Graphene-based Interconnect Exploration for Large SRAM Caches for Ultra-Scaled Technology Nodes

Journal:	*Transactions on Electron Devices*
Manuscript ID	TED-2022-08-2097-R
Manuscript Type:	Regular
Date Submitted by the Author:	25-Aug-2022
Complete List of Authors:	Pei, Zhenlin; The University of Texas at Arlington, Electrical Engineering Mayahinia, Mahta; Karlsruhe Institute of Technology Liu, Hsiao-Hsuan; IMEC, Tahoori, Mehdi; Karlsruher Institut fur Technologie, Institute for Nanotechnology Catthoor, Francky; IMEC, Tokei, Zsolt; IMEC Pan, Chenyun; University of Texas at Arlington, Electrical Engineering
Area of Expertise:	Graphene, SRAM, Design/technology Co-optimization, Performance Benchmarking, Heterogeneous Interconnect, Subarray

Graphene-based Interconnect Exploration for Large SRAM Caches for Ultra-Scaled 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 Cu interconnect thanks to their great electric properties. In this paper, an interconnect-memory co-design framework is developed to efficiently optimize various graphene-based interconnect technologies. Four interconnect materials and two 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.**

***Keywords*—Graphene, heterogeneous interconnect, SRAM, benchmarking, design/technology co-optimization.**

I. Introduction

The technology research becomes more “interconnect-centric” at sub-10 nm nodes due to the large resistance of traditional Cu interconnect caused by the ever-increasing size effect and impact of barrier thickness on wire and via resistance [1-4]. To address the challenges of the interconnect, enormous research efforts have been performed, including those proposing new interconnect materials, such as Cobalt, Graphene, and Ruthenium [5, 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) [7, 8]. 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 (i) large contact resistance, leading to a limited usage for very short interconnects at local levels, and (ii) a limited number of available graphene layers due to the fabrication process, which leads to a large resistance and limits 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 [9]. In addition, the interconnect of an SRAM array consists of wordlines (WLs), 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.

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 [10-12]. However, predictive models are known to have a limited accuracy due to their extrapolation for key device-level parameters. The predictive technology model 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 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 paper, 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 trade-offs 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 [10], 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 advantage of graphene, we propose two heterogeneous interconnect design schemes, and different interconnect materials are used at ***intra-subarray*** level (e.g. BLs and WLs) and ***inter-subarray*** level (e.g. H-trees). Key trade-offs 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 below.

- An efficient interconnect/cache co-design framework is developed by incorporating realistic device technology and subarray design for ultra-scaled technology nodes.
- Four promising graphene-based interconnect options are benchmarked against Cu counterparts, and a large design

Manuscript received in August 2022. This work is supported by IMEC.

Zhenlin Pei (zhenlin.pei@mavs.uta.edu) and Chenyun Pan are with the Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX 76010, USA. 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 with the Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven (KU Leuven), 3000 Leuven, Belgium. Zsolt Tokei is with IMEC, 3001 Leuven, Belgium.

space is explored to maximize the potential benefits of graphene for the cache-level performance.

- We propose two heterogeneous interconnect design schemes to fully utilize graphene and quantify the impact of the number of graphene layers on the cache-level performance.
- 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.

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) Copper (Cu) for the baseline, (2) graphene-capped Ruthenium (Ru), (3) graphene-capped Cu, and (4) thick graphene [10, 13-21].

For the baseline Cu interconnect, its resistivity model follows existing work with a side wall specularity of 0.5 and a grain boundary reflectivity of 0.5 that are calibrated based on experimental data [17, 22]. For general graphene-based interconnects, the current flowing through a single-layer graphene is obtained by the Landauer formula [23], which is a function of the effective mean-free-path (MFP) of graphene. The MFP depends on several factors, including the graphene edge roughness and substrate material property. In the previous work [10], the MFP has been fitted based on the mobility extracted based on experimental data by the semiclassical equation (1) [24, 25]

$$\sigma = en\mu = \frac{2e^2}{h}\left(\sqrt{\pi n}\cdot MFP\right) \qquad (1)$$

where n is carrier density, σ is conductivity, and μ is mobility. At $n \approx 4.85 \times 10^{10} cm^{-2}$, the MFP is set as $460nm$ based on the existing work [14], which scales up with the increase of width due to the edge scattering. Assuming side contacts are used, the graphene contact resistance is given by $R_{con}/(WN_{layer})$, where W is the interconnect width, N_{layer} is the number of graphene layers, and R_{con} is the contact resistance of 100 Ω·µm [21].

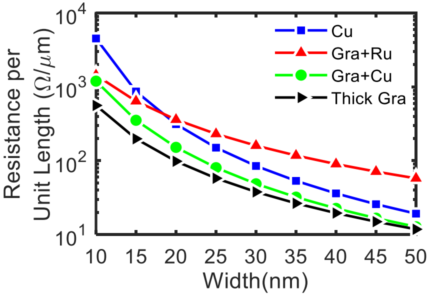

Fig. 1. Resistance per unit length versus width for four interconnect materials with an aspect ratio of 1.

For graphene-capped Ru, the resistance per unit length is extracted based on experimental data for different thicknesses [15]. 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 [18-20]. To compare different interconnect materials under a given aspect ratio of 1, the resistance per unit length is shown in Fig. 1, where thick graphene provides the best resistance at a small width thanks to its large MFP. The capacitance per unit length of the interconnect is extracted by Synopsys Raphael for various interconnect geometry [26]. 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 [10, 11], and the circuit schematic is shown in Fig. 2.

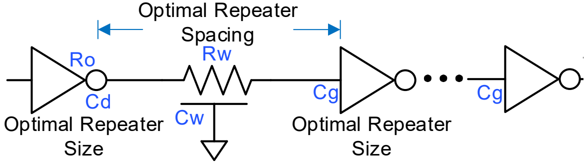

Fig. 2. Circuit model for interconnects with optimal repeater insertion. R_o is the the repeater output resistance, C_d and C_g are device drain and gate capacitance, respectively, R_w and C_w are interconnect resistance and capacitance, respectively, and addional contact and quantum resistance R_{con}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

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

The resistance per unit length shown in Fig. 1 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 intra-subarray 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, namely using a combination of graphene-capped Cu and thick graphene for inter- or intra-subarray levels depending on the width and aspect ratio.

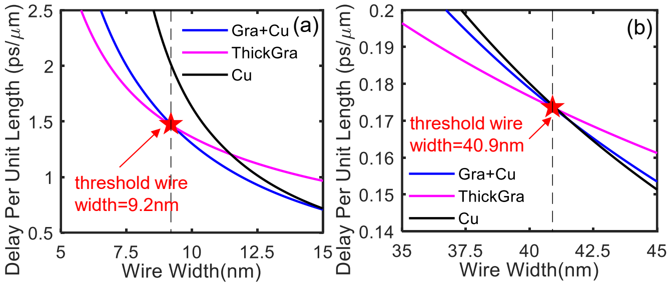

Fig. 3. Delay per unit length under the optimal repeater size and spacing versus interconnect width for graphene-capped Cu and thick graphene under the aspect ratio of 1 with a max number of graphene layers of (a) 10 and (b) 100.

Based on the interconnect repeater insertion model described in the previous subsection, Fig. 3 shows the delay per unit length versus the interconnect width under the optimal repeater insertion with a max number of graphene layers of 10 and 100. For the thick graphene with 10 layers, a threshold interconnect 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 max number of graphene layers increases to 100 as shown in Fig. 3 (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 interconnect width at intra- and inter-subarray levels. If the width is below (or above) the threshold, thick graphene (or graphene-capped Cu) will be used for that level.

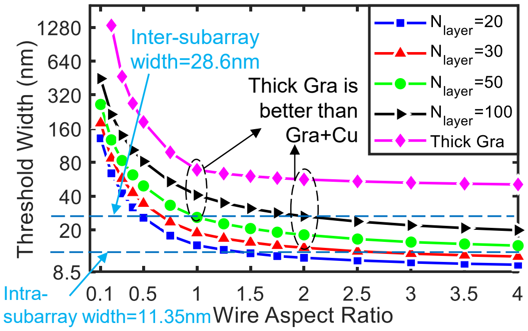

Fig. 4. Threshold width versus aspect ratio of interconnects with optimal repeater insertion at the max number of graphene layers.

In Fig. 3, only two different max numbers of graphene layers, 10 and 100, are investigated under a fixed aspect ratio of 1, and in Fig. 4, we sweep the aspect ratio and 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. 4, designers can obtain valuable information regarding how to choose the best material based on the interconnect geometry and number of available graphene layers. For example, two dashlines in Fig. 4 show the intra- and inter-subarray interconnect widths. For intra-subarray 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 inter-subarray interconnects with an aspect ratio of 1, the minimum number of graphene layers to outperform graphene-capped Cu counterpart increases to 50.

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 trade-off 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 two heterogeneous design schemes, where the intra- and inter-subarray level interconnects use different combinations of the interconnect materials, to maximize cache-level performance.

C. Subarray Models

To enable a fast, accurate, and flexible analysis of the latency and energy dissipation of large cache modules, we have developed a high-level equation-based model for the SRAM-based . 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. Device parameters, including gate capacitance, drain capacitance, ON current, supply voltage, temperature-dependent leakage current, and threshold voltage are extracted by Cadence Spectre and Synopsys HSPICE simulations [27, 28]. The overall methodology of our proposed high-level modeling for the energy and latency are shown in Fig. 5.

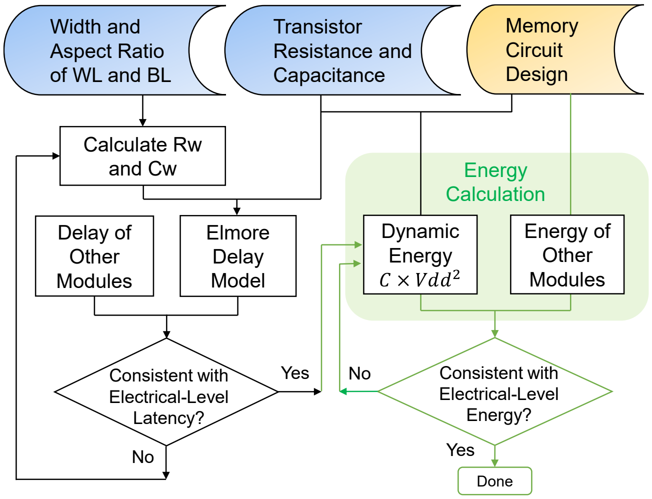
Width and Aspect Ratio of WL and BL
Transistor Resistance and Capacitance
Memory Circuit Design
Calculate Rw and Cw
Energy Calculation
Delay of Other Modules
Elmore Delay Model
Dynamic Energy $C \times Vdd^2$
Energy of Other Modules
Consistent with Electrical-Level Latency?
Yes
No
No
Consistent with Electrical-Level Energy?
Yes
Done

Fig. 5. Design methodology for the high-level modeling of the latency and energy of the SRAM subarray.

1) Modeling of the Subarray-Level Latency

The first step toward such high-level modelling is approximating the subarray-level latency. In this regard, we model the read and write latencies as follows:

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

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

where t_{decode}, t_{WL}, t_{BL}, and $t_{bit-flip}$ are the decoder latency, charging latency for the wordline (WL) and bitline (BL), and bit-flip latency, respectively. The calculation of the decoder latency is fully aligned with mature high-level memory simulators [29]. 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 the high-level modeling, besides accuracy, speed and flexibility are also essential. In this regard, our latency approximation only covers the dominant latency terms, and it is in good agreement with the electrical level simulation.

Based on the induced RC load on the WL and BL, the charging latency 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 a contribution to the capacitive load, which is characterized based on electrical-level simulations. In this work, we have modeled the BL and WL as an RC network. Hence, the latency term corresponding to WL and BL charge latency can be calculated through the Elmore delay equation. For calculating the interconnect parasitic resistance and capacitance, we have reused the equation from the existing work [29]. 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 [22].

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 non-negligible 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 [29].

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) and 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 the higher-level design exploration. Therefore, high-level simulation of 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. 5, the terms of the main energy contributions are determined accurately. Once the high-level model of the energy closely converges to the energy through the electrical simulation, *fitting parameters* (A_{read} and A_{write}) can be applied to the high-level model of the energy. The following equations show the model for the write and read energy:

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

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

where A_{write} and A_{read} are the fitting parameters for write and read energy, respectively. $E_{decoder}$, $E_{timingCtrl(write)}$, $E_{timingCtrl(read)}$, E_{WL}, E_{BL}, $E_{bit-flip}$, $E_{WriteDriverEN}$, E_{SA}, E_{colMux}, $E_{outLatch}$, and $E_{leakage}$ are the decoder energy, energy for the write/read timing control, wordline (WL), bitline (BL), bit-flip, write driver enable, sense amplifier, column Mux, and output latch, and leakage energy, respectively.

D. *Cache-Level Memory Models*

An open-source and well-known simulator, CACTI, is adopted to optimize the SRAM array [11, 30]. CACTI sweeps the cache organization parameters to obtain optimal parameters for target metrics the user 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. The original CACTI has been already validated by SPICE simulation and reported data from the commercial caches for Intel 65nm L3 cache and Sun SPARC 90nm L2 cache [30]. With a validated cache simulator, various interconnect configurations and organization parameters can be efficiently explored at the early design stage with a 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. In this work, critical trade-offs 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 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 graphene-capped Cu and thick graphene are adopted for intra- and inter-subarray levels, respectively) and four cache sizes (i.e., 0.5MB, 2MB, 16MB, and 128MB) in the case study.

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 are investigated for intra- and inter-subarray wires to maximize the SRAM array-level performance.

Under different intra-subarray interconnect aspect ratio assumptions, the breakdown bar charts of access time and energy for different interconnect widths are shown in Fig. 6. 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 performance. The scaling factor of 1 corresponds to an intra-subarray interconnect width of 11.35 nm. Here, the default cache size is 16MB with 4 banks using the subarray with 128 columns, and the associativity is 2 under the inter-subarray 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. 6 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. 6 (a), when the intra-subarray interconnect width is small, the delay improves with the increase of the width thanks 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 **(i)** limited by the access time if the intra-subarray interconnect width is too small or large or **(ii)** limited by the energy if the width is too large. Note that for the energy per access in Fig. 6 (b), the y axis is shown in log scale due to the large span in different energy components.

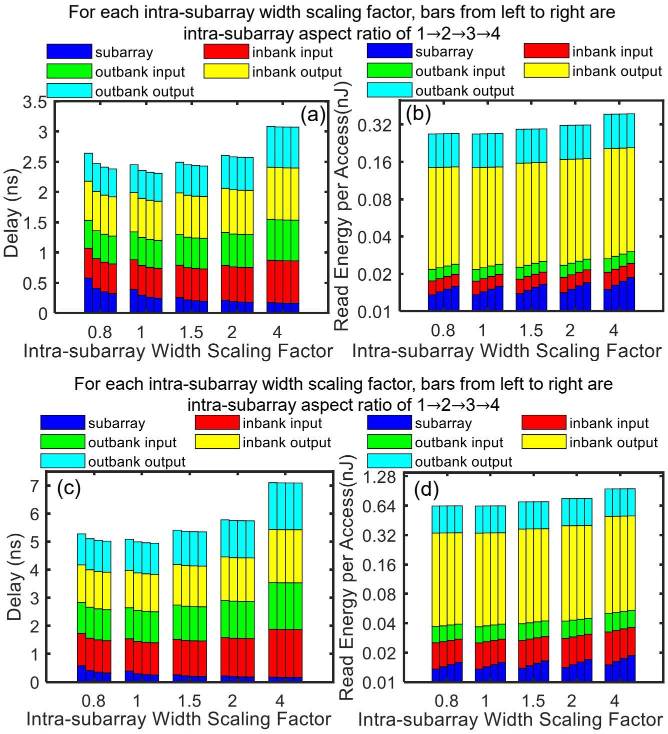

Fig. 6. (a) access time and (b) read energy (power delay product) per access breakdown bar chart versus intra-subarray width scaling factor for a variety of aspect ratio using thick graphene interconnects. Here, the cache size is 16MB with associativity of 2 with 4 banks using subarray with 128 columns under the inter-subarray interconnect aspect ratio and width scaling factor of 1. (c) and (d) show the access time and read energy (power delay product) per access for a 128MB cache with the same configurations as (a) and (b).

Fig. 6 (c) and (d) show the delay and energy for a larger cache size of 128 MB. Compared to Fig. 6 (a) and (b), the overall trend is similar, except for the fact that the delay and energy contributions from the H-tree inside the bank increase due to the increase of the bank size. To take both delay and energy into account, the EDP versus the intra-subarray interconnect width scaling factor for different aspect ratios is shown in Fig. 7 (a) and (c), where an optimal width exists to minimize the EDP. Under the consideration of array area, optimal intra-subarray width and aspect ratio of interconnects exists to minimize the EDAP, as shown in Fig. 7 (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. 8 shows the cache-level EDP for various graphene-based interconnect options under different assumptions in the max number of graphene layers. Here, two heterogeneous interconnect design schemes (black and pink curves) are investigated, namely using two interconnect materials combinations for intra- and inter-subarray wires.

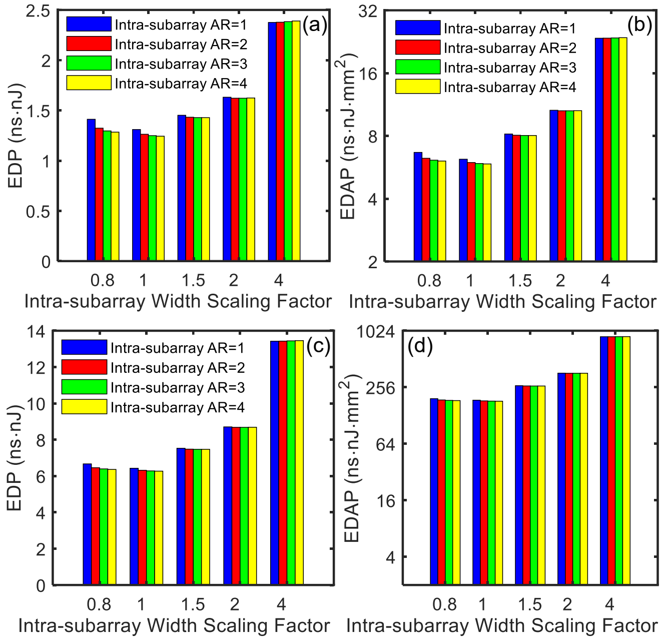

Fig. 7. (a) EDP and (b) EDAP versus intra-subarray graphene interconnect width scaling factor for a variety of aspect ratio. Here, the cache size is 16MB with associativity of 2 with 4 banks using subarray with 128 columns under the inter-subarray interconnect aspect ratio and width scaling factor of 1. (c) and (d) show the EDP and EDAP for a 128MB cache with the same configurations as (a) and (b).

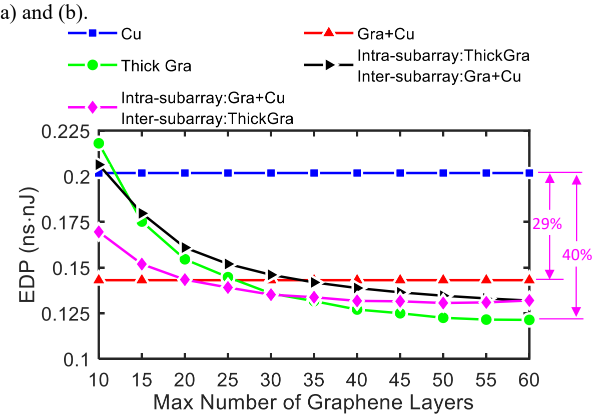

Fig. 8. EDP comparison of cache using Cu, graphene-capped Cu, thick graphene, and heterogeneous interconnect design schemes versus the max number of graphene layers for a cache size of 0.5MB for an associativity of 2 with 4 banks using the subarray with 128 columns under the optimal intra- and inter-subarray interconnect aspect ratio and width.

In Fig. 8, when the number of graphene layers is large, using graphene for both intra- and inter-subarray interconnects provides the best performance due to the small delay per unit length as shown in Fig. 3 and Fig. 4. 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 inter-subarray interconnects (pink curve), the SRAM can provide the best performance when there is ~25 layers of graphene layers. This is because the aspect ratio of intra-subarray interconnects is much larger than the inter-subarray interconnects due to the narrow bitline and wordline width. The cache can overcome the limitation of the available number of graphene layers by using graphene-capped Cu for intra-subarray interconnects while at the same time taking advantage of the low resistance and capacitance of thick graphene for the inter-subarray 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 inter-subarray interconnects; heterogeneous interconnects using thick graphene only for inter-subarray interconnects provide the best EDP when the available number of graphene layers is between 20 and 30, as shown in the pink curve. When the number of graphene layers is above 30, 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 60 graphene layers are available.

C. *Impact of Cache Size on Cache Performance*

To quantify the impact of cache size at the array-level performance, Fig. 9 shows the optimal delay and energy for four materials under the optimal intra-/inter-subarray interconnect aspect ratio and width. In general, the delay and energy increase with the increase of 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 thanks to 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, as shown in Fig. 10 (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.

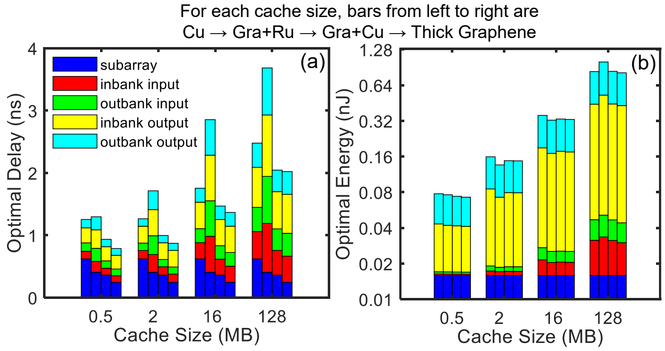

Fig. 9. Optimal (a) access time and (b) read energy (power delay product) per access breakdown bar chart versus cache size under associativity of 2 with 4 banks using subarray with 128 columns under the optimal intra- and inter-subarray interconnect aspect ratio and width.

To better visualize the relative performance of different materials for different cache sizes, Fig. 10 (c) and (d) show 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 advantages for the intra-subarray level interconnects compared to the inter-subarray level H-tree interconnects.

To visualize the optimal interconnect design parameters, Fig. 11 shows the optimal inter-subarray 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. 11 (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.

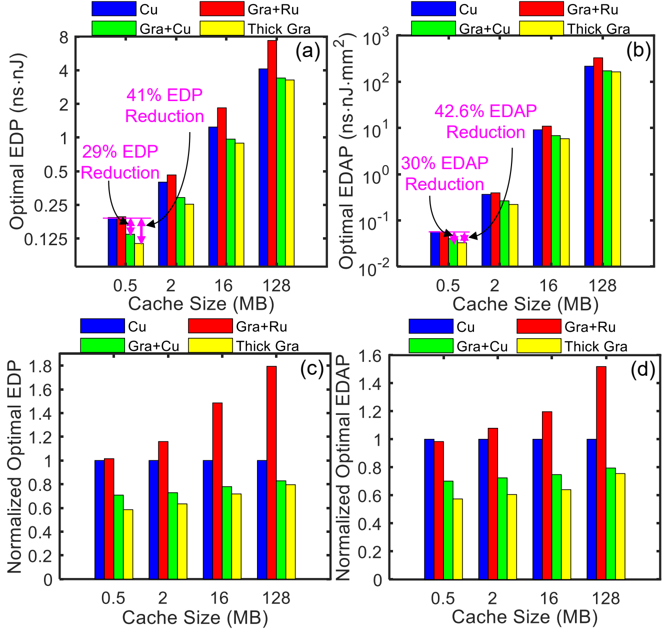

Fig. 10. Optimal (a) EDP and (b) EDAP and normalized (c) EDP and (d) EDAP versus cache size for four material options for an associativity of 2 with 4 banks using subarray with 128 columns under the optimal intra- and inter-subarray interconnect aspect ratio and width.

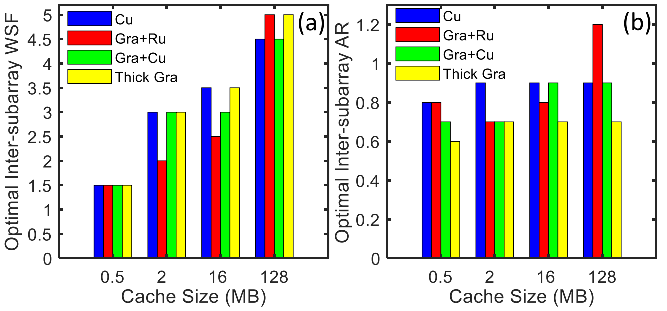

Fig. 11. Optimal inter-subarray interconnect (a) width scaling factor and (b) aspect ratio versus cache size with an associativity of 2 with 4 banks using subarray with 128 columns under the optimal intra- and inter-subarray interconnect aspect ratio and width.

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 inter-subarray 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 inter-subarray interconnects with a large aspect ratio to maximize the cache-level performance.

REFERENCES

[1] R. Brain, "Interconnect scaling: Challenges and opportunities," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 9.3. 1-9.3. 4. DOI: 10.1109/IEDM.2016.7838381

[2] G. Bonilla, N. Lanzillo, C.-K. Hu, C. Penny, and A. Kumar, "Interconnect scaling challenges, and opportunities to enable system-level performance beyond 30 nm pitch," in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 20.4. 1-20.4. 4. 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 Transactions on Electron Devices,* vol. 62, pp. 3938-3944, 2015. DOI: 10.1109/TED.2015.2487888

[4] K. Cho, H. Choi, I. J. Jung, J. Oh, T. W. Oh, K. Kim, G. Kim, T. Choi, C. Sim, and T. Song, "SRAM write-and performance-assist cells for reducing interconnect resistance effects increased with technology scaling," *IEEE Journal of Solid-State Circuits,* vol. 57, pp. 1039-1048, 2022. DOI: 10.1109/JSSC.2021.3138785

[5] S. Dutta, S. Kundu, A. Gupta, G. Jamieson, J. F. G. Granados, J. Bömmels, C. J. Wilson, Z. Tőkei, and C. Adelmann, "Highly scaled ruthenium interconnects," *IEEE Electron Device Letters,* vol. 38, pp. 949-951, 2017. DOI: 10.1109/LED.2017.2709248

[6] M. H. van der Veen, K. Vandersmissen, D. Dictus, S. Demuynck, R. Liu, X. Bin, P. Nalla, A. Lesniewska, L. Hall, and K. Croes, "Cobalt bottom-up contact and via prefill enabling advanced logic and DRAM technologies," in *2015 IEEE International Interconnect Technology Conference and 2015 IEEE Materials for Advanced Metallization Conference (IITC/MAM)*, 2015, pp. 25-28. DOI: 10.1109/IITC-MAM.2015.7325605

[7] 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, 2017. DOI: 10.1016/j.carbon.2017.01.089

[8] 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 Transactions on Electron Devices,* vol. 62, pp. 1530-1536, 2015. DOI: 10.1109/TED.2015.2409875

[9] M. K. Gupta, P. Weckx, P. Schuddinck, D. Jang, B. Chehab, S. Cosemans, J. Ryckaert, and W. Dehaene, "A comprehensive study of nanosheet and forksheet SRAM for beyond N5 node," *IEEE Transactions on Electron Devices,* vol. 68, pp. 3819-3825, 2021. DOI: 10.1109/TED.2021.3088392

[10] Z. Pei, F. Catthoor, Z. Tokei, and C. Pan, "Beyond-Cu Intermediate-Length Interconnect Exploration for SRAM Application," *IEEE Transactions on Nanotechnology,* 2022. DOI: 10.1109/TNANO.2022.3157952

[11] R. Balasubramonian, A. B. Kahng, N. Muralimanohar, A. Shafiee, and V. Srinivas, "CACTI 7: New tools for interconnect exploration in innovative off-chip memories," *ACM Transactions on Architecture and Code Optimization (TACO),* vol. 14, pp. 1-25, 2017. DOI: 10.1145/3085572

[12] Predictive Technology Model (PTM), available online at https://ptm.asu.edu, 2012.

[13] T. Nogami, "Overview of interconnect technology for 7nm node and beyond-New materials and technologies to extend Cu and to enable alternative conductors," in *2019 Electron Devices Technology and Manufacturing Conference (EDTM)*, 2019, pp. 38-40. DOI: 10.1109/EDTM.2019.8731225

[14] S. Achra, X. Wu, V. Trepalin, T. Nuytten, J. Ludwig, S. Brems, V. Afanas'ev, C. Huyghebaert, B. Soree, M. Heyns, I. Asselberghs, and Z. Tőkei, "Characterization of Interface Interactions between Graphene and Ruthenium," presented at the IEEE International Interconnect Technology Conference (IITC), San Jose, California, USA, 2020. DOI: 10.1109/IITC47697.2020.9515595

[15] S. Achra, I. Asselberghs, X. Wu, S. Brems, C. Huyghebaert, B. Sorée, M. Heyns, and Z. Tokei, "Graphene-Ruthenium hybrid interconnects," presented at the IEEE International Interconnect Technology Conference (IITC), Brussels, Belgium, 2019.

[16] X. Zhang, H. Huang, R. Patlolla, W. Wang, F. W. Mont, J. Li, C.-K. Hu, E. G. Liniger, P. S. McLaughlin, and C. Labelle, "Ruthenium interconnect resistivity and reliability at 48 nm pitch," in *2016 IEEE International Interconnect Technology Conference/Advanced Metallization Conference (IITC/AMC)*, 2016, pp. 31-33. DOI: 10.1109/IITC-AMC.2016.7507650

[17] C. Pan and A. Naeemi, "A Proposal for a Novel Hybrid Interconnect Technology for the End of Roadmap," *Electron Device Letters, IEEE,* vol. 35, pp. 250-252, 2014. DOI: 10.1109/LED.2013.2291783

[18] H. C. Lee, M. Jo, H. Lim, M. S. Yoo, E. Lee, N. N. Nguyen, S. Y. Han, and K. Cho, "Toward near-bulk resistivity of Cu for next-generation nano-interconnects: Graphene-coated Cu," *Carbon,* vol. 149, pp. 656-663, 2019. DOI: https://doi.org/10.1016/j.carbon.2019.04.101

[19] S. Achra, X. Wu, V. Trepalin, T. Nuytten, J. Ludwig, V. Afanas' ev, S. Brems, B. Sorée, Z. Tokei, and M. Heyns, "Metal induced charge transfer doping in graphene-ruthenium hybrid interconnects," *Carbon,* vol. 183, pp. 999-1011, 2021. DOI: https://doi.org/10.1016/j.carbon.2021.07.070

[20] T. Yu, E.-K. Lee, B. Briggs, B. Nagabhirava, and B. Yu, "Bilayer graphene/copper hybrid on-chip interconnect: A reliability study," *IEEE transactions on nanotechnology,* vol. 10, pp. 710-714, 2010. DOI: 10.1109/TNANO.2010.2071395

[21] W. S. Leong, H. Gong, and J. T. Thong, "Low-contact-resistance graphene devices with nickel-etched-graphene contacts," *ACS nano,* vol. 8, pp. 994-1001, 2014. DOI: https://doi.org/10.1021/nn405834b

[22] I. Ciofi, A. Contino, P. J. Roussel, R. Baert, V.-H. Vega-Gonzalez, K. Croes, M. Badaroglu, C. J. Wilson, P. Raghavan, and A. Mercha, "Impact of wire geometry on interconnect RC and circuit delay," *IEEE Transactions on Electron Devices,* vol. 63, pp. 2488-2496, 2016. DOI: 10.1109/TED.2016.2554561

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

[24] A. Contino, I. Ciofi, R. Baert, X. Wu, I. Asselberghs, U. Celano, C. J. Wilson, Z. Tokei, G. Groeseneken, and B. Soree, "Circuit Delay and Power Benchmark of Graphene against Cu Interconnects," presented at the IEEE International Interconnect Technology Conference (IITC), Brussels, Belgium, 2019.

[25] K. I. Bolotin, K. Sikes, Z. Jiang, M. Klima, G. Fudenberg, J. Hone, P. Kim, and H. Stormer, "Ultrahigh electron mobility in suspended graphene," *Solid State Communications,* vol. 146, pp. 351-355, 2008. DOI: 10.1016/j.ssc.2008.02.024

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

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

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

[29] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 31, pp. 994-1007, 2012. DOI: 10.1109/TCAD.2012.2185930

[30] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, "CACTI 5.1," HP Laboratories, Palo Alto. 2008.