AI Hardware innovation for sustainable society

Shintaro Yamamichi

IBM Semiconductors, IBM Research -Tokyo, Kawasaki 212-0032, Japan

SHINTYM@jp.ibm.com

Abstract

Despite the continuous advances in semiconductor technology, the computational demands for AI processing are growing at an even faster rate, necessitating breakthroughs in semiconductor devices and packaging structures specialized for AI processing. In particular, the power consumption of computers required for AI processing is increasing explosively every year, posing a major challenge in achieving a sustainable society. Focusing on the fact that most neural network calculations in AI processing are multiply-accumulate operations, various low-power AI hardware have been proposed, to address the issue of how to efficiently perform multiply-accumulate operations by arranging processors and memory. This presentation will introduce the development status of three types of AI hardware. First, we will introduce AIU Spyre[1], which can be manufactured using existing foundry processes and achieves low power consumption in data centers by optimizing the calculation accuracy of the calculation unit and memory architecture. Next, we will introduce AIU NorthPole[2], which achieves significant power savings when processing relatively small AI models at the edge by subdividing on-chip memory and multiply-accumulate units into an array. Finally, we will introduce AIU Analog[3], which achieves low power consumption through analog calculation processing by physically representing the multiply-accumulate operation itself as the sum of the currents flowing through the resistive elements of non-volatile memory.

References

1 C. Berry, Hot Chips 2024.

2 D. S. Mohda, *et al.*, Science **382**, 329-335 (2023)

3 S. Ambrogio *et al.*, Nature **620**, 768-775 (2023)

Shintaro Yamamichi received his M.E. and Ph. D. degrees in electrical engineering, from Kyoto University, Japan in 1989 and 2002, respectively. He was involved in semiconductor R&D in NEC and Renesas Electronics. He was also a visiting industrial fellow at University of California, Berkeley in 1997. In 2013, he joined IBM, Science and Technology team in Tokyo. From 2016, He led the research projects, including quantum computer installation, AI hardware, advanced packaging, and material informatics. From 2023, he is the Director of IBM Semiconductors Japan, IBM Research

-Tokyo, leading circuit design, chiplet technology, fab automation and data science activities, as well as the collaboration with research partners in global.