Challenges and Trends of In-Memory Computing Using Emerging Memory Devices

SeokJu Yun*

Samsung Advanced Institute of Technology, Korea

In-memory computing (IMC) is an emerging non-von Neumann architecture where computation is performed in the memory array itself. Recently, IMCs are being widely explored to minimize power consumption in data movement from massive multiply-and-accumulate (MAC) which dominates the workload of deep neural network inference in artificial intelligence (AI) applications. Various emerging memory devices and new array architectures are being competitively applied for IMC developments. To improve the energy efficiency, cell density, and computing accuracy of the IMC, high-performance memory devices such as fast read and write speed, high density, and high reliability, etc., are required. Nonvolatile-memory (NVM) IMC is suitable for low-power edge-AI devices requiring neural network (NN) parameter storage in the power-off mode, rapid response to device wake-up, and high energy efficiency for MAC operations. Among nonvolatile memories, magnetic random-access memory (MRAM) has demonstrated promising developments for IMC. Compared with other NVMs, the MRAM-IMC presents better repeatability, CMOS compatibility, and high endurance. However, MRAM-IMC suffers from inherently low tunnel magnetoresistance (TMR) and low resistance of MRAM which results in poor signal margin and higher power consumption, respectively. Resistive random-access memory (ReRAM) is a feasible candidate for the IMC due to its manufacturability and appropriate ON/OFF resistance ratio. In recent research, ReRAM-based IMC macros have demonstrated MAC operations in 8b input and 8b weight with high memory density and energy efficiency. To support sufficiently large on-chip memory and multi-level cell (MLC) functions, IMC based on embedded flash memory (eFlash) and phase change memory (PCM) have been presented. However, eFlash or PCM-based IMCs tend to present accuracy results in simpler neural network inference, implemented using differential cells with the low accuracy of multilevel weights and low area efficiency. NVM-based IMC generally employs analog voltage on the bit-line (BL) to generate memory cell current and accumulate on BL for MAC operations. However, this approach increases energy consumption by necessitating the use of the direct current in the cell array. In addition, they require power-hungry analog-to-digital converters (ADCs) deteriorating the computing latency, and energy efficiency. A typical approach to solve this issue is reduction output precision, but it limits the output ratio (output precision/ideal output precision) and degrades inference accuracy when dealing with complex applications and/or datasets. Although these analog IMCs (AIMCs) present higher energy efficiency than digital accelerators, inconsistent performance, low accuracy caused by transistor analog circuit variation, low reliability, and poor flexibility are still challenging. Static Random Access Memory (SRAM) digital IMCs (DIMC) are being noticeably attracted attention due to sufficient accuracy and flexibility for various input and weight bit widths, while also benefiting from technology scaling. Recently, DIMC-based DNN accelerators have presented remarkable improvements in energy and area efficiency (TOPS/W and TOPS/mm2) while implementing 5nm and 4nm CMOS processes. Although SRAM has a larger bit cell size than other embedded memory, research and developments of the SRAM DIMC are being accelerated since it shows much higher energy efficiency and throughput per area than AIMC from operating lower supply operation and using scale-down process.

*Corresponding author	SeokJu Yun
Affiliation	Samsung Advanced Institute of Technology
E-mail address	sjwannabi@gmail.com