Semiconductor Innovations for the AI Era

Dr. Vijay Narayanan

IBM Fellow, T. J. Watson Research Center, Yorktown Heights, NY USA

Artificial intelligence (AI) is now pervasive – augmenting our capabilities and enriching our experiences – but it was less than a decade ago that the first key breakthroughs in deep learning were made. Tremendous progress has since been made in expanding AI applications as well as the accuracy of AI models, often by generating massive models that are trained on large datasets. However, this explosive growth in model size and the concomitant increase in required compute is unsustainable without significant semiconductor innovations across the hardware stack from materials and devices at the transistor level up through packaging. In this talk, materials advances needed to sustain continued CMOS scaling will be discussed including advanced transistor gate stacks, novel interconnect materials, and next-generation photoresists for high-NA photolithography. In addition, architectural progress in devices that harness the third dimension will be reviewed and will be shown to be critical to propel transistor density scaling [1]. In looking beyond transistor performance, a novel non-von Neumann computational approach will be introduced that envisions artificial neural networks mapped to arrays of non-volatile memory (NVM) elements. These NVM elements act as artificial synapses and encode the weights of a neural network [2] that execute analog compute operations in-memory, thereby enabling significant power performance benefits. It will also be shown that co-optimization of materials, algorithms and architecture is needed to unlock the promise of analog in-memory compute. Lastly, enhanced connectivity and scalability using a chiplet approach will be described to allow for seamless integration of disparate compute components [3]. Indeed, novel compute technologies combined with heterogeneous integration techniques that address key bandwidth challenges will be needed to power the AI of tomorrow.

References

1. N. Loubet *et al.*, *VLSI* **(2017)**. H. Jagannathan *et al.*, IEDM **(2021)**.
2. G. W. Burr *et al.*, *Adv. Phys.: X*, 2:1, 89-124 **(2017)**.
3. K. Sakuma et al., *IEEE ECTC* **(2021)**. M. Farooq *et al.*, *IEEE ECTC* **(2022)**.