Jiho Kim

E-mail: jihokimhi@gmail.com

Summary

Computer System Architect & Researcher specializing in High-Performance AI Infrastructure. Deep understanding of computer architecture, memory subsystem, storage, GPUs and system software demonstrated by Top-tier publications at **ISCA, MICRO and HPCA.** By translating these fundamental architectural insights into the networking domain, I have hands-on industry experience in developing **FPGA-based RDMA systems to accelerate AI workloads.** Currently, working on memory subsystem research to enable memory centric AI inference system including DRAM, HBM and 3D Stacked Memory.

Skills

Specialization: RDMA (RoCEv2), AI Infrastructure Network, High-performance Interconnect.
Computer Architecture: GPU, SSD, System Software, HW-SW Co-design, Performance Modeling, System Debugging
Simulators: gpgpu-sim, simpleSSD, booksim, htsim
Tools/Languages: C/C++, CUDA, Python, SystemC, Verilog, FPGA
AI-Augmented Engineering: SONiC feature integration pipelines, FPGA-based HW module generation.

Experience

SK Hynix Gyeonggi-do, South Korea
Technical Leader (TL), Memory System Research *Aug. 2026 – Current*

- **3D Stacked DRAM Research:** Realizing High-Performance AI Inference through 3D Stacked DRAM Architecture. Accounting fora practical constraints of DRAM architecture and technologies, we propose a 3D Stacked DRAM memory subsystem that delivers TB/s-scale memory bandwidth, exceeding what HBM class memory can provide

Mangoboost Seoul, South Korea
System Architect, AI Infra TF *Nov. 2025 – Jun. 2026*

- **RDMA–SONiC Collaboration :** Researching cross-stack optimizations between RDMA (RoCEv2) endpoints and SONiC-based switch fabric for AI training/inference workloads. Prototyping switch-side support for advanced RDMA behaviors (congestion control, packet spary) and building automation pipelines for feature integration and validation.

Mangoboost Seoul, South Korea
System Architect, RDMA Hardware Team *May. 2023 – Oct. 2025*

- **RDMA Selective Retransmission:** Implemented RDMA SACK to selectively retransmit dropped packets. **2.42x** bandwidth (or goodput) has been increased by minimizing redundant retransmissions in the network compared to Go-back-N
- **Packet Spraying and Out-of-Order delivery:** Delivered **2x** performance improvement for AI collective communication by implementing packet-level adaptive routing on Leaf-Spine topologies to resolve QP load imbalances.
- **400G FPGA-based RNIC:** Optimized RDMA pipeline using SystemC on Xilinx Versal-based Mangoboost cards, achieving a 400Gbps line rate. Doubled bidirectional RNIC throughput by re-architecting data paths and optimizing hardware resource utilization (LUT, BRAM, and URAM) post-Versal porting.
- **Sample Card Release:** Design and Implementation of high-performance RoCEv2 on Xilinx Alveo Ultrascale+ (U55C and U45N) smartNIC platforms. Optimized throughput up to 200Gbps to support AI and NVMe-over-Fabrics (NVMe-oF) workloads.Achieved a **15%** performance improvement for small-sized WQEs by implementing Inline CMD mechanisms. Integrated Fast Memory Registration (FMR) to mitigate kernel-level latency and prevent bottlenecks in NVMe-oF environments. Resolved complex hardware deadlock conditions through system-level debugging using ILA (Integrated Logic Analyzer.)

Samsung Memory Hwaseong, South Korea
Intern, Advanced Solution Team *Dec. 2022 – Feb. 2023*

- **SmartSSDs:** Enabling communication between SmartSSD 2.0s using Aurora IP for two VMK180 Versal Xilinx FPGAs

Education

Ph.D, School of Electrical Engineering, KAIST, Aug. 2018 – Feb. 2024 **(Advisor: John Kim)**
M.S.E, Depart. of Electronic and Electrical Engineering, Hongik University, Mar. 2015 – Aug. 2017
B.S, Depart. of Electronic and Electrical Engineering, Hongik University, Mar. 2009 – Feb. 2015

Selected Publications

1st Author, "Decoupled SSD: Rethinking SSD Architecture through Network-based Flash Controllers", **ISCA 2023**
1st Author, "Networked SSD: Flash Memory Interconnection Network for High-Bandwidth SSD", **MICRO 2022**
1st Author, "Navigator: dynamic multi-kernel scheduling to improve GPU performance",**DAC 2019**
1st Author, "Improving GPU multitasking efficiency using dynamic resource sharing", **CAL 2018**
Author, "Uncovering Real GPU NoC Characteristics: Implications on Interconnect Architecture", **MICRO 2024**
Author, "Ghost Arbitration: Mitigating Interconnect Side-Channel Timing Attacks", **MICRO 2024**
Author, "Network-on-Chip Microarchitecture-based Covert Channel in GPUs", **MICRO 2021**
Author, "Trident: A Hybrid Correlation-Collision GPU Cache Timing Attack for AES Key Recovery", **HPCA 2021**