Approximate Computing Based Tensor Processing Unit for High-Performance AI Compute

Mohammed E. Elbtity
iCAS Lab, University of South Carolina
Introduction

• Data movements between memories and CPU/GPU are expensive.

• Neural networks (NN) algorithms can be highly parallelized.

• Systolic arrays (or TPUs) can be considered a promising solution to accelerate the NN algorithms.

• NN models are tolerant to quantization and approximation.

• We can utilize the area, power, and performance of approximate PEs while maintain comparable accuracy.
TPU with Approximate MAC

Controller

Input Memory (IFMAP) -> FIFO -> DEMUX -> PE -> PE -> PE

Weight Memory

DEMUX

Output Memory (OFMAP)
Results (Accuracy)

- Comparable accuracy can be achieved!
Results (Area)

- The average Area reduction is 55.3%.
Results (Power)

- The average Power reduction is 41.3%.
Results (Timing)

- 5.2x TOPS/mm² improvement.

- 4.4x TOPS/W improvement.