Optimizing Two-Dimensional Convolution Accelerators for Area, Energy, and Flexibility

David Paz[1], Andy Wright[2]
Advisor - Prof. Arvind[3]
[1]Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA
[2, 3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA

Two-dimensional convolution calculations have direct applications in image processing, filtering and pattern detection. Highly optimized hardware accelerators are capable of increasing performance on very specific computational tasks by large factors over sequential software applications. These accelerators are often optimized for instructions per clock cycle. However, one important aspect of hardware accelerator design that is often overlooked relates to application flexibility. For instance, Convolution Neural Network (CNN) accelerators are designed with a predefined number of accelerator cores and application specific tasks that may not be modified dynamically and can be expensive.

This study aims to develop energy efficient and flexible two-dimensional convolutional accelerators ideal for IoT and smaller devices to provide significant performance gains over sequential computations and flexibility over application-specific accelerators such as CNN accelerators. Bluespec System Verilog (BSV) has been used to develop the accelerator by implementing pipelined Full Binary Tree structures. These structures have been fully incorporated into a three-stage pipelined RISC-V processor, and they are capable of interacting directly with memory and exploit temporal locality to maximize performance with minimal energy and area cost, while providing users with fine-grain control of their kernel and matrices.

MSRP GENERAL