Dr. Arman Roohi

NCMN Seminar Series

Enabling Efficient and Reliable Edge Computing: From Device to Architecture

November 11, 2020 - 4 p.m.

Online via Zoom
Meeting ID: 926 3266 8007
URL: https://unl.zoom.us/j/92632668007

Abstract

Benefits of alternatives to von-Neumann architectures for emerging applications such as neuromorphic computing and Internet-of-Thing (IoT) include avoidance of the processor-memory bottleneck, reduced energy consumption, and area-sparing computation. Viable solutions to the challenge of designing these emerging computing systems span the interrelated fields of machine learning, computer architecture, and the potential to leverage the complementary characteristics of emerging device technologies. This talk covers two of the most significant applications, including energy harvesting systems and big data processing.

Energy-harvesting-powered computing offers intriguing and vast opportunities to transform the landscape of IoT devices dramatically. These devices require drastically reduced energy consumption such that they can operate using only ambient sources of light, thermal, etc. If lightweight embedded computing could be realized with free and/or inexhaustible sources of energy, new classes of maintenance-free, compact, and inexpensive computing applications would become possible. As a new foundational computing approach to operate within the energy constraints, it is proposed to innovate Intermittent-Robust Computation (IRC) leveraging the non-volatility inherent in post-CMOS switching devices.

The foundations of IRC are advanced from the ground up by extending Spintronics device models to realize reconfigurable gates logic approaches and libraries, that leverage intrinsic-non-volatility to realize middleware-coherent, intermittent computation without checkpointing, or micro-tasking and energy overheads vital to IoT. The synthesis and optimization procedures, as design methodology, instantiate the developed library cells within standard Register Transfer Language specifications to generate power-failure resilient VLSI implementations. Another highly used application is deep Convolutional Neural Network (CNN), which has shown impressive performance for computer vision, e.g., image recognition tasks, achieving close to human-level perception rates. The ability of conventional computing platforms to support memory-oriented computing for processing large datasets is hindered due to exiting limitations either in the device, i.e., power wall, or architecture, i.e., memory wall. Moreover, the processing demands of high-depth CNNs spanning hundreds of layers face severe challenges in terms of memory and computation resources, which is crucial for resource-limited IoT nodes. This issue has been motivating the development of alternative approaches in both SW/HW domains to improve conventional CNN efficiency. Therefore, developing an optimized in-memory processing accelerator for convolutional layers via algorithm and hardware co-design approach will be discussed.