Yu **Zhu**

4th-year PhD Student, ETH Zurich
ML Systems • AI Infra • Hardware Acceleration • Vector Search
Intern • Summer 2026
(+1) 408-797-9473 | yu.zhu@inf.ethz.ch | Yu Zhu | Google Scholar

Education

ETH Zurich — *Zurich, Switzerland*
PhD Student in Computer Science — *June 2022 - Present*
- Objective: GPU/ML Systems Intern — focusing on storage, network, and accelerator-based ML pipelines.
- Expected graduation: 2027.

ETH Zurich — *Zurich, Switzerland*
M.S. in Electrical Engineering and Information Technology — *September 2019 - March 2022*

Southeast University — *Nanjing, China*
B.E. in Electronic Science and Engineering — *September 2015 - June 2019*

Skills

Programming	Python, C/C++, Verilog, High Level Synthesis
Accelerators	CUDA (GPUDirect RDMA), FPGA (Vitis/Vivado HLS)
Networking	RDMA (RoCE/IB), UCX/libfabric, gRPC, DAOS
ML Systems	PyTorch, Tensorflow, Nsight Compute/Systems, NCCL, RAPIDS, DALI, RAG
Languages	English, Chinese

Internship

GPU-direct Object Storage System in LLM Era [5] — *HPE, California, US*
Intern — *June 2025 - August 2025*
- **SmartNIC offload:** Fully offloaded the DAOS datapath to SmartNICs (NIC-resident fast path, on-NIC crypto/compression), separating control (gRPC) and RDMA data planes to remove host mediation.
- **Impact at scale:** Demonstrated higher aggregate throughput, lower tail latency, and drastically reduced host CPU utilization under incast vs. TCP/host-resident baselines on a multi-node cluster.
- **Recognition/roadmap:** Work accepted at SC'25 RESDIS; roadmap includes multi-tenant QoS enforced on the NIC and end-to-end zero-copy to GPUs via GPUDirect RDMA.

Projects

Retrieval Augmented Generation over Distributed Object Storage — *Zurich, Switzerland*
Ongoing Project — *September 2025 – Present*
- Designing storage-aware vector indexing structures that operate directly on immutable objects in systems such as S3, Ceph, MinIO.
- Developing freshness and bounded-staleness consistency models to support timely RAG over eventually consistent object stores.
- Building cost-aware vector retrieval pipelines optimized for object-store billing models, reducing request count and bandwidth consumption.
- Exploring near-data acceleration using SmartNICs and FPGAs for partial retrieval, coarse filtering, and vector distance evaluation.

Streaming ETL for Online Recommender Model Training — *Zurich, Switzerland*
Ongoing Project — *July 2024 - Present*
- Designed a network-attached FPGA **ETL engine** that maps software ETL operators to reconfigurable hardware pipelines; supports streaming from memory/disk/network and multiple concurrent pipelines.
- Co-designed an **FPGA–GPU** data path to stream preprocessed batches directly into GPUs, eliminating CPU bottlenecks and sustaining end-to-end online training throughput.
- Demonstrated up to 85× higher ETL throughput vs. CPU clusters and up to 17× over GPU ETL on industry-scale datasets, reducing preprocessing latency and cost.

Quantized Vision Transformer for Feature Extractor

Zurich, Switzerland

Master Thesis Supervisor

July 2024 - January 2025

- Implemented the pre-trained Vision Transformer (ViT) efficiently on a local board and distributed FPGA cluster with the support of High-bandwidth Memory (HBM) and efficient network communication (TCP/IP, RDMA).
- Explored the architecture optimization, such as Flash Attention, to maximize the performance of ViT and combined with traditional preprocessing pipelines like Resize() (Nearest Neighbor or Bilinear Interpolation) to adapt to different sizes of input images.

Preprocessing Pipelines for Computer Vision

Zurich, Switzerland

Master Thesis Supervisor

February 2024 - August 2024

- Offloaded computation-intensive preprocessing tasks for Computer Vision models (ResNet, SimCLR) into FPGA, including Crop(), Resize(), Flip().
- Constructed a benchmark in contrast to common optimization libraries in CPU (tf.data) and GPU (DALI).
- Outperformed Xilinx Vision Library (X+ML pipeline) which targets for edge devices only.

In-Network Preprocessing for Recommender System [4]

Zurich, Switzerland

Initiator & Leader

April 2023 - June 2024

- Offloaded preprocessing into FPGA and aimed to mitigate the **speed mismatch** between online preprocessing in CPU and training in GPU.
- Proposed network-based solution where the FPGA can fully saturate the input bandwidth and serve for dataset larger than local memory.
- Outperformed 128-core CPU by 39∼68x and defeated Nvidia A100 GPU (accelerated by RAPIDS Suite) by 8∼17x.

Distributed Inference for Recommender Systems [1][3]

Zurich, Switzerland

Core Member

February 2021 - March 2023

- Optimized memory-bound embedding layer and computation-bound fully-connected layers for DLRM inference on FPGA.
- Applied different communication protocols for multi-node design, including 100Gbps **TCP/IP**, hardware **MPI** and **RDMA**.
- Reached an acceleration of 28x when compared with an optimized CPU baseline and 7x with a single FPGA implementation.

Hardware Acceleration for Vector Search [2]

Zurich, Switzerland

Core Member

March 2022 - August 2022

- Implemented Hierarchical-Navigable-Small-World (**HNSW**) accelerator for Approximate-Nearest-Neighbor-Search (**ANNS**) on FPGA.
- Optimized the dataflow by data partitioning, meta-info flow, edge pre-fetching, iteration overlapping and memory interleaving.
- Accelerated the vector searching process by 37x over CPU, 23x over local FPGA and 5x over GPU.

Publications

2021 [1] **Yu Zhu**, Zhenhao He, Wenqi Jiang, Kai Zeng, Jingren Zhou, and Gustavo Alonso. “Distributed recommendation inference on fpga clusters.” FPL’21.

2023 [2] Wenqi Jiang, Shigang Li, **Yu Zhu**, Johannes de Fine Licht, Zhenhao He, Runbin Shi, Cedric Renggli et al. “Co-design hardware and algorithm for vector search.” SC’23.

2024 [3] Zhenhao He, Dario Korolija, **Yu Zhu**, Benjamin Ramhorst, Tristan Laan, Lucian Petrica, Michaela Blott, and Gustavo Alonso. “ACCL+: an FPGA-Based Collective Engine for Distributed Applications.” OSDI’24.

2024 [4] **Yu Zhu**, Wenqi Jiang, and Gustavo Alonso. “Multi-Tenant SmartNICs for In-Network Preprocessing of Recommender Systems.” ArXiv preprint (2024).

2025 [5] **Yu Zhu**, Aditya Dhakal, Pedro Bruel, Gourav Rattihalli, Yunming Xiao, Johann Lombardi, Dejan Milojicic. “An RDMA-First Object Storage System with SmartNIC Offload.” SC’25 RESDIS Workshop.