Seong Hoon Seo

CONTACT INFORMATION

Building 944, NPRC Center
Gwanak-ro 1, Gwanak-gu
Seoul 08826, Republic of Korea

Voice: +82 10-5346-9045
E-mail: andyseo247@snu.ac.kr
Website: https://seonghoonseo.github.io

RESEARCH INTERESTS

Computer architecture and systems for deep learning, efficient on-device LLM inference, heterogeneous computing, processing-in-memory, quantization

EDUCATION

Seoul National University Mar. 2020 - Present
Ph.D. Candidate, Department of Computer Science and Engineering

- Advisor: Prof. Jae W. Lee

Seoul National University Mar. 2015 - Feb. 2020
College of Liberal Studies

- Bachelor of Science in Computer Science and Engineering
- Bachelor of Arts in Economics

Daewon Foreign Language High School Mar. 2012 - Feb. 2015

- Majored in English and Spanish

PUBLICATIONS

MemSOS: OS-Guided Selective Memory Mirroring
Junghoon Kim, Jongheon Jeong, Seokwon Moon, **Seong Hoon Seo**, Yeonhong Park, Jinkyu Jeong, Nam Sung Kim, and Jae W. Lee, The 32nd IEEE International Symposium on High-Performance Computer Architecture **(HPCA)**, January 2026.

DP-LLM: Runtime Model Adaptation with Dynamic Layer-wise Precision Assignment
Sangwoo Kwon, **Seong Hoon Seo**, Jae W. Lee, and Yeonhong Park, The 39th Annual Conference on Neural Information Processing Systems **(NeurIPS)**, December 2025.

FACIL: Flexible DRAM Address Mapping for SoC-PIM Cooperative On-device LLM Inference
Seong Hoon Seo, Junghoon Kim, Donghyun Lee, Seonah Yoo, Seokwon Moon, Yeonhong Park, and Jae W. Lee, The 31st IEEE International Symposium on High-Performance Computer Architecture **(HPCA)**, March 2025.

A 40nm 5.6TOPS/W 239GOPS/mm^2 Self-Attention Processor with Sign Random Projection-based Approximation
Seong Hoon Seo*, Soosung Kim*, Sung Jun Jung, Sangwoo Kwon, Hyunseung Lee, and Jae W. Lee, The 48th IEEE European Solid-State Circuits Conference **(ESSCIRC)**, September 2022.

ELSA: Hardware-Software Co-design for Efficient, Lightweight Self-Attention Mechanism in Neural Networks
Tae Jun Ham*, Yejin Lee*, **Seong Hoon Seo**, Soosung Kim, Hyunji Choi, Sung Jun Jung, and Jae W. Lee, The 48th ACM/IEEE International Symposium on Computer Architecture **(ISCA)**, June 2021.

Accelerating Genomic Data Analytics With Composable Hardware Acceleration Framework
Tae Jun Ham, David Bruns-Smith, Brendan Sweeney, Yejin Lee, **Seong Hoon Seo**, U Gyeong Song, Young H. Oh, Krste Asanovic, Jae W. Lee, and Lisa Wu Wills, IEEE Micro, vol. 41, no. 3, pp. 42-49, 1 May-June 2021.
Special Issue on Top Picks from the 2020 Computer Architecture Conferences

MERCI: Efficient Embedding Reduction on Commodity Hardware via Sub-Query Memoization
Yejin Lee, **Seong Hoon Seo**, Hyunji Choi, Hyoung Uk Sul, Soosung Kim, Jae W. Lee, and Tae Jun Ham, The 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS)**, April 2021.

Genesis: A Hardware Acceleration Framework for Genomic Data Analysis
Tae Jun Ham, David Bruns-Smith, Brendan Sweeney, Yejin Lee, **Seong Hoon Seo**, U Gyeong Song, Young H. Oh, Krste Asanovic, Jae W. Lee, and Lisa Wu Wills, The 47th ACM/IEEE International Symposium on Computer Architecture **(ISCA)**, June 2020.

*denotes equal contributions

Patents

Device for accelerating self-attention operation in neural networks (KR102886499B1)
Inventor: Yejin Lee, Tae Jun Ham, **Seong Hoon Seo**, Soosung Kim, Hyunji Choi, Jae W. Lee, and Sung Jun Jung

Honors and Awards

- **Qualcomm Innovation Fellowship Korea** Nov. 2025
- **Encouragement Prize in the 31st Samsung Humantech Paper Awards** Jan. 2025
  FACIL: Flexible DRAM Address Mapping for SoC-PIM Cooperative On-device LLM Inference
  (Acceptance Rate: 116/3152 $\approx$ 3.7%)
- **IEEE Micro Top Picks** Jan. 2021
  Genesis: A Hardware Acceleration Framework for Genomic Data Analysis
  Selected as one of the 12 Top Picks from the 2020 Computer Architecture Conferences

Teaching Experience

Logic Design (Instructor: Prof. Jae W. Lee), Seoul National University Mar. - Jun. 2020
Teaching Assistant

- Undergraduate course taught in English, including weekly lab sessions held by the TAs

Community Service

- **Student Volunteer**, International Symposium on Code Generation and Optimization 2022
- **Student Volunteer**, IEEE Micro Special Issue on Top Picks from the 2022 Computer Architecture Conferences

Undergraduate Research Internships

Architecture and Code Optimization Lab Jul. 2018 - Feb. 2020
Seoul National University (Prof. Jae W. Lee)

- Contributed to ELSA (ISCA'21) and Genesis (ISCA'20)

Machine Intelligence and Pattern Analysis Lab Jan. 2018 - Feb. 2018
Seoul National University (Prof. Nojun Kwak)

- Explored input preprocessing for KITTI dataset

COMPUTER SKILLS

- **Languages:** Python, C/C++, Chisel, Verilog
- **Frameworks:** PyTorch
- **Tools:** Git, Docker, NVIDIA Nsight Systems/Compute, Intel VTune Profiler
- **Platforms and Hardware:** Linux, NVIDIA GPU/Jetson, AMD/Xilinx FPGA

LANGUAGES

English (fluent), Korean (native), Spanish (basic proficiency)

Last Updated: December 23, 2025