Zhe Zhou

Room 512, science#5 building, Peking University, Beijing,100871

+8618811317969 • zhou.zhe@pku.edu.cn

I am currently a third-year Ph.D student of *Center For Energy-efficient Computing and Applications (CECA)* at Peking University, supervised by *Prof. Guangyu Sun*. My research interests include **near-memory processing, domain-specific accelerator design, deep-learning algorithms, and edge computing systems**.

Education

- **Peking University** — **Beijing**
  Computer Science, Ph.D Student — *2019–2024 (expected)*
  Research Interests: Near-Memory Computing, Computer Architecture, Deep Learning, Edge Computing

- **Peking University** — **Beijing**
  Earth Science, Bachelor — *2015–2019*
  Double Major: Computer Science

Award

- **2021: Merit Student of Peking University**
- **2016: Excellent Social Work Award of Peking University**

Industrial Experience

- **Alibaba DAMO Academy (Machine Intelligence Laboratory)** — **Beijing, China**
  Research Intern, Big Model Acceleration — *May 2021 – Jan 2022*
  I researched efficient Transformer-based algorithms.

- **Alibaba DAMO Academy (T-HEAD Semiconductor)** — **Shanghai and Beijing, China**
  Research Intern, Intelligent Graph Processing — *July 2020 – January 2021*
  I researched efficient GNN (Graph Neural Networks) algorithms and accelerators. I proposed to compress and accelerate GNNs with block-circulant weight matrices. Both algorithm-level and hardware-level experiments demonstrated the effectiveness of the proposed solution.

- **Advanced Institute of Information Technology (AIIT)** — **Hangzhou, China**
  Research Intern, Edge Computing — *April 2019 – April 2020*
  I studied and developed an anomaly detection system, which is able to detect the anomaly of machines by analysing the vibration signal with an LSTM network. It includes low-power FPGAs (IoT side) to run the LSTM model, an edge server to perform online training.

- **Otureo Inc** — **Beijing, China**
  Embedded Software Engineer — *August 2017 – February 2018*
  I developed a real-time (30 fps) CNN based face-detection algorithm on a Hi-Silicon 3519A embedded platform with ARM Compute Library.

Technical and Personal Skill

- **Programming Language:** C/C++, Python, OpenCL, CUDA.
- **Hardware Design Language:** Vivado HLS, Verilog,Chisel.
- **Deep Learning Framework** Pytorch, Tensorflow, Caffe.
- **Simulation Tools** Zsim, Intel PinTool, Ramulator
- **Others:** Linux, Docker, LaTeX

Publication

Zhe Zhou, Xuechao Wei, Jiejing Zhang, and Guangyu Sun. Pets: A unified framework for parameter-efficient transformers serving. In *Usenix ATC*, 2022. Acceptance ratio: 16% **(CCF-A)**.

Zhe Zhou, Junlin liu, Guangyu Sun, and Zhenyu Gu. Energon: Towards efficient acceleration of transformers using dynamic sparse attention. In *TCAD*, 2022, **(CCF-A)**.

Zhe Zhou, Bizhao Shi, Zhe Zhang, Guangyu Sun, and Guojie Luo. Blockgnn: Towards efficient gnn acceleration with block-circulant weight matrices. In *Design Automation Conference (DAC)*, 2021, **(CCF-A)**.

Zhe Zhou, Xintong Li, Xiaoyang Wang, Zheng Liang, Guangyu Sun, and Guojie Luo. Hardware-assisted service live migration in resource-limited edge computing systems. In *Design Automation Conference (DAC)*, 2020, **(CCF-A)**.

Zhe Zhou(*), Bingzhe (*) Wu, Zheng Liang, Guangyu Sun, Chenren Xun, and Guojie Luo. Saface: Towards scenario-aware face recognition via edge computing system. In *HotEdege*, 2020 (* denotes equal contribution).

Zhe Zhou, Xintong Li, and Guangyu Sun. Accelerate service live migration in resource-limited edge computing systems. In *ArchEdge*, 2019.

Xiaoyang Wang, **Zhe Zhou**, Guangyu Sun, Jidong Zhai, and Peng Han. Edge-stream: a stream processing approach for distributed applications on a hierarchical edge-computing system. In *SEC*, 2020.

Nelson Spencer, Khalil Wassim, Kim Sangyun, Di Jia, **Zhe Zhou**, Zhihang Yuan, and Guangyu Sun. Rapid configuration of asynchronous recurrent neural networks for asic implementations. In *HPEC*, 2021.