Last updated Sept 2018

409 Auburn Street #2
Ithaca, New York
14850

+1 (607) 379-7088
rz252@cornell.edu

Ritchie Zhao

Education

PhD Candidate in Electrical and Computer Engineering – Cornell University
Aug 2014 - Present
Advisor: **Zhiru Zhang**

BS in Electrical and Computer Engineering – University of Toronto
Sept 2009 – May 2014
Cumulative GPA 3.92, Graduated with honors

Research

Aug 2017 – May 2018

Explored algorithmic co-design of DNN models for specialized hardware, focusing on efficient building blocks for convolutional nets. *Work in progress*

May 2016 – May 2017

Created an accelerator for binarized neural networks (BNNs) on an embedded FPGA platform (Xilinx Zedboard). Leveraged algorithmic changes and novel hardware constructs to optimize for performance and resource usage. Achieved 15x speedup over embedded GPU with negligible loss of accuracy. The same BNN accelerator was integrated into Celerity, an academic ASIC taped-out to silicon at TSMC 16nm process. *Published in [2,3]*

Jan 2016 – Apr 2016

Explored the idea of an embedded DSL which can express common high-level synthesis (HLS) optimizations using basic language features. Modified the Halide DSL to encode Vivado HLS pragmas and generate synthesizable code. *Work in progress*

Jun 2015 – Jan 2016

Improved HLS for complex data structures (e.g. heap, hash table) by mapping to a decoupled architectural template. Latency-insensitive design exploits task-level parallelism between method calls and main program. *Published in [5]*

Aug 2014 – May 2015

Worked on cross-layer optimizations in FPGA-targeted HLS. Developed an approach for joint pipeline scheduling and LUT-mapping using mixed-ILP and implemented it in LLVM. *Published in [7]*

Industry Experience

Research Intern – Microsoft Research Redmond
May 2017 – Aug 2017 and *May 2018 – Aug 2018*

- Worked on Project Brainwave under Eric S. Chung and Doug Burger. Primary research topic was tuning neural networks for efficient deployment on the Catapult FPGA fabric
- Investigated and prototyped novel quantization techniques for DNNs on Brainwave, enabling DNN serving at very low precision
- Helped to migrate several internal DNN models from Tensorflow to FPGA

Extreme Blue Technical Intern – IBM Toronto Labs
May 2014 – Aug 2014

- Worked on IBM Dash, a mathematical programming language (similar to Matlab) which enables parallelism for hardware acceleration through language features and IR
- Helped develop the GPU device code backend for the Dash compiler, which eventually obtained 16x speedups on financial applications such as Monte Carlo Black-Scholes
- Created presentations and webpages to pitch the project to IBM executives. Team was invited to present at the IBM headquarters in Armonk

Engineering Intern – Altera Corporation
May 2012 – Aug 2013

- Worked with the timing team to build and maintain timing models for Altera's IV and V families of FPGAs. Performed engineering and end-to-end testing for the Timing Analyzer in Altera's Quartus II Design Suite
- Independently worked on features such as piecewise-linear voltage waveform propagation, per-node clock uncertainty, and timing legality checks
- Created software patches for a key Altera customer while coordinating with field engineers

Publications

[1] E. Chung, J. Fowers, K. Ovtcharov, M. Papamichael, A. Caulfield, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman, M. Abeydeera, L. Adams, H. Angepat, C. Boehn, D. Chiou, O. Firestein, A. Forin, K. S. Gatlin, M. Ghandi, S. Heil, K. Holohan, A. El Husseini, T. Juhasz, K. Kagi, R. K. Kovvuri, S. Lanka, F. van Megen, D. Mukhortov, P. Patel, B. Perez, A. G. Rapsang, S. K. Reinhardt, B. D. Rouhani, A. Sapek, R. Seera, S. Shekar, B. Sridharan, G. Weisz, L. Woods, P. Y. Xiao, D. Zhang, **R. Zhao**, and D. Burger. "Serving DNNs in Real Time at Datacenter Scale with Project Brainwave". *IEEE Micro*, Mar. 2018

[2] S. Davidson, S. Xie, C. Torng, K. Al-Hawaj, A. Rovinski, T. Ajayi, L. Vega, C. Zhao, **R. Zhao**, S. Dai, A. Amarnath, B. Veluri, P. Gao, A. Rao, G. Liu, R. K. Gupta, Z. Zhang, R. G. Dreslinski, C. Batten, and M. B. Taylor. "The Celerity Open-Source 511-Core RISC-V Tiered Accelerator Fabric: Fast Architectures and Design Methodologies for Fast Chips". *IEEE Micro*, Mar. 2018

[3] **R. Zhao**, W. Song, W. Zhang, T. Xing, J.-H. Lin, M. Srivastava, R. Gupta, and Z. Zhang. "Accelerating Binarized Convolutional Neural Networks with Software-Programmable FPGAs". *Int'l Symp. on Field-Programmable Gate Arrays (FPGA)*, Feb. 2017

[4] S. Dai, **R. Zhao**, G. Liu, S. Srinath, U. Gupta, C. Batten, and Z. Zhang. "Dynamic Hazard Resolution for Pipelining Irregular Loops in High-Level Synthesis". *Int'l Symp. on Field-Programmable Gate Arrays (FPGA)*, Feb. 2017

[5] **R. Zhao**, G. Liu, S. Srinath, C. Batten, and Z. Zhang. "Improving High-Level Synthesis with Decoupled Data Structure Optimization". *Design Automation Conference (DAC)*, Jun. 2016

[6] M. Tan, G. Liu, **R. Zhao**, S. Dai, and Z. Zhang. "ElasticFlow: A Complexity-Effective Approach for Pipelining Irregular Loop Nests". *Int'l Conf. on Computer-Aided Design (ICCAD)*, Nov. 2015

[7] **R. Zhao**, M. Tan, S. Dai, and Z. Zhang. "Area-Efficient Pipelining for FPGA-Targeted High-Level Synthesis". *Design Automation Conference (DAC)*, Jun. 2015

Teaching

ECE2300 - Introduction to Digital Logic
Spring 2016
PhD Teaching Assistant, Cornell University