Reconfigurable Machine Learning Accelerator Circuits

Himanshu Kaul
Intel Corporation

Performance and power limiting data computations in machine learning applications can benefit immensely from the improved energy and area efficiency of special-purpose accelerators. However, these accelerators need to support reconfigurability in number formats, precisions, data sparsity, and type of computations while maximizing both area and energy efficiency across the increasing variety of machine learning workloads. Reconfigurability typically incurs significant area and energy overhead. This presentation will describe reconfigurable accelerator circuits targeted for k-nearest-neighbor (kNN) and matrix multiply computations, critical for machine learning applications. These accelerators, each fabricated in 14nm tri-gate CMOS, minimize reconfiguration overhead, exploit higher energy efficiency of low-precision and sparse data computation, and employ ultra-low voltage optimized circuits to enable peak energy efficiency operation at near-threshold voltages.

Determining the closest match among high-dimensional vectors within a large vector database for kNN computations results in high compute cost. The kNN accelerator, reconfigurable for either Manhattan or Euclidean distance, employs data-adaptive precision to improve energy efficiency. A majority of vectors are eliminated without costly full-precision computation, with as-needed precision refinement to guarantee kNN accuracy of closely matched vectors, enabling up to 5.2× higher throughput. Adaptive precision compute uses partial distance compute circuits, 2b window-based sort, and MSB-to-LSB based selective distance refinement. Furthermore, state tracking circuits within the accelerator reduce search space by 16×, minimizing the energy to find the next nearest neighbor. The 128 x 128-D kNN accelerator occupies 0.333mm^2 area and achieves 21.5M vectors/s with 3.37nJ/vector or 9.7TOPS/W measured at 750mV, 25°C (14nm CMOS). Robust near-threshold circuits operating across 360mV-850mV enable 1.23nJ/vector peak efficiency at 390mV.

Recent matrix-multiply accelerators are optimized for specific algorithms and workloads, limiting computation to only integer/floating-point numbers or only dense/sparse matrices. Support for both multiple-precision integer/floating-point formats as well as dense and sparse matrices is critical for multi-purpose processor-integrated and discrete accelerators. A 4x4 node reconfigurable matrix-multiply accelerator, scalable to larger matrices through tiling or iteration, uses reconfigurable routers in an output-stationary systolic array and two cycle reconfigurable multiply-accumulate (MAC) pipeline. Each MAC uses a unified datapath that can be reconfigured for signed/unsigned INT8/INT16/FP16 multiplications and wider INT24/INT48/FP32 accumulators to support differing performance, numerical range and precision requirements. Lower-precision INT8 mode operates with 4× higher throughput, reconfiguring each node from 1x1 to 2x2 matrix. The accelerator occupies 0.024mm^2 in 14nm CMOS, with dense matrix measurements at 750mV demonstrating energy efficiency ranging from 0.6TFLOPS/W in FP16 mode to 2.9TOPS/W in INT8 mode. Peak energy efficiency increases by a further 4× to 11.3TOPS/W at a near-threshold voltage of 280mV. When reconfigured for sparse matrices, the blocking and handshake-based routing circuits with multiple-node broadcast enable 8.8× higher energy-efficiency with 20% nonzero sparse matrices.