

Seong Hoon Seo

CONTACT INFORMATION	Haedong Advanced Engineering 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 <i>Ph.D., Department of Computer Science and Engineering</i> • Thesis: Memory and Storage-Centric Techniques for Efficient On-Device LLM Inference • Advisor: Prof. Jae W. Lee	Mar. 2020 – Aug. 2026
	Seoul National University <i>B.S. and B.A., College of Liberal Studies</i> • Bachelor of Science in Computer Science and Engineering • Bachelor of Arts in Economics	Mar. 2015 – Feb. 2020
	Daewon Foreign Language High School • Majored in English and Spanish	Mar. 2012 – Feb. 2015
EXPERIENCE	AI Institute, Seoul National University <i>Postdoctoral Researcher</i> • Supervisor: Prof. Jae W. Lee • This position also fulfills my mandatory military service	Sep. 2026 – Aug. 2027 (expected)
	Qualcomm AI Research <i>Engineering Intern, QuantCore Systems Team</i> • Contributed to low-bit quantization of LLMs/VLMs for mobile deployment, focusing on sensitivity analysis and mixed-precision assignment • Manager: Jaeseong You	Feb. – Jul. 2026
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/mm² 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)

Inventors: 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

ACADEMIC SERVICE	• Program Committee: HPCA 2027, MICRO 2026 • Journal Reviewer: ACM TACO (2026) • Student Volunteer: CGO 2022, IEEE Micro Top Picks 2022	
UNDERGRADUATE RESEARCH INTERNSHIPS	Architecture and Code Optimization Lab <i>Seoul National University (Prof. Jae W. Lee)</i> • Contributed to ELSA (ISCA'21) and Genesis (ISCA'20) Machine Intelligence and Pattern Analysis Lab <i>Seoul National University (Prof. Nojun Kwak)</i> • Explored input preprocessing for KITTI dataset	Jul. 2018 – Feb. 2020 Jan. – Feb. 2018
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: September 9, 2026