CCF Transactions on High Performance Computing (2021) 3:221–223
https://doi.org/10.1007/s42514-021-00077-6

EDITORIAL

Editorial for the special issue on large-scale AI in classical HPC environment and AI for science

Wei Xue[1] · Haohuan Fu[1] · Weile Jia[2] · Guangming Tan[2]

Published online: 20 September 2021

Recent progress on large-scale machine learning (ML) and deep learning (DL) have demonstrated great potential in both traditional artificial intelligence (AI) applications in computer science (such as natural language processing, knowledge engineering and computer vision) and AI-enabled applications in scientific domains (such as AlphaFold2). Such a trend makes AI an important high performance computing application, and ignites the innovative opportunity in the field of scientific discovery. To enable scalable AI in a high performance computing environment, we are facing quite a few challenges in architecture, software ecology and application innovation. This issue focuses on the novel ideas, methods, as well as efforts for resolving the above challenges, and to adapt HPC to new computing paradigms of knowledge and data convergent science discovery.

We have eight invited papers selected for this special issue based on a peer-review procedure, which cover several different aspects that relate to the software, algorithm and application challenges mentioned above.

The first part of the special issue focuses on the software methods and tools for facilitating AI and Big Data applications, which is the fundamental challenge on development and deployment of these emerging applications on high performance systems. We have two papers that discuss the design and development of AI toolkits on Sunway Supercomputers and the storage innovation for large-scale bioinformatics application, respectively.

- Dr. Xin Liu and her colleagues present the new-generation AI toolkits for the leading Sunway supercomputers, which consist of the optimized DNN library SWDNNv2, the lightweight and scalable deep learning framework SWMind, and the SWPyTorch, a reformed version of PyTorch. The toolkits manage to support Python-compatible interface and high-performance and mixed-precision training, which exploit the potential of the unique heterogenous many-core architecture of Sunway supercomputers for efficient neural network computing.
- The paper written by Zhiyang Ding et al. tackles the I/O issue of data-intensive applications and proposes the data processing method based on active storage technique, which refactors the parallel application into compute-intensive and data-intensive tasks and offloads the data-intensive tasks to high performance storage nodes. Taking the famous Bioinformatics program mpiBLAST as an example, the proposed method is combined with performance model-based design space exploration to reduce the execution time by up to 50% on a cluster system.

The second part of the special issue, consisting of three research papers, focuses on algorithmic innovations for improving both scalability and efficiency of large-scale AI applications. With the increased level of parallelism and the increased level of heterogeneity of high performance computers, and the increased scale of application problems, algorithmic improvements have become a major challenge for successfully deploying large-scale AI systems in modern HPC Environment.

- The paper written by Prof. Yang You and his collaborators focuses on the weak scaling issue of kernel ridge regression, which is one of the fundamental methods in machine learning. Several interesting Divide-and Con-

✉ Wei Xue
xuewei@tsinghua.edu.cn

Haohuan Fu
haohuan@tsinghua.edu.cn

Weile Jia
jiaweile@ict.ac.cn

Guangming Tan
tgm@ict.ac.cn

[1] Tsinghua University, Beijing 100084, China

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

quer algorithms have been proposed for different trade-offs between accuracy and speed, and have achieved the weak scaling efficiency of up to 92% for a theoretical speedup of 4096.
- The paper written by Hao Bai et al. focuses on high performance connected component analysis, which is widely used in the Graph500 benchmark and a number of graph analytic applications. The elaborated data structure of adjacency information and the optimized union find algorithm have been presented for the Tianhe Supercomputer, and the performance of graph connected component analysis has been improved by 5 times, compared to those with both BFS and DFS algorithms.
- Prof. Weifeng Liu and his team present their new attempt in the third paper of this part to optimize LU and Cholesky factorizations of dense matrices on the Cambricon AI accelerator. By exploiting various optimization methods, the remarkable factorization performance improvement has been achieved in both half and single precisions.

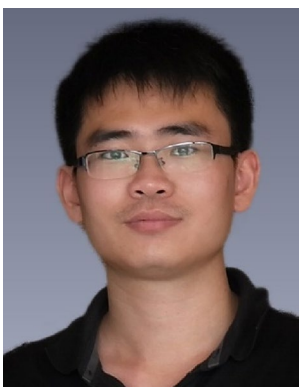
The third part of the special issue has three showcases of application innovations across different areas, which demonstrate the recent achievements in the rapidly evolving research area of AI for Science.

- The paper written by Dr. Ka Chun Cheung et al. falls in the hot topic of solving partial differential equation with machine learning technique. Two important methods, the physics informed neural network and the Fourier neural operator, are in-depth introduced and discussed in this paper. And the partial differential equation solver NVIDIA SimNet has been presented and demonstrated in detail.
- The paper written by Prof. Yuedong Yang and his team targets on improving both the accuracy and robustness of bladder cancer prognosis prediction by using multi-omics data and the transfer learning method. Moreover, a light-weight and accurate risk prediction model has been introduced with mRNA data through XGBoost algorithm for both better clinical utility and identification of the real cancer-related genes.
- The third paper in this part and also the last paper in the special issue is a survey paper on the hybrid approach (HPC+AI) for Earth System Science. Written by Dr. Simon See and Jeff Adie from NVIDIA, the paper manages to identify the important application areas, major challenges, and potential solutions for this hybrid type of approach.

We would like to take this chance to thank all the authors and the reviewers for their splendid contributions to this special issue of CCF THPC. Only with their great efforts, we are able to put together the eight research papers that discuss different topics, and present different ideas that help to bridge large-scale AI applications and the classical HPC Environment, and also help to drive the continuous development of AI for Science.

Wei Xue is an Associate Professor at Department of Computer Science and Technology in Tsinghua University, China. He is the director of High Performance Computing Institute and a joint faculty in Department of Earth System Science. His research interests include high performance computing and uncertainty quantification. As one of team leaders, he received the 2016 and 2017 Gordon Bell Prizes and finalist of 2018 Gordon Bell Prize. He is a senior member of CCF and a member of IEEE and ACM.

Haohuan Fu is a Professor in the Ministry of Education Key Laboratory for Earth System Modeling, and Department of Earth System Science in Tsinghua University, where he leads the research group of High Performance Geo-Computing (HPGC). He is also the deputy director of the National Supercomputing Center in Wuxi, leading the research and development division. His research work focuses on providing both the most efficient simulation platforms and intelligent data analysis platforms for geoscience applications, leading to two consecutive winning of the ACM Gordon Bell Prizes in 2016 and 2017.



Weile Jia is an Associate Professor at Institute of Computing Technology, Chinese Academy of Sciences. His research interests lie in high performance computing, artificial intelligence (AI) and first-principle calculations. He received the 2020 ACM Gordon Bell Prize by applying HPC and AI techniques in ab initio molecular dynamics.

Guangming Tan is a Professor from the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include parallel programming and algorithm, domain-specific architecture, and bioinformatics. He has published papers including conference/journals like SC, PLDI, PPoPP and the IEEE Transactions on Parallel and Distributed Systems.