


# Weihao Cui

🌐 [raphael-hao.top](http://raphael-hao.top)    [github.com/Raphael-Hao](https://github.com/Raphael-Hao)    [linkedin.com/in/weihao-cui-b5a625a5](https://www.linkedin.com/in/weihao-cui-b5a625a5)  
☎ (+86) 18621816508   ✉ [raphaelhao@outlook.com](mailto:raphaelhao@outlook.com)   ✉ [weihao@sjtu.edu.cn](mailto:weihao@sjtu.edu.cn)  
📍 No.800 Dongchuan Road, Minhang District, Shanghai, China  
📅 Apr. 1992, Nantong, Jiangsu, China



**Bio.** I obtained my Ph.D. degree at Department of Computer Science and Engineering (CSE), Shanghai Jiao Tong University (SJTU), China, working under the guidance of Prof. Quan Chen and Prof. Minyi Guo 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 DNN compiler for the dynamic neural network, resource management of accelerator that supports spatial multitasking, efficient DNN inference system for latency reduction, co-location of multiple DNN services on GPUs, and so on.
- **Systems for Cloud Computing**, such as GPU serverless computing, scheduling of latency-critical and best-effort jobs in datacenters with accelerators, and so on.

Currently, I am more focused on various approaches to optimizing the AI system, especially for large language models and dynamic neural networks.

## Education

- 
- |                     |   |
|---------------------|---|
| Sep 2020 – Sep 2023 | <b>Ph.D.</b> in <i>Shanghai Jiao Tong University</i> (SJTU), Shanghai<br>Emerging Parallel Computing Center (EPCC)<br><i>Advisor : Prof. Quan Chen</i>                    |
| Sep 2017 – Apr 2020 | <b>Master</b> in <i>Shanghai Jiao Tong University</i> (SJTU), Shanghai<br>Emerging Parallel Computing Center (EPCC)<br><i>Advisor : Prof. Quan Chen</i><br>GPA : 3.73/4.0 |
| Sep 2011 – Jun 2015 | <b>Bachelor of Engineering</b> in <i>Shanghai Jiao Tong University</i> (SJTU), Shanghai<br>Majors in Ocean Engineering  |

## Publications

- 
- Binghao Chen, Han Zhao, **Weihao Cui**, Yifu He, Shulai Zhang, Quan Chen, Zijun Li, Minyi Guo. “Maximizing the Utilization of GPUs Used by Cloud Gaming through Adaptive Co-location with Combo” **SoCC 2023**.
  - Han Zhao, **Weihao Cui**, Quan Chen, Jingwen Leng, Deze Zeng, Minyi Guo. “Improving Cluster Utilization through Adaptive Resource Management for DNN and CPU Jobs Co-location” **TC 2023**
  - Yangjie Zhou, Yaoxu Song, Jingwen Leng, Zihan Liu, Weihao Cui, Zhendong Zhang, Cong Guo, Quan Chen, Li Li, Minyi Guo. “AdaptGear : Accelerating GNN Training via Adaptive Subgraph-Level Kernels on GPUs.” **CF 2023**
  - **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**.
  - Han Zhao, **Weihao Cui**, Quan Chen, Minyi Guo. “ISPA : Exploiting Intra-SM Parallelism in GPUs via Fine-grained Resource Management.” **TC 2022**.
  - 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.” **ICS 2022**.
  - **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.” **ATC 2022**.
  - 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.” **HPCA 2021**.
  - **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.” **SC 2021**.
  - Han Zhao, **Weihao Cui**, Quan Chen, Jieru Zhao, Jingwen Leng, and Minyi Guo. “Exploiting Intra-SM Parallelism in GPUs via Persistent and Elastic Blocks.” **ICCD 2021**
  - Wei Zhang, Quan Chen, Ninxing Zheng, **Weihao Cui**, Kaihua Fu, and Minyi Guo. “Towards QoS-awareness and Improved Utilization of Spatial Multitasking GPUs.” **TC 2021**.
  - **Weihao Cui**, Quan Chen, Han Zhao, Mengze Wei, Xiaoxin Tang, and Minyi Guo. “E<sup>2</sup>bird : Enhanced Elastic Batch for Improving Responsiveness and Throughput of Deep Learning Services.” **TPDS 2021**.
  - 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.” **ICDCS 2020**.

- > **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.” **ICCD 2019**.
- > 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.” **ICS 2019**.

## Skills

Programming Skills : **C/C++, CUDA, Python**,  $\text{\LaTeX}$ , Git, Go.

## Projects & Experiences

March 2018 December 2018	<b>GPU sharing technology based on QoS awareness, SJTU &amp; Huawei, C/C++/Python</b> Features : <ul style="list-style-type: none"><li>&gt; Bare metal/container environment : support multiple users and multiple processes/threads to share a GPU single GPU resource.</li><li>&gt; Provide reasonable GPU QoS mechanisms, such as priority, end-to-end latency, and percentage of the resource.</li><li>&gt; Fully compatible with existing applications, such as various Deep Learning frameworks, without code modification and recompilation, transparent to the application software.</li></ul> <div><span>GPU</span> <span>Deep Learning</span> <span>QoS</span> <span>Resource Management</span></div>
July 2019 September 2019	<b>Performance Investigation of CNI(Container Network Interface) plugin of K8S, Alibaba Group, C/Shell</b> Responsibilities : <ul style="list-style-type: none"><li>&gt; Investigate new tracing technology in Linux kernel — eBPF (extended Berkeley Package Filter).</li><li>&gt; Investigate the CNI plugin of K8S implemented by eBPF — cilium.</li><li>&gt; Performance test between cilium and other K8S CNI plugins including calico, canal, flannel.</li></ul> <div><span>eBPF</span> <span>cilium</span> <span>CNI plugin</span> <span>K8S</span></div>
May 2020 December 2020	<b>Efficient Neural Architecture Search, Tencent, Python/Shell</b> Responsibilities : <ul style="list-style-type: none"><li>&gt; Investigate the hardware-aware neural architecture search.</li><li>&gt; Explore efficient approaches for accelerating the neural architecture search.</li></ul> <div><span>NAS</span> <span>Deep Learning</span></div>
December 2021 December 2022	<b>AI System for Dynamic Neural Network, MSRA Shanghai, Python/C++/CUDA</b> Responsibilities : <ul style="list-style-type: none"><li>&gt; Investigate dynamic neural networks.</li><li>&gt; Optimization of the training/inference of dynamic neural network.</li></ul> <div><span>Dynamic Neural Network</span> <span>Deep Learning</span></div>

## Honors & Awards

**Feb 2023** Microsoft Stars of Tomorrow.  
**Sep 2021** National Fellowship of Shanghai Jiao Tong University.  
**May 2020** Tencent rhino-bird elite training program.

(last update : 11 Oct. 2023)