

Sam Son

sam.son@berkeley.edu | [Google Scholar](#) | [LinkedIn](#)

RESEARCH INTERESTS

I design and analyze systems to make large-scale computing infrastructure more cost-efficient. My current work addresses a deployment question for large-scale LLM serving: *given serving scenarios, what infrastructure (hardware and interconnect), configuration (parallelism), or system-level idea (job colocation) delivers the lowest cost at a target SLO?* My recent projects approach this question through high-fidelity performance modeling of LLM serving and the design of GPU-sharing techniques for MoE inference clusters. Previously, I worked on CXL memory tiering and CPU-cycle harvesting, both aimed at extracting more useful work from existing server hardware.

Keywords: LLM Serving Infrastructure, GPU Cluster Management, Capacity Planning, Operating Systems.

EDUCATION

University of California, Berkeley

Berkeley, CA

Ph.D. in Electrical Engineering and Computer Sciences

Aug 2022 – Expected Dec 2026

Advisors: Scott Shenker and Sylvia Ratnasamy

Seoul National University M.S. in Computer Science & Engineering (Advisor: Jae W. Lee)

Mar 2020 – Aug 2022

POSTECH B.S. in Convergence IT Engineering, *Summa Cum Laude*

Mar 2014 – Feb 2020

RESEARCH & INDUSTRY EXPERIENCE

University of California, Berkeley

Berkeley, CA

Graduate Student Researcher

Aug 2022 – Present

- **GPU sharing for MoE inference clusters (current):** Designing GPU-sharing mechanisms that solves low cluster-level GPU utilization in MoE LLM serving. I address stragglers under wide expert parallelism and performance estimation under KV-cache/expert-routing imbalance.
- **GPU Interconnect analysis for LLM serving:** Quantified the cost-performance impact of GPU network design on MoE serving, guiding scale-up cluster design.
- **Server memory & CPU efficiency:** Built CXL tiered-memory management and CPU stall-cycle harvesting to get more useful work out of existing servers.

Meta Platforms, Inc.

Menlo Park, CA

Research Scientist Intern, AI-Systems Co-design

May 2025 – Aug 2025

- Co-developed a high-fidelity LLM inference performance model for capacity planning. Evaluated the benefit of NVL72 and communication-computation overlap techniques

Meta Platforms, Inc.

Remote

Part-time Student Researcher, Meta Superintelligence Lab

Aug 2025 – Dec 2025

- Spatial GPU sharing to improve fleet-level resource efficiency of large-scale generative AI inference.

Meta Platforms, Inc.

Menlo Park, CA

Research Scientist Intern, AI-Systems Co-design

May 2024 – Aug 2024

- Contributed to the CXL prototype evaluation aimed at lowering datacenter DRAM cost through memory tiering.

SELECTED PUBLICATIONS

Rethinking Network Topologies for Cost-Effective Mixture-of-Experts LLM Serving | *arXiv preprint, 2026*

- J. Choi[†], **S. Son**[†], S. Choi, H. Kim, Y. S. Shao, S. Shenker, S. Ratnasamy, B. Nikolic. [arXiv:2605.00254](https://arxiv.org/abs/2605.00254)
- Quantified the performance–cost tradeoff of GPU network topologies (scale-up vs. scale-out) for MoE LLM serving, informing which interconnect to deploy for a given workload.

Extending Managed Runtimes for Object-Level CXL Memory Management | *EuroSys 2027 (to appear)*

- **S. Son**, Z. Luo, W. Zhang, S. Ratnasamy, S. Shenker.
- Designed and implemented object-level CXL memory management for managed runtimes, making CXL tiered memory a cost-effective substitute for local DRAM.

Harvesting Memory-bound CPU Stall Cycles in Software with MSH | *OSDI 2024*

- Z. Luo, **S. Son**, S. Ratnasamy, S. Shenker.
- Recovered CPU cycles otherwise wasted on memory stalls via software hyperthreading, raising server utilization without the rigidity of hardware SMT.

Efficient Microsecond-scale Blind Scheduling with Tiny Quanta | *ASPLOS 2024*

- Z. Luo, **S. Son**, D. Bali, E. Amaro, A. Ousterhout, S. Ratnasamy, S. Shenker.
- Built a microsecond-scale CPU scheduler using compiler-assisted cooperative preemption, cutting scheduling overhead for latency-critical services.

FlashNeuron: SSD-Enabled Large-Batch Training of Very Deep Neural Networks | *FAST 2021*

- J. Bae, J. Lee, Y. Jin, **S. Son**, S. Kim, H. Jang, T. J. Ham, J. W. Lee.
- SSD-based tensor offloading that relieves GPU memory pressure, enabling larger training batches.

[†] Equal contribution. See Google Scholar for the complete publication list.

SELECTED HONORS & AWARDS

Korean Foundation for Advanced Studies (KFAS) Scholarship	2022 – 2027
Best Master’s Thesis Award , Seoul National University CSE	Aug 2022
Best TA Award , College of Engineering, Seoul National University	Jun 2020

TECHNICAL SKILLS

Programming Languages: C, C++, Python, Go, Shell
ML/GPU Systems: PyTorch, CUDA, vLLM, NCCL, Nsight Systems/Compute
Infrastructure Analysis: Performance modeling & simulation, cost/TCO analysis
Systems: Linux Kernel, GDB, ftrace, perf