CCF Transactions on High Performance Computing (2024) 6:113–114
https://doi.org/10.1007/s42514-024-00189-9

EDITORIAL

Editorial for the special issue on heterogenous computing

Shanjiang Tang[1] · **Yusen Li**[2]

Published online: 23 April 2024

In the current era of AI and Big Data, an increasing and significant amount of computing power is needed for many applications and algorithms such as AIGC models, face detection, autonomous driving and atmosphere simulation. Recently, there is a significant amount of interest among the community in improving AI and big data applications with heterogenous computing, which refers to a computing system using different types of computing cores such as GPU, NPU, ASIC, DSP and FPGA. It can improve the performane and enery efficiency by dispatching different workloads to processors that are designed for specialized processing and specific purposes. This issue aims to cover challenges that can hamper efficiency and utilization for AI and big data applications on heterogenous computing systems, such as efficient utilization of the raw hardware, I/O management, task scheduling, etc.

We have seven invited papers selected for this special issue based on a peer-review process, which covers a variety of aspects related to heterogenous computing mentioned above.

The first part of special issue focuses on the AI applications with heterogenous computing, including benchmarks for AI processors and heterogenous accelerator for recommender system.

- The paper written by Xiao et al. (2024) develops an AI benchmark named AIBench specially for Huawei Ascend AI processors. It can evaluate the performance of each computation unit of the AI processor, including the matrix unit and vector unit. It can also quantify the data transmission bandwidth among the buffer within the chip. Experiments using the AIBench benchmark shows that the Ascend 910 AI chip can achieve up to 216 TFLOPs for float16 data for the matrix unit and 3390 GFLOPS for the vector unit.
- The paper written by Shen et al. (2024) proposes a resistive random accessed memory (ReRAM) based processing-in-memory (PIM) accelerator named ReGCNR for GCN-based recommendation. It fits with large-size embedding table and user-item graph with 3-dimensional (3-D) stacked heterogeneous ReRAM and maximizes the efficiency of the execution pipeline using a joint degree mapping schema. The performance can be improved by assembling a well-coordinated pipeline and hardware scheduling design.

The second part of special issue focuses on the big data applications with heterogenous computing, including graph data processing, streaming data processing, data compression, non-uniform data sampling.

- The paper written by Lu et al. (2023) presents a large-scale heterogenous computing framework called AutoNUSC for non-uniform sampling two-dimensional convolution applications. It can simplify the programming for heterogenous computing systems such as CPU + GPU and CPU + DSP by abstracting and encapsulating the parallel execution process of non-uniform sampling two-dimensional convolution, including task scheduling, data division, node communication, fault-tolerant recovery, etc. Experiments show that it can reduce the burden of users in programming NUSC applications and can achieve up to 339 times within a single node compared to the serial program.
- The paper written by Liu et al. (2024) considers the load balancing optimization issue in the presence of skewed data streams (i.e., heterogenous workload). It proposes an adaptive Key-Splitting algorithm named FlexD that achieves dynamic adaption of key separation limits for streaming data processing in a multi-node cluster. It can alleviate the load imbalance problem by dividing keys among downstream operators. The authors implement

✉ Shanjiang Tang
tashj@tju.edu.cn

Yusen Li
liyusen@nbjl.nankai.edu.cn

[1] Tianjin University, Tianjin, China

[2] Nankai University, Tianjin, China

it on Apache Storm and experimental results show that it achieves a good balance between load balancing and aggregation cost.

- The paper written by Li et al. (2023) focuses on cross-platform porting issue for GPU-based graph computation in face of different vendor's GPUs with varied software stacks. It proposes a large-scale graph computing framework called OneGraph for multiple types of accelerators that can manage multiple heterogeneous devices at the same time by borrowing idea from intel's oneAPI and allows users' codes run on different GPU platforms with no code modification. Experimental results show that it can achieve an average speedup of 3.3x over the state-of-art baseline.
- The paper written by Sun et al. (2024) proposed a FPGA-based acceleration architecture for Apache Spark operators, including K-means, PageRank, and sorting. It explores the classic divide-and-conquer paradigm to accelerate these Spark operators with multiple FPGA processing units, and directly shuffle intermediate results to destination servers for aggregation based on FPGA's RDMA networks. It also adopts the pipelining, loop unrolling, FPGA BRAM partitioning and a data execution model to maximize tasks/data partitioning on each FPGA. Experimental results that the proposed system outperforms the native Spark by about 3.5–112x.
- The paper written by Liu et al. (2023) considers the text analytics directly on compression (TADOC) for DCU platform, a new Chinese domestic accelerator from Sugon. It proposes a compressed data direct computing technology called D-TADOC for Chinese datasets on DCU without decompression and can visualize it. Experimental results show that D-TADOC can outperform its baseline on the CPU by about 40.5x.

Finally, we would like to take this chance to thank all the authors and reviewers for their efforts and contributions to this special issue of CCF THPC.

References

Li, S., Zhu, J., Han, J., et al.: OneGraph: A cross-architecture framework for large-scale graph computing on GPUs based on oneAPI. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00172-w

Liu, Y., Zhang, F., Pan, Z., et al.: CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00153-z Compressed data direct computing for Chinese dataset on DCU

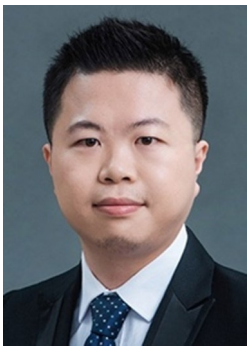
Liu, G., Wang, Z., Zhou, A.C., et al.: Adaptive key partitioning in distributed stream processing. CCF Trans. HPC. (2024). https://doi.org/10.1007/s42514-023-00179-3

Lu, Y., Yu, C., Xiao, J., et al.: A large-scale heterogeneous computing framework for non-uniform sampling two-dimensional convolution applications. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00148-w

Shen, X., Huang, Y., Zheng, L., et al.: A heterogeneous 3-D stacked PIM accelerator for GCN-based recommender systems. CCF Trans. HPC. (2024). https://doi.org/10.1007/s42514-024-00180-4

Sun, Y., Liu, H., Liao, X., et al.: FPGA-based acceleration architecture for Apache Spark operators. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00158-8

Xiao, Y., Wang, Z.K.: AIBench: A Tool for Benchmarking Huawei ascend AI processors. CCF Trans. HPC. (2024). https://doi.org/10.1007/s42514-024-00187-x



Shanjiang Tang received the PhD degree from School of Computer Engineering, Nanyang Technological University, Singapore in 2015, and the MS and BS degrees from Tianjin University (TJU), China, in Jan 2011 and July 2008, respectively. He is currently an associate professor in College of Intelligence and Computing, Tianjin University, China. His research interests include parallel computing, cloud computing, big data analysis, and machine learning.

Yusen Li received the Ph.D. degree from Nanyang Technological University in 2014. He is currently a Professor with the Department of Computer Science and Security, Nankai University, China. His research interests include scheduling, load balancing, and other resource management issues in distributed sys- tems and cloud computing.