Zhe Zhou

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

+8618811317969 • zhou.zhe@pku.edu.cn • pkuzhou.github.io

I am a fourth-year Ph.D. student at Peking University's Center for Energy-efficient Computing and Applications (**CECA**), where I am supervised by Prof. Guangyu Sun. My research interests primarily lie in the areas of computer architecture, near-data processing, domain-specific accelerators, deep learning systems, and heterogeneous computing. I pay particular attention to addressing the **Memory Wall** problem, which is especially challenging for emerging memory-intensive workloads. Through my work, I have published in top-tier computer architecture/system conferences including **HPCA**, **ATC**, **DAC**, **PACT**, and **TCAD**.

Education

- **Peking University** — **Beijing**
  Computer Science, Ph.D Student — *2019–2024 (expected)*
- **Peking University** — **Beijing**
  Earth Science, Bachelor — *2015–2019*

  Double Major: Computer Science

Honor & Award

- **2022: ByteDance Scholarship (10 positions in China)**
- **2022: China National Scholarship (top 2%)**
- **2022: Academic Innovation Award of Peking University (top 1%)**
- **2021: Merit Student of Peking University**
- **2016: Excellent Social Work Award of Peking University**

Industrial Experience

- **Microsoft Research Asia (Networking Group)** — **Beijing, China**
  Research Intern, Near-Memory Processing — *Aug 2022 –*

  I focus on how to optimize the emerging Compute Express Link (CXL) based memory disaggregation systems.

- **Alibaba DAMO Academy (Machine Intelligence Laboratory)** — **Beijing, China**
  Research Intern, Big Model Acceleration — *May 2021 – Jan 2022*

  I researched efficient Transformer algorithms and inference serving systems. I proposed PetS, a unified framework for parameter-efficient transformers serving. This framework can improve the serving capacity by 26x and speeds up a multi-task transformers serving system by 50%. PetS was published in the USENIX ATC'22 conference.

- **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. BlockGNN is suitable for many compute-intensive GNNs and can achieves up to 8x speedup compared to baseline accelerators.

- **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. This project has won the first place in the "New infrastructure application scenario innovation competition of Zhe Jiang province".

Teaching

- **Teaching Assistant**. Computer Architecture Practice, 2020 fall, 2019 fall.
  In this class, we teach students how to design, optimize and implement DNN accelerators with Vivado HLS and PYNQ FPGAs.

Technical and Personal Skill

- **Programming Language:** C/C++, Python, OpenCL, CUDA.
- **Hardware Design Language:** Vivado HLS, Verilog, Chisel.
- **Used Simulation Tools:** FireSim, Zsim, Ramulator, DRAMSim, SniperSim, ChampSim, QEMU
- **Others:** Linux, Docker, LaTeX

Publication

Zhe Zhou, Cong Li, Fan Ynag, and Guangyu Sun. Dimm-link: Enabling efficient inter-dimm communication for near-memory processing. In ***HPCA***, 2023. **(CCF-A)**.

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, Cong Li, Xuechao Wei, and Guangyu Sun. Gnnear: Accelerating full-batch training of graph neural networks with near-memory processing. In ***PACT***, 2022, **(CCF-B)**.

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)**.

Cong Li, **Zhe Zhou**, Li Xingchen, and Guangyu Sun. Nmexplorer: An efficient exploration framework for dimm-based near-memory tensor reduction. In ***DAC***, 2023. **(CCF-A)**.

Xiaoyang Wang(*), **Zhe Zhou**(*), and Guangyu Sun. Fd-cnn: a frequency-domain fpga acceleration scheme for cnn-based image processing applications. In ***TECS***, 2021 **(CCF-B)**(* denotes equal contribution).

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.

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.

Nelson Spencer, Yun Kim Sang, Di Jia, **Zhe Zhou**, Zhihang Yuan, and Guangyu Sun. Reconfigurable asic implementation of asynchronous recurrent neural networks. In *ASYNC*, 2021.