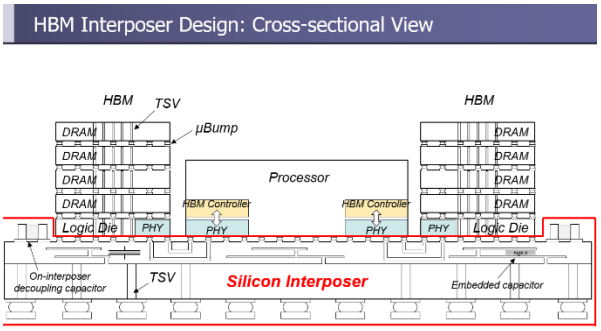
HBM-HBF Memory Architecture and Interconnection Innovations for Next Generation Agentic AI with Token Economics

Joungho Kim
KAIST, Daejon, South Korea

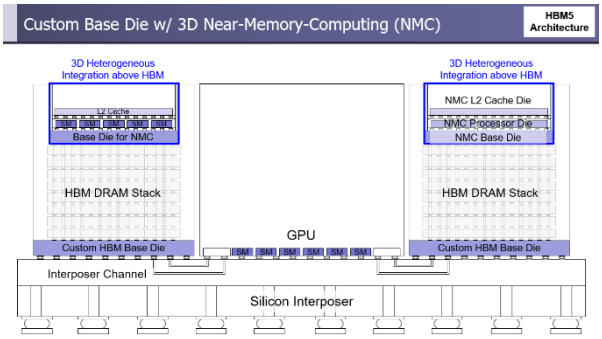
Corresponding e-mail: joungho@kaist.ac.kr

Artificial intelligence is advancing beyond the eras of generative AI and agentic AI, and is now entering the age of Physical AI. The training and inference of these future, ultra-scale AI systems require unprecedented computational power and memory performance. In particular, memory bandwidth and capacity have emerged as decisive factors that define an AI system's service capability, quality, and scalability. As a result, AI computing will increasingly converge into ultra-large AI data centers, and innovation in memory technologies—rather than GPUs—will become the central driver of technological progress.

In this presentation, we will examine HBM (High Bandwidth Memory) and HBF (High Bandwidth Flash), the high-bandwidth, high-capacity memory technologies essential for building and operating next-generation ultra-large AI systems. Drawing on AI model characteristics and system-level analysis, we will introduce an optimized HBM–HBF integrated memory architecture. We will also discuss HBDF (High Bandwidth Dynamic Flash) as an emerging component of this ecosystem. Finally, we will present the long-term technology roadmap for HBM–HBF.



< Cross-sectional View of Silicon Interposer based HBM Module >



< 3D Integrated NMC-HBM Architecture above HBM DRAM >

- By integrating a NMC processor die and cache die above HBM, the proposed 3D NMC-HBM achieves high performance and energy efficient computing through dedicated TSV interconnection and power supply network.