基於類比運算之下世代人工智慧晶片之關鍵技術

Analog Computing Based Artificial Intelligent Chip Techniques

鄭桂忠 [1]，張孟凡 [1]，謝志成 [1]，呂仁碩 [1]，謝秉璇 [1]

[1]National Tsing-Hua University

E-mail: kttang@ee.nthu.edu.tw

As emerging applications such as machine learning (ML) and big-data processing, convolution neural network (CNN) is widely utilized on many image classification applications. However, typical edge devices based on the von Neumann architecture have the processing elements (PEs) separated from the memory devices, such that PEs must frequently access data via the memory bus. Considerable read latency, high parasitic load on the data bus, and limited bandwidth for memory access in movement of data from memory to PEs greatly increase the overall latency and energy consumption which is known as the Von Neumann bottleneck. To address the above issue, this project is expected to a forward-looking computing system architecture and chip design method based on analog computing for mobile edge devices to develop next-generation artificial intelligence chips. This computing system architecture includes computing-in-memory (CIM), computing-in-sensor (CIS), neuromorphic computing (NC). Computing-in-memory operations use SRAM and ReRAM as carriers, and in-sensor operations will use CMOS image sensor as carriers. Our team plans to develop low-voltage, low-power, and neural-like artificial intelligence based on analog operations of the chip. The research results will bring considerable influence and impact to the market. The neuro-like intelligent vision system chip of mobile devices has many applications in security monitoring, automated robots, drone detection, and smart manufacturing. It's expected to make considerable contributions to academic research, national development, and economic markets.

References

[1] X. Si, ., **R.-S. Liu**, **C.-C. Hsieh**, **K.-T. Tang**, **M.-F. Chang**, "A Twin-8T SRAM Computation - In - Memory Macro for Multiple-bits CNN-Based Machine Learning," IEEE International Solid-State Circuits Conference (ISSCC) Dig.Tech. Papers, pp. 396-398, Feb. 2019

[2] **K.-T. Tang**, ...**R.-S. Liu**, **C.-C. Hsieh**, **M.-F. Chang**, "Considerations of Integrating Computing-In-Memory and Processing-In-Sensor into Convolutional Neural Network Accelerators for Low-Power Edge Devices", Symposium on VLSI Technology, June 2019

[3] W.-H. Chen, **R.-S. Liu, C.-C. Hsieh, K.-T. Tang, M.-F. Chang**,"A 65nm 1Mb Nonvolatile Computing – in - Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp. 494-495, Feb 2018

[4] J. W Su, **R.-S. Liu**, **C.-C. Hsieh**, **K.-T. Tang**, **M.-F. Chang**, "A 28nm 64Kb Inference-Training Two-Way Transpose Multibit 6T SRAM Compute-in-Memory Macro for AI Edge Chips", ISSCC, pp. 240-242, 2020.

[5] J.W Su, **R.-S. Liu**, **C.-C. Hsieh**, **K.-T. Tang**, **M.-F. Chang**,, "A 28nm 384kb 6T-SRAM Computation – in - Memory Macro with 8b Precision for AI Edge Chips", ISSCC, pp.250-251, 2021.

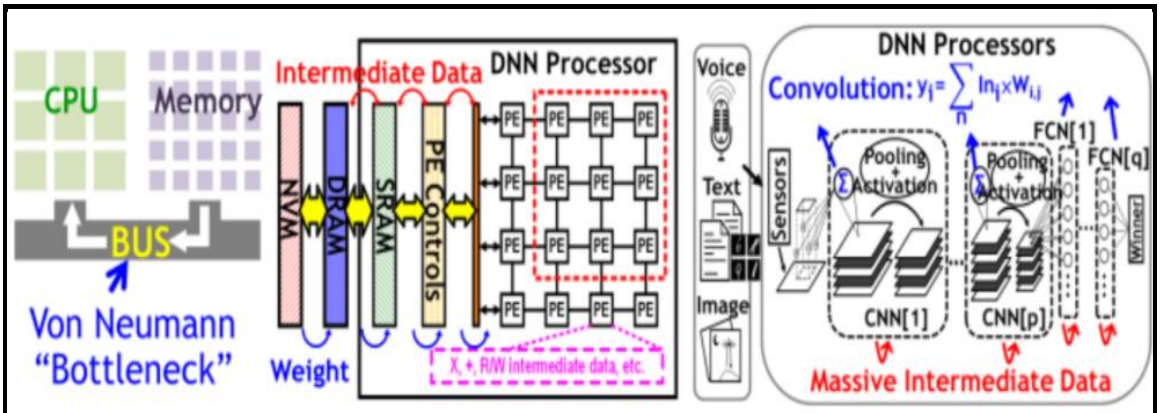

Fig 1. Von Neumann Bottleneck

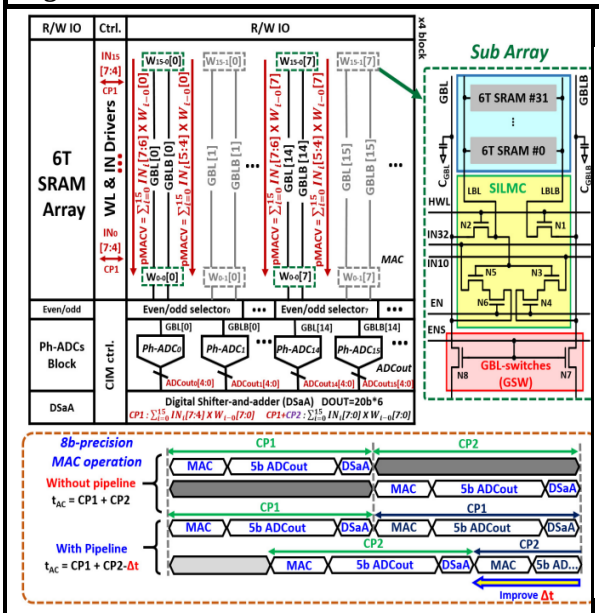

Fig 2: Proposed macro structure and data flow

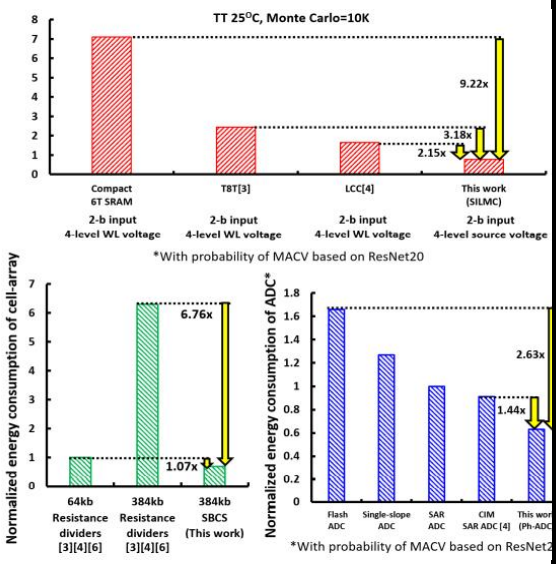

Fig 3: Simulated performance

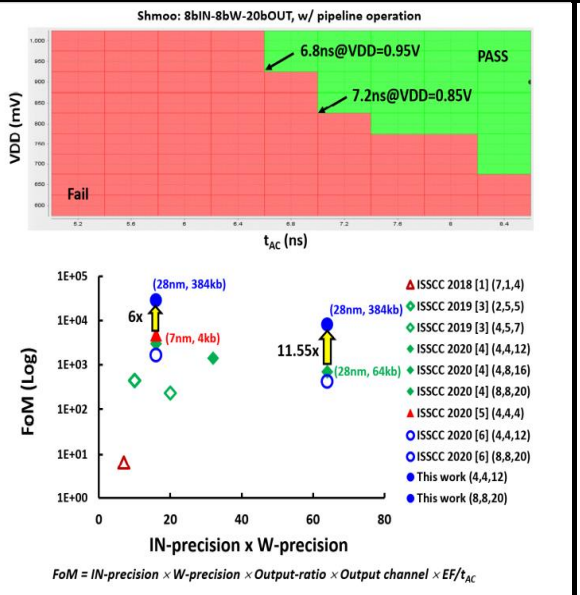

Fig 4: Measurement results

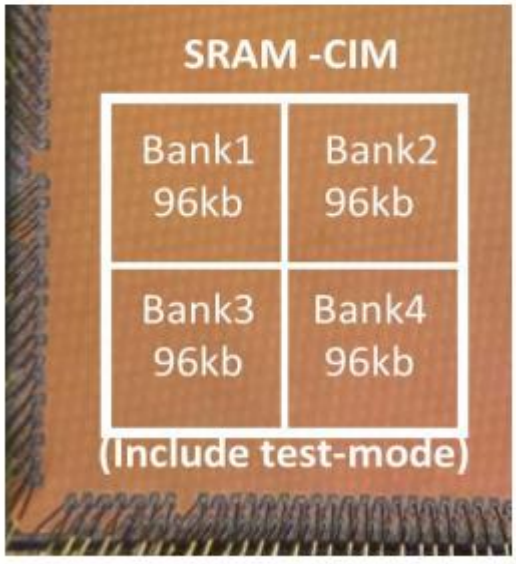

Fig 5: Die photo