Yu Zhu

(+41) 78-220-5686 | yu.zhu@inf.ethz.ch | personal web

Research

- My interests focus on computer systems, data management and reconfigurable hardware.
- Currently I am working on hardware acceleration in distributed systems.

Education

ETH Zurich — *Zurich, Switzerland*

PhD student in Computer Science — *Jul. 2022 - Present*

ETH Zurich — *Zurich, Switzerland*

M.S. in Electrical Engineering and Information Technology — *Sep. 2019 - May 2022*

Southeast University — *Nanjing, China*

B.E. in Electronic Science and Engineering — *Sep. 2015 - Jun. 2019*

Technical University of Munich — *Munich, Germany*

Exchange student in Electrical and Computer Engineering — *Oct. 2018 - Mar. 2019*

Publication

[1] Wenqi Jiang, Shigang Li, **Yu Zhu**, Johannes de Fine Licht, Zhenhao He, Runbin Shi, Cedric Renggli, Shuai Zhang, Torsten Hoefler, Gustavo Alonso. "Hardware Specialization for Vector Similarity Search." Submitted.

[2] **Yu Zhu**, Zhenhao He, Wenqi Jiang, Kai Zeng, Jingren Zhou, and Gustavo Alonso. "Distributed recommendation inference on fpga clusters." In 2021 31st International Conference on Field-Programmable Logic and Applications (FPL), pp. 279-285. IEEE, 2021.

Projects

Recommendation Inference on Large FPGA Clusters — *Zurich, Switzerland*

Ongoing Project, supervised by Prof. Gustavo Alonso — *Jul. 2022 - Present*

- Build common FPGA library for different recommendation models in a distributed method.

Graph based Approximate-Nearest-Neighbor-Search on FPGA [1] — *Zurich, Switzerland*

Master Thesis, supervised by Prof. Gustavo Alonso — *Nov. 2021 - Apr. 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, edges prefetching, iteration overlapping and memory interleaving.
- Built priority queues with different methods and selected the best design to integrate into hardware to reduce **II** of continuous insertion.
- Evaluated the throughput on 1M 128-dim SIFT dataset by tuning the number of HBM banks per kernel and the number of kernel replications, where the best design of FPGA outperformed a 32-core CPU server by 10% in terms of QPS.

Aggregation Group-by on FPGAs — *Zurich, Switzerland*

Semester Project, supervised by Prof. Gustavo Alonso — *May. 2021 - Sep. 2021*

- Designed and implemented hash-based group-by aggregation for **high cardinality** (4 HBMs were used, each HBM supported 4M cardinality).
- Took advantage of Content-Addressable-Memory (CAM) as cache to do preaggregation and avoid read-after-write hazard for off-chip memory.
- Avoided concatenating local hash tables in the final stage for **scalability** by partitioning input key-value tuples into different aggregation engines according to LSB of corresponding hash values.
- Evaluated the throughput on three datasets (uniform, hot-key, zipf) and generated software baseline in Spark SQL with 4 CPU cores. The number of input tuples was 64M and each key-value pair was 16B. When the cardinality was high, like 1M, hot-key distribution in my design performed the best, the throughput is about **6**× when compared with CPU; for uniform/zipf distribution, the acceleration of throughput was about **3**×.

Distributed Recommendation Inference on FPGA Clusters [2] — *Zurich, Switzerland*

Semester Project, supervised by Prof. Gustavo Alonso — *Oct. 2020 - Apr. 2021*

- Applied deep neural networks in personalized recommendation systems on FPGA by optimizing the memory-bound embedding layer and computation-bound fully-connected layers.
- Reduced the bottleneck of memory access by utilizing HBM and fully explored the potential of computation in FPGA cluster which is connected via **100Gbps** hardware network stack.
- Four-node cluster reached **7.68**× speedup in throughput compared with single FPGA and although the network transmission introduced extra latency, the overall latency was even smaller due to less computation.

High-Performance Signal Generator — *Nanjing, China*

Bachelor Thesis — *Oct. 2018 - Jun. 2019*

- Adopted an optimization method for high speed 48-bit **DDS**(Direct Digital Synthesizer) phase accumulator in FPGA to design a high-performance signal generator module based on the deep analysis of DDS.
- Combined hign-speed **SRAM** with **ROM** to improve the waveform storage depth of the generator module and utilized ultra low distortion and high speed 16-bit **D/A** convertor to design low-pass filter with elliptic function.

Precision Time Base Module *Nanjing, China*

Extracurricular research *Mar. 2018 - Sep. 2018*

- Adopted equal precision frequency measurement algorithm to complete the frequency measurement of external trigger signal, and completed the conversion calculation of delaying time and phasing shift offset word parameters.
- Employed DDS chip AD9914 to achieve high-precision step-shift clock generation to generate an accurate clock signal with adjustable frequency and phase, and applied SPI communication protocol to configure register and achieve 40KHz step delay pulse signal output.

Others

Programming C/C++, Python, Matlab, Verilog, HLS