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 - 2026.06

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, CUTLASS, PyTorch, TVM, Megatron-LM, Docker, Kubernetes, Redis, Vitis/Vivado, AWS, GCP

EXPERIENCES

Machine Learning Engineer Intern | **Alibaba Cloud**, PAI Team | *LLM, ML System, CUDA/C++, Python* 2025.03 - 2025.08

- Worked in Platform for AI (PAI) team by providing supports for optimizing QWen-3 LLM pre-training tasks
  - Developed low precision (**FP8/MXFP8**) operators and training framework. Implemented FP8/MXFP8 related CUDA kernels, achieved **70%-80% MBU** for quantization and **60%-70% MFU** for FP8 GEMM computation on H800
  - Profiled end-to-end FP8 training trace, proposed optimizations including **Kernel fusion**/**Triton kernel redesign**/**Pipeline design**, reduced **host and device side latency** and achieved **10%-15% acceleration** compared to BF16 training
  - Analyzed and benchmarked 'Sweet Spot' of FP8 training under different training workload and parallel strategy
  - Designed a suite of experiments and micro-benchmarks to **understand accuracy loss** during FP8 training, and proposed an infrastructure to automatically **finetune training hyperparameters** for stable FP8 training of different models
  - Developed **token dispatch and computation pipeline** in MoE layer and achieved **~2.5% acceleration** each iteration

Research Assistant | **ETHz**, Systems Group | *LLM, ML System, ML Compiler, CUDA/C++, 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 PyTorch Compiler and TVM
- Shien Z.*, **Guan F.*,** Gustavo A. "Fast Ternary Large Language Model Inference with Sparse Addition-based SpMM" (submit)
  - Implemented high-performance Ternary Sparse Matrix Multiplication (Ternary SpMM) kernel for efficient LLM Inference
  - Achieved **10x** speedup compared with cuSPARSE and **1.5x-5x** speedup over 86% sparsity compared with cuBLAS
  - Integrated kernel into attention layer, and achieved **>110 token/s** generation throughput with Llama-3-1B on RTX-3080
  - Saved **>60%** memory usage and **>35%** power consumption during Llama-3-1B inference on Consumer-level GPU

Software Engineer Intern | **ABB**, S2 Group | *Compiler, LLVM, C++, Python* 2024.03 - 2024.09

- Worked on program analysis with LLVM and Intermediate Representation (IR) for automated program parallelization
  - Balz M., **Guan F**. "PTSAnalysis-A Static Program Analysis System to Capture Inter-Function Data Dependency" (ICPS'25)
  - 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 control-flow (CFG) and value-flow (VFG) graph from LLVM IR to extract data dependency, improved accuracy from **~50% to ~87%** and enabled analysis on complex control system with >20k lines of code
  - Implemented **LLVM Analysis Pass** that extracts data flow during compilation to accelerate and improve next-stage analysis
  - Developed IDE extension with light-weight language server in C++ to support low latency analysis in source code
  - Integrated the tool into CI/CD process at ABB for industry system development, and will open source for public use

Backend Engineer Intern | **HouQi Tech** (Start-up) | *Cloud, Kubernetes, Docker, Redis, Golang, C++* 2023.03 - 2023.05

- Provided low-latency unstructured data management as a **micro-service** with **Milvus,** a high-performance vector database
- Used Redis as intermediate storage in vector search to support low latency (20ms) ranking algorithm for search results

Undergraduate Researcher | **UIUC**, FAST Lab | *Near-Storage Computing, Xilinx FPGA, HLS, C++* 2022.02 - 2023.02

Advisor: Prof. Nam Sung Kim

- Developed benchmark framework 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

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 proved support of long sequence for Llama up to **64k** on 4-nodes cluster with 4 GPUs

Inspection on Reliability of Distributed Training on Adversarial Network [Thesis] 2024.09 - 2024.12

- Developed a framework to manipulate network topology and setup distributed training over **SLURM** 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 configs