Cross-Layer Design Space Exploration of NVM-based Caches for Deep Learning

Ahmet Inci[1], Mehmet Meric Isgenc[1], Diana Marculescu[1,2]

[1]Carnegie Mellon University [2]The University of Texas at Austin

I. Introduction

Over the last decade, the performance boost achieved through CMOS scaling has plateaued, necessitating sophisticated computer architecture solutions to gain higher performance in computing systems while maintaining a feasible power density. These objectives, however, are concurrently challenged by the limitations of the performance of memory resources [1]. In contrast to the initial insight of Dennard on power density [2], deep CMOS scaling has exacerbated static power consumption, causing the heat density of ICs to reach catastrophic levels unless properly addressed [3].

As computers suffer from memory and power related limitations, the demand for data-intensive applications has been on the rise. With the increasing data deluge and recent improvements in GPU architectures, deep neural networks (DNNs) have achieved remarkable success in various tasks such as image recognition and object detection by utilizing inherent massive parallelism of GPU platforms. However, DNN workloads continue to have large memory footprints and significant computational requirements to achieve higher accuracy. Thus, DNN workloads exacerbate the memory bottleneck which degrades the overall performance of the system. To this end, while deep learning (DL) practitioners focus on model compression techniques [4], system architects investigate GPU architectures to overcome the memory bottleneck problem and improve the overall system performance [5]. We note the current trend of GPU architectures is towards increasing last-level cache capacity. Our analysis shows that conventional SRAM technology incurs scalability problems as far as power, performance, and area (PPA) is concerned [6]. Non-volatile memory (NVM) technology is one of the most promising solutions to tackle memory bottleneck problem for data-intensive applications. However, because much of emerging NVM technology is not available for commercial use, there is an obvious need for a framework to perform design space exploration for these emerging NVM technologies for DL workloads.

In this work, we present *DeepNVM++* [7], an extended and improved framework [8] to characterize, model, and optimize NVM-based caches in GPU architectures for deep learning workloads. Without loss of generality, we demonstrate our framework for spin-transfer torque magnetic random access memory (STT-MRAM) and spin-orbit torque magnetic random access memory (SOT-MRAM), keeping in mind that it can be used for any NVM technology, GPU platform, or deep learning workload. Our cross-layer analysis framework incorporates both circuit-level characterization aspects and the memory behavior of various DL workloads running on an actual GPU platform. *DeepNVM++* enables the evaluation of *power, performance, and area* of NVMs when used for last-level (L2) caches in GPUs and seeks to exploit the benefits of this emerging technology to improve the performance of deep learning applications.

We present both *iso-capacity* and *iso-area* performance and energy analysis for systems whose last-level caches rely on conventional SRAM and emerging STT-MRAM and SOT-MRAM technologies. To perform *iso-capacity* analysis, we carry out extensive memory profiling of various deep learning workloads for both training and inference on existing GPU platforms. For the *iso-area* analysis, existing platforms cannot be used for varying cache sizes, so we rely on architecture-level simulation of GPUs to quantify and better understand last-level cache capacity and off-chip memory accesses. In both cases, our framework automatically combines resulting memory statistics with circuit and microarchitecture-level characterization and analysis of emerging NVM technologies to gauge their impact on DL workloads running on future GPU-based platforms.

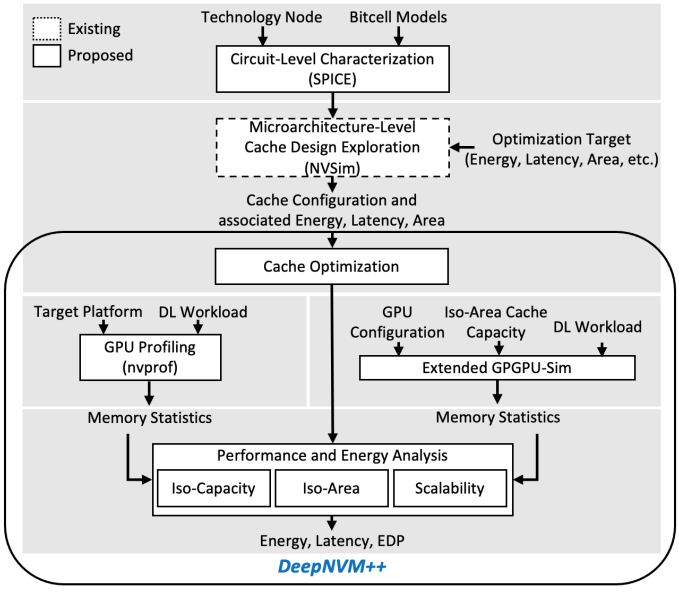

Fig. 1: Overview of the cross-layer analysis flow

We also perform a scalability analysis and compare SRAM, STT-MRAM, and SOT-MRAM for various cache capacities in terms of area, latency, and energy results. Next, we evaluate and show how NVM-based caches behave in terms of performance and energy when compared to conventional SRAM-based caches for DL workloads in a scalability analysis. Our comprehensive cross-layer framework is demonstrated on STT-/SOT-MRAM technologies and can be used for the characterization, modeling, and analysis of any NVM technology for last-level caches in GPUs for DL applications.

II. Related Work and Paper Contributions

Although 16nm has become a commonplace technology for high-end customers of foundries, an intriguing inflection point awaits the electronics community as we approach the end of the traditional density, power, and performance benefits of CMOS scaling. To move beyond the computing limitations imposed by staggering CMOS scaling trends, MRAM has emerged as a promising candidate.

STT bitcells [9] use a magnetic tunnel junction (MTJ) pillar as their core storage element and an additional access transistor to enable read and write operations. Although STT bitcells offer non-volatility, low read latency, and high endurance, the write current is also high, which increases power consumption. To this end, SOT bitcells have been proposed to overcome the write current challenges by isolating the read and write paths. Because the read disturbance errors are much less likely in SOT bitcells, both read and write access devices can be tuned in accordance with the lower current requirements. The read and write current requirements of STT and SOT bitcells can have a crucial impact on the eventual MRAM characteristics because they affect the CMOS access transistors, bitcell area, and peripheral logic. Thus, a comparison of these bitcells and the traditional SRAM merits a meticulous analysis that take these factors into account.

Prior work has proposed effective approaches to overcome the shortcomings of emerging NVM technologies such as using hybrid SRAM and NVM-based caches that utilize the complementary features of different memory technologies, relaxing non-volatility properties to reduce the high write latency and energy, and implementing cache replacement policies for higher level caches such as L1 caches and register files. However, NVM technology appear to be a better choice for lower level caches such as L2 or L3 caches due to its long write latency and high cell density. Higher level L1 caches are latency-sensitive and optimized for performance, whereas last-level caches are capacity-sensitive and optimized for a high hit rate to reduce off-chip memory accesses. Therefore, NVM-based caches provide a better use case for replacing SRAM in last-level caches due to their high cell density when compared to SRAM-based caches. To this end, we evaluate power, performance, and area of NVM technology when used for last-level caches in GPU platforms.

While prior work has shown the potential of NVM technologies for generic applications to some extent, there is a need for a cross-layer analysis framework to explore the potential of NVM technologies in GPU platforms, particularly for DL workloads. The most commonly used modeling tool for emerging NVM technologies is *NVSim* [10], a circuit-level model for performance, energy, and area estimation. However, *NVSim* is not sufficient to perform a detailed cross-layer analysis for NVM technologies for DL workloads since it does not take architecture-level analysis and application-specific memory behavior into account. In this paper, we incorporate *NVSim* with our cross-layer modeling and optimization flow including novel architecture-level iso-capacity and iso-area analysis flow to perform design space exploration for conventional SRAM and emerging NVM caches for DL workloads. This paper makes the following contributions:

1) **Circuit-level bitcell characterization.** We perform detailed circuit-level characterization combining a commercial 16nm CMOS technology and prominent STT [9] and SOT [11] models from the literature to iterate through our framework in an end-to-end manner to demonstrate the flexibility of *DeepNVM++* for future studies.
2) **Microarchitecture-level cache design exploration.** We use *NVSim* [10] to perform a fair comparison between SRAM, STT-MRAM, and SOT-MRAM by incorporating the circuit-level models developed in 1) using 16nm technology and choosing the best cache configuration for each of them.
3) **Iso-capacity analysis.** To compare the efficacy of MRAM caches to conventional SRAM caches, we perform our novel iso-capacity analysis based on *actual platform profiling* results for the memory behavior of various DNNs by using the *Caffe* framework on an NVIDIA 1080 Ti GPU (implemented in 16nm technology) for the ImageNet dataset.
4) **Iso-area analysis.** Because of their different densities, we compare SRAM and NVM caches in an iso-area analysis to quantify the benefits of higher density of NVM technologies on DL workloads running on GPU platforms. Since existing platforms do not support resulting iso-area cache sizes, we extend the *GPGPU-Sim* [12] to run DL workloads and support larger cache capacities for STT-MRAM and SOT-MRAM.
5) **Scalability analysis.** Finally, we perform a thorough scalability analysis and compare SRAM, STT-MRAM, and SOT-MRAM in terms of power, performance, and area to project and gauge the efficacy of NVM and SRAM-based caches for DL workloads as cache capacity increases.

To the best of our knowledge, putting everything together, *DeepNVM++* is the *first comprehensive framework* for cross-layer characterization, modeling, and analysis of emerging NVM technologies for deep learning workloads running on GPU platforms. Our results show that in the iso-capacity case, STT-MRAM and SOT-MRAM achieve up to $3.8\times$ *and* $4.7\times$ *energy-delay product reduction* and $2.4\times$ *and* $2.8\times$ *area reduction* compared to SRAM baseline, respectively. In the iso-area case, STT-MRAM and SOT-MRAM achieve up to $2\times$ *and* $2.3\times$ *energy-delay product reduction* and accommodate $2.3\times$ *and* $3.3\times$ *cache capacity* compared to SRAM, respectively. We also perform a scalability analysis and show that STT-MRAM and SOT-MRAM achieve orders of magnitude EDP reduction when compared to SRAM for large cache capacities. Our novel framework can be used to further explore the feasibility of emerging NVM technologies for DL applications for different design choices such as technology nodes, bitcell models, DL workloads, cache configurations, optimization targets, and target platforms.

ACKNOWLEDGMENT

This research was supported in part by NSF CSR Grant No. 1815780.

REFERENCES

[1] W. A. Wulf and S. A. McKee, "Hitting the memory wall: Implications of the obvious," *SIGARCH Comput. Archit. News*, 23(1):20-24, 1995.
[2] R. H. Dennard and et al., "Design of ion-implanted mosfet's with very small physical dimensions," *JSSC*, 9(5):256-268, 1974.
[3] A. K. Coskun, T. S. Rosing, and K. Whisnant, "Temperature aware task scheduling in mpsocs," in *DATE*, 2007, pp. 1–6.
[4] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2016.
[5] A. Inci, E. Bolotin, Y. Fu, G. Dalal, S. Mannor, D. Nellans, and D. Marculescu, "The architectural implications of distributed reinforcement learning on cpu-gpu systems," *arXiv preprint arXiv:2012.04210*, 2020.
[6] M. Chang, P. Rosenfeld, S. Lu, and B. Jacob, "Technology comparison for large last-level caches (l3cs): Low-leakage sram, low write-energy stt-ram, and refresh-optimized edram," in *HPCA*, 2013, pp. 143–154.
[7] A. Inci, M. M. Isgenc, and D. Marculescu, "DeepNVM++: cross-layer modeling and optimization framework of non-volatile memories for deep learning," *arXiv preprint arXiv:2012.04559*, 2020.
[8] A. F. Inci, M. M. Isgenc, and D. Marculescu, "DeepNVM: a framework for modeling and analysis of non-volatile memory technologies for deep learning applications," in *DATE*, 2020, p. 1295–1298.
[9] J. Kim, A. Chen, B. Behin-Aein, S. Kumar, J. Wang, and C. H. Kim, "A technology-agnostic mtj spice model with user-defined dimensions for stt-mram scalability studies," in *CICC*, Sept 2015, pp. 1–4.
[10] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *TCAD*, 31(7):994-1007, 2012.
[11] M. Kazemi, G. E. Rowlands, E. Ipek, R. A. Buhrman, and E. G. Friedman, "Compact model for spin–orbit magnetic tunnel junctions," *IEEE Transactions on Electron Devices*, 63(2):848-855, 2016.
[12] A. Bakhoda, G. L. Yuan, W. Fung, and et al., "Analyzing cuda workloads using a detailed gpu simulator," in *ISPASS*, 2009, pp. 163–174.