SURIM OH

soh31@ucsc.edu

RESEARCH INTEREST

My research interests lie broadly in computer architecture and systems. I have recently been working on the decoupled frontend and instruction prefetching in modern processors to improve the CPU performance for datacenter workloads with large instruction footprints. I have worked on the topic of ASIC for LLMs during the internships.

EDUCATION

University of California, Santa Cruz *Santa Cruz, CA, USA*
PhD Student in Computer Science and Engineering *Sep 2020 - Present*
· Advisor: Professor Heiner Litz, Total GPA: 4.0/4.0

Seoul National University *Seoul, South Korea*
Master of Science in Computer Science and Engineering *Feb 2015 - Feb 2017*
· Advisor: Professor Bernhard Egger, Total GPA: 3.52/4.0
· Thesis: *Hierarchical Manycore Resource Management Framework using Control Processors* [pdf]

Sogang University *Seoul, South Korea*
Bachelor of Science in Computer Science and Engineering *Feb 2011 - Feb 2015*
· Total GPA: 3.70/4.0 (95.8/100), Summa Cum Laude
· Exchange student at **Northern Arizona University** *AZ, USA, Spring 2014*

WORK EXPERIENCE

Meta *Sunnyvale, California, USA*
ASIC Engineer Intern, Architecture *June 2025 - Sep 2025*

Meta *Sunnyvale, California, USA*
ASIC Engineer Intern, Architecture *June 2024 - Sep 2024*

Akeana *San Jose, California, USA*
PhD Intern *June 2023 - Sep 2023*

SAP Labs Korea *Seoul, South Korea*
Developer *Jan 2018 - Sep 2020*

Hyundai Motor Company R&D Division *Hwaseong, South Korea*
Engineer *Feb 2017 - Jan 2018*

SKILLS

C, C++, Python, Bash, Scarab simulator, Intel Perf, DynamoRIO, Verilog, Magic, OpenCL, Intel PCM, LLVM, HSIM-TQSIM (System C-based) manycore simulator

PUBLICATIONS

HSCC 2025: Paul K. Wintz, Yasin Sonmez, Paul Griffoen, Mingsheng Xu, **Surim Oh**, Heiner Litz, Ricardo G. Sanfelice, Murat Arcak. SHARC: Simulator for Hardware Architecture and Real-time Control. *In Proceedings of the 28th ACM International Conference on Hybrid Systems: Computation and Control (HSCC)*, 2025. [pdf]

HotInfra 2024: **Surim Oh**, Eric Qin, Yang Yang, Mengchi Zhang, Raj Parihar, Ashish Pandya. LLM Inference Performance on Chiplet-based Architectures and Systems. *for Poster Presentation In the 2nd Workshop on Hot Topics in System Infrastructure (HotInfra) Co-located with SOSP*, 2024. [pdf]

ISCA 2024: **Surim Oh**, Mingsheng Xu, Tanvir Ahmed Khan, Baris Kasikci, Heiner Litz. UDP: Utility-Driven Fetch Directed Instruction Prefetching. *In the 51th International Symposium on Computer Architecture (ISCA)*, 2024. [pdf]

M.S. Thesis: **Surim Oh**. Hierarchical Manycore Resource Management Framework using Control Processors. *Seoul National University*, Seoul, South Korea, February 2017. [pdf] [slides]

TPDS 2019: Younghyun Cho, **Surim Oh**, and Bernhard Egger. Performance Modeling of Parallel Loops on Multi-Socket Platforms using Queueing Systems. In *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, in press, available online, September 2019. [pdf]

MULTIPROG 2017: Younghyun Cho, **Surim Oh**, and Bernhard Egger. Cooperative Parallel Runtimes for Multicores. *Presented at the 10th International Workshop on Programmability and Architectures for Heterogeneous Multicores*, January 2017. [pdf]

CATC 2016: **Surim Oh**, Younghyun Cho, and Bernhard Egger. Efficient Resource Management for Many-cores with Centralized L2 Caches using Distributed Control Processors. *Presented at the 7th Compiler, Architectures and Tools Conference*, September 2016. [pdf] [slides]

PACT 2016: Younghyun Cho, **Surim Oh**, and Bernhard Egger. Online Scalability Characterization of Data-parallel Programs on Many Cores. In *Proceedings of the 25th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, September 2016. [pdf]

JSSPP 2016: Younghyun Cho, **Surim Oh**, and Bernhard Egger. Adaptive Space-shared Scheduling for Shared-memory Parallel Programs. *Presented at the 20th Workshop on Job Scheduling Strategies for Parallel Processing, May 2016. In Lecture Notes in Computer Science (LNCS)*, Volume 10353, July 2016. [pdf]

PATENTS

US Patent 2018A: Bernhard Egger, **Surim Oh**, Younghyun Cho, Dong-hoon Yoo. Method of processing OpenCL Kernel and Computing Device Therefor. *US Patent 20180181443A1*, June 2018. Worldwide applications in KR, EP, CN, JP including US.

US Patent 2018B: Bernhard Egger, Younghyun Cho, **Surim Oh**. Dong-hoon Yoo. Computing devices and methods of allocating power to plurality of cores in each computing device. *US Patent 20180246554A1*, August 2018. Worldwide applications in KR, CN including US.

RESEARCH AND PROJECT EXPERIENCE

University of California, Santa Cruz *Santa Cruz, CA, USA*
PhD Student *Sep 2020 - Present*

- Studied a state-of-the-art **Fetch Directed instruction Prefetching (FDIP) on a CPU microprocessor** simulator, Scarab, and the performance impact of FDIP on frontend-bound applications with large instruction footprints.
- Introduced Utility-Driven FDIP (UDP) by learning the performance impact of running-ahead distance of FDIP and the usefulness of instruction cache lines for optimal running-ahead distance and filtering mechanisms - *ISCA 2024.*

Meta *Sunnyvale, CA, USA*
ASIC Engineer Intern, Architecture *Jun 2024 - Sep 2024*

- Performance projection of LLM Inference on NVIDIA H100, B200, and two different versions of MTIAs by exploiting an open-source hardware evaluation framework, LLMCompass.
- Detailed analysis on chiplet-based architecture has been submitted as a paper to a workshop HotInfra'24 co-located with SOSP 2024.

Akeana *San Jose, CA, USA*
PhD Intern *Jun 2023 - Sep 2023*

- Designed and implemented a trace cache storing consecutively executed basic blocks on an internal RISC-V ISA simulator on top of Spike.
- The designed trace cache with branch history and confidence counter obtains performance gain by expanding fetch bandwidth.

SAP Labs Korea *Seoul, South Korea*
Developer *Jan 2018 - Sep 2020*

- Contributed to table replication technology of **SAP HANA DBMS** to scale out mixed OLTP/OLAP workloads.
- Designed and implemented replication log formats and protocols for querying the logs from other DB systems to be compatible with other general DB systems with minimum source side cost.

Hyundai Motor Company R&D Division *Hwaseong, South Korea*
Engineer *Feb 2017 - Jan 2018*

- Contributed to **Vehicle data monitoring system** that collects in-vehicle data from many distributed ECUs and store the data in a remote centralized server.
- Designed and implemented a merge tool for video data obtained from testing cars to develop Advanced Driver Assistance Systems technology.

Computer System And Platform Laboratory, Seoul National University *Seoul, South Korea*
Graduate Researcher *Feb 2015 - Feb 2017*

- Worked on **resource management techniques on manycore SoC architecture** for simultaneously running OpenCL applications, distributing/minimizing runtime scheduling overhead and power consumption. The technique exploits architecture support of tiny/hierarchical control processors on top of a TQSIM-HSIM (Timed QEMU-based & SystemC-based) manycore simulator. - *CATC 2016, M.S. Thesis, US Patent 2018A, US Patent 2018B.*
- Collaborated with four other research groups from SNU and Samsung Advanced Institute of Technology and presented at HumanTech Paper Award organized by Samsung (not awarded).
- Developed a tool that queries HW performance counters related to the NUMA interconnection network in Intel/AMD systems for an **analytical performance model on NUMA architecture** based on queueing theory to estimate the resource utilization on multiprocessor systems for parallel programs - *PACT 2016, TPDS 2019.*

AWARD

SnuMAP: 2nd prize in 10th Open Source Software World Challenge *Seoul, South Korea*
SNU Manycore Profiler for Big-Data with SNU team *December 2016*

TEACHING EXPERIENCE

Advanced Computer Architecture (CSE220), UC Santa Cruz *Fall 2024*
Teaching Assistant
Computer Architecture (CSE120), UC Santa Cruz *Fall 2023, Winter 2023, Fall 2021*
Teaching Assistant
Computer Architecture, Seoul National University *Spring 2016, Fall 2015*
Teaching Assistant

REFERENCES

Prof. Heiner Litz (University of California, Santa Cruz)
E-Mail: hlitz@ucsc.edu