Sharc Lab
Prof. Callie Hao

Exploiting HBM for CNN Acceleration on FPGAs

Multiple Kernel Execution using PYNQ-for-XRT

Akshay Kamath, Callie Hao, Sharc Lab, School of Electrical and Computer Engineering, Georgia Tech

Georgia Tech | Center for Research into Novel Computing Hierarchies

Georgia Tech | School of Electrical and Computer Engineering
College of Engineering

Motivation

FPGAs are ideal for energy-efficient CNN acceleration.

Accelerating large-scale models is highly challenging.

- E.g. ResNet-50 with HD (720 x 1280) inputs
- Off-chip memory access is the primary bottleneck!

Typical approach to acceleration is via tiling of inputs.

- Intra-tile parallelization limited by on-chip buffers
- Inter-tile parallelization limited by off-chip memory

HBM-based FPGAs have independent memory banks.

- Facilitates multiple tile-kernel execution in parallel!

PYNQ-for-XRT framework allows quick prototyping.

- Python-based host programming over OpenCL

Tile Parallelization

Intra-tile parallelization involves optimizing convolution computation.

Inter-tile parallelization is achieved through multiple kernel instantiation.

- Each kernel is connected to different HBM memory banks

Host program pseudo-code

```
for each set of N tiles:
    Sync input buffers from host to device memory
    Start execution of all kernels in parallel
    Wait for all kernels to finish execution
    Sync output buffers from device memory to host
```

Kernel (device) pseudo-code

```
Load input feature map tile from off-chip memory
Load weights and bias from off-chip memory
Perform convolution on-chip w/wo ReLU
Store output feature map tile to off-chip memory
```

Experimental Results

Demonstrated on CRNCH Lab's Alveo U50 FPGA

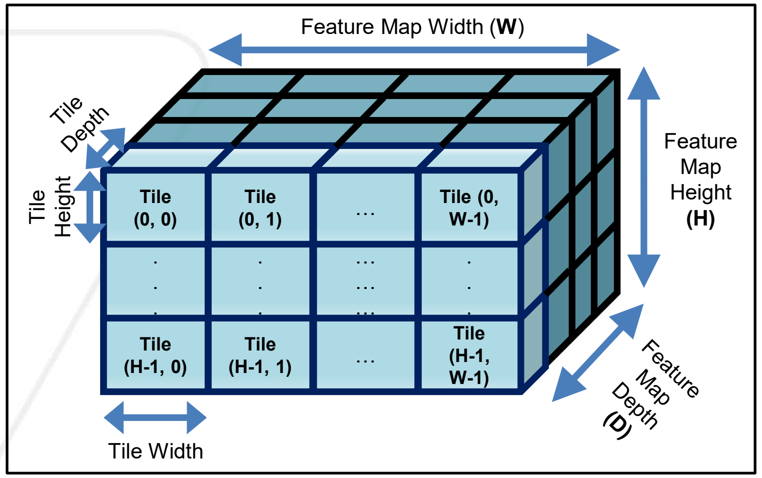
- Includes x86 CPU connected to 32 HBM banks
- Kernel developed using Xilinx Vitis HLS

Setup for a sample ResNet50 3x3 convolution layer with BDD100K dataset	
IO Tile Dimensions	64 * 23 * 20
Weight Blocks Dimensions	64 * 64 * 3 * 3
Fixed-point Precision	32 bits
No. of Kernel Instances	4

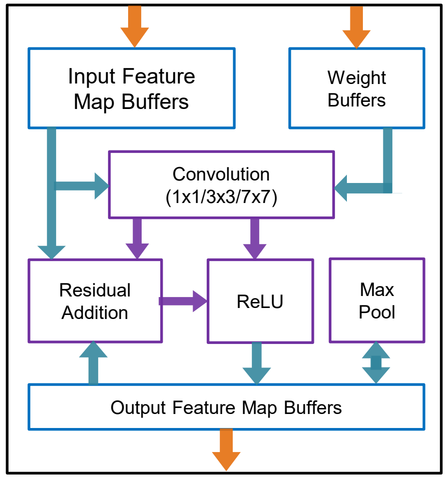
Configuration	Avg. Latency (ms)
Single kernel	5.959
4 kernels (sequential)	23.935
4 kernels (parallel)	6.857

Parallel execution, thus, results in **3.5x** speedup!

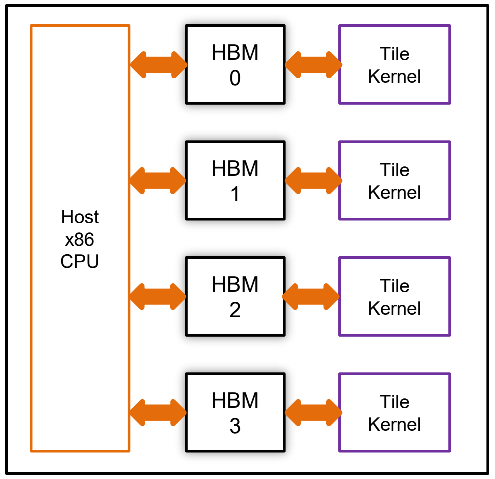
- Less than ideal 4x speedup due to overhead of synchronization of data for 4 kernels



Tiling-based Convolution for Acceleration of CNN Inference on FPGA


Input Feature Map Buffers
Weight Buffers
Convolution (1x1/3x3/7x7)
Residual Addition
ReLU
Max Pool
Output Feature Map Buffers

HLS Convolution Tile Kernel Architecture



System-architecture for multiple kernel execution

Georgia Tech