SEONJIN NA

✉ seonjin.na@gatech.edu | 🌐 https://seonjinna.github.io/ | GitHub | LinkedIn

SUMMARY

Seonjin Na is a **postdoctoral fellow** at **Georgia Institute of Technology**, focusing on GPU/NPU Architecture, Large Language Model (LLM) Inference Optimizations with HW/SW co-design, and Secure Architecture for GPU/NPU.
Keywords: **LLM Efficiency, GPU/NPU Architecture, HW/SW Co-design, Performance Profiling**

TECHNICAL SKILLS

Data Analysis and Profiling Tool: Pandas, Matplotlib, Intel VTune, Linux Perf, NVIDIA Nsight Profiler
Machine Learning Framework: Pytorch, vLLM, DeepSpeed, HuggingFace, Intel Extension for Pytorch
Architectural Simulator: GPGPU-Sim, MGPUSim, ScaleSim, Gem5, ChampSim, Macsim
Programming Language and Others: Python, C, C++, Go, CUDA, NVBit

WORK EXPERIENCE

Georgia Institute of Technology: Postdoctoral Fellow June 2023 – Present

- Advised by Prof. Hyesoon Kim and collaborated with Prof. Tushar Krishna.

Microsoft Research Asia: Research Intern Mar 2019 – June 2019

- Advised by Lintao Zhang and Yunxin Liu.

EDUCATION

Doctor of Philosophy, School of Computing Feb. 2023
Korea Advanced Institute of Science and Technology (KAIST)
Advisor: Jaehyuk Huh

Master of Science, School of Computing Feb. 2018
Korea Advanced Institute of Science and Technology (KAIST)
Advisor: Jaehyuk Huh

Bachelor of Science, Computer Science and Engineering Feb. 2016
Sogang University
Summa Cum Laude

RESEARCH EXPERIENCE

Georgia Institute of Technology: Postdoctoral Fellow June 2023 - Present

- **Flexible LLM Inference Serving System:** Proposed a flexible system for LLM inference that minimizes latency by leveraging both CPU and GPU hardware in offloading-based inference, using a performance estimator and extensions from HuggingFace and Intel Extension for PyTorch [2].
- **CPU LLM Inference Characterization:** Investigated the opportunity of CPU computation in offloading-based LLM inference and identified optimal CPU configurations to improve CPU LLM inference performance, using hardware performance counters and analysis with the VTune profiler [4].
- **Efficient Address Translation for MCM-GPU Systems:** Analyzed the performance impact of address translation in Multi-Chip Module (MCM) GPU systems and proposed coalesced address translation scheme with page-mapping adjustments to enhance performance [5].
- **Sampling-Based Fast GPU Simulation:** Designed a sampling-based faster GPU simulation method, reducing execution time needed for running various GPU workload simulations with low error rate [6].
- **Efficient GPU Memory Safety Mechanism:** Addressed memory safety concerns on GPUs, such as buffer overflows, by proposing a hardware-based fine-grained GPU memory safety mechanism with minimal impact on performance [3].
- **Efficient Memory Protection for System-on-Chip System:** Investigated security mechanisms in heterogeneous SoC architectures (CPU, GPU, NPU) and developed a multi-granular memory protection scheme that dynamically adapts to diverse memory access patterns [1].

KAIST: Graduate Research Assistant Mar 2018 – Feb 2023

- **Efficient Secure Communication for Trusted Multi-GPU Systems:** Investigated the performance overhead of securing CPU-GPU and GPU-GPU data transfers in multi-GPU systems and proposed a dynamic, batched metadata management scheme to minimize the performance impact of it, using the MGPUSim simulator [7].
- **Low-Overhead GPU Memory Protection:** Proposed an efficient memory protection scheme that leverages observed memory update behaviors. The evaluation was conducted using the GPGPU-Sim and NVBit [11].
- **Accelerating DNN Training on NPUs:** Developed dataflow transformations on NPUs to maximize data reuse of scratchpad memory (SPM) in the backward pass of DNN training, improving inter-operation data locality and reducing training latency [8].
- **Extending Trusted Execution Environments on NPUs:** Proposed a trusted NPU architecture with hardware changes and proposed efficient data protection tailored for NPUs using Scale-Sim simulator [9],[10].

PUBLICATIONS

[1] Sunho Lee, **Seonjin Na**, Jeongwon Choi, Jinwon Pyo, Jaehyuk Huh, **Unified Memory Protection with Multi-granular MAC and Integrity Tree for Heterogeneous Processors**, *International Symposium on Computer Architecture* **(ISCA)**, 2025.

[2] **Seonjin Na**, Geonhwa Jeong, Byunghoon Ahn, Aaron Jezghani, Jeffrey Young, Christopher J. Hughes, Tushar Krishna, Hyesoon Kim, **FlexInfer: Flexible LLM Inference with CPU Computations** 🔗, *Conference on Machine Learning and Systems* **(MLSys)**, 2025.

[3] Jaewon Lee, Euijun Chung, Saurabh Singh, **Seonjin Na**, Yonghae Kim, Jaekyu Lee, Hyesoon Kim, **Let-Me-In: (Still) Employing In-pointer Bounds metadata for Fine-grained GPU Memory Safety** 🔗, *IEEE International Symposium on High-Performance Computer Architecture* **(HPCA)**, Mar 2025.

[4] **Seonjin Na**, Geonhwa Jeong, Byunghoon Ahn, Jeffrey Young, Tushar Krishna, Hyesoon Kim, **Understanding Performance Implications of LLM Inference on CPUs** 🔗, *IEEE International Symposium on Workload Characterization* **(IISWC)**, Sep 2024.

[5] Yuan Feng, **Seonjin Na**, Hyesoon Kim, Hyeran Jeon, **Barre Chord: Efficient Virtual Memory Translation for Multi-Chip-Module GPUs** 🔗, *International Symposium on Computer Architecture* **(ISCA)**, June 2024.

[6] Euijun Chung, **Seonjin Na**, Hyesoon Kim, **Allegro: GPU Simulation Acceleration for Machine Learning Workloads** 🔗, *MLArchSys Workshop in the International Symposium on Computer Architecture* **(MLArchSys)**, June 2024.

[7] **Seonjin Na**, Jungwoo Kim, Sunho Lee, Jaehyuk Huh, **Supporting Secure Multi-GPU Computing with Dynamic and Batched Metadata Management** 🔗, *IEEE International Symposium on High-Performance Computer Architecture* **(HPCA)**, Mar 2024.

[8] Jungwoo Kim, **Seonjin Na**, Sanghyeon Lee, Sunho Lee, Jaehyuk Huh, **Improving Data Reuse in NPU On-chip Memory with Interleaved Gradient Order for DNN Training** 🔗, *IEEE/ACM International Symposium on Microarchitecture* **(MICRO)**, Oct 2023.

[9] Sunho Lee, **Seonjin Na**, Jungwoo Kim, Jaehyuk Huh, **Tunable Memory Protection for Secure Neural Processing Units** 🔗, *IEEE International Conference on Computer Design* **(ICCD)**, Oct 2022.

[10] Sunho Lee, Jungwoo Kim, **Seonjin Na**, Jaehyuk Huh, **TNPU: Supporting Trusted Execution with Tree-less Integrity Protection for Neural Processing Unit** 🔗, *IEEE International Symposium on High-Performance Computer Architecture* **(HPCA)**, Mar 2022.

[11] **Seonjin Na**, Sunho Lee, Yeonjae Kim, Jongse Park, Jaehyuk Huh, **Common Counters: Compressed Encryption Counters for Secure GPU Memory** 🔗, *IEEE International Symposium on High-Performance Computer Architecture* **(HPCA)**, Mar 2021.

AWARDS & HONORS

Outstanding Post-doctoral Research Award 2025
College of Computing at Georgia Institute of Technology (Georgia Tech)

MICRO 2024 Ph.D Forum 2024
2024 IEEE/ACM International Symposium on Microarchitecture (MICRO)

National Scholarship 2016-2023
Korea Advanced Institute of Science and Technology (KAIST)

Summa Cum Laude 2016
Sogang University

Gold Prize 2015
ACM-ICPC Asia Daejeon Regional Contest 4th place

Honorable Mention 2013
ACM-ICPC Asia Daejeon Regional Contest 13th place

Academic Scholarship, 8 semesters 2012-2015
Sogang University

PATENTS

[P1] Jaehyuk Huh, Sunho Lee, **Seonjin Na**, **Apparatus and Method for Providing Secure Execution Environment for NPU**, Patent US 12045337, United States, 2024

[P2] Jaehyuk Huh, **Seonjin Na**, Jungwoo Kim, Sunho Lee, **Dynamic One-time Pad Table Management for Secure Multi-GPU Communication**, Patent KR 1020230055347, South Korea, 2023

[P3] Jaehyuk Huh, Jungwoo Kim, **Seonjin Na**, Sanghyeon Lee, Sunho Lee, **Improving the Utilization of NPU On-chip Memory with Computation Rearrangement for DNN Training**, Patent KR 1020230055346, South Korea, 2022

[P4] Jaehyuk Huh, Sunho Lee, **Seonjin Na**, **Hardware-based Security Architecture for Trusted Neural Processing Unit**, Patent KR 1020220055977, South Korea, 2022

[P5] Jaehyuk Huh, **Seonjin Na**, Sunho Lee, Yeonjae Kim, and Jongse Park, **Efficient Encryption Method and Apparatus for Hardware-based Secure GPU Memory**, Patent KR 1023652630000, South Korea, 2022

ACADEMIC SERVICES

Technical Program Committee

- Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2026
- General Purpose Processing on Graphics Processing Units (GPGPU) 2025
- International Parallel & Distributed Processing Symposium (IPDPS) 2025
- International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC) 2024

Journal Reviewer

- ACM Transactions on Computer Systems (TOCS) 2025
- ACM Transactions on Architecture and Code Optimization (TACO) 2025
- IEEE Computer Architecture Letter (CAL) 2025
- ACM Transactions on Computer Systems (TOCS) 2024
- ACM Transactions on Architecture and Code Optimization (TACO) 2024 x 4
- IEEE Transactions on Dependable and Secure Computing (TDSC) 2023
- IEEE Computer Architecture Letter (CAL) 2023

Workshop/Tutorial Chair

- IEEE International Symposium on Workload Characterization (IISWC) 2025

Travel Grant Chair

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

Web Chair

- IEEE Computer Society TCuARCH
- Vortex Workshop and Tutorial at MICRO 2024

Artifact Evaluation Committee

- IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) 2025
- ACM European Conference on Computer Systems (EuroSys) 2025
- Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2025
- IEEE/ACM International Symposium on Microarchitecture (MICRO) 2024
- USENIX Symposium on Operating Systems Design and Implementation (OSDI) 2024
- USENIX Annual Technical Conference (ATC) 2024
- International Symposium on Computer Architecture (ISCA) 2024