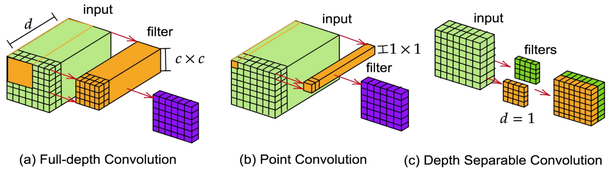
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
R&D SHOWCASE 2021
Technology, Social Impact

An FPGA Overlay for CNN Inference with Fine-Grained Flexible Parallelism

Abstract



(a) Full-depth Convolution (b) Point Convolution (c) Depth Separable Convolution

- A CNN is predominantly formed of the convolution operation, wherein a given input volume is convolved with a set of filters to generate an output volume.
- Accelerating CNN inferencing on FPGAs is extremely crucial to meet performance requirements.
- In this work, we propose an FPGA Overlay for CNN that can accelerate different types of convolutions.

Overview

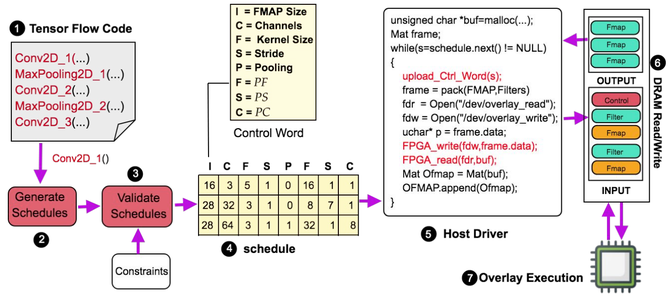


Figure 1: An overview of our framework. The CPU host processes the tensor flow specification and controls the CNN over on the FPGA through a set of control words.

Hardware Architecture

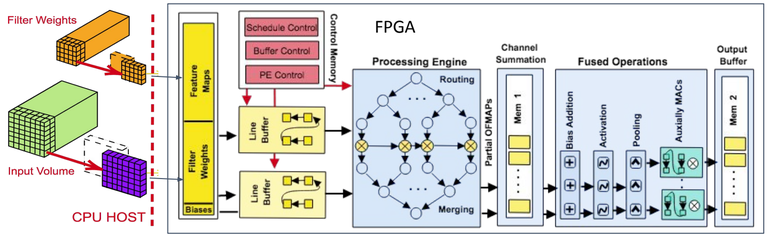
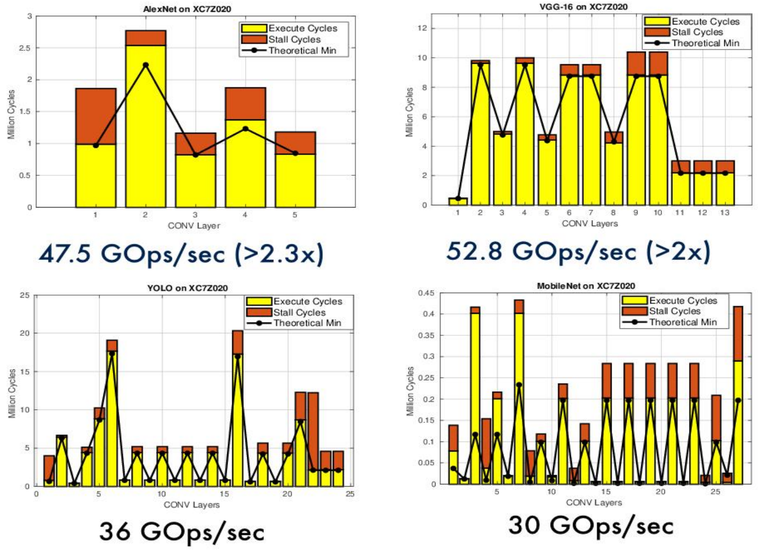


Figure 2: Micro-architecture of our overlay. The host streams filter weights and feature maps to the FPGA. Configurations of our overlay is set by the control memory.

- The host processes a CNN layer-by-layer.
- Each layer is further divided into batches. This batching is done so as to maximize DSP utilization, data reuse and minimize number of FPGA invocations.
- Our overlay supports fine-grained, flexible parallelism wherein every layer can be processed with different degrees of parallelism.
- The overlay is soft-reconfigured using control words to process a batch with maximum parallelism over the overlay.
- Our architecture is fully pipelined and operates with fewer stalls, resulting in low-latency execution.
- We propose a set of constraints over the degrees of parallelism. The satisfaction of these constraints ensures that the compute-to-memory overlap inside our accelerator is maximized during the processing of a layer.

Evaluation

- We studied the effectiveness of our hardware by accelerating AlexNet, VGG16, YOLO and MobileNet CNNs targeting a Zynq FPGA.
- The chosen nets have a mix of different types of convolution layers and filter sizes, presenting a good variation in model size and structure.



Authors: Ziaul Choudhury, Shashwat Shrivastava and Suresh Purini

Research Center Name: Computer Systems Group