Taeyoung Kong

www.kongty.com linkedin.com/in/kongty kongty0329@gmail.com

Summary

ML hardware architect focused on scale-up networks and LLM parallelism trade-offs for AI accelerators. Build analytical and cycle-level performance models that drive perf/TCO analysis, performance optimization, and HW/SW co-design.

Education

Stanford University, Stanford, CA
Ph.D. in Electrical Engineering (Advisor: Prof. Mark Horowitz) 2024
M.S. in Electrical Engineering 2020
Seoul National University, Seoul, Korea | *B.S. in Electrical and Computer Engineering* 2017

Experience

Google, Sunnyvale, CA | *Machine Learning Hardware Architect* Jan 2024 – Present

- Lead TPU 8t/8i collective performance on torus/Dragonfly topologies, co-optimizing network routing and collective algorithms and guiding kernel engineers: all-to-all at 99% of roofline; over 2× faster all-gather/reduce-scatter.
- Lead development of a flow-based scale-up network simulator that generates production routing and collective algorithms and serves as the roofline reference for silicon correlation.
- Drive HW/SW co-design of the next-generation TPU scale-up network (ICI), evaluating topologies, perf/TCO of future network technologies, and collective algorithm trade-offs across LLM training and serving scenarios.
- Build analytical performance models of LLM pretraining parallelism strategies to identify compute, memory, and communication bottlenecks and find perf/TCO-optimal TPU architectures.
- Develop a cycle-level ICI performance simulator, 99% correlated with silicon, to explore network design spaces, estimate collective performance, and identify microarchitecture bottlenecks.

Stanford University, Mark Horowitz Research Group | *Ph.D. Researcher* Sep 2017 – May 2024

- Co-designed CGRA hardware and a scheduler that dynamically selects dataflows for concurrent DNNs; built an analytical CGRA performance model to evaluate schedules: up to 1.24× throughput, 61% lower latency (WDDSA'22).
- Led CGRA memory system design, analyzing performance and power efficiency across image processing and ML workloads with different dataflows, using a hardware generator DSL for design-space exploration (TECS'22).
- Contributed to tape-outs of two generations of the AHA programmable CGRA accelerator for image processing and ML in GF 12nm and TSMC 16nm (JSSC'23, JSSC'25).
- Developed an efficient video object detector in PyTorch that uses a tracker to exploit temporal correlation and process only regions of interest, cutting operations 5.1–8.7× at equal accuracy (MLSys'19).

Ph.D. Internships | *Meta, Samsung, Xilinx* Summers 2019 – 2021

- **Meta** (2021): Built a design-space exploration tool for floorplan and microarchitecture of ML accelerator MAC arrays.
- **Samsung** (2020): Built a numerics-aware analytical NPU model to compare PPA across dataflows and numeric formats.
- **Xilinx** (2019): Built a C++ controller for dynamic QoS of on-chip network (NoC) traffic in Versal ACAP.

FCI (Silicon Motion), South Korea | *Digital Design Engineer* Apr 2012 – Jun 2014

- Designed and verified RTL and implemented C++ functional models for 3G/4G transceivers in Samsung Galaxy phones.

Skills

Languages: C++, Python, SystemVerilog **Tools & Frameworks**: PyTorch, XProf

Selected Publications

[JSSC'25, VLSI'24] *Onyx: A 12nm Programmable Accelerator for Dense and Sparse Applications.* K. Koul, . . . , **T. Kong**, et al.
[JSSC'23, VLSI'22] *Amber: A 16nm System on Chip with a Coarse-Grained Reconfigurable Array for Flexible Acceleration of Dense Linear Algebra.* K. Feng*, **T. Kong***, et al. (*equal contribution)
[TECS'22] *AHA: An Agile Approach to the Design of Coarse-Grained Reconfigurable Accelerators and Compilers.* K. Koul, J. Melchert, . . . , **T. Kong**, et al.
[WDDSA'22] *Hardware Abstractions and Hardware Mechanisms to Support Multi-Task Execution on Coarse-Grained Reconfigurable Arrays.* **T. Kong**, K. Koul, P. Raina, M. Horowitz, C. Torng.
[MLSys'19] *CaTDet: Cascaded Tracked Detector for Efficient Object Detection from Video.* H. Mao, **T. Kong**, W. J. Dally.