NVM-based Analog Compute for AI Acceleration

Nicole Saulnier, IBM Research

The field of AI is rapidly evolving due to exponential growth of data, algorithmic innovations, and ever improved compute capability. Deep Learning based AI algorithms are being deployed for this purpose and run on specialized AI hardware such as Graphical Processing Units (GPU). However, GPUs consume significant power since data transfer occurs from memory to processor. Resistive Processing Units (RPU) envision artificial neural networks mapped to arrays of non-volatile memory (NVM) elements that execute operations in-memory and constant time, thereby enabling significant power performance benefits. This talk will discuss the various devices that could be used to build such an analog AI compute system. Additionally, some of the shortcomings of these devices as well as materials, algorithmic and architectural innovations to address them will be discussed.

Nicole Saulnier is currently the manager of AI Materials and Process Integration at IBM Research in Albany, NY. She earned her Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University where she studied computational modeling of single photon emitting devices. She joined IBM in 2011 as a member of the Advanced Patterning team and became an expert in lithography solutions including multi-patterning lithography, EUV lithography, OPC, and source mask optimization for 14nm technology and beyond. In 2017, Nicole became a manager in the Semiconductor Materials and Process Technology Research department with a focus on emerging technologies. Since the launch of the AI Hardware Center, Nicole and her team now work solely on Analog-AI technologies including PCM and RRAM.