CCF Transactions on High Performance Computing (2024) 6:1–2
https://doi.org/10.1007/s42514-024-00182-2

EDITORIAL

Special issue of HPCChina 2023

Yunquan Zhang[1] · **Guangming Tan**[1] · **Liang Yuan**[1]

Published online: 22 February 2024

HPCChina is the annual conference established in 2005 by the Technical Committee on High Performance Computing (TCHPC) of China Computer Federation (CCF). The HPCChina is the leading venue in China for presenting high-quality original research in all fields related to high performance computing. In 2023, the conference received a total of 123 submissions and accepted 70 papers based on a strict peer-review procedure. The special issue invited eight papers of high quality. They can be divided into two kinds, algorithm research or system study. The first four papers focus on parallel algorithms in numerical methods. The second four papers study scheduling strategy, algorithm selector, communication protocol and performance evaluation, all from the system perspective. We provide a short summary of each paper as follows.

Yidong Chen et al. (Chen et al. 2023) propose a parallel ADMM (alternating direction method of multipliers) -based algorithm called block splitting proximal ADMM (BSPADMM). BSPADMM computes the problem by using sparse matrix–vector multiplication, without communication between processors. Experimental results on three datasets of varying scales show that BSPADMM outperforms state-of-the-art ADMM techniques.

Haoyuan Zhang et al. (Zhang et al. 2023) propose and implement a mixed-precision Block-ISAI preconditioner for solving linear systems from multiphysics areas. By leveraging FP32 computing, their approach accelerates the sparse matrix–vector product kernel while maintaining satisfactory accuracy. Meanwhile, an efficient, warp-based GPU implementation for Block-ISAI preconditioner with Tensor core acceleration is proposed. Experimental results shows noteworthy speedup.

Yang Wang et al. (Wang et al. 2023) propose a high-performance parallel direct implementation of dilated convolutions on multi-core DSPs in a CPU-DSP heterogeneous prototype processor, which can effectively capture the data locality in dilated convolutions. The experimental results demonstrate that the direct implementation achieves much better performance than GEMM-based ones on multi-core DSPs for all the tested layers.

Dazheng Liu et al. (Liu et al. 2023) propose an optimized scheme that overlaps computation with communication based on grouping levels for the semi-Lagrangian interpolation scheme of Yin-he Global Spectral model (YHGSM). Experimental results show that the scheme can reduce the running time for the semi-Lagrangian scheme by 12.5% and effectively reduce the communication overhead of the model, improving the efficiency of YHGSM.

Yueyuan Zhou et al. (Zhou et al. 2023) propose FILL, a resource scheduling system designed for co-running multiple GROMACS jobs. FILL employs space partitioning technology to effectively allocate hardware resources and facilitates collaborative scheduling of CPU and GPU resources. Experimental results validate the effectiveness of FILL in optimizing system throughput for multiple GROMACS simulations.

Lu Bai et al. (Bai et al. 2023) propose ConvDarts, a fast and exact optimal algorithm selector for all convolution configurations (parameters of convolutional operations). ConvDarts reduces the training time of classical deep learning networks and also reduces the required memory space. ConvDarts provides more possibilities for the training of network models in resource-constrained environments.

Hang Cao et al. (Cao et al. 2023) propose an efficient elastic RDMA Protocol (eRDMA) to enabling RDMA's merits for HPC applications in the cloud. eRDMA applys the direct data movement (DDM) of cloud infrastructure processing Unit (CIPU), overlay of virtual private cloud (VPC), and compatibility for RDMA verbs to fully utilize the elastic resources with the features of RDMA network for

✉ Guangming Tan
tgm@ict.ac.cn

Yunquan Zhang
zyq@ict.ac.cn

Liang Yuan
yuanliang@ict.ac.cn

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

HPC scenarios in the cloud. The effectiveness of eRDMA is demonstrated by various experimental results across different platforms for many HPC and general TCP applications.

Zhengxian Lu et al. (Lu et al. 2023) present a performance evaluation of dispatching and mapping mechanisms in different deep learning frameworks by examining their kernel function efficiency and operator dispatching mechanisms. The evaluation results demonstrate the device utilization capability of five frameworks, namely PyTorch, TensorFlow 1, TensorFlow 2, MXNet, and PaddlePaddle, and reveal the potential for further optimizing the training performance of deep learning frameworks.

References

Bai, L., Ji, W., Li, Q., et al.: ConvDarts: A fast and exact convolutional algorithm selector for deep learning frameworks. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00167-7

Cao, H., Xu, C., Han, Y., et al.: An efficient cloud-based elastic RDMA protocol for HPC applications. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00170-y

Chen, Y., Pan, J., Han, Z., et al.: BSPADMM: Block splitting proximal ADMM for sparse representation with strong scalability. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00164-w

Liu, D., Liu, W., Pan, L., et al.: Optimization of the parallel semi-lagrangian scheme to overlap computation with communication based on grouping levels in YHGSM. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00163-x

Lu, Z., Du, C., Jiang, Y., et al.: Quantitative evaluation of deep learning frameworks in heterogeneous computing environment. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00168-6

Wang, Y., Wang, Q., Pei, X., et al.: High performance dilated convolutions on multi-core DSPs. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00166-8

Zhang, H., Ma, W., Yuan, W., et al.: Mixed-precision block incomplete sparse approximate preconditioner on Tensor core. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00165-9

Zhou, Y., Ren, Z., Shao, E., et al.: FILL: A heterogeneous resource scheduling system addressing the low throughput problem in GROMACS. CCF Trans. HPC. (2023). https://doi.org/10.1007/s42514-023-00169-5