Memory Subsystems for AI: Bridging Performance and Energy from Cloud to Edge

Dongkyun Kim*
SK hynix, Icheon, South Korea

*Corresponding e-mail: dongkyun.kim@sk.com

Abstract: In the AI era, the scale of large language models (LLMs) is growing at an unprecedented pace, driving a rapid increase in the required compute capability of server platforms. Yet the expected explosion of AI applications suggests that server-only approaches may not be sufficient to meet latency, scalability, cost, and energy constraints. As a result, AI inference is increasingly being distributed across the entire computing continuum—from data centers to client PCs and mobile devices—accelerating the rise of on-device AI and motivating active hardware–software co-optimization.

This keynote examines memory as the critical bottleneck and key enabler across server, client, and mobile systems, and discusses how the memory subsystem must evolve to deliver higher bandwidth, lower energy per bit, and more efficient data movement. In particular, I will present a system-level perspective on emerging memory-centric solutions—such as wide-I/O high-bandwidth memory for edge platforms, Processing-in-Memory (PIM), and Compute-in-Memory (CIM)—highlighting their benefits and limitations, design trade-offs, and practical deployment considerations. Finally, I will outline development directions and architectural opportunities for building scalable and energy-efficient AI computing through next-generation memory subsystems.