Tianchu **Ji**

tianchu.ji@stonybrook.edu | chickenjohn | tianchu-ji

Education

Stony Brook University — *Stony Brook, NY*

PH.D. CANDIDATE, COMPUTER ENGINEERING — *Sept. 2016 - Present*

Huazhong University of Science and Technology — *Wuhan, China*

B.E., IC DESIGN AND INTEGRATED SYSTEMS — *Sept. 2012 - June 2016*

Publications

- **T. Ji**, S. Jain, M. Ferdman, P. Milder, H. A. Schwartz, and N. Balasubramanian, "On the Distribution and Sparsity of Attention within Transformers," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Aug. 2021.
- Y.Shen, **T. Ji**, M. Ferdman, and P. Milder, "Argus: an End-to-End Framework for Accelerating CNNs on FPGAs", in *IEEE Micro*, 2019.
- Y.Shen, **T. Ji**, M. Ferdman, and P. Milder, "A scalable memory interconnect for many- port DNN accelerators and wide DRAM controller interfaces", in *28th International Conference on Field Programmable Logic and Applications(FPL).*, 2018.

Research Projects

Inference-time Sparse Attention in the Transformers and its Hardware Acceleration — *Stony Brook, NY*

AHGO, COMPAS AND LUNR LAB@STONY BROOK UNIVERSITY — *Sep. 2020 - Present*

- Discovering the high sparsity in the self-attention mechanism
- Pruning **80%** of the sparsity with less than **1.0%** accuracy drop, no retraining.
- **3-bit** inference-time quantization.
- Utilizing the sparsity in the hardware accelerator for the transformers model.

Long Short-Term Memory Neural Network Acceleration on FPGA — *Stony Brook, NY*

AHGO LAB AND COMPAS LAB@STONY BROOK UNIVERSITY — *Aug. 2018 - May 2020*

- a highly scalable and resource-efficient FPGA hardware acceleration on FPGA
- a cycle-accurate latency analyzer helping analyzing delay of LSTM computation
- a resource-aware latency optimizer which generates the optimized accelerator

High Resolution Time-to-Digital Converter on FPGA — *Stony Brook, NY*

AHGO LAB AND COMPAS LAB@STONY BROOK UNIVERSITY — *May 2019 - April 2020*

- **1.9-ps** RMS resolution at **590MHz** on Xilinx Ultrascale+
- Double-Edge triggered carry chain for better performance

Scalable memory interconnect for many-port DNN accelerators on FPGA — *Stony Brook, NY*

AHGO LAB AND COMPAS LAB@STONY BROOK UNIVERSITY — *Jan. 2017 - July 2018*

- An interconnect is implemented between DNN accelerator and DRAM controller with wide data width without unnecessary flexibility
- Our design can both be scalable and reach high throughput.
- Our design reduces LUT and FF use by **4.7x** and **6.0x**, and improves frequency by 1.8x.

Working Experience

Amazon Web Services (AWS) Redshift AQUA Team — *East Palo Alto, CA*

SDE INTERN — *May. 2020 - Aug. 2020*

- Achieving **1.29x** DDR4 throughput by removing the auto-refresh overhead
- Offline DDR Retention Time profiling and online DDR access checking

Skills

- C/C++, Python, Verilog, SystemVerilog, SpinalHDL, Chisel, Tensorflow, PyTorch, Huggingface
- Static Timing Analysis, manual route&placement
- Xilinx Spartan-6, Virtex-7, Virtex-Ultrascale+