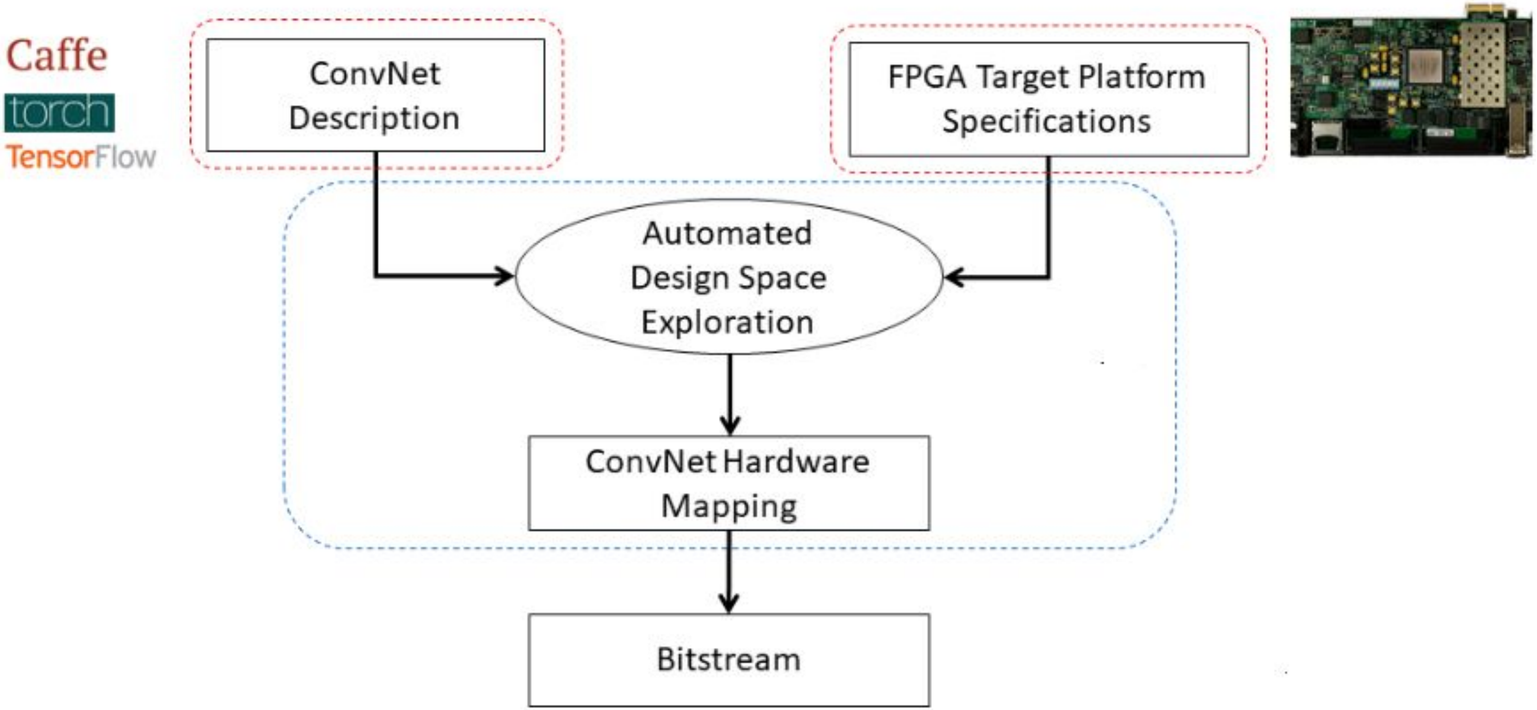
Toolflows for Mapping Convolutional Neural Networks on FPGAs

Aayush Bansal
Anshul Bansal

What does a toolflow do?
→ It maps the NN model to the hardware architecture



ConvNet Description

FPGA Target Platform Specifications

Automated Design Space Exploration

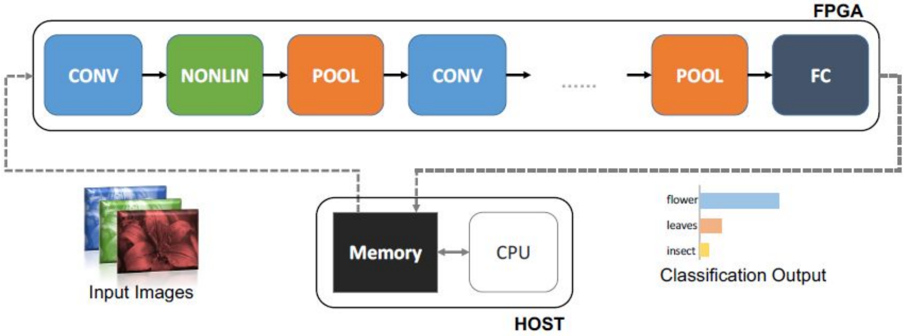
ConvNet Hardware Mapping

Bitstream

Hardware Architectures

1. **Streaming Architecture:**

- One h/w block per layer
- Blocks are chained to form a pipeline
- Increased efficiency due to pipelining

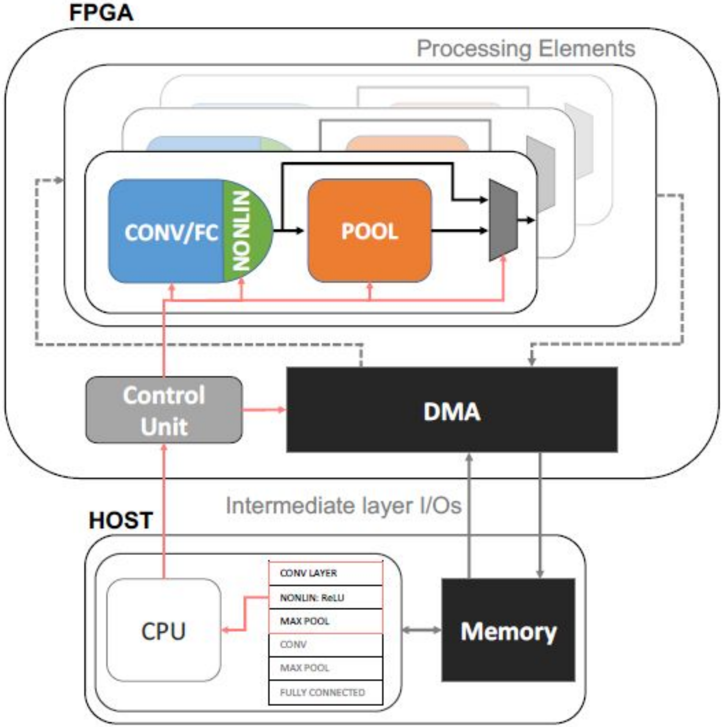


→ fpgaConvNet, DeepBurning, Haddoc2, AutoCodeGen, FINN

Hardware Architectures

2. **Single Computation Engine :**

- Executes the layers sequentially
- Same bitstream can target many CNNs



→ AngelEye, Alamo, DnnWeaver, Caffine, FP-DNN, Snowflake, SysArrayCell, FFTCodeGen

- Some of these toolflows are not compatible with Intel FPGAs, therefore we will not talk about them
  → Caffine, AngelEye, fpgaConvNet, AutoCodeGen, DeepBurning, FINN,

- Haddoc2 requires the weights to be stored on-chip. Also it does not support partial unrolling

- Toolflows under consideration:
  → FP-DNN, FFTCodeGen, DNNWeaver, ALAMO

FP-DNN

Architecture : Single Computation Engine

Interface : Tensorflow

NN Models : CNN, RNN, DNN

Devices : Intel Standalone

Precision : FXP and FP

Design Space Exploration: Algorithmic

- Consist of a Matrix Multiplication (MM) Engine
- Usage of double buffers
- Reuses FPGA resources across layers
- 16-bit FXP representation
- Can target large scale CNNs

FFTCodeGen

Architecture : Single Computation Engine

Interface : Proprietary

NN Models : CNN, DNN

Devices : Intel HARP

Precision : FXP and FP

Design Space Exploration: Roofline and Analytical Model

- Target the Intel HARP architecture
- Partitions the workload between FPGA and CPU
- Performs convolutions in frequency domain
- Optimised for high throughput applications
- Uniform quantization and scaling across all layers
- Outperforms every toolflow present

DNNWeaver

Architecture : Single Computation Engine

Interface : Caffe

NN Models : CNN, DNN

Devices : Intel and Xilinx

Precision : FXP(Dynamic)

Design Space Exploration: Custom Search Algorithm

- High degree of portability
- Based on parameterised architectural template
- Consist of PU with each PU having an array of PE
- Input CNN is mapped to dataflow based representation
- Focus on throughput and employs batch processing
- Support dynamic quantization

ALAMO

Architecture : Single Computation Engine

Interface : Caffe

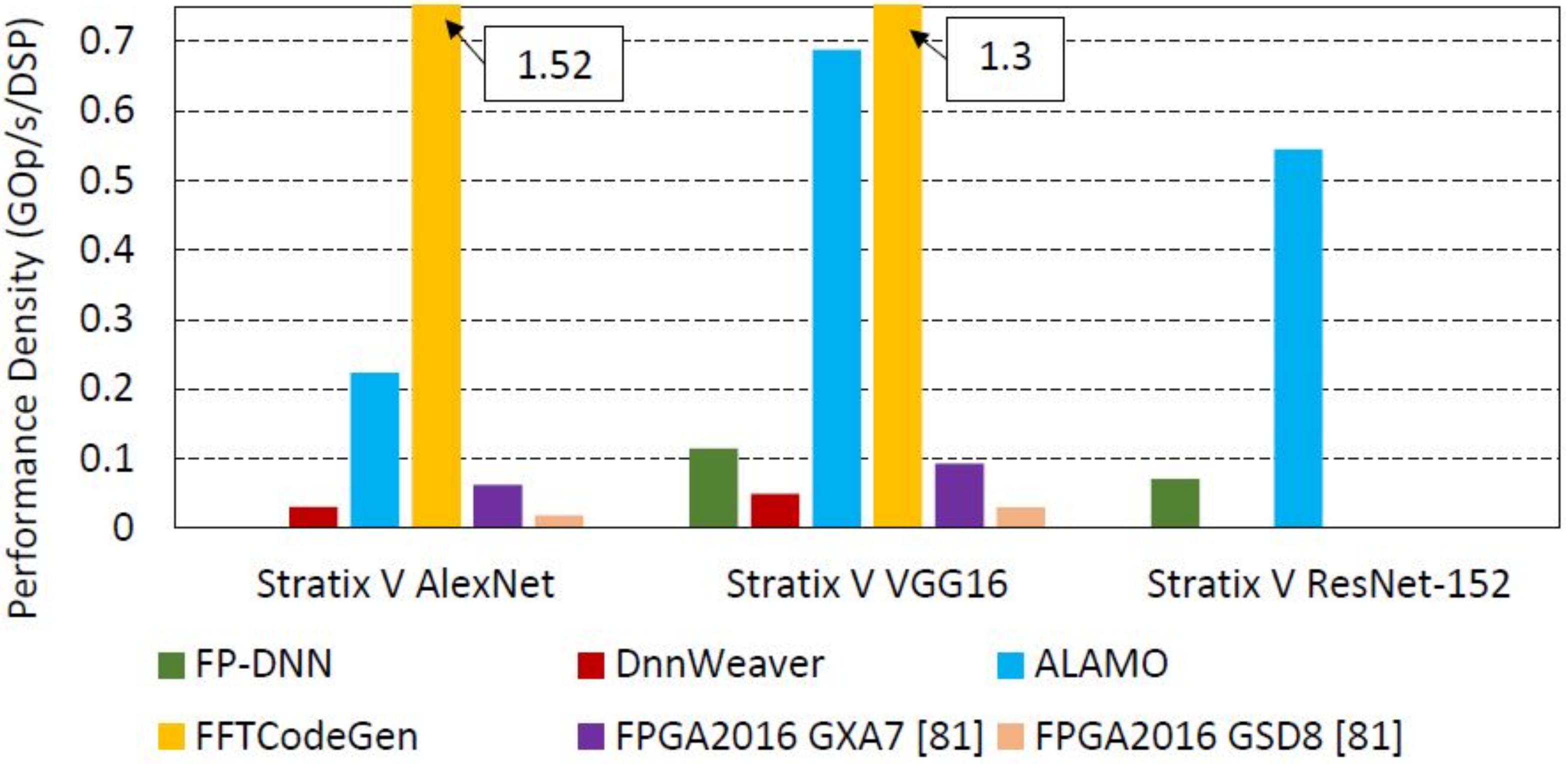
NN Models : CNN, DNN

Devices : Intel SoC and Standalone

Precision : FXP(Dynamic)

Design Space Exploration: Heuristic

- Support Intel Standalone and SoC platform
- Layers are scheduled sequentially
- Instantiates only the required h/w blocks
- The compiler determines the unroll factor
- Batch size is 1, throughput and latency is co-optimised
- Designed to combine high throughput and low-latency applications
- Support Dynamic quantization


Performance Density (GOp/s/DSP)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.52
1.3
Stratix V AlexNet
Stratix V VGG16
Stratix V ResNet-152
FP-DNN
DnnWeaver
ALAMO
FFTCodeGen
FPGA2016 GXA7 [81]
FPGA2016 GSD8 [81]