

# SURIM OH

soh31@ucsc.edu

## RESEARCH INTEREST

---

My research interests lie broadly in computer architecture and hardware-software interface. I have recently been working on the topic of fetch-directed instruction prefetching in modern processors to improve the CPU performance for datacenter workloads with large instruction footprints and on the topic of ASIC for LLMs.

## EDUCATION

---

<b>University of California, Santa Cruz</b> <i>PhD Student in Computer Science and Engineering</i> · Advisor: Professor Heiner Litz, Total GPA: 4.0/4.0	<i>Santa Cruz, CA, USA</i> <i>Sep 2020 - Present</i>
<b>Seoul National University</b> <i>Master of Science in Computer Science and Engineering</i> · Advisor: Professor Bernhard Egger, Total GPA: 3.52/4.0 · Thesis: <i>Hierarchical Manycore Resource Management Framework using Control Processors</i> [pdf]	<i>Seoul, South Korea</i> <i>Feb 2015 - Feb 2017</i>
<b>Sogang University</b> <i>Bachelor of Science in Computer Science and Engineering</i> · Total GPA: 3.70/4.0 (95.8/100), Summa Cum Laude · Exchange student at <b>Northern Arizona University</b>	<i>Seoul, South Korea</i> <i>Feb 2011 - Feb 2015</i>  <i>AZ, USA, Spring 2014</i>

## WORK EXPERIENCE

---

<b>Meta</b> <i>ASIC Engineer Intern, Architecture</i>	<i>Sunnyvale, California, USA</i> <i>June 2024 - Sep 2024</i>
<b>Akeana</b> <i>PhD Intern</i>	<i>San Jose, California, USA</i> <i>June 2023 - Sep 2023</i>
<b>SAP Labs Korea</b> <i>Developer</i>	<i>Seoul, South Korea</i> <i>Jan 2018 - Sep 2020</i>
<b>Hyundai Motor Company R&amp;D Division</b> <i>Engineer</i>	<i>Hwaseong, South Korea</i> <i>Feb 2017 - Jan 2018</i>

## SKILLS

---

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

## PUBLICATIONS

---

**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**

PhD Student

*Santa Cruz, CA, USA*

*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**

ASIC Engineer Intern, Architecture

*Sunnyvale, CA, USA*

*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**

PhD Intern

*San Jose, CA, USA*

*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

Developer

Seoul, South Korea

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

Engineer

Hwaseong, South Korea

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

Graduate Researcher

Seoul, South Korea

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  
SNU Manycore Profiler for Big-Data with SNU team

Seoul, South Korea

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