Abstract

The analysis of big-data at exa-scale (10^{18} bytes/s or flops) has introduced the emerging need to reexamine the existing hardware platform that can support memory-oriented computing. A big-data-driven application requires huge bandwidth with maintained low-power density. The most widely existed data-driven application is machine learning in big data storage system, as the most exciting feature of future big-data storage system is to find implicit pattern of data and excavate valued behavior behind. However, to handle big image data at exa-scale, there is a memory wall that has long memory access latency as well as limited memory bandwidth.

The recent emerging RRAM can provide non-volatile memory storage but also intrinsic computing for matrix-vector multiplication, which is ideal for low-power and high-throughput data analytics accelerator performed in memory. However, the existing RRAM-crossbar based computing is mainly assumed as a multi-level analog computing, whose result is sensitive to process non-uniformity as well as additional overhead from AD-conversion and I/O. In this work, we explore the matrix-vector multiplication accelerator on a binary RRAM-crossbar with adaptive 1-bit-comparator based parallel conversion. Moreover, a distributed in-memory computing architecture is also developed with according control protocol. Both memory array and logic accelerator are implemented on the binary RRAM-crossbar, where logic-memory pair can be distributed with protocol of control bus. Experiment results have shown that compared to the analog RRAM-crossbar, the proposed binary RRAM-crossbar can achieve significant area-saving with better calculation accuracy. Moreover, significant speed-up can be achieved for matrix-vector multiplication in the neuron-network based machine learning such that the overall training and testing time can be both reduced respectively. In addition, large energy saving can be also achieved when compared to the traditional CMOS-based out-of-memory computing architecture. We also develop an RRAM based oscillator network for L2-norm calculation. With RRAM-crossbar and coupled-RRAM-oscillator, machine learning applications can be accelerated with better energy-efficiency and smaller area-overhead.