnect optimization under the Elmore delay model," in *Proc. 31st Annu. Conf. Design Autom. Conf. (DAC)*, 1994, pp. 387–391, doi: 10.1145/196244.196430.

[41] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, "CACTI 5.1," in *HP Laboratories*. Santa Clara, CA, USA: Palo Alto, 2008.