Kyushick Lee

+1 (347) 925 1315 | kyushick@gmail.com | github.com/kyule7 | https://www.linkedin.com/in/kyushick-lee/ | Date: Mar. 6, 2026

Software engineer and computer architect with 10+ years of experience in memory and computing system architecture, parallel programming, system resilience, with a recent focus on LLM inferencing/training.

PROFESSIONAL EXPERIENCE

Microsoft, Redmond, WA, USA

Senior Software Engineer - Azure Hardware Architecture (AHA), AI Frameworks — *Aug. 2021 - Present*

Building kernels, runtime libraries and integrated an LLM serving stack on Maia ASIC accelerators for large-scale inference.

- Designed the Maia host/device programming model and runtime semantics (kernel launch, streams/events, synchronization).
- Delivered key components of the Maia SDK (host runtime, bindings) and integrated Maia into PyTorch and ONNX Runtime.
- Implemented device/stream control management to achieve asynchrony and concurrency with low control overhead.
- Developed memory and scratchpad allocators and abstractions to shard and replicate tensors across and within devices.
- Built a link/load system for device binaries to enable JIT workflows and integrate Maia into Triton and MSCCL.
- Partnered with OpenAI to integrate Maia kernels and KV-cache/graph into their inference stack, validated and deployed Maia-enabled (first-gen) LLM inferencing on Azure, and delivered the Maia-powered GitHub Copilot demo at Ignite 2023.
- Owned custom MoE kernels for the adopted MAI model and fused MoE kernels tuned for Maia, and end-to-end validation.
- Built Generic Vector Kernel framework for programmability/extensibility, replacing legacy ops for coverage/performance.
- Productized GVK-based kernel authoring (runtime support + binding pattern) and demonstrated fused vector ops (e.g., MoE).

Software Engineer II - Azure Hardware Architecture (AHA) — *Aug. 2019 - Aug. 2021*

- Built kernel and collective libraries for an FPGA training accelerator; integrated with ONNX Runtime for distributed training.
- Designed a hardware abstraction layer for FPGA control, enabling flexible low-level resource access.
- Built a simulator modeling I/O processing, command serialization, and kernel launch/data transfer.
- Developed a header-only checkpointing library to preserve and restore application state across host and device.
- Created a configuration tool to manage the simulator and FPGA software stack across multiple FPGA SKUs.
- Established parallel pipelines (CTest, PyTest), hardware tests in self-hosted labs; profiling plugins for performance regression.

University of Texas at Austin, USA — *Advisor: Dr. Mattan Erez*

Research Assistant - Locality Parallelism and Hierarchy (LPH) group — *Aug. 2013 – Aug. 2019*

- Developed Containment Domains runtime system/analytical model/tools alternative to existing checkpoint systems.
- Enabled and evaluated Containment Domains in MPI, CUDA, Legion programs.

Intel, Hillsboro, OR, USA — Mentor: Dr. Suresh Srinivas

Graduate Engineering Intern - Open Source Technology Center — *May. 2018 – Aug. 2018*

- Characterized front-end bottlenecks of executing Node.js using perf, and optimized code layout based on profiles.

Nvidia Research, Austin, TX, USA — Mentor: Dr. Steve Keckler

Research Intern - Architecture Research Group — May. 2017 – Aug. 2017

- Studied the resilience trend in GPU-dense systems, and architectural support for scalable GPU checkpointing.

Nvidia Research, Santa Clara, CA, USA — Mentor: Dr. Steve Keckler

Research Intern - Architecture Research Group — May. 2016 – Aug. 2016

- Developed transparent checkpointing system for CUDA programs, and evaluated it with CUDA HPC applications.

Lawrence Livermore National Laboratory, Livermore, CA, USA — Mentor: Dr. Greg Bronevetsky

Research Intern - Center for Applied Scientific Computing — Jun. 2014 – Aug. 2014

- Developed the analysis tool that predicts the performance and characterizes the effect of soft errors in an application.

Seoul National University, South Korea — Mentor: Dr. Jung Ho Ahn

Research Intern - Scalable Computer Architecture Laboratory — Jan. 2012 – Jun. 2013

- Design and modeling of high-radix crossbar switch.

EDUCATION

The University of Texas at Austin - *Ph.D. in Electrical and Computer Engineering* — GPA: 3.7 / 4.0 — Aug 2019

Hanyang University, Seoul, South Korea - *B.S. in Electrical and Computer Engineering* — GPA: 4.2 / 4.5 — Feb. 2013

SKILLS

Languages: C/C++, Python, MPI, CUDA, OpenMP, Pthread, Verilog, JavaScript, PHP, Lua

Tools: gdb, DDT, Visual Studio, CMake, LLVM, Pin, Perf, VTune, Virtuoso, Vivado, HSPICE

Selected Publications

- **Kyushick Lee**, Michael Sullivan, Siva Kumar Sastry Hari, Timothy Tsai, Stephen W. Keckler, Mattan Erez, "GPU Snapshot: Checkpoint Offloading for GPU-Dense Systems", International Conference on Supercomputing (ICS), June 2019
- **Kyushick Lee**, Michael Sullivan, Siva Kumar Sastry Hari, Timothy Tsai, Stephen W. Keckler, Mattan Erez, "On the Trend of Resilience for GPU-Dense Systems", Dependable Systems and Networks (DSN), **best paper** in SELSE, June 2019
- Karthik Murthy, Mike Bauer, **Kyushick Lee**, Alex Aiken, Mattan Erez et al, "Resilience À la carte: Application Tailored Resilience in Legion" Poster Symposium at PSAAP II review, Palo Alto, CA, Dec. 2018
- Michael Sullivan, Ikhwan Lee, Jinsuk Chung, **Kyushick Lee**, Mattan Erez et al, "Containment Domains Semantics version 0.2", Technical Report TR-LPH-2014001, 2014