**ABSTRACT**

Artificial neural networks (ANNs) are a pivotal component of highly successful modern Artificial Intelligence (AI) applications, such as OpenAI's ChatGPT. FPGA’s inherent parallelism and efficiency provide an opportunity to provide quicker computation of ANNs than CPUs or GPUs, while consuming much less power, both in data center and edge computing environments. A custom instruction component for the Nios II soft processor and supporting software were developed to allow easy, configurable generation and deployment of high-performance ANN inference on FPGA devices. The custom instruction performs integer matrix-vector multiplications, with each column of the product computed in parallel with pipelined multiply accumulate units. Nios II embedded computer systems were created both with and without the generated custom instruction. FPGA device resource utilization, power consumption, and execution time were evaluated for each system when performing ANN-based image classification on the MNIST handwritten digit database. Speedups of over 300x were achieved when using the custom instruction, at the expense of significantly higher FPGA resource utilization and a minimal increase in power consumption.