Design of an Energy-Efficient Neuron Circuit with Temporal Encoding for Capacitive Coupling Based Compute-In Memory Technology

Jung Nam Kim[1,3], Minsuk Koo[2,3,4a] and Yoon Kim[1,3,4a]
[1]Department of Electrical and Computer Engineering, University of Seoul,
[2]School of Advanced Fusion Studies, University of Seoul,
[3]Center for Semiconductor Research, University of Seoul,
[4]IM Electronics co.,
E-mail : wjdska0012@uos.ac.kr

***Abstract*— Matrix-vector multiplication (MVM) is a core operation in large language models (LLMs), and compute-in-memory (CIM) technologies offer a promising path to overcome data movement bottlenecks. Among them, capacitive coupling principle-based CIM (CCP-CIM) enables low-power operation by eliminating static current paths. In this work, we propose an energy-efficient neuron circuit optimized for CCP-CIM. The design features a cascaded input stage for enhanced transconductance linearity, as well as feedback-assisted control and overflow/underflow detection to reduce unnecessary digital conversions. Furthermore, the discharge rate is dynamically adjustable through analog biasing, enabling flexible control of the neuron's response range. Implemented in TSMC 28 nm CMOS, the proposed design achieves up to 2.15× less energy consumption than a conventional neuron circuit. This work supports scalable and adaptive analog inference for edge AI applications.**

***Keywords*— Capacitive Coupling, Compute-In Memory, Matrix-Vector Multiplication**

I. INTRODUCTION

Large language model (LLM)-based generative artificial intelligence (AI) applications have sparked intense interest in the underlying computing infrastructure. Most modern LLMs employ multi-head attention (MHA) blocks [1] to construct their encoder and decoder layers. Recent advancements further incorporate model distillation [2] and retrieval-augmented generation (RAG) [3] techniques to achieve state-of-the-art performance.

At their core, LLM computations are predominantly composed of multiply-and-accumulate (MAC) operations. While analog compute-in-memory (CIM) technologies [4–9] have garnered increasing attention as a promising approach for accelerating such operations, general-purpose GPU (GPGPU)-based digital computing platforms remain the industry standard. This persistence stems from the intrinsic advantages of digital design: process-agnostic reusability, established design flows, and inherently precise computation.

However, digital-centric computing is not without its limitations. In conventional digital systems, arithmetic logic units (ALUs) and memory modules are physically separated and connected via bandwidth-limited buses. Consequently, both computation speed and energy efficiency are bottlenecked by data movement constraints.

CIM architectures address this challenge by leveraging the intrinsic parallelism of memory arrays to accelerate matrix-vector multiplication (MVM). In CIM systems, matrix elements are mapped to memory states, while input vectors are applied to input terminals. A single memory access yields an analog signal—either current or voltage—that directly represents the result of the MVM operation.

CIM can be broadly categorized into two types based on their physical operating principles: Kirchhoff's current law (KCL)-based CIM [4–6] and capacitive coupling principle (CCP)-based CIM [7–9]. As summarized in Table 1, these two approaches differ in how they represent matrix and vector quantities. KCL-CIM represents matrix elements as conductances, input vectors as voltages, and produces current-mode outputs. In contrast, CCP-CIM represents matrix elements using capacitance, applies differential input voltages, and generates voltage-mode outputs.

CCP-CIM is inherently more energy-efficient than KCL-CIM, as it avoids continuous current flow paths and thus minimizes static power consumption. Capacitive coupling is also better suited for low-leakage and low-power circuit designs, making it an ideal candidate for energy-constrained AI accelerators. Furthermore, CCP-CIM produces MVM outputs as analog voltages within a fixed voltage headroom, which simplifies circuit design under low-voltage conditions and mitigates dynamic range management issues.

Conventional CCP-CIM implementations typically rely on analog-to-digital converters (ADCs)—such as successive

a. Corresponding author; koo@uos.ac.kr, yoonkim82@uos.ac.kr

Manuscript Received Jun. 12, 2025, Revised Sep. 17, 2025, Accepted Sep. 17, 2025

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

TABLE I. Comparisons of KCL and CCP Type

Type	KCL	CCP
Matrix element	Conductance	Capacitance
Input vector	Voltage level	Voltage difference
Output vector	Current level	Voltage level

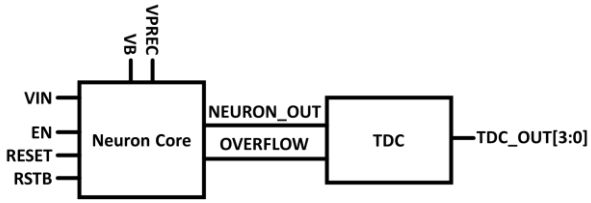

Figure 1. Neuron circuit architecture.

approximation register (SAR) ADCs [7] or Flash ADCs [8]—to digitize the analog output. Although these ADCs provide high precision, they often introduce considerable static power consumption, which undermines the low-power advantage of CCP-CIM architectures.

To overcome this challenge, we propose a low-power 4-bit neuron circuit operating in the temporal domain. The proposed neuron digitizes analog voltage outputs using a temporal-to-digital converter (TDC), and additionally enables control of output signal timing via analog biasing. This design allows for adjustment of the signal window of interest. Our prior work [9] demonstrated the feasibility of energy-efficient, temporally encoded neuron circuits for CCP-CIM. Building on that foundation, we further enhance energy efficiency by integrating overflow and underflow detection mechanisms.

II. DESIGN METHODOLOGY

Figure 1 shows the configuration of the proposed neuron circuit. A single neuron circuit consists of a neuron core and TDC. Detailed explanations of each block are provided in the following subsections. The design was implemented using the TSMC 28 nm logic CMOS process.

A. Neuron circuit design

Figure 2 presents a schematic comparison between the previous and proposed neuron core designs. The control signals include RSTB, EN, RESET, and INIT, while VIN, VPREC, and VB are analog input signals. In particular, the VIN signal is directly connected to the memory array, where the CCP-based analog MVM operation takes place. As a result, the analog MVM output—represented as a voltage level—is fed into the neuron core via VIN. During inference, the VIN voltage swings both positively and negatively around the precharged level. The detailed inference scheme utilizing the memory array has been described in our previous work [9].

Figure 3 shows the timing diagram of the neuron circuit.

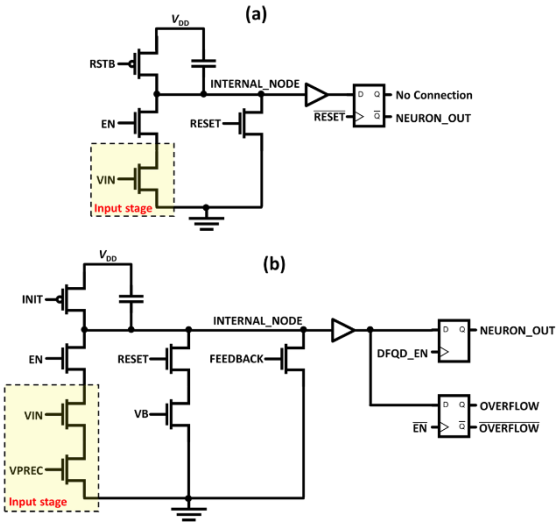

Figure 2. Comparison of neuron core schematics: (a) prior design and (b) proposed design.

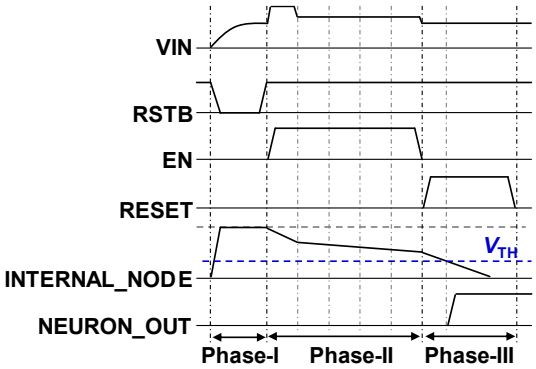

Figure 3. Neuron operation timing diagram.

The operation of the neuron circuit consists of three phases: Phase-I, Phase-II, and Phase-III. In the Phase-I, the circuit is reset by either the RSTB or INIT signal, which charges the INTERNAL_NODE to V_{DD}. In the Phase-II, the EN signal is asserted (set high), enabling the neuron circuit to integrate the VIN input. In the Phase-III, the EN signal is deasserted (set low), and the RESET signal is asserted (set high), causing the INTERNAL_NODE to discharge toward V_{SS} through a transistor controlled by the RESET signal. When the voltage of the INTERNAL_NODE drops below the threshold voltage of the amplifier, the amplifier is triggered and the neuron output (NEURON_OUT) is generated. This output is subsequently passed to the temporal-to-digital converter (TDC) stage.

The discharge rate during Phase-II is tuned by the reference voltage VPREC, while its dynamic range is bounded by VPREC$\pm\Delta$VIN, which depends on the MAC output vector values. This ensures that the INTERNAL_NODE voltage spans a sufficient range before Phase-III, allowing the TDC to reliably quantize the result within the 10 ns cycle budget at 100 MHz operation.

The previous neuron design, shown in Figure 2(a), was fabricated using the TSMC 180 nm CMOS logic process. The input stage consisted of a single transistor. Since the transistors controlled by the EN and RESET signals act as electrical switches, their effect on the signal path could be

neglected. In this earlier design, the short-channel effect (SCE) was mitigated by employing transistors with sufficient channel length.

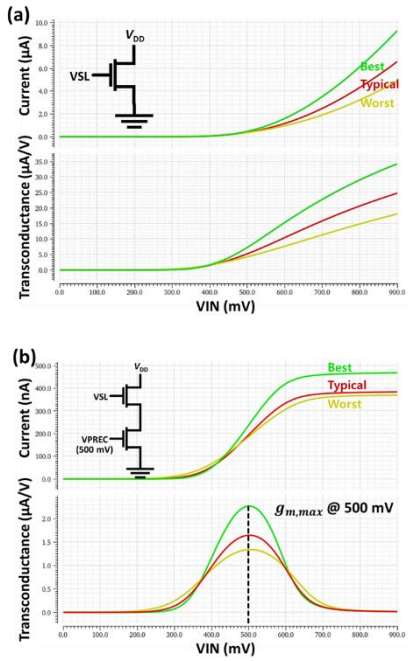

Figure 4. Current and transconductance characteristics of the input stage for (a) the prior neuron circuit and (b) the proposed neuron circuit

The proposed design introduces architectural enhancements over the previous version: improvement of input stage transconductance linearity, adjustable discharge rate in Phase-III a feedback input path, and an overflow detection mechanism. Figure 2(b) illustrates the proposed neuron circuit designed in the TSMC 28 nm CMOS logic process. In advanced technology nodes, SCE becomes more pronounced compared to older nodes. To address this issue, the input stage and reset path have been modified to include two cascaded transistors.

This modification not only suppresses SCE but also improves the linearity of the input stage′s (g_m). Figure 4 compares the simulated transfer characteristics and transconductance of (a) the previous design and (b) the proposed design. The previous design exhibits a monotonically increasing g_m, which impairs accurate mapping between input voltage and output current, thereby degrading the linearity of the neuron core.

In contrast, the proposed design achieves a peak g_m near a designated bias voltage of VPREC and then decreases. This behavior is well aligned with the inference operation. By setting VPREC equal to the precharge voltage of the memory array, a plateau region in g_m is utilized, ensuring minimal variation across the input voltage range relevant to analog MAC outputs.

In addition, the proposed design incorporates a cascaded transistor configuration in the reset path, which allows the discharge rate of the INTERNAL_NODE to be modulated via the VB bias voltage. This capability introduces a dynamic control mechanism over the neuron's operating range. By adjusting the discharge rate, the neuron can effectively shift the point of interest within the analog MVM output range.

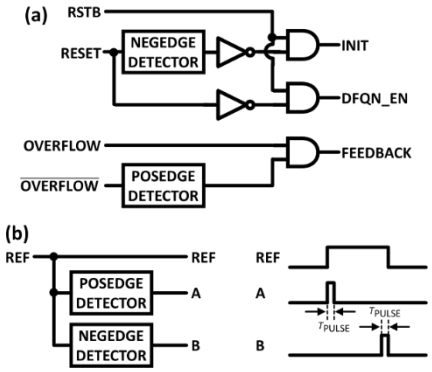

Figure 5. (a) Input signal generate logics for the proposed neuron core. (b) Timing diagram of the edge detectors (posedge/negedge).

Specifically, when the integrated result of the MVM operation is too small to trigger the amplifier within the standard discharge profile, increasing the VB bias accelerates the discharge, allowing even smaller MVM outputs to reach the amplifier's threshold and produce a valid output. Conversely, lowering the VB bias slows down the discharge, making the neuron more selective and responsive only to larger MVM results. This tunability enables dynamic range operation, allowing the neuron circuit to adaptively focus on specific regions of the input distribution—either amplifying low-range signals or filtering out smaller contributions to prioritize higher analog weights.

Another key improvement is the addition of a feedback control path and an overflow detection mechanism, which interfaces with the subsequent TDC stage. In the prior design, overflow detection was absent, leading to unnecessary activations of the TDC and considerable energy dissipation. To address this, the proposed design generates an overflow signal that both triggers TDC response and generates feedback to the neuron core.

Figure 5 illustrates the control signal generation circuit for the proposed neuron core, including the overflow-induced feedback logic. Upon detection of overflow, the TDC clamps its output and a feedback signal pulls down the INTERNAL_NODE of the neuron core for a short duration (< 1 ns). This prevents the inverter-based comparator from remaining in its high-current region, thereby significantly improving energy efficiency. This technique is also applied to the generation of the INIT signal. When the RESET signal is turned off, a short-duration INIT pulse is generated, which briefly pulls up the INTERNAL_NODE of the neuron core.

B. Temporal-to-digital converter (TDC) design

Figure 6 (a) shows the schematic of the 4-bit TDC block used in the proposed neuron circuit, which converts temporally encoded signals into digital outputs and incorporates overflow detection logic. The timing resolution, corresponding to the least significant bit (LSB), is approximately 700 ps, which is determined by the delay element. This resolution is sufficient to cover the 10 ns conversion window corresponding to a single 100 MHz system clock cycle, and it is independent of workload characteristics since the analog MAC accumulation is completed during Phase-II prior to the voltage-to-time conversion.

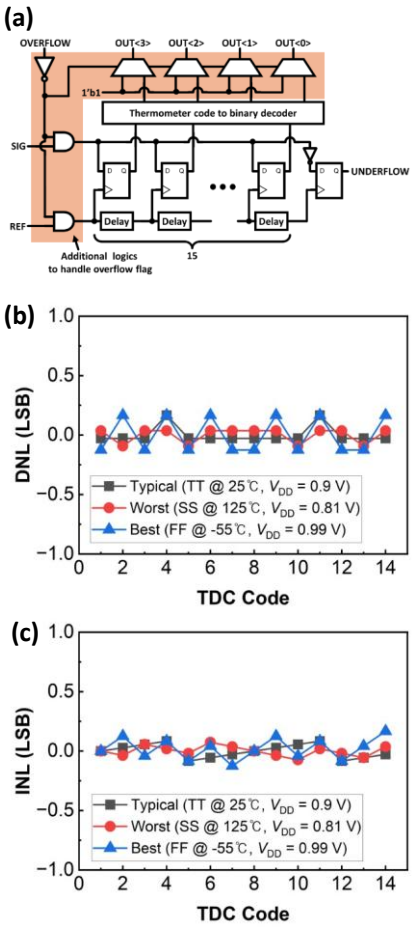

Figure 6. (a) Schematic of temporal-to-digital converter for proposed neuron circuit. (b) DNL and (c) INL measurement results obtained from post-layout simulation.

When an overflow signal is asserted, the TDC bypasses its normal evaluation path and forces its output to the maximum digital value, thereby preventing unnecessary digital conversions and improving overall system energy efficiency. The TDC design is thus tightly coupled with the neuron core to ensure both accuracy and power-aware operation.

Figure 6 (b) and (c) shows nonlinearity (DNL) and integral nonlinearity of TDC. Linearity was evaluated on a post-layout simulation. Across three representative PVT corners—Typical (TT, 25 °C, 0.9 V), Worst (SS, 125 °C, 0.81 V), and Best (FF, −55 °C, 0.99 V)—the measured DNL/INL remained within ±0.17 LSB, ensuring robust quantization.

III. RESULTS AND DISCUSSIONS

Figure 7 shows the simulation results of the proposed neuron circuit. While the EN signal remains high, the VIN signal—driven by the MVM result—varies continuously. The neuron core integrates this input, leading to a voltage drop at the INTERNAL_NODE whose slope is proportional to the MVM result.

Following the deassertion of the EN signal, the RESET signal is asserted, initiating a linear discharge at the

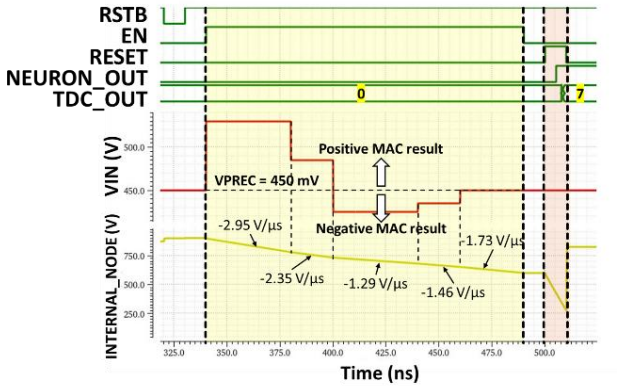

Figure 7. Schematic simulation results

INTERNAL_NODE. Once the voltage crosses the trip point

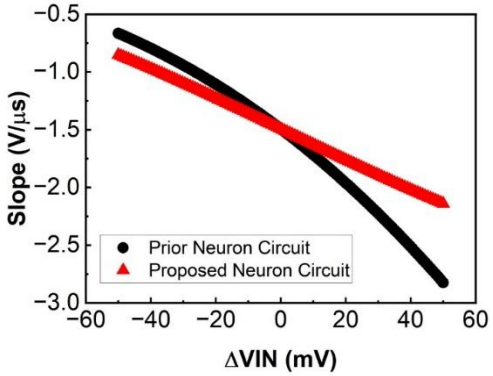

Figure 8. Discharge rate comparisons between prior neuron circuit and proposed neuron circuit.

of the inverter-based comparator, the NEURON_OUT signal is generated and then digitized by the TDC block.

For energy consumption analysis, the operation of the neuron circuit is categorized into three distinct states: underflow, normal, and overflow.

− In the underflow condition, the INTERNAL_NODE voltage fails to reach the trip point of the comparator before the RESET signal is deasserted, and thus no output pulse is generated.

− In the normal condition, the INTERNAL_NODE reaches the trip point after the EN signal is deasserted but before the RESET signal is deasserted, resulting in valid output.

− In the overflow condition, the INTERNAL_NODE reaches the trip point before the EN signal is deasserted, triggering premature activation of the output.

Figure 8 compares the discharge rate linearity between the prior neuron circuit and the proposed neuron circuit. The slope represents the discharge rate of the INTERNAL_NODE during Phase-II operation, while ΔVIN denotes the input voltage deviation from VPREC, as shown in Figure 7. The R^2 values of the prior and proposed circuits are 0.98526 and 0.99985, respectively. This improvement corresponds to approximately a 99% reduction in the residual error, demonstrating that the proposed neuron circuit achieves a near-linear analog input–output transfer

characteristic.

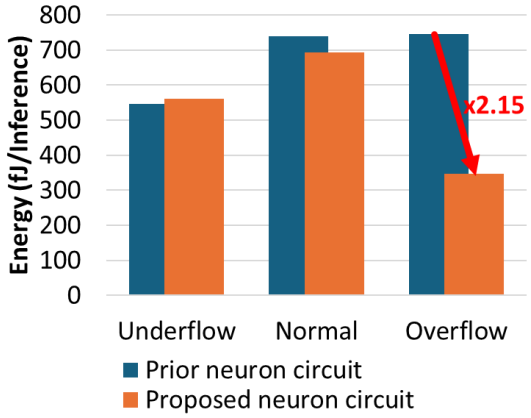

Figure 9. Energy comparison of neuron circuits

Figure 9 compares the total energy consumption of the prior and proposed neuron circuits for each operating condition. For a fair comparison and to exclude process scaling effects, both designs were implemented using the same 28 nm technology node. In the underflow condition, the prior design exhibits slightly better energy efficiency, as the additional logic blocks in the proposed design introduce overhead in this case.

However, in both the normal and overflow conditions, the proposed neuron circuit demonstrates improved energy efficiency. Under the overflow condition, the bypass path effectively reduces energy consumption by avoiding unnecessary operation of the TDC block. In the normal condition, the pull-up technique applied to the INTERNAL_NODE also contributes to energy savings. A detailed energy consumption across conditions is summarized in Table II.

TABLE II. Summary of Energy Consumption

	Operation condition	Average energy consumption (fJ/operation)		
		Total	Neuron Core	TDC
Prior neuron design	Underflow	546.7	120.5	426.2
	Normal	739.3	273.0	466.3
	Overflow	746.2	258.8	487.5
Proposed neuron design	Underflow	561.7	118.2	443.5
	Normal	694.3	206.2	488.1
	Overflow	346.3	227.4	118.9

IV. CONCLUSION

This work presents an energy-efficient neuron circuit tailored for CCP-CIM systems. The proposed design integrates a plateaued transconductance (g_m) input stage with a temporal-domain output stage using a TDC, along with overflow/underflow detection and feedback-assisted signal control to reduce unnecessary energy consumption. By modulating the analog bias in the reset path, the neuron dynamically shifts its point of interest within the analog MVM output range. This tunable sensitivity enables more effective activation under varying inference conditions, particularly when batched inputs cause small but non-negligible overflow.

The proposed circuit was implemented in a TSMC 28 nm logic CMOS process. Simulation results demonstrate substantial energy savings, especially under overflow conditions, enabled by TDC bypass logic and overflow feedback to the comparator node. Although the design incurs marginal overhead under underflow scenarios, its overall energy efficiency and adaptability make it well suited for low-power, edge-level analog AI inference systems.

ACKNOWLEDGMENT

This research was supported by National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(RS-2023-00217673). The EDA tool was supported by the IC Design Education Center(IDEC), Korea.

REFERENCES

[1] A. Vaswani, et. al., “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, 30, 2017.
[2] G. Hinton, et. al., “Distilling the knowledge in a neural network,” arXiv:1503.02531, Mar. 2015.
[3] P. Lewis, et. al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv:2005.11401, Apr. 2021.
[4] X. Li et al., "First Demonstration of Homomorphic Encryption using Multi-Functional RRAM Arrays with a Novel Noise-Modulation Scheme," *Int. Electron Devices. Meet.*, 2022.
[5] M. Kim et al., "A 3D NAND Flash Ready 8-Bit Convolutional Neural Network Core Demonstrated in a Standard Logic Process," *IEEE Int. Electron Devices. Meet.*, 2019.
[6] J. -M. Hung et al., "An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4-21.6TOPS/W for Edge-AI Devices," *IEEE Int. Solid-State Circuits Conf.*, 2022.
[7] B. Zhang et al., "A 177 TOPS/W, Capacitor-based In-Memory Computing SRAM Macro with Stepwise-Charging/Discharging DACs and Sparsity-Optimized Bitcells for 4-Bit Deep Convolutional Neural Networks," *IEEE Cust. Integr. Circuits Conf.*, 2022.
[8] Z. Jiang, S. Yin, J. -S. Seo and M. Seok, "C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism," *IEEE J. Solid-State Circuits*, vol. 55, no. 7, Jul. 2020.
[9] J.N Kim, et. al., "Voltage-Summation-Based Compute-in-Memory Technology with Capacitive Synaptic Devices," *Adv. Intell. Syst.*, 2500028, Feb. 2025.

Jung Nam Kim received B.S. degree in Physics from Soongsil University, Seoul, South Korea, in 2020. He is currently pursuing Ph.D. degree in electrical and computer engineering at the University of Seoul.

Minsuk Koo received B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 2007, the M.S. degree in electrical engineering from Seoul National University in 2009, and Ph.D degree in Electrical and Computer Engineering from Purdue University in 2020. From 2009 to 2012, he was a Senior Engineer with RadioPulse Inc., Seoul, South Korea, where he was involved in the development of ZigBee transceiver and SoC products. From 2020 to 2024, he was with the Department of Computer Science and Engineering at Incheon National University. Since 2024, he has been with the School of Advanced Fusion Studies and AI Semiconductor at the University of Seoul.

Yoon Kim received B.S. and Ph.D. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 2006 and 2012, respectively. From 2012 to 2015, he was a Senior Engineer with Samsung Electronics Company Ltd., South Korea. In 2018, he joined the University of Seoul and became an Associate Professor in 2020.