

Seunghyun Moon

ASSISTANT PROFESSOR · DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

Konkuk University, 05029, Seoul, South Korea

☎ +82 10-9911-5787 | ✉ smoon@konkuk.ac.kr | 🌐 www.linkedin.com/in/shmoon94

Research Interest

From classic discriminative models to state-of-the-art generative models, our research group develops semiconductor systems capable of powering a diverse range of emerging applications, currently specified in AI. We invent hardware-efficient algorithms and design unprecedented hardware architectures that break new ground in performance and efficiency. Ultimately, through rigorous silicon implementation, we demonstrate the reliability and feasibility of our proposed algorithm-hardware integration, paving the way for the era of on-device AI applications.

Education

Pohang University of Science and Technology (POSTECH)

B.S. IN DEPARTMENT OF ELECTRICAL ENGINEERING

- Advisor: Prof. Hong Jeong
- GPA: 4.03/4.3, Summa Cum Laude (Ranked 1st in Dept. of Electrical Engineering)

Pohang, Republic of Korea

Mar. 2012 – Feb. 2016

Pohang University of Science and Technology (POSTECH)

M.S./PH.D. IN DEPARTMENT OF CONVERGENCE IT ENGINEERING

- Advisor: Prof. Jae-Yoon Sim
- Research Topic: Energy-Efficient Deep Learning Accelerator IC Design

Pohang, Republic of Korea

Mar. 2016 – Feb. 2023

Experience

Memory System R&D, SK Hynix

RESEARCH INTERN

- Deep Learning Task Development

Icheon, Republic of Korea

June. 2017 – Aug. 2017

Pohang University of Science and Technology (POSTECH)

POSTDOCTORAL RESEARCHER AT INSTITUTE OF ARTIFICIAL INTELLIGENCE

- Advisor: Prof. Jae-Yoon Sim
- Research Topic: Energy-Efficient Deep Learning Accelerator IC Design

Pohang, Republic of Korea

Mar. 2023 – Aug. 2023

Columbia University

POSTDOCTORAL RESEARCHER IN DEPARTMENT OF ELECTRICAL ENGINEERING

- Supervisor: Prof. Mingoo Seok
- Research Topic: Deep Learning Accelerator IC Design Considering External Memory Access

New York, NY, USA

Sep. 2023 – Jan. 2025

Konkuk University

ASSISTANT PROFESSOR IN DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

- Intelligent System Architecture Lab
- Field of Research: Digital VLSI and System Design for Intelligent Applications

Seoul, Republic of Korea

Mar. 2025 –

Awards, Fellowships, & Grants

2014 – 2015 **Undergraduate Fellowship**
Oh-hyun Kwon (Personnel, Vice CEO of Samsung)

2022 **Semiconductor Design Contest – Industry Special Award (Samsung)**
Korea Semiconductor Industry Association (KSIA)

Teaching Experience

Spring 2025	Logic Circuits AI Accelerator Design	Konkuk University
Fall 2025	Computer Architecture AI Digital Integrated Circuits SoC Architecture	Konkuk University
Spring 2026	Logic Circuits AI Accelerator Design	Konkuk University

Presentations

INVITED TALKS

- [1] “Research Directions in AI Hardware Design,” in Undergraduate Technical Seminar, Konkuk University. Jan. 2025.
- [2] “T-REX: A 68-567 $\mu\text{s}/\text{token}$, 0.41-3.95 $\mu\text{J}/\text{token}$ Transformer Accelerator with Reduced External Memory Access and Enhanced Hardware Utilization in 16nm FinFET,” in ISSCC 2025 Review Workshop, Solid-State Circuit Society Seoul Chapter, Yonsei University. May. 2025.
- [3] “Making Transformers Fast: Recent Hardware Trends,” in BK21 Four Seminar, School of Computer Science and Electrical Engineering, Handong Global University. Sep. 2025.
- [4] “T-REX: Hardware-Software Co-Optimized Transformer Accelerator with Reduced External Memory Access and Enhanced Hardware Utilization,” in Invited Session, International System-on-Chip Conference (*ISOCC*), Busan, Oct. 2025.
- [5] “Two Challenges in AI Accelerators,” in Research Seminar, Department of Electronic Engineering, Hanyang University. Oct. 2025.

CONTRIBUTED PRESENTATIONS

- [1] **S. Moon** et al. “An 8.9–71.3 TOPS/W Deep Learning Accelerator for Arbitrarily Quantized Neural Networks,” in IEEE International Symposium on Integrated Circuits and Systems (*ISICAS*), Bordeaux, France, 2022.
- [2] **S. Moon** et al. “A 127.8TOPS/W Arbitrarily Quantized 1-to-8b Scalable-Precision Accelerator for General-Purpose Deep Learning with Reduction of Storage, Logic, and Latency Waste,” in IEEE International Solid-State Circuits Conference (*ISSCC*), San Francisco, CA, 2023.
- [3] **S. Moon** et al. “T-REX: A 68-567 $\mu\text{s}/\text{token}$, 0.41-3.95 $\mu\text{J}/\text{token}$ Transformer Accelerator with Reduced External Memory Access and Enhanced Hardware Utilization in 16nm FinFET,” in IEEE International Solid-State Circuits Conference (*ISSCC*), San Francisco, CA, 2025.
- [4] **S. Moon** et al. “PersASR: A 2.3- $\mu\text{J}/\text{frame}$ 96.8%-accurate TinyML Automatic Speech Recognition Processor with Data Augmentation-Driven Personalization in 16-nm FinFET,” in IEEE Custom Integrated Circuits Conference (*CICC*), Seattle, WA, 2026.

Academic Service

PEER REVIEW

More than 10 papers as invited reviewer at IEEE Transactions on Very Large Scale Integration Systems (*TVLSI*), IEEE Transactions on Circuits and Systems for Artificial Intelligence (*TCASAI*), IEEE Transactions on Circuits and Systems I: Regular Papers (*TCAS-I*), IEEE Journal of Solid-State Circuits (*JSSC*), IEEE Open Journal of the Solid-State Circuits Society (*OJ-SSCS*)

More than 10 papers at IEEE Biomedical Circuits and Systems (*BIOCAS*), IEEE Asia Pacific Conference on Circuits and Systems (*APCCAS*)

Skills

Tools – VCS, Verdi, Questa Sim, Model Sim, Design Compiler, IC Compiler, IC Compiler II, Innovus, PrimeTime, Tempus, Formality, StarRCXT, Virtuoso, Calibre DRC/LVS/PEX, Vivado/ISE for FPGA Design, PADS for PCB Design

Languages – Verilog, C/C++, Python, MATLAB, Latex

Research & Educational Projects

May. 2025 - **KU Research Grant**

Apr. 2026 Konkuk University

Real-Time and Energy-Efficient Video Object Detection Hardware for Autonomous Vehicles

Mar. 2026 - **Engineering Research Center (ERC)**

Feb. 2029 National Research Foundation of Korea (NRF) under Ministry of Science and ICT (MSIT)

Scalable Quantum Computing Technology Convergence Platform Center

Publications

INTERNATIONAL CONFERENCES

- [1] B. Kim, J. Park, **S. Moon**, K. Kang, and J.-Y. Sim, “Configurable Energy-Efficient Lattice-Based Post-Quantum Cryptography Processor for IoT Devices,” in IEEE European Solid-State Circuits Conference (**ESSCIRC**), pp.525-528, Sep. 2022.
- [2] **S. Moon**, H.-G. Mun, H. Son, and J.-Y. Sim, “A 127.8TOPS/W Arbitrarily Quantized 1-to-8b Scalable-Precision Accelerator for General-Purpose Deep Learning with Reduction of Storage, Logic, and Latency Waste,” in IEEE International Solid-State Circuit Conference (**ISSCC**), pp.330-331, Feb. 2023.
- [3] H.-G. Mun, H. Son, **S. Moon**, J. Park, B. Kim, and J.-Y. Sim, “A 28 nm 66.8 TOPS/W Sparsity-Aware Dynamic-Precision Deep-Learning Processor,” IEEE Symposium on VLSI Circuits (**VLSI**), C16-1, Jun. 2023.
- [4] K.-J. Lee, B. Kim, H.-G. Mun, **S. Moon** and J.-Y. Sim, “Joint Optimization of Cache Management and Graph Reordering for GCN Acceleration,” in ACM/IEEE International Symposium on Low Power Electronics Design (**ISLPED**), Aug. 2023.
- [5] B. Zhang, **S. Moon**, and M. Seok, “A 1-TFLOPS/W, 28-nm Deep Neural Network Accelerator featuring Online Compression and Decompression and BF16 Digital In-Memory-Computing Hardware,” in IEEE Custom Integrated Circuits Conference (**CICC**), Apr. 2024.
- [6] **S. Moon**, M. Li, G. Chen, P. Knag, R. Krishnamurthy, and M. Seok, “T-REX: A 68-567 μ s/token, 0.41-3.95 μ J/token Transformer Accelerator with Reduced External Memory Access and Enhanced Hardware Utilization in 16nm FinFET,” in IEEE International Solid-State Circuit Conference (**ISSCC**), pp.406-407, Feb. 2025.
- [7] C.-T. Lin, **S. Moon**, P. Huang, and M. Seok, “EPIC: a Sub-6mW In-Memory Computing-Based RISC-V Microcontroller Unit with On-Chip Training Support for TinyML,” in IEEE European Solid-State Electronics Research Conference (**ESSERC**), Sep. 2025.
- [8] **S. Moon***, J. Park*, C.-T. Lin, S. H. Cho, G. Chen, P. Knag, R. Krishnamurthy, and M. Seok, “PersASR: A 2.3- μ J/frame 96.8%-accurate TinyML Automatic Speech Recognition Processor with Data Augmentation-Driven Personalization in 16-nm FinFET,” in IEEE Custom Integrated Circuits Conference (**CICC**), Apr. 2026.

INTERNATIONAL JOURNALS

- [1] **S. Moon**, K.-J. Lee, H.-G. Mun, B. Kim and J.-Y. Sim, “An 8.9–71.3 TOPS/W Deep Learning Accelerator for Arbitrarily Quantized Neural Networks,” in IEEE Transactions on Circuits and Systems II: Express Briefs (**TCAS-II**), vol. 69, no. 10, pp. 4148-4152, Oct. 2022.
- [2] K.-J. Lee, **S. Moon** and J.-Y. Sim, “A 384G Output NonZeros/J Graph Convolutional Neural Network Accelerator,” in IEEE Transactions on Circuits and Systems II: Express Briefs (**TCAS-II**), vol. 69, no. 10, pp. 4158-4162, Oct. 2022.
- [3] H.-G. Mun, **S. Moon**, B. Kim, K.-J. Lee, and J.-Y. Sim, “Bottleneck-Stationary Compact Model Accelerator with Reduced Requirement on Memory Bandwidth for Edge Applications,” in IEEE Transactions on Circuits and Systems I: Regular Papers (**TCAS-I**), vol. 70, no. 2, pp. 772-782, Feb. 2023.

- [4] **S. Moon**, H.-G. Mun, H. Son, and J.-Y. Sim, “Multipurpose Deep-Learning Accelerator for Arbitrary Quantization With Reduction of Storage, Logic, and Latency Waste,” in IEEE Journal of Solid-State Circuits (*JSSC, Invited*), vol. 59, no. 1, pp. 143-156, Jan. 2024.
- [5] **S. Moon**, and E. Kwon, “DeltaTrack: Flow-Driven Multiple Object Tracking Accelerator with Variable LSB Approximation for Real-Time and Energy-Efficient Video Analytics,” in IEEE Transactions on Circuits and Systems II: Express Briefs (*TCAS-II*), Dec. 2025.
- [6] **S. Moon**, M. Li, G. Chen, P. Knag, R. Krishnamurthy, and M. Seok, “T-REX: Hardware-Software Co-Optimized Transformer Accelerator with Reduced External Memory Access and Enhanced Hardware Utilization,” in IEEE Journal of Solid-State Circuits (*JSSC, Invited*), Jan. 2026.

PREVIOUS FIELD OF RESEARCH

- [1] **S. Moon**, H. Jeon, I Seo, and W.-S. Han, “Development of a Graph-based Visualization Tool for Fraud Detection,” in Proceedings of the Korea Information Processing Society Conference, pp. 781-784, 2017.
- [2] J. Lee, **S. Moon**, K. Kim, D. Kim, S. Cha, and W.-S. Han, “Parallel replication across formats in SAP HANA for scaling out mixed OLTP/OLAP workloads,” in Proceedings of the VLDB Endowment 10 (12), pp. 1598-1609, 2017.
- [3] J. Lee, W.-S. Han, H. Na, C. Park, K. Kim, D. Kim, J. Lee, S. Cha and **S. Moon**, “Parallel replication across formats for scaling out mixed OLTP/OLAP workloads in main-memory databases,” in The VLDB Journal 27, pp. 421-444, 2018.