Guanshujie Fu

(+41) 077-2181-702 | fuguan@ethz.ch | Portfolio | LinkedIn

EDUCATIONS

ETH Zürich (ETHz) M.S. in Info Technology-ITET GPA: **5.25**/6.00 2023.09 - Present
University of Illinois, Urbana Champaign (UIUC) B.S. in Computer Engineering-ECE GPA: **3.85**/4.00 2019.09 - 2023.06

- Graduated with *High Honors*, Dean's List (2020&2021&2022)

SKILLS

Programming: *C/C++, Go, Python, Java, Assembly, SystemVerilog, JavaScript, P4, Haskell, MATLAB*
Frameworks/Tools: *CUDA/ROCm, PyTorch, TVM, Docker, Kubernetes, Redis, Vitis/Vivado, AWS, GCP*

EXPERIENCES

Research Assistant | ETHz, Systems Group | *LLM, ML System, ML Compiler, CUDA, Python* 2024.09 - Present
Advisor: Benjamin Ramhorst, Dr. Shien Zhu, Prof. Gustavo Alonso

- Working on machine learning compiler benchmarking tool by adding backend support for TVM
- Fast Ternary Large Language Model Inference with Sparse Addition-based SpMM on Edge Device (**ATC'25** submitted)
  - Implemented high-performance **Sparse Matrix computation kernel** and integrated as PyTorch extension for LLM inference
  - Achieved over **10x** speedup compared with cuSPARSE and **1.5x-5x** speedup over 86% sparsity compared with cuBLAS
  - Saved **>50%** memory usage and **>35%** power consumption compared with vanilla LlaMa-3.2-1B inference on Desktop GPU
  - Integrated kernel into attention layer, and achieved **>1100 token/s** generation throughput with LlaMa-3.2-1B on Desktop GPU

Software Engineer Intern | ABB Ltd | *Software Analysis, Compiler, LLVM, C++, Python* 2024.03 - 2024.09
Supervisor: S2 Group-Philip Sommer, Balz Maag

- Worked on program analysis with **LLVM** and **Intermediate Representation (IR)** for automated program parallelization
  - Delivered **three** presentations to Stakeholders from different business units and published an industry-oriented paper on ICPS'25 as main author, which opened a new direction for code analysis program and its application in development
  - Implemented **static analysis algorithm** on complex value-flow graph (VFG) generated from IR to extract data dependency, improved dependency accuracy up to **87%** compared with original implementation
  - Developed IDE extension with light-weight language server in C++ to support in-time analysis in source code
  - Integrated the tool into CI/CD process at ABB for its industry motor control system, and will be open sourced for public use

Backend Engineer Intern | HouQi Tech Co. Ltd | *Cloud, Kubernetes, Docker, Redis, Golang, C++* 2023.03 - 2023.05

- Developed Golang-based vector operation APIs using Milvus, enabling fast processing of multiple concurrent requests
- Provided low-latency **unstructured data management** as a micro-service within a larger cloud platform framework
- Used Redis as intermediate storage in vector search to support low latency (20ms) ranking algorithm for search results
- Developed a RTSP video stream pulling/pushing scheme capable of decoding and converting video data into OpenCV Mat format within 30ms. Integrated with a face detection algorithm for efficient processing

Undergraduate Researcher | UIUC, FAST Lab | *Near-Storage Computing, Xilinx FPGA, HLS, C++* 2022.02 - 2023.02
Advisor: Professor Nam Sung Kim

- Developed benchmark programs to assess SmartSSD performance across various targeted metrics in computer system
- Implemented and optimized *data compression algorithm* (Run Length Encoding and LZ77) using HLS C++ in SmartSSD
- Provided asynchronous memory page compression mechanism for utilizing SmartSSD as a page cache expander
- Implemented *data-intensive* database key value filter applications using HLS stream data and C++ to SmartSSD

PROJECTS

Integration of Sequence Parallelism in Nanotron 2024.09 - 2024.12

- Contributing to the open-source distributed transformer model training framework Nanotron
- Integrated DeepSpeed Ulysses **sequence parallel** algorithm into Nanotron code base to support long sequence training
- Tested to validate correctness and benchmarked to prove support of long sequence for LlaMa up to **64k** on 4-nodes cluster

Inspection on Distributed Training on Adversarial Network 2024.09 - 2024.12

- Developed a framework for manipulating network topology with traffic control and configure distributed training over cluster
- Provided automatic benchmark and manipulation for different backends (gloo/NCCL) over different hardware (CPU/GPU)
- Benchmarked **distributed data parallel** (DDP) and **fully shared data parallel** (FSDP) stability over different network config
- Collected and formulated network effectiveness on the training of different transformer model over different datasets

Convolutional Layer Forward-pass Acceleration [Repo] 2022.06 - 2022.08

- Optimized forward pass of convolution layer of LeNet-5 with CUDA C/C++ to run it efficiently on GPU
- Used optimization methods including Matrix Unrolling, Kernel Fusion, Streaming and Reduction
- Used NVIDIA Nsight Systems to analyze and optimize, reduced the inference operation time **from ~170ms to ~70ms**

A Unix-like OS Kernel Design [Repo] 2022.03 - 2022.05

- Led a team to design and implement an OS kernel resembling Linux with basic and advanced features in C and Assembly
- The kernel includes file system, virtual memory, process management & scheduling, interrupts & exceptions and etc.
- Designed a high-resolution (60fps, 800*600 resolution) graphic user interface with standard VGA capable