Memory Technology enabling the future computing systems

Paolo Fantini

Micron Technology Inc., via Trento 26 - Vimercate, 20871, ITALY

More than 2.5 exabytes of data are created every day demonstrating that we entered in the Data Economy Era, unlocked by the development of more and more advanced memory and storage technologies. The exponential growth of the annual volume of data generated in the global Datasphere (Fig. 1) is predicted to be strongly fueled by the more and more pervasive growth of Artificial Intelligence (AI) and 5G, forming a powerful duo of the data economy.
AI fundamentally re-defines the insights we can deliver through data.
5G, at variance, is unleashing data movement, asking for low latency and enhanced mobile bandwidth.

This new scenario demands that also the underlying computing system infrastructure is re-architected from the ground up to optimize for *data centricity*.

The access to these vast pools of data is exacerbating the limitations of the speed and energy that is consumed during the transfer of data between the memory and the CPU. In other words, memory, the enabler of the Data Economy, is now becoming the bottleneck. Thus, new approaches, inspired to the human mind, are essential for creating more efficient computing systems evolving towards a data-centricity and extending to the intelligent edge. Definitively, the integration of Compute and Memory can address the “Memory Bottleneck” by means of the:

1. Insertion of a “tightly” compute coupled layer in the Memory Hierarchy
2. Moving compute primitives onto the memory die
3. Merging the compute and memory with in-memory Neural Network fabrics

Emerging Memories and Memory Abstraction are likely further enablers for a solution of the “Memory Bottleneck”.
Fig. 2 reports the prismatic decomposition of the AI semantic, meaning: Neural Network accelerators, Deep Learning, Matrix-Vector-Multiplications, Brain-inspired neural Network using spikes to transmit and store information, In-memory Logic, Associative memory and storage,…

The talk will highlight as the breakthroughs in interconnect technology is pivotal to optimize the link between Host and Memory and, thus, to fulfil the various In-Memory-Computing applications.

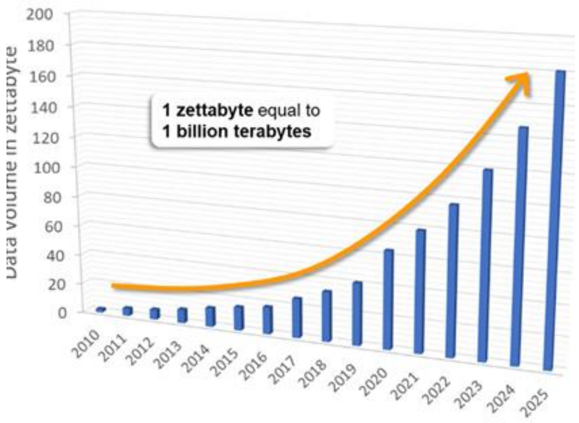

Figure 1: Annual Size of Global Datasphere over time (in ZB) and its projection up to 2025. *Source: IDC Global DataSphere report 2021.*

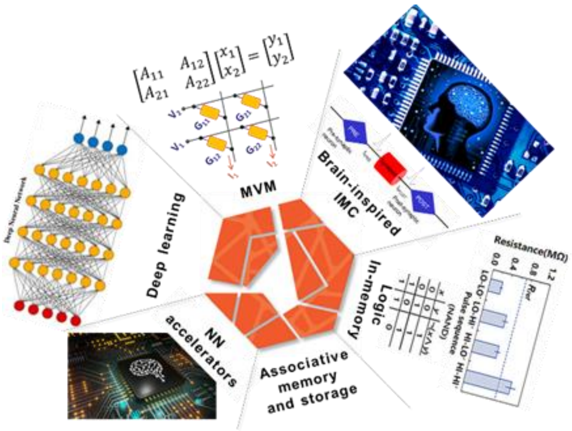

Figure 2: Various In-Memory Computing applications required for a computational memory.

[*] *corresponding author e-mail: pfantini@micron.com*