SUNHO LEE

https://github.com/myshlee417 https://myshlee417.github.io myshlee417@gmail.com
KAIST, Daejeon, Republic of Korea (+82)10-3617-1820

RESEARCH INTERESTS

I am interested in a secure and efficient architecture of hardware accelerators (such as GPU and NPU).

My research objective is to design high-performance accelerators with security guarantees. To achieve this goal, my recent studies focus on 1) *hardware security* and 2) *performance improvement* of hardware accelerators.

Hardware Security of Accelerators: As accelerators are widely used in mission-critical tasks, the importance of security gets larger. Although I extended Trusted Execution Environment to GPU and NPU in previous works, countless security weaknesses still remain. Therefore, I aim to increasing the security level to resist unintended operations.

Performance Improvement of Accelerators: Since machine learning requires speedy processing, the performance improvement of accelerators is crucial. Hence, I consider both hardware and software to enhance parallelism or cut down unnecessary procedures. In a recent publication, I proposed the fine-grained scheduling algorithm in GPU by leveraging Multi-Process Service. I set the further reduction of the execution time as a future research direction.

From these two sub-goals, I target combining a trusted system with a high-performance accelerator design. It is expected to protect users from accidents (caused by attackers or extreme environments) within a reasonable latency.

EDUCATION

KAIST, Daejeon, Republic of Korea *Mar 2021 -*
Ph.D. Student, School of Computing
Advisor: Jaehyuk Huh

KAIST, Daejeon, Republic of Korea *Mar 2019 - Feb 2021*
Master of Science, School of Computing
Advisor: Jaehyuk Huh
Thesis: *Hardware Security Techniques for Trusted Machine Learning Accelerators*

Yonsei University, Seoul, Republic of Korea *Mar 2015 - Feb 2019*
Bachelor of Science, Computer Science

PUBLICATIONS

- Jungwoo Kim, Seonjin Na, Sanghyeon Lee, **Sunho Lee**, and Jaehyuk Huh, "Improving Data Reuse in NPU On-chip Memory with Interleaved Gradient Order for DNN Training", *the 56th IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, October 2023
- Seungho Lee, **Sunho Lee**, Jaehyuk Huh, and Sejin Kwon, "Proposal of Aerospace-informatics by Design of Ramjet Inlet Using Machine Learning", *the 2023 Aerospace Europe Conference (**AEC**) joint event between the 10th European Conference for Aerospace Sciences (**EUCASS**) and the 9th Council of European Aerospace Societies (**CEAS**)*, July 2023
- **Sunho Lee**, Seonjin Na, Jungwoo Kim, Jongse Park, and Jaehyuk Huh, "Tunable Memory Protection for Secure Neural Processing Units", *the 40th IEEE International Conference on Computer Design (**ICCD**)*, October 2022
- Seungbeom Choi, **Sunho Lee**, Yeonjae Kim, Jongse Park, Youngjin Kwon, and Jaehyuk Huh, "Serving Heterogeneous Machine Learning Models on Multi-GPU Servers with Spatio-Temporal Sharing", *the 2022 USENIX Annual Technical Conference (**USENIX ATC**)*, July 2022
- **Sunho Lee**, Jungwoo Kim, Seonjin Na, Jongse Park, and Jaehyuk Huh, "TNPU: Supporting Trusted Execution with Tree-less Integrity Protection for Neural Processing Unit", *the 28th IEEE International Symposium on High-Performance Computer Architecture (**HPCA**)*, Feburary 2022
- Seonjin Na, **Sunho Lee**, Yeonjae Kim, Jongse Park, and Jaehyuk Huh, "Common Counters: Compressed Encryption Counters for Secure GPU Memory", *the 27th IEEE International Symposium on High-Performance Computer Architecture (**HPCA**)*, Feburary 2021

PATENTS

- **[Pending]** Jaehyuk Huh, Jungwoo Kim, Seonjin Na, Sanghyeon Lee, and **Sunho Lee**, "Improving the Utilization of NPU On-chip Memory with Computation Rearrangement for DNN Training", *Korean Patent*
- **[Pending]** Jaehyuk Huh, Seonjin Na, Jungwoo Kim, and **Sunho Lee**, "Dynamic One-time Pad Table Management for Secure Multi-GPU Communication", *Korean Patent*
- **[Pending]** Jaehyuk Huh, **Sunho Lee**, and Seonjin Na, "Apparatus and Method for Providing Secure Execution Environment for NPU", *US Patent* (with Samsung Electronics)
- **[Pending]** Jaehyuk Huh, Seungbeom Choi, **Sunho Lee**, Yeonjae Kim, Youngjin Kwon, Jongse Park, "Machine Learning Inference Time-spatial SW Scheduler Based on Multiple GPU", *Korean Patent*
- **[Pending]** Jaehyuk Huh, **Sunho Lee**, and Seonjin Na, "Hardware-based Security Architecture for Trusted Neural Processing Unit", *Korean Patent* (with Samsung Electronics)
- **[10-2365263-0000]** Jaehyuk Huh, Seonjin Na, **Sunho Lee**, Yeonjae Kim, and Jongse Park, "Efficient Encryption Method and Apparatus for Hardware-based Secure GPU Memory", *Korean Patent*

RESEARCH EXPERIENCES

KAIST, Daejeon, Republic of Korea *Mar 2019 -*
Ongoing Researches at CASYS (Computer Architecture and SYStem) Lab
Advisor: Jaehyuk Huh

Accelerator Hardware-based Security
- Memory protection optimization for GPU: Common counters for duplicate counters (Published in **HPCA 2021**)
- Memory protection optimization for multi-tenant GPU
- Trusted execution environment for NPU: Tensor-granularity counters (Published in **HPCA 2022**)
- Memory protection optimization for NPU: Partial memory protection (Published in **ICCD 2022**)
- Side-channel attack protection for NPU
- Dynamic secure-granularity management for heterogeneous processors

Accelerator Performance
- Multi-tenancy support for a multi-GPU system: Time and spatial sharing (Published in **USENIX ATC 2022**)
- Multi-tenancy support for NPU: Shared resources management
- On-chip memory management for training NPU: Access order rearrangement (Published in **MICRO 2023**)

Yonsei University, Seoul, Republic of Korea *Sep 2017 - June 2018*
Undergraduate Research Intern at ELC (Embedded systems Languages and Compilers) Lab
Advisor: Bernd Burgstaller

Parallelism
- Accelerating big-data streaming engine: Multi-thread and shared-memory
- Parallelization of SFA (Simultaneous Deterministic Finite Automata) construction: MPI and Huang's algorithm

RECOGNITION

KAIST, Daejeon, Republic of Korea
Outstanding Teaching Assistant Award - CS311 Computer Organization *Spring 2022, Fall 2019*

Yonsei University, Seoul, Republic of Korea
Dean's List *Spring 2015, Spring 2018*
Undergraduate Capstone Project Award (Third Place) - Project Leader *Spring 2018*
Title: *Cloud SFA: Parallel Construction of Simultaneous Deterministic Finite Automata in Distributed System*

Samsung Electronics, Hwaseong, Republic of Korea
Best Paper Award (Third Place) *Summer 2022*
Title: *TNPU: Supporting Trusted Execution with Tree-less Integrity Protection for Neural Processing Unit*

PARTICIPATION

uArch (in conjunction with **ISCA 2022**), New York City, United States of America
Student Panel
Life in Grad School *June 2022*

SKILLS

Programming Languages	C, C++, Python
NPU Simulators	SCALE-Sim, MAESTRO, Gemmini
GPU Programming	CUDA, MPS
Multi-core CPU Programming	MPI, OpenMP
Machine Learning Frameworks	Pytorch, Tensorflow

TEACHING EXPERIENCES

KAIST, Daejeon, Republic of Korea
Teaching Assistant

CS230 System Programming	*Fall 2021*
CS311 Computer Organization	*Spring 2022, Spring 2021, Fall 2019*
CS211 Digital System and Lab	*Spring 2019*

KAIST Education Center, Daejeon, Republic of Korea
Mentor & Lecturer

Seocho AI College	*Summer 2019, Summer 2021*
Python for Beginners	*Summer 2022, Summer 2021*