Weihao **Cui**

github.com/Raphael-Hao linkedin.com/in/weihao-cui-b5a625a5
(+86) 18621816508 raphaelhao@outlook.com weihao@sjtu.edu.cn
No.800 Dongchuan Road, Minhang District, Shanghai, China
Apr. 1992, Nantong, Jiangsu, China

Bio. I am currently a 3rd-year Ph.D. Student at Department of Computer Science and Engineering (CSE), Shanghai Jiao Tong University (SJTU), China, working under the guidance of Prof. Quan Chen on **AI System and Cloud Computing**.

Research interests. My Ph.D. research work covers a range of issues :

- **Systems for Deep Learning on Accelerators**, such as resource management of accelerator that supports spatial multitasking, efficient DNN inference system for latency reduction, co-location of multiple DNN services on GPUs, and DNN compiler for the dynamic neural network.
- **Resource Management for Cloud Computing**, such as scheduling of latency-critical and best-effort jobs in datacenters with accelerators, distributed memory object cache system in heterogeneous computer systems consisting of HBM (High Bandwidth Memory).

Currently, I am interested in various approaches to optimizing the AI system.

Education

Sep 2020 – **Mar 2023(expected)**	**Ph.D. Student** in *Shanghai Jiao Tong University* (SJTU), Shanghai Emerging Parallel Computing Center (EPCC) *Advisor : Prof. Quan Chen*
Sep 2017 – Apr 2020	**Master** in *Shanghai Jiao Tong University* (SJTU), Shanghai Emerging Parallel Computing Center (EPCC) *Advisor : Prof. Quan Chen* GPA : 3.73/4.0
Sep 2011 – Jun 2015	**Bachelor of Engineering** in *Shanghai Jiao Tong University* (SJTU), Shanghai Majors in Ocean Engineering

Preprint Publications

- **Weihao Cui**,Chunyu Xue, Han Zhao, Quan Chen, Minyi Guo. "Aodos : Affinity-aware Orchestration and Deterministic Operator Overlap for Simultaneous DNN Services in the GPU Cluster." submitted to TOCS.

Accepted Publications

- **Weihao Cui**, Zhenhua Han, Lingji Ouyang, Yichuan Wang, Ningxin Zheng, Lingxiao Ma, Yuqing Yang, Fan Yang, Jilong Xue, Lili Qiu, Lidong Zhou, Quan Chen, Haisheng Tan, Minyi Guo. "Optimizing Dynamic Neural Networks with Brainstorm" *OSDI 2023. CCF-A*
- Han Zhao, **Weihao Cui**, Quan Chen, Minyi Guo. "ISPA : Exploiting Intra-SM Parallelism in GPUs via Fine-grained Resource Management." *IEEE Transactions on Computers. 2022. CCF-A.*
- Shulai Zhang, **Weihao Cui**, Quan Chen, Zhengnian Zhang, Yue Guan, Jingwen Leng, Chao Li, Minyi Guo. "PAME : Precision-Aware Multi-Exit DNN Serving for Reducing Latencies of Batched Inferences." *ACM International Conference on Supercomputing. 2022. CCF-B.*
- **Weihao Cui**, Han Zhao, Quan Chen, Hao Wei, Zirui Li, Deze Zeng, Chao Li, Minyi Guo. "DVABatch : Diversity-aware Multi-Entry Multi-Exit Batching for Efficient Processing of DNN Services on GPUs." *USENIX Annual Technical Conference. 2022. CCF-A.*
- Han Zhao, **Weihao Cui**, Quan Chen, Youtao Zhang, Yanchao Lu, Chao Li, Jingwen Leng, Minyi Guo. "Tacker : Tensor-CUDA Core Kernel Fusion for Improving the GPU Utilization while Ensuring QoS." *International Symposium on High-Performance Computer Architecture. 2021. CCF-A*
- **Weihao Cui**, Han Zhao, Quan Chen, Ningxin Zheng, Jingwen Leng, Jieru Zhao, Zhuo Song, Tao Ma, Yong Yang, Chao Li, and Minyi Guo. "Enable Simultaneous DNN Services Based on Deterministic Operator Overlap and Precise Latency Prediction." *In The International Conference for High Performance Computing, Networking, Storage and Analysis. 2021. CCF-A.*
- Han Zhao, **Weihao Cui**, Quan Chen, Jieru Zhao, Jingwen Leng, and Minyi Guo. "Exploiting Intra-SM Parallelism in GPUs via Persistent and Elastic Blocks." *In 2021 IEEE 39th International Conference on Computer Design (ICCD), IEEE, 2021. CCF-B*
- Wei Zhang, Quan Chen, Ninxing Zheng, **Weihao Cui**, Kaihua Fu, and Minyi Guo. "Towards QoS-awareness and Improved Utilization of Spatial Multitasking GPUs." *IEEE Transactions on Computers. 2021. CCF-A.*
- **Weihao Cui**, Quan Chen, Han Zhao, Mengze Wei, Xiaoxin Tang, and Minyi Guo. "E^2bird : Enhanced Elastic Batch for Improving Responsiveness and Throughput of Deep Learning Services." *IEEE Transactions on Parallel and Distributed Systems 32, no. 6 (2020) : 1307-1321. CCF-A.*
- Han Zhao, **Weihao Cui**, Quan Chen, Jingwen Leng, Kai Yu, Deze Zeng, Chao Li, and Minyi Guo. "CODA : Improving Resource Utilization by Slimming and Co-locating DNN and CPU Jobs." *In 2020 IEEE 40th International Conference on Distributed Com-*

puting Systems (ICDCS), pp. 853-863. IEEE, 2020. CCF-B.

- **Weihao Cui**, Mengze Wei, Quan Chen, Xiaoxin Tang, Jingwen Leng, Li Li, and Mingyi Guo. "Ebird : Elastic batch for improving responsiveness and throughput of deep learning services." *In 2019 IEEE 37th International Conference on Computer Design (ICCD), pp. 497-505. IEEE, 2019. CCF-B*.
- Wei Zhang, **Weihao Cui**, Kaihua Fu, Quan Chen, Daniel Edward Mawhirter, Bo Wu, Chao Li, and Minyi Guo. "Laius : Towards latency awareness and improved utilization of spatial multitasking accelerators in datacenters." *In Proceedings of the ACM International Conference on Supercomputing, pp. 58-68. 2019. CCF-B*.

Skills

Programming Skills : **C/C++, CUDA, Python**, Go, LaTeX, Git.

Projects & Experiences

March 2018 – December 2018 | **GPU sharing technology based on QoS awareness, SJTU & Huawei, C/C++/Python**

Features :

- Bare metal/container environment : support multiple users and multiple processes/threads to share a GPU single GPU resource.
- Provide reasonable GPU QoS mechanisms, such as priority, end-to-end latency, and percentage of the resource.
- Fully compatible with existing applications, such as various Deep Learning frameworks, without code modification and recompilation, transparent to the application software.

GPU | Deep Learning | QoS | Resource Management

July 2019 – September 2019 | **Performance Investigation of CNI(Container Network Interface) plugin of K8S, Alibaba Group, C/Shell**

Responsibilities :

- Investigate new tracing technology in Linux kernel — eBPF (extended Berkeley Package Filter).
- Investigate the CNI plugin of K8S implemented by eBPF — cilium.
- Performance test between cilium and other K8S CNI plugins including calico, canal, flannel.

eBPF | cillium | CNI plugin | K8S

May 2020 – December 2020 | **Efficient Neural Architecture Search, Tencent, Python/Shell**

Responsibilities :

- Investigate the hardware-aware neural architecture search.
- Explore efficient approaches for accelerating the neural architecture search.

NAS | Deep Learning

December 2021 – Now | **AI System for Dynamic Neural Network, MSRA Shanghai, Python/C++/CUDA**

Responsibilities :

- Investigate dynamic neural networks.
- Optimization of the training/inference of dynamic neural network.

Dynamic Neural Network | Deep Learning

Languages

English : Reading ●●●●● CET-4 : 609
Listening ●●●●○ CET-6 : 478
Speaking ●●●○○

Honors & Awards

Fall 2021 National Fellowship of Shanghai Jiao Tong University.
May 2020 Tencent rhino-bird elite training program.

(last update : 8 Oct. 2022)