CCF Transactions on High Performance Computing (2022) 4:367–369
https://doi.org/10.1007/s42514-022-00133-9

EDITORIAL

Editorial for the special issue on memory architectures and systems for modern applications

Guangyu Sun[1] · Liang Shi[2] · Jingtong Hu[3]

Published online: 18 January 2023

Modern data-centric applications, such as artifical artificial intelligence (AI) workloads, graph data analysis, IoT and mobile systems, have emerged and been widely used in our daily life. These applications keep raising the requirements of large capacity, high performance, low power consumption, and high reliability for the whole memory hierarchy, which may be difficult to be satisfied by traditional memory architectures and systems. On the other hand, with the rapid advancement of memory technologies, various emerging memory technologies have been proposed to mitigate the problems mentioned above. These emerging memory technologies have the advantages of high density, low standy power, etc. However, they also face the challenges of programming overhead, limited lifetime, and reliablility issues. Thus, to leverage the uniques features and handle the limitations of these emerging memory technologies, we are expecting innovations in memory architecture and system designs.

We have eight invited papers selected for this special issue based on a peer-review procedure, which cover several different topics that relate to the non-volatile memory systems, computing-in-memory architecture, and domain-aware memory optimization techniques.

Two papers in the first part of the special issue focus on the emerging non-volatile memory, which can provide large memory capacity, long-term data durability, low power consumption for data center applications. One paper (Islam et al. 2022) studies the performance of commercialized NVM device from the storage data structures' perspective. The other one (Chen et al. 2022) proposes a thread scheduling approach to optimize the thread scheduling problem on NVM nodes with the help of reinforcement learning method.

- Researchers (Islam et al. 2022) from Prof. Dong Dai's group study the performance of Intel Optane DC Persistent Memory (Optane DC PMEM) from the storage data structures' perspective. They focus on studying the performance characteristics of the low-level indexing data structures and the high-level graph storage data structures using the new pmemids_bench, which includes seven commonly used indexing data structures and two popular graph data structures implemented in four persistent modes and four parallel modes. Evaluations on real Optane DC-based platform reveal nine observations that cover various aspects of Optane DC programming, providing useful reference for developers to design their persistent applications.
- Prof. Lei Liu et al. address that the exploration space for thread scheduling in the hybrid memory systems using DRAM and NVM is expanding rapidly (Chen et al. 2022). Existing schedulers may not provide efficient scheduling solutions in such complicated cases. To solve the problem, they propose a thread scheduling approach called Smart Scheduler, by leveraging a reinforcement learning method. The experimental results show that the proposed Smart Scheduler can converge faster than rule-based algorithms and scheduling domain methods and reduce program execution time by up to 59.9%. It also outperforms rule-based algorithms and scheduling domain methods by 4.1% and 19.1% in quality of service latency.

The second part of the special issue, consisting of three research papers, focuses on the emerging computing-in-memory/storage architectures. These computing architectures can provide high data access bandwidth, high computing performance, and high energy-efficiency for data-centric

✉ Guangyu Sun
gsun@pku.edu.cn

[1] School of Integrated Circuits, Peking University, Beijing 100871, China

[2] School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

[3] Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260, USA

applications, such as deep neural networks, graph data analysis, and IoT or edge systems.

- The paper (Guo et al. 2022) written by Prof. Xin Si and his collaborators addresses the problem that most Computing-In-Memory (CIM) works lack configurability regardless of custom demands. To tackle this issue, they propose a 28 nm 128 Kb configurable CIM architecture based on voltage coupling (VCCIM) and a CIM-based modeling and predicting (CIMMP) method. The computing macro based on this architecture can achieve an energy efficiency of 12.1~17.6 TOPS/W and 71.70~72.01% inference accuracy when applied to a VGG-16 network CIFAR- 100 data set.
- Prof. Wang Kang and his team (Luo et al. 2022) leverage coupled magnetic tunnel junctions (MTJs), which are driven by the interplay of field-free spin orbit torque (SOT) and spin transfer torque (STT) effects, to realize two different stateful CIM paradigms for ternary MAC operations. Based on both paradigms, they further demonstrate the highly parallel array structures to implement a memory array, which support both memory access and CIM for ternary neural networks (TNNs). Experimental results show that the area overhead for CIM is only about 0.8% of the memory array. The advantage of this design in power consumption is illustrated in comparison with the CPU, GPU and other state-of-the-art works.
- In the third paper (Zhou et al. 2022), Prof. Jie Zhang et al. present a survey on storage-accelerator, with respect to the system designs, architectural innovations, and application-level optimizations. These accelerators are normally proposed to tackling with the well-known challenge that the main memory in the traditional computing system cannot satisfy the requirements of the emerging large-scale applications in terms of computing power and memory capacity. The survey would aid the development of the research community and inspire the researchers, who are interested in the relevant areas.

The third part of the special issue covers three memory or storage design innovations, which are optimized for various modern applications, including neuromorphic computing, graph neural networks (GNN) processing, and storage system in consumer devices.

- The paper (Yang et al. 2022) written by Prof. Weixia Xu and his team focuses on the memory organization in emerging neuromorphic processors. Based on the characteristics of the brain and Spiking Neural Networks (SNNs), they propose a set-associative memory organization (SAMO) and a compressed SRAM memory organization (CMAM) for loose and tight coupling structures in SNN to construct an area-efficient memory organization for generalized neuromorphic architectures. Experiments show that the methods use less chip area and consume less power than the CAM implementation in related work by 23.4–75.8% and 21.2–75.7%, while bringing minor processor performance overhead.
- Prof. Yue Dai et al. observe that although a mixed-precision feature quantization method can address the memory access overhead of GNN processing, the linear approximation and computation complexity become the main constraints for the overall GNN accuracy and performance (Dai et al. 2022). They propose segmented quantization to partition the feature range into segments conduct efficient mixed-precision computing between quantized feature and full precision weights. The technique helps to achieve high inference accuracy while maintaining low computation complexity. The experiments show that up to 5% average accuracy and up to 6.8 × performance improvements can be achieved over the state-of-the-art GNN accelerators.
- The third paper (Xu et al. 2022) in this part and also the last paper in the special issue is about the lightweight distributed file system proposed on consumer devices. Prof. Liang Shi and his team propose several practical optimization solutions for a lightweight distributed file system. Experiments on real devices show that the average access latency can be reduced by 29.7% with swap-based client-side persistent caching. Cross-device prefetching reduces around 33% access latency in the best case. Average cache synchronization latency is reduced by 13.7% and the worst synchronization latency is reduced by 63.7% with write-back scheduling.

We would like to thank all the authors for their innovative research work. And we appreciate the efforts from all reviewers and editors. Only with their great contributions, we are able to put together the eight interesting research papers in this special issue of *CCF Transactions on High Performance Computing*. These papers not only bridge the memory/storage architecture design and the requirements from modern applications, but also inspire further investigations in the related fields.

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

Chen, Y., Qiu, K., Chen, L., et al.: Smart scheduler: an adaptive NVM-aware thread scheduling approach on NUMA systems. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00110-2

Dai, Y., Tang, X., Zhang, Y.: An efficient segmented quantization for graph neural networks. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00121-z

Guo, A., Xue, C., Chen, X., et al.: VCCIM: a voltage coupling based computing-in-memory architecture in 28 nm for edge AI applications. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00111-1

Islam, A.A.R., York, C., Dai, D.: A performance study of optane persistent memory: from storage data structures' perspective. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00123-x

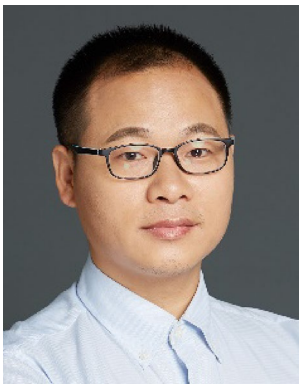
Luo, L., Liu, D., Zhang, H., et al.: SpinCIM: spin orbit torque memory for ternary neural networks based on the computing-in-memory architecture. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00108-w

Xu, Y., Li, H., Wang, H., et al.: Practical optimizations for lightweight distributed file system on consumer devices. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00132-w

Yang, Z., Wang, L., Wang, Y., et al.: Lotus: a memory organization for loose and tight coupling neurons in neuromorphic architecture. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00113-z

Zhou, Z., Yi, S., Zhang, J.: Survey on storage-accelerator data movement. CCF Trans. HPC (2022). https://doi.org/10.1007/s42514-022-00112-0

Guangyu Sun is currently an Associate Professor in the School of Integrated Circuits at Peking University. He received his B.S. and M.S. degrees from Tsinghua University, Beijing, in 2003 and 2006, respectively, and his Ph.D. degree from the Pennsylvania State University in 2011. His Ph.D. thesis, "Exploring Memory Hierarchy Design with Emerging Memory Technologies", received the 2012 EDAA outstanding dissertation award. His research interests include design and automation for computer architecture, cross-layer co-optimization, emerging memory technologies, etc. He has published 150+ journals and refereed conference papers on DAC, ISCA, MICRO, HPCA, IEEE TCAD, etc. He has won the DAC Under-40 Innovators Award, CCF-IEEE CS Young Computer Scientists Award, Microsoft Research Asia Collaborative Research Award, CCF-Intel Young Faculty Researcher Program, and the best paper awards three times. He has been serving as the PI and co-PI on many research grants from NSFC. His research work has also been supported by Alibaba DAMO Academy, Huawei, Baidu, AMD, Intel, MSRA, etc. Dr. Sun is an active volunteer in design automation and computer architecture communities. He was the general co-chair of NVMSA2021 and the TPC co-chair of NVMSA2020, RTCSA2019, APPT2017, and NAS2012. He has served as a program committee member and a track chair for over 20 conferences in these areas, including DAC, ICCAD, MICRO, HPCA, etc. He is an associate editor of ACM JETC.



Liang Shi received the B.E. degree from Xi'an University of Post and telecommunication, Xi'an China in July 2008, and the Ph.D. degree from the University of Science and Technology of China and City University of Hong Kong, in July 2013. He is currently a full professor with the School of Computer Science and Technology, East China Normal University. Before that, he was an associate professor at School of Computer Science, Chongqing University from 2013 to 2018. His research interests include storage system, operating systems, embedded systems, and distributed systems. He has published over 120 research papers in peer-reviewed journals (e.g., TC, TCAD, TOS and TPDS) and conferences (e.g., FAST, ATC, HPCA, MICRO, DAC). His works have received best paper award from NVMSA, best paper award nomination from ASPDAC, and ISLPED. He is also the recipient of Shanghai Excellent Young Researcher Fellowship in 2022 (Also called Qi-Ming-Xing Plan). He proposed a series of methods on optimizing the performance, lifetime and reliability of storage systems and embedded systems through near storage computing, controller algorithm design and software-hardware co-design. He has served in technical program committees for international conferences and workshops (e.g., DAC, MSST, CODES-ISSS, ASPDAC) and as a reviewer in high quality journals (e.g., TECS, TOS, TCAD and TPDS).

Jingtong Hu is currently an Associate Professor in the Department of Electrical and Computer Engineering at University of Pittsburgh, Pittsburgh, PA, USA. Before that, he was an Assistant Professor at Oklahoma State University from 2013 to 2017. He received his Ph.D. in Computer Science from University of Texas at Dallas in 2013 and his B.E. in Computer Science and Technology from Shandong University, China in 2007. His current research interests include hardware/software co-design for machine learning algorithms, on-device AI, and embedded systems. His works have received Donald O. Pederson Best Paper Award from IEEE Transactions on Computer-Aided Design of Circuits and Systems and 5 best paper nominations from DAC, ASP-DAC, and ESWEEK, etc. He is also the recipient of University of Pittsburgh William Kepler Whiteford Faculty Fellowship, Oklahoma State University Outstanding New Faculty Award, Air Force Summer Faculty Fellowship, and ACM SIGDA Meritorious Service Award. He has served on the technical program committee of many international conferences such as DAC, DATE, ASP-DAC, ESWEEK, CPS-IoT Week, etc. He served as a guest editor for Sensors, IEEE Transactions on Computers, ACM Transactions on Cyber-Physical Systems, ACM Transactions on Embedded Systems, and is currently serving as executive committee member and education chair for ACM SIGDA, associate editor for IEEE Embedded Systems Letters, the Journal of Systems Architecture: Embedded Software Design, and ACM Transactions on Cyber-Physical Systems.