Xuhao Luo

9500 Gilman Drive, La Jolla, CA 92093
(858) 295-9996 ⋄ x3luo@eng.ucsd.edu

Education

University of California San Diego Sep 2019 - Mar 2021
M.S. in Computer Science, Department of Computer Science and Engineering
GPA: 3.78/4.0

University of Science and Technology of China (USTC) Sep 2015 - Jun 2019
B.S. in Applied Physics, School of Physical Sciences
Major in Microelectronics and Solid State Electronics

Research Interests

Operating System, Networking, Computer Architecture, Heterogeneous Computing

Research Experience

An Asynchronous Executor for Distributed ML System Jun 2020 - Sep 2020
Research Project at Microsoft Research Asia

- Designed and implemented an asynchronous executor for task scheduling and dispatching on multiple hardwares.
- Designed and implemented CUDA-based high-performance inter-GPU communication channel for distributed ML within large-scale GPU cluster.
- Multi-GPU collective operation(AllReduce, AllGather, Broadcast) throughput outperforms Nvidia NCCL under the same system setting.

An FPGA-based Disaggregated Memory System Sep 2019 - Jun 2020
Research Project at UCSD *Supervisor: Prof. Yiying Zhang*

- Working on FPGA-based distributed remote memory system for system resource disaggregation.
- Designed and implemented a go-back-N based full reliable network stack on both FPGA and host Linux server to enable high-performance reliable network communication between host and FPGA, as well as connection management for communication across multiple FPGAs and host servers. Using kernel-bypass to achieve high-throughput and low-latency.
- Achieved sub-10us latency and 10Gbps(limited by hardware interface) throughput.

Past Projects

Design and Implementation of HLS Based Quantized Neural Network Accelerator Jan 2019 - May 2019
Graduation Project *Supervisor: Prof. Xi Jin*

- Studied the 8-bit quantization algorithm, including the quantization algorithm, the dequantization algorithm and the implementation of the 8-bit quantized convolution.
- Designed and implemented a general 8-bit quantized convolution module on Xilinx Virtex FPGA, which achieved high parallelization through array architecture, and realized memory access optimization through data reuse.
- Developed the TensorFlow C++ API for the hardware accelerator using OpenCL. Used this accelerator to accelerate the ResNet-50 CNN and achieved a speedup of 5.17x and a memory usage reduction of 66% compared with the CPU TensorFlow implementation on Xeon E5 2686.

Binary Neuron Network (BNN) Acceleration using HLS Jul 2018 - Sep 2018
Summer Internship at Cornell University *Supervisor: Prof. Zhiru Zhang*

- Designed and implemented a BNN accelerator for LeNet-5 for MNIST handwritten digits recognition.
- Applied multiple methods to improve the performance of the accelerator including parallelization, pipelining, line buffer, task-level parallelism and batch processing.

· Implemented the accelerator on Zedboard, ZC706, and AWS EC2 F1. Achieved speedups of 33x(580fps), 88x(1543fps) and 114x(2170fps) compared with the software implementation baseline on Intel Xeon 5420 CPU.

Skills

Language	C/C++, Python, Go, Rust, OpenCL, Verilog, HTML, JavaScript, MATLAB
Tools/Framework	TensorFlow, Docker, Zookeeper, LLVM, Google Test