Jia Wei

DoB: 01/28/1997 / Male / Xi'an, Shaanxi, China/ Contact: weijia4473@stu.xjtu.edu.cn; (+86)15235339399
Personal Page: https://www.zhihu.com/people/wei-jia-29-17

EDUCATION

Northwest University Xi'an, China
Bachelor of Engineering in *Internet of Things* 09/2015 - 07/2019
Major Courses: Computer Architecture, Computer Network, Data Structure, Database, etc.,
Scholarships: First Class Scholarship (#2 in my major across 58)

Xi'an Jiaotong University Xi'an, China
Ph.D. Candidate of *Computer Science and Technology* 09/2019-12/2024
Supervised by Xingjun Zhang, Professor, Dean of School of Computer Science and Technology
Major Courses: Neural Networks, Machine Learning, Distributed Systems, Parallel Computing, etc.,
Scholarships: National Scholarship (Top 3), Huawei Scholarship

University of Alberta Edmonton, Canada
Research Award Recipient (Joint Ph.D.) 06/2023-11/2024
Supervised by Witold Pedrycz, Professor, Royal Society of Canada Fellow, IEEE Fellow

ACADEMIC PROJECTS

"Modelling data-intensive supercomputer storage systems" 07/2021 - Present
Complete the Data-intensive Supercomputer Survey and publish it in a well-known journal and conference.
Design a model for evaluating Data-intensive Supercomputers.

"Overcome GPU memory wall" 07/2021 - Present
Propose to selectively transfer immediate **feature maps** to **NVMe devices** (SSD etc.) via Nvidia GPU Direct Storage API.
Design and implement a GPU data read/write API with better performance than Torch. load/save.
Give the total storage capacity requirements upon specific processors and memory.

"Use CSDs to improve the preprocessing performance of deep learning" 07/2021 - Present
Utilize **Compute Storage Devices** (CSDs) to finish the data preprocessing such as image normalization and image transposition.
Design a joint hardware-software programming framework to accelerate the data pre-processing process on the CSD.
Implement a management module on the ARM core combined with many acceleration modules on the FPGA.

"Joint optimal scheduling algorithm and heterogeneous job management" Chinese Natural Science foundation 03/2021-03/2023
Propose a general model named MC^2 to evaluate High Performance Computing (HPC) energy consumption.
Use an improved reinforcement learning algorithm to achieve better task scheduling in HPC (Minimize energy consumption while maintaining QoS).

"Distributed Cloud Platform Construction" 10/2019-10/2022
Build a cloud platform with high robustness and low latency via kubernetes, dockers, and glusterfs.
Implement a digital simulation system to verify the capabilities and performance of our platform.

"Development of parallel computing software for E-class computers" Chinese National Key R&D Program 09/2016-03/2022
Achieve high scalability and performance of parallel software, based on vectorization, data chunking, computation and communication overlap, load balancing, and other techniques.
Port and optimize the Pytorch platform to the Tianhe-3 E-class supercomputer prototype.

TEACHING EXPERIENCE

"Assembly Language (undergraduate)" Xi'an Jiaotong University (XJTU) Teacher Assistant (TA) 09/2020 – 01/2021
Tutor over 100 undergraduate students through course labs and answer questions.
Review and correct student assignments.

"Fuzzy Sets in Human-Centric Systems (graduate)" University of Alberta (UofA) Teacher Assistant (TA) 09/2023 – 01/2024
Assist teachers in preparing course materials.
Answer course questions.

"Tutor undergraduate students on their final projects" 09/2021 – 05/2022
Propose topics related to the deployment of deep learning and optimization on Tianhe-3 Prototype.
Help students identify innovative ideas and writing frameworks.
Assist students with experimental validation and thesis writing.

SELECTED PUBLICATIONS

[1] **Wei J,** Zhang X, Wang L, et al. Fastensor: Optimise the Tensor I/O Path from SSD to GPU for Deep Learning Training[J]. ACM Transactions on Architecture and Code Optimization, 2023, 20(4): 1-25. (CCF A, With HiPEAC 2024 Main Paper Track Presentation)

[2] **Wei J**, Zhang X, Pedrycz W, Ding W. Dynamic Fuzzy Sampler for Graph Neural Networks [J]. IEEE Transactions on Fuzzy Systems, 2024. (CCF B, Impact Score: 10.7)

[3] **Wei J**, Zhang X, Zhuo Z, et al. Leader Population Learning Rate Schedule[J]. Information Sciences, 2023, 623: 455-468. (CCF B, Impact Score: 8.233)

[4] **Wei J**, Zhang X. How Much Storage Do We Need for High Performance Server[C]//2022 IEEE 38th International Conference on Data Engineering (ICDE). IEEE, 2022: 3221-3225. (CCF A)

[5] **Wei J**, Zhang X, Ji Z, et al. DPLRS: Distributed Population Learning Rate Schedule[J]. Future Generation Computer Systems, 2022, 132: 40-50. (Impact Score: 7.187)

[6] **Wei J**, Zhang X, Ji Z, et al. Deploying and scaling distributed parallel deep neural networks on the Tianhe-3 prototype system[J]. Scientific Reports, 2021, 11(1): 1-14. (Impact Score: 4.379)

[7] **Wei J**, Chen M, Wang L, et al. Status, challenges and trends of data-intensive supercomputing[J]. CCF Transactions on High Performance Computing, 2022: 1-20. (CCF C)

[8] **Wei J**, Zhang X, Wang L, et al. MC2 energy consumption model for large-scale distributed data parallel training of deep neural networks[J]. Computer Research and Development, 2023. (CCF Chinese Rank A)

[9] Wei Z, Zhang X, Ji Z, Li J, **Wei J**. Revisit and Benchmarking of Automated Quantization Towards Fair Comparison[J]. IEEE Transactions on Computers, 2023.

[10] Wei Z, Zhang X, Li J, Ji Z, **Wei J**. BenQ: benchmarking automated quantization in deep neural network accelerators [C]. Design, Automation, and Test in Europe (DATE), 2022.

[11] Li J, Zhang X, **Wei J**, et al. GARLSched: Generative adversarial deep reinforcement learning task scheduling optimization for large-scale high performance computing systems[J]. Future Generation Computer Systems, 2022. (Impact Score: 7.187)

[12] Ji Z, Zhang X, Li J, **Wei J**, Wei Z. EP4DDL: addressing straggler problem in heterogeneous distributed deep learning[J]. The Journal of Supercomputing, 2022: 1-18. (Impact Score 2.474)

[13] Ji Z, Zhang X, Wei Z, Li J, **Wei J**. A tile-fusion method for accelerating Winograd convolutions[J]. Neurocomputing, 2021, 460: 9-19. (Impact Score 5.719)

[14] Li J, Zhang X, Wei Z, **Wei J**, Ji Z. Energy-aware task scheduling optimization with deep reinforcement learning for large-scale heterogeneous systems[J]. CCF Transactions on High Performance Computing, 2021, 3(4): 383-392.

UNDER REVIEW

[1] **Wei J**, Zhang X, Pedrycz W. BEND: Bagging Deep Learning Training Based on Efficient Neural Network Diffusion[J]. arXiv preprint arXiv:2403.15766, 2024.

[2] **Wei J**, Zhang X, Pedrycz W, et al. Advancing Distributed Deep Learning with Ecosphere Learning Rate Schedule [J].

[3] **Wei J**, Zhang X, Pedrycz W, et al. Homophone-Based Chinese Natural Language Data Augmentation [J].

[4] **Wei J**, Zhang X, Pedrycz W, et al. Dual-pronged deep learning preprocessing on heterogeneous platforms with CPU, GPU and CSD.

[5] **Wei J**, Zhang X, Pedrycz W. NKD: Neural Network Diffusion-based Efficient Multi-Teacher Knowledge Distillation [C]. AAAI 2025 Second Phase, 2024.

REFERRER

Witold Pedrycz, Professor, University of Alberta, Royal Society of Canada Fellow, IEEE Fellow

Xingjun Zhang, Professor, Xi'an Jiaotong University, Dean of School of Computer Science and Technology

SKILLS/INTERESTS

Programming Languages: **Python**(Proficient), C++, C, JAVA, Matlab, etc.
Tools & Frameworks: **Pytorch**(Proficient), Git, SVN, Tensorflow, Latex, Keras, SQL, Jupyter Notebook, Linux Operations,etc.
Hobbies: Saxophone, Badminton, Basketball, Blogging, etc.