

SEONJIN NA

Postdoctoral Fellow at Georgia Tech

Phone: +1 404-259-3240

Email: seonjin.na@gatech.edu

Github: <https://github.com/seonjinna>

Website: <https://seonjinna.github.io>

RESEARCH INTERESTS

My research interests lie in GPU architecture, security, and accelerated systems for machine learning. During my Ph.D. studies, I focused on developing a secure architecture aimed at extending trusted execution environments (TEEs) to accelerators such as GPUs and NPUs, with an emphasis on reducing the performance overhead associated with security features. Currently, I am actively engaged in extending my research field to address various challenges in the GPU architecture, security, and accelerating large language models.

EMPLOYMENT

Georgia Institute of Technology

June. 2023 - present

Postdoctoral Fellow, School of Computer Science

Supervisor: Hyesoon Kim

Microsoft Research Asia

Mar. 2019 - June. 2019

Visiting Fellow

Supervisors: Lintao Zhang & Yunxin Liu

EDUCATION

KAIST

Mar. 2018 - Feb. 2023

Doctor of Philosophy, School of Computing

Advisor: Jaehyuk Huh

KAIST

Mar. 2016 - Feb. 2018

Master of Science, School of Computing

Advisor: Jaehyuk Huh

Sogang University

Mar. 2012 - Feb. 2016

Bachelor of Science, Computer Science

Summa Cum Laude

PUBLICATIONS

- **Seonjin Na**, Geonhwa Jeong, Byunghoon Ahn, Jeffrey Young, Tushar Krishina, Hyesoon Kim, "Understanding Performance Implications of LLM Inference on CPUs", *IEEE International Symposium on Workload Characterization (IISWC)*, Sep 2024.
- Euijun Chung, **Seonjin Na**, Hyesoon Kim, "Allegro: GPU Simulation Acceleration for Machine Learning Workloads", *MLArchSys Workshop in the 51th International Symposium on Computer Architecture (MLArchSys Workshop in ISCA)*, June 2024.
- Yuan Feng, **Seonjin Na**, Hyesoon Kim, and Hyeran Jeon, "Barre Chord: Efficient Virtual Memory Translation for Multi-Chip-Module GPUs", *the 51th International Symposium on Computer Architecture (ISCA)*, June 2024.

- **Seonjin Na**, Jungwo Kim, Sunho Lee, and Jaehyuk Huh, "Supporting Secure Multi-GPU Computing with Dynamic and Batched Metadata Management", *the 30th IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, March 2024.
- 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.
- 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.
- 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)*, February 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)*, February 2021.

RESEARCH EXPERIENCES

Characterizing LLM inference on latest CPUs

Feb. 2024 - Present

- Investigated the performance implications of LLM inference on latest CPU hardware.
- Compared LLM inference performance on CPUs with that on GPUs across various configurations.
- Contributed as **first author**.
- **Published in IISWC 2024**

Accelerating GPU Simulation for Machine Learning Workloads

April. 2024 - Present

- Investigated the limitations of prior studies designed for GPU simulation acceleration.
- Proposed a statistical-based sampling scheme to reduce GPU simulation time significantly.
- Contributed as **second author**.
- **Published in MLArchSys Workshop in ISCA 2024**

Efficient Row-hammer Attack Mitigation

Sep. 2023 - Present

- Investigated the limitations of previous row-hammer mitigation mechanisms.
- Proposed row-hammer mitigation with lower metadata overhead compared to previous works.
- Contributed as **second author**.
- **Work-in-progress**

SSD Aware GPU Thread Block Scheduling

Sep. 2023 - Present

- Investigated the problem of previous GPU memory safety mechanisms.
- Proposed power-side channel attack mitigation mechanisms for GPUs.
- Contributed as **second author**.
- **Under Review**

Efficient GPU Address Translation for MCM-GPU Systems

Sep. 2023 - Present

- Investigated performance implications of address translation on MCM-GPU system.
- Proposed efficient address translation mechanism for the MCM-GPU system.
- Contributed as **second author**.
- **Published in ISCA 2024**

GPU Power Side-channel Attack Mitigation

June. 2023 - Present

- Investigated the problem of previous GPU memory safety mechanisms.

- Proposed power-side channel attack mitigation mechanisms for GPUs.
- Contributed as **second author**.
- **Under Review**

Hardware-based Efficient GPU Memory Safety

June. 2023 - Present

- Investigated the problem of previous GPU memory safety mechanisms.
- Proposed a practical GPU memory safety mechanism with low-performance overhead.
- Contributed as **fourth author**.
- **Under Review**

Privacy-aware ML Program Cloning

June. 2023 - Present

- Investigated the privacy-aware tracing mechanism to prevent DNN model extraction attacks.
- Contributed as **first author**.
- **Work-in-progress**

NPU Side-channel Attack Protection

Jan. 2022 - Present

- Investigated the side-channel attack-based vulnerability of execution on NPUs.
- Contributed as **second author**
- **Work-in-progress**

Dynamic Secure-granularity Management for Heterogeneous System

Jan. 2022 - Present

- Investigated the performance impacts of data protection techniques on heterogeneous processors.
- Contributed as **second author**
- **Under Review**

Efficient On-chip Memory Management and Scheduling on NPUs

Dec. 2021 - Jul. 2023

- Analyzed the data dependency of tensor computations in DNN training.
- Proposed mechanisms to maximize data reuse in NPU on-chip memory.
- Contributed as **second author**
- **Published in MICRO 2023**

Efficient Data Protection Technique for Multi-GPU Systems

Sep. 2021 - Present

- Investigated the performance impact of prior data protection methods on a multi-GPU system.
- Analyzed the communication characteristics of GPU workloads.
- Proposed an efficient data protection technique to minimize the secure communication overhead.
- Contributed as **first author**.
- **Published in HPCA 2024**

Efficient Memory Protection Mechanism for Secure NPU

Sep. 2020 - March. 2021

- Investigated a significant performance degradation of CPU memory protection schemes on NPUs.
- Proposed selective memory protection and multi-granular counter mode encryption techniques.
- Contributed as **second author**.
- **Published in ICCD 2022**

Trusted NPU Architecture

Sep. 2019 - Sep. 2021

- Extended the existing CPU TEE design to isolate the NPU execution context from OS.
- Proposed a tree-less integrity protection by exploiting a tensor-based NPU execution model.
- Contributed as **third author**.
- **Published in HPCA 2022**

Efficient Memory Protection Mechanism for Secure GPU Memory

Sep. 2017 - Sep. 2020

- Analyzed memory update behaviors of GPU benchmarks and real-world applications using the NVbit.
- Proposed efficient GPU data protection scheme to significantly reduce performance degradation.

- Contributed as **first author**.
- Published in **HPCA 2021**

PATENTS

Dynamic One-time Pad Table Management for Secure Multi-GPU Communication

- Jaehyuk Huh, Seonjin Na, Jungwoo Kim, Sunho Lee
- Korea Patent; Pending

Improving the Utilization of NPU On-chip Memory with Computation Rearrangement for DNN Training

- Jaehyuk Huh, Jungwoo Kim, Seonjin Na, Sanghyeon Lee, Sunho Lee
- Korea Patent; Pending

Apparatus and Method for Providing Secure Execution Environment for NPU

- Jaehyuk Huh, Sunho Lee, Seonjin Na
- US Patent (With Samsung Electronics); Pending

Hardware-based Security Architecture for Trusted Neural Processing Unit

- Jaehyuk Huh, Sunho Lee, Seonjin Na
- Korean Patent (With Samsung Electronics); Filling Date: 2021/07/23;

Efficient Encryption Method and Apparatus for Hardware-based Secure GPU Memory

- Jaehyuk Huh, Seonjin Na, Sunho Lee, Yeonjae Kim, and Jongse Park
- Korea Patent; Filling Date: 2020/11/23; Issued Date: 2022/02/16

AWARDS AND HONORS

Best Paper Award

2022

- TNPU: Supporting Trusted Execution with Tree-less Integrity Protection for Neural Processing Unit, 3th place

National Scholarship

Mar. 2016 - 2023 Feb

- KAIST

Summa CumLaude

Feb. 2016

- Sogang University

Gold Prize

Nov. 2015

- The 2015 ACM-ICPC Asia Daejeon Regional Contest 4th place

Honorable Mention

Nov. 2013

- The 2013 ACM-ICPC Asia Daejeon Regional Contest 13th place

Academic Scholarship, 8 semesters

Mar. 2012 - Sep. 2015

- Sogang University

ACADEMIC SERVICES

Reviewer

- International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) 2024
- IEEE Transactions on Dependable and Secure Computing 2023
- IEEE Computer Architecture Letter 2023

Program Committee

- International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) 2024

Travel Grant Chair

- International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2025

Artifact Evaluation Committee

- EuroSYS 2025, ASPLOS 2025, USENIX ATC 2024, USENIX OSDI 2024, ISCA 2024

Web Chair

- IEEE Computer Society TCuARCH

TEACHING EXPERIENCE

KAIST

- CS510 Computer Architecture: Spring 2020
- CS230 System Programming: Fall 2016, Spring 2017, Fall 2018, Fall 2020
- CS311 Computer Organization: Fall 2019

Sogang University

- Introduction to C Programming: Winter 2014

SKILLS

- **Programming Languages** : C/C++, Go, CUDA, Python, Java
- **Library/Frameworks** : NVBit, Pytorch, Tensorflow
- **Simulators** : GPGPU-Sim, MGPU-Sim, Gem5, Gem5-gpu, Scale-Sim, SST, ChampSim, Macsim