Yu Zhu

(+41) 78-220-5686 | yuzhuyu@ethz.ch | personal web (more details of projects)

Education

ETH Zurich *Zurich, Switzerland*

M.S. in Electrical Engineering and Information Technology *Sep. 2019 - Present*

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] **Zhu, Yu**, et al. "Distributed Recommendation Inference on FPGA Clusters." International Conference on Field-Programmable Logic and Applications (FPL 2021). 2021.

Projects

Graph based Approximate-Nearest-Neighbor-Search on FPGA

Zurich, Switzerland

Master Thesis, supervised by Prof. Gustavo Alonso *Nov. 2021 - Present*

- Implemented Hierarchical-Navigable-Small-World (**HNSW**) accelerator for Approximate-Nearest-Neighbor-Search (**ANNS**) on FPGA.
- Optimized the dataflow by prefetching to hide memory latency and batching computation to fully utilize memory bandwidth.
- Built an efficient priority queue for parallel comparison with size **128** to reduce the initiation interval of continuous insertion.
- Evaluated the throughput on 1M 128-dim SIFT dataset. When utilizing 8 HBM ports to access data in parallel, the performance of FPGA was comparable to CPU running with 4 threads.

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 [1]

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

Languages English, Chinese