3DIC 2024
Sendai, Japan

INTERNATIONAL 3D SYSTEMS INTEGRATION CONFERENCE (3DIC)
September 25-27, 2024
Sendai, Japan

Keynote Speech III
14:45-15:25, September 25, 2024

IBM NorthPole: Brain-Inspired Neural Inference Architecture Intertwining Compute with Memory

Dr. Takanori Ueda
IBM Research - Tokyo
Staff Research Scientist

Abstract: The rapid advancement of artificial neural networks is barely sustained by commodity computing based on the von Neumann architecture, introduced in the 1940s. While the distinct boundary between processor and memory offers significant flexibility in semiconductor manufacturing and system design, the communication between processor and memory results in performance and power overheads, known as the von Neumann bottleneck. In contrast, artificial neural networks are modeled after biological brains that have no clear boundary between compute and memory. Shouldn't processors also evolve towards a more brain-like architecture to achieve the exceptional power efficiency of the brain?

IBM NorthPole is a brain-inspired neural inference architecture that blurs the boundary by intertwining compute and memory on a chip. Once a trained AI model is fully loaded into the on-chip memory, inference results for new incoming data are calculated exclusively within the chip without accessing off-chip memory, thus overcoming the von Neumann bottleneck. The software, co-designed with the NorthPole architecture, provides an end-to-end toolchain to optimize neural networks for the architecture. The first NorthPole chip, manufactured using Global Foundries 12nm technology, compared to a GPU fabricated on a comparable 12nm process node, achieved 25 times higher energy efficiency, 5 times higher space efficiency, and 22 times lower latency.

CV: Dr. Takanori Ueda is a researcher in the semiconductor division at IBM. His current work spans from the enablement of new semiconductor technologies to AI chip designs. He is a core member of the NorthPole project team, where he has made essential contributions, particularly in the areas of verification and physical design of the computing core. Dr. Ueda began his career in the semiconductor field in 2018, coinciding with the launch of the NorthPole project. Prior to this, his primary focus was on accelerating business workloads including microservices, database applications, and natural language processing, through his expertise in software optimization techniques.

Before joining IBM, Dr. Ueda conducted research on relational database systems. He earned his Ph.D. from Waseda University in 2013. His work has been recognized with multiple awards from academic societies.