Zhihao Shu

University of Georgia . School of Computing
(551)331-7226 | Zhihao.Shu@uga.edu

Education

University of Georgia - SoC Aug. 2023 – Present
PhD in Computer Science | ***GPA**: 3.98/4.0* *Athens, GA*

NYU - Courant Institute of Mathematical Sciences Aug. 2021 – May 2023
M.S. Computer Science | ***GPA**: 3.72/4.0* *New York, NY*

University of Wisconsin - Madison Aug. 2019 – May 2021
B.S. Computer Science, Math minor | ***GPA**: 3.88/4.0* *Madison, WI*

- **Honors**: Dean's List

University of Delaware Aug. 2018 – May 2019
B.S. Computer Science, Game Design minor | ***GPA**: 3.7/4.0* *Newark, DE*

- **Honors**: Dean's List

Experience

Software Engineer Intern Jun 2021 – Aug. 2021
Robotrak *Nanjing, China*

- Developed a mobile camera application integrated with an external optical device to capture retinal (fundus) images and videos using smartphone cameras.
- Enabled self-assessment of retinal health via mobile devices, improving accessibility for eye screening and telemedicine use cases.

Software Engineer Intern Jun 2020 – Aug. 2020
Robotrak *Nanjing, China*

- Built a medical-assistive app for ophthalmologists to plan and automate laser eye surgeries.
- Implemented laser point mapping, energy adjustment, and sequencing features to control the laser treatment machine.
- Enhanced precision and safety of ophthalmic laser procedures through interactive app-based surgical planning.

Research Interests

- Real-time Machine Learning
- Mobile and Edge Computing
- Parallel and High-Performance GPU Computing
- Compiler and System Co-design for Deep Learning Acceleration

Research Experience

- **Optimizing Large Language Models on Mobile GPUs** – Enhanced `llama.cpp` kernel performance by **2×** through OpenCL kernel fusion, vectorized computation, and KV cache compression; reduced model memory usage via SVD-based weight reconstruction while maintaining inference accuracy.
- **FlashMem** – Developed **FlashMem**, achieving up to **10×** lower memory usage through hierarchical storage and asynchronous prefetching; designed overlapped computation–loading pipelines and optimized GPU kernel scheduling for efficient on-device inference.
- **SmartMem** – Optimized GPU data layout and memory access patterns for efficient execution with kernel fusion and adaptive scheduling across diverse mobile architectures.
- **Real-time Core-Periphery ViT on Mobile Devices** – Proposed an algorithm–system co-design approach that jointly optimizes model sparsity and GPU execution layout, achieving real-time Vision Transformer inference on mobile GPUs.

Publications

1. **Zhihao Shu**, Md Musfiqur Rahman Sanim, Hangyu Zheng, Kunxiong Zhu, Miao Yin, Gagan Agrawal, Wei Niu. *FlashMem: Supporting Modern DNN Workloads on Mobile with GPU Memory Hierarchy Optimizations.* In Submission: ASPLOS 2026.

2. **Zhihao Shu**, Xiaowei Yu, Zihao Wu, Wenqi Jia, Yinchen Shi, Miao Yin, Tianming Liu, Dajiang Zhu, Wei Niu. *Real-time Core-Periphery Guided ViT with Smart Data Layout Selection on Mobile Devices.* NeurIPS 2024.

3. Xuan Shen, Zhao Song, Yufa Zhou, Bo Chen, Yanyu Li, Yifan Gong, Kai Zhang, Hao Tan, Jason Kuen, Henghui Ding, **Zhihao Shu**, Wei Niu, Pu Zhao, Yanzhi Wang, Jiuxiang Gu. *LazyDiT: Lazy Learning for the Acceleration of Diffusion Transformers.* AAAI 2025.

4. Xuan Shen, Zhenglun Kong, Changdi Yang, Zhaoyang Han, Lei Lu, Peiyan Dong, Cheng Lyu, Chih-hsiang Li, Xuehang Guo, **Zhihao Shu**, Wei Niu, Miriam Leeser, Pu Zhao, Yanzhi Wang. *EdgeQAT: Entropy and Distribution Guided Quantization-Aware Training for the Acceleration of Lightweight LLMs on the Edge.* arXiv.

5. Wei Niu, Md Musfiqur Rahman Sanim, **Zhihao Shu**, Jiexiong Guan, Xipeng Shen, Miao Yin, Gagan Agrawal, Bin Ren. *SmartMem: Layout Transformation Elimination and Adaptation for Efficient DNN Execution on Mobile.* ASPLOS 2024.

6. Gen Li, **Zhihao Shu**, Jie Ji, Minghai Qin, Fatemeh Afghah, Wei Niu, Xiaolong Ma. *Data Overfitting for On-Device Super-Resolution with Dynamic Algorithm and Compiler Co-Design.* ICLR 2024.