embedded VISION SUMMIT 2018

Exploiting Reduced Precision for Machine Learning on FPGAs

XILINX

Kees Vissers

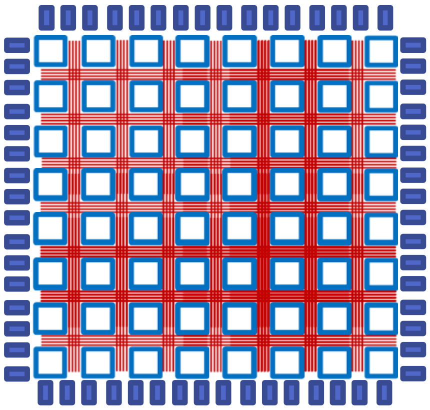
May 22, 2018

Introduction

Xilinx

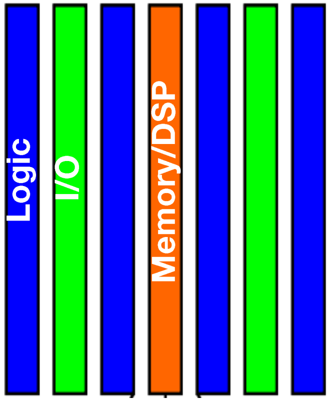
- What is an FPGA
- Zynq MPSoC products for the embedded vision market
- Neural Networks on Zynq MPSoC products
- Benefits of Reduced Precision Neural Networks on Zynq MPSoCs
- Programming environments

FPGA technology over time



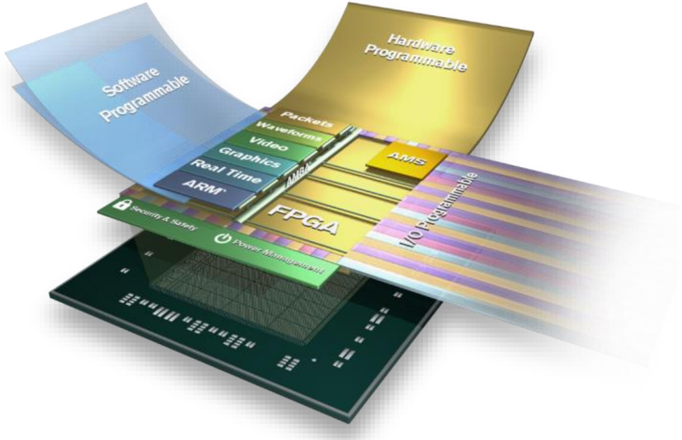
LUTs

Basic bit-oriented logic
4-input, 6-input Lookup table



Columns

Basic bit-oriented logic +
Word-oriented Multiply-accumulate
Word-oriented Memory



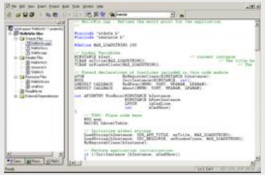
Processors + FPGA

Complete Processor systems
Dedicated programming environments
Heterogeneous MPSoC

Your program becomes a configuration that sets table values and switches via synthesis, Place and Route tools.

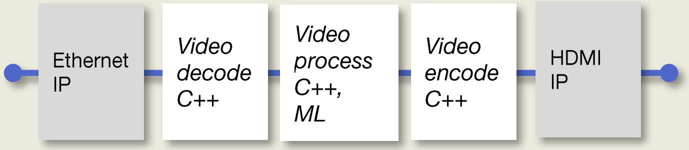
FPGA programming: dataflow and memory model

SDx Environments



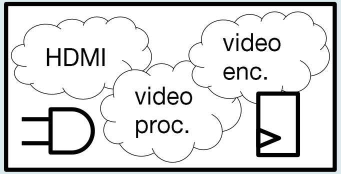
SW Programmability, Host code with Accelerator code, OpenCL, C/C++

HLx Editions



High Level Synthesis (HLS), C/C++/OpenCL with Vivado IPI

VIVADO



Traditional HW design, Verilog or VHDL

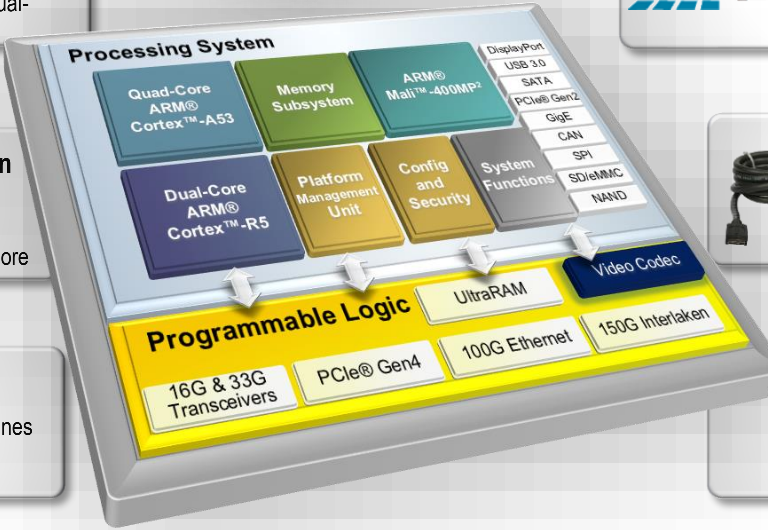
Zynq Ultrascale+ MPSOC

Memory Subsystem
High Bandwidth
Low Latency

ARM Cortex R5 **Real-Time Processors**
32-bit Dual-Core

mali **Graphics Processor**
ARM Mali-400MP2

ARM Cortex A53 **Application Processor**
64-bit
Dual/Quad-Core



High Speed Peripherals
Key Interfaces

Fabric Acceleration
Customizable Engines
High Speed Connectivity

H.265 HEVC High Efficiency Video Coding
H.264 MPEG-4/AVC
Video Codec
8K4K (15fps)
4K2K (60fps)

Platform & Power Management
Granular Power Control
Functional Safety

Configuration & Security Unit
Anti-Tamper & Trust
Industry Standards

Zynq® UltraScale+™ MPSoC: ZU7EV

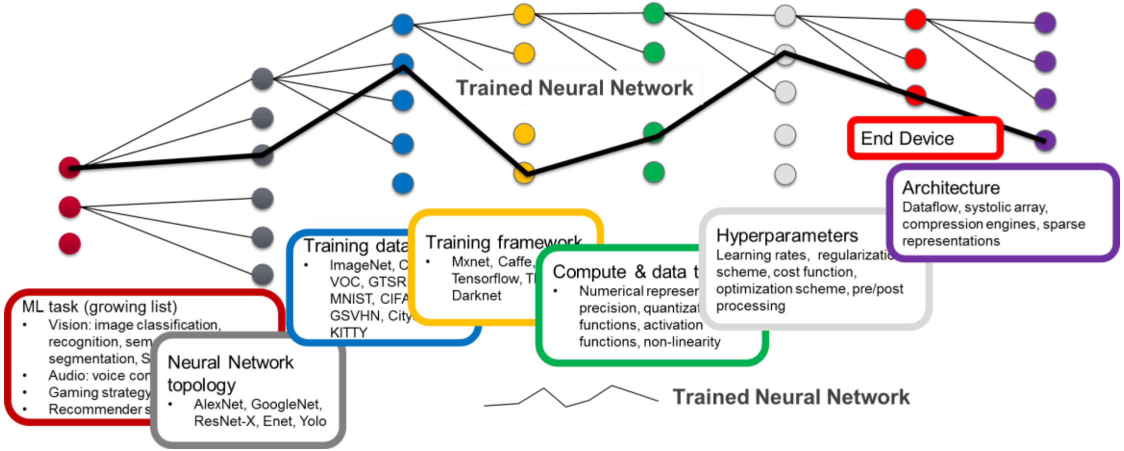
- **Quad-core** ARM® Cortex™-A53 MPCore™ up to 1.5GHz
- **230 K LUTs**
- **11 Mb Block Ram**
- **27 Mb Ultra Ram**
- **1728 DSP slices**
- **Dedicated Video Encoder – Decoder hard block**
- **2 PCIexpress hard blocks (Gen3 x 16/ Gen4 x 8)**

Neural Networks

Neural Networks

- Exploit the Programmable Logic part of the MPSoC
- **Reduced Precision** is showing great promise
- The **trade-off** between precision and accuracy or error rate is essential.
- This presentation will show you the **pareto optimal** solutions!

Multi-dimensional exploration space



Architecture
Dataflow, systolic array, compression engines, sparse representations

ML task (growing list)
- Vision: image classification, recognition, sem segmentation, S
- Audio: voice con
- Gaming strategy
- Recommender s

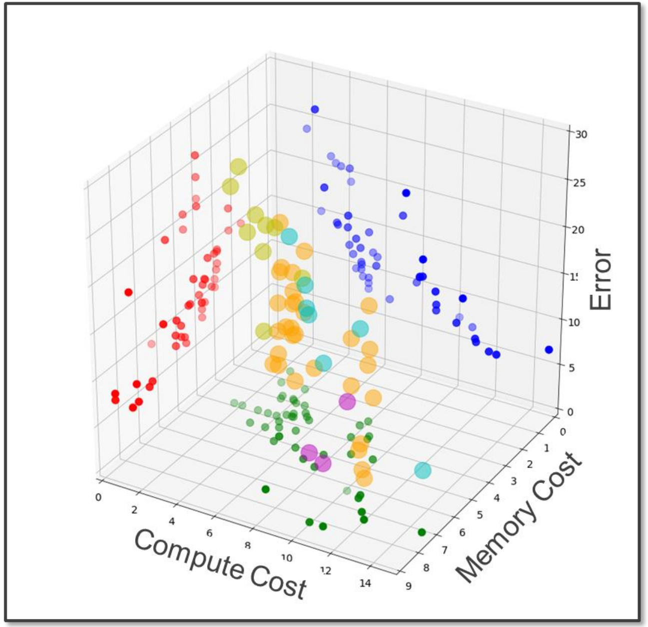
Neural Network topology
- AlexNet, GoogleNet, ResNet-X, Enet, Yolo

Training data
- ImageNet, C VOC, GTSR MNIST, CIFA GSVHN, City KITTY

Training framework
- Mxnet, Caffe, Tensorflow, T Darknet

Compute & data t
- Numerical represer precision, quantizat functions, activation functions, non-linearity

Hyperparameters
Learning rates, regularizatio scheme, cost function, optimization scheme, pre/post processing

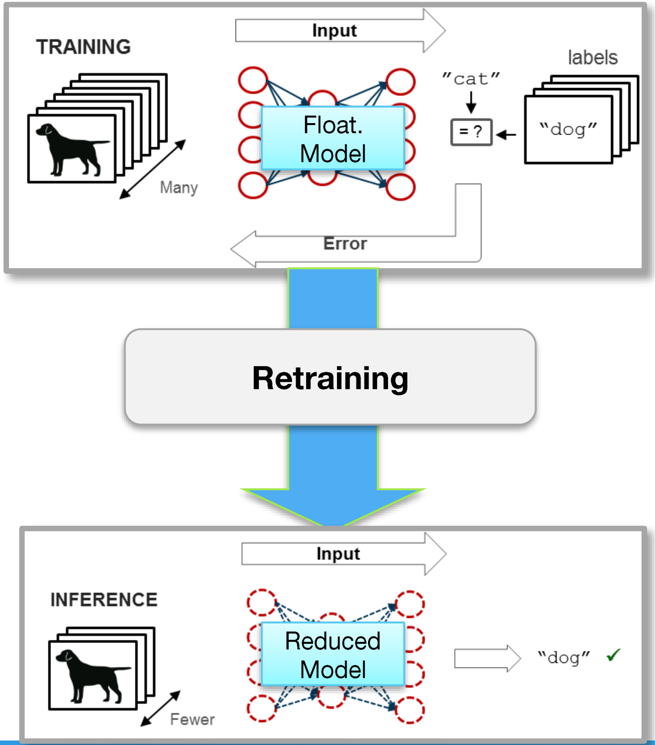


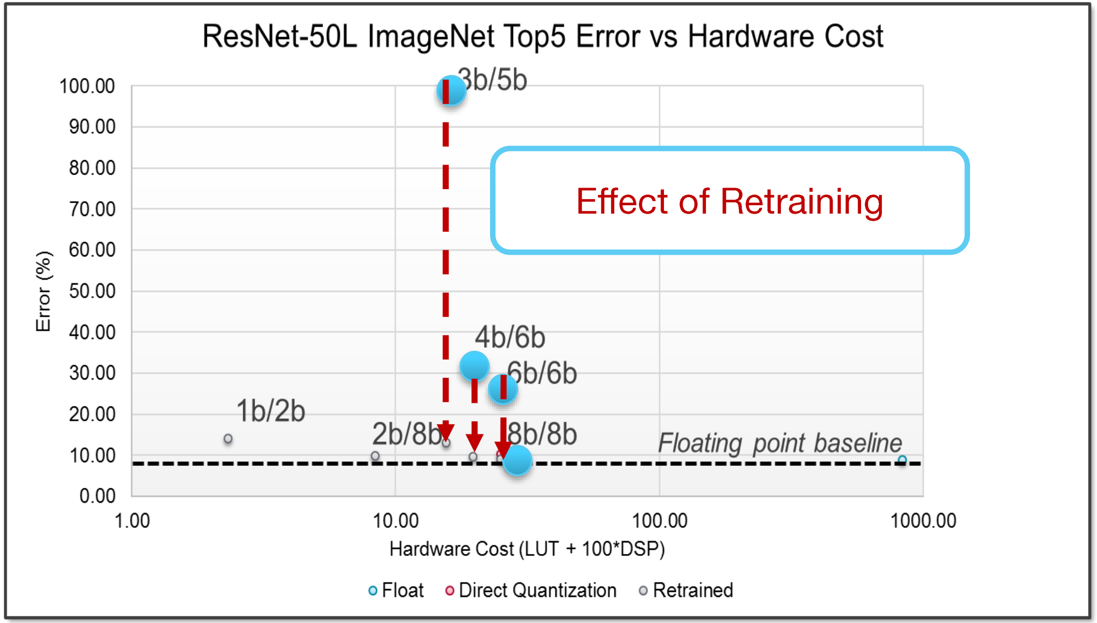
ImageNet Classification: Error, compute cost, memory requirements, topology

Each combination yields to a different point in the multi-dimensional design space:
error, cost, throughput, latency, power

Data analysis required to understand the compromises and find optimal solutions

Training environment used for Reduced Precision





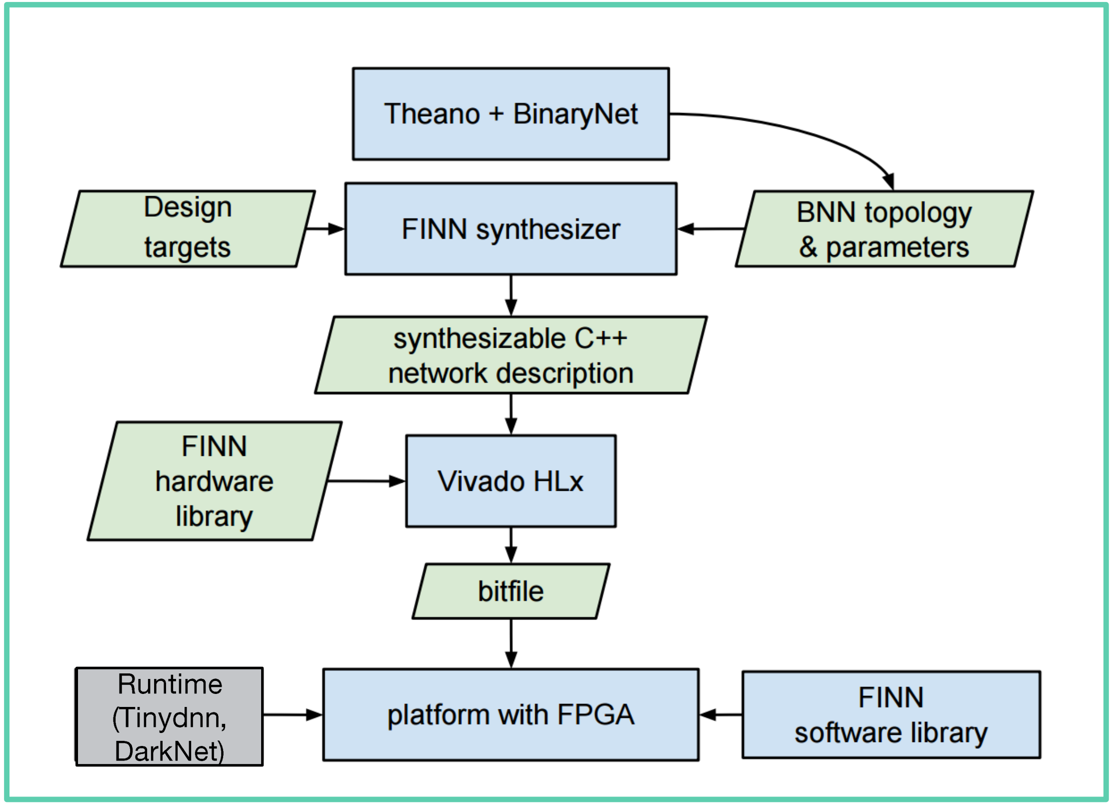
Notation: 3b/5b: 3 bit weights/ 5 bit activation

- **<8bit: retraining**

Finn research framework used for mapping on ZU7EV

theano TensorFlow™ Caffe

- All code in C/C++
- Hardware Library is all HLS code
- Can execute on CPU and FPGA - No RTL needed



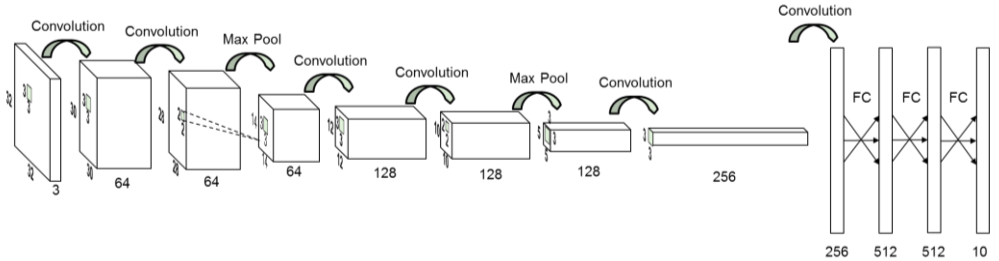
Study of accuracy versus cost

- Several Networks were studied in detail, including **retraining**.
  - CNV on Cifar10
  - Resnet50 on IMAGENET

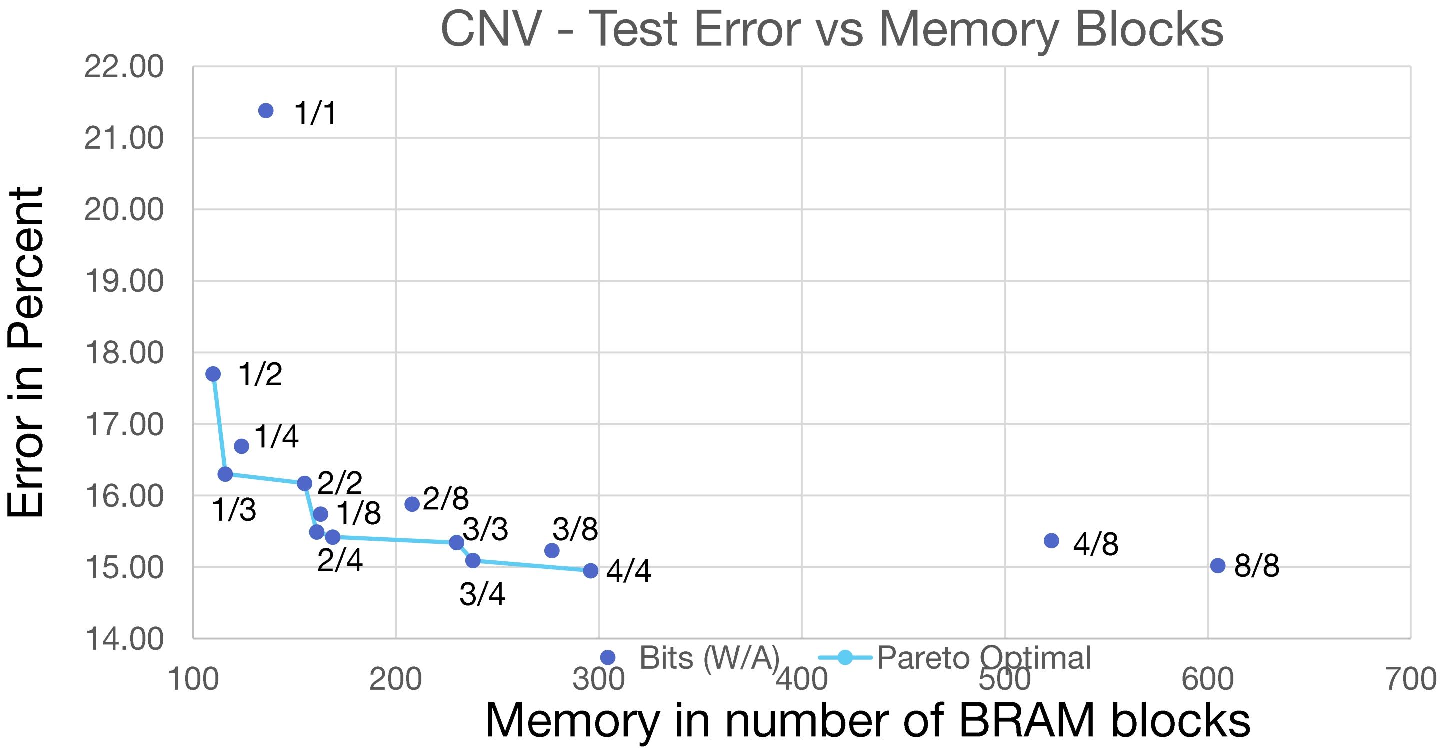
CNV on CIFAR-10

CNV on CIFAR-10

- Topology;
  - Number of layers: 2 (3x3) Conv + Max Pool +
    2 (3x3) Conv + Max Pool + 2 Convolutional + 3 FC
  - Compute requirement: 112.5 MOPS/Frame
  - # Parameters: 1.54 M
- Parallelism adapted for BRAMs usage at 8 bits
  - Fixed parallelism for each precision
  - Estimated performances:
    - 385 FPS @ 200 MHz
    - Latency: 13 ms
    - Throughput: 43 GOPS

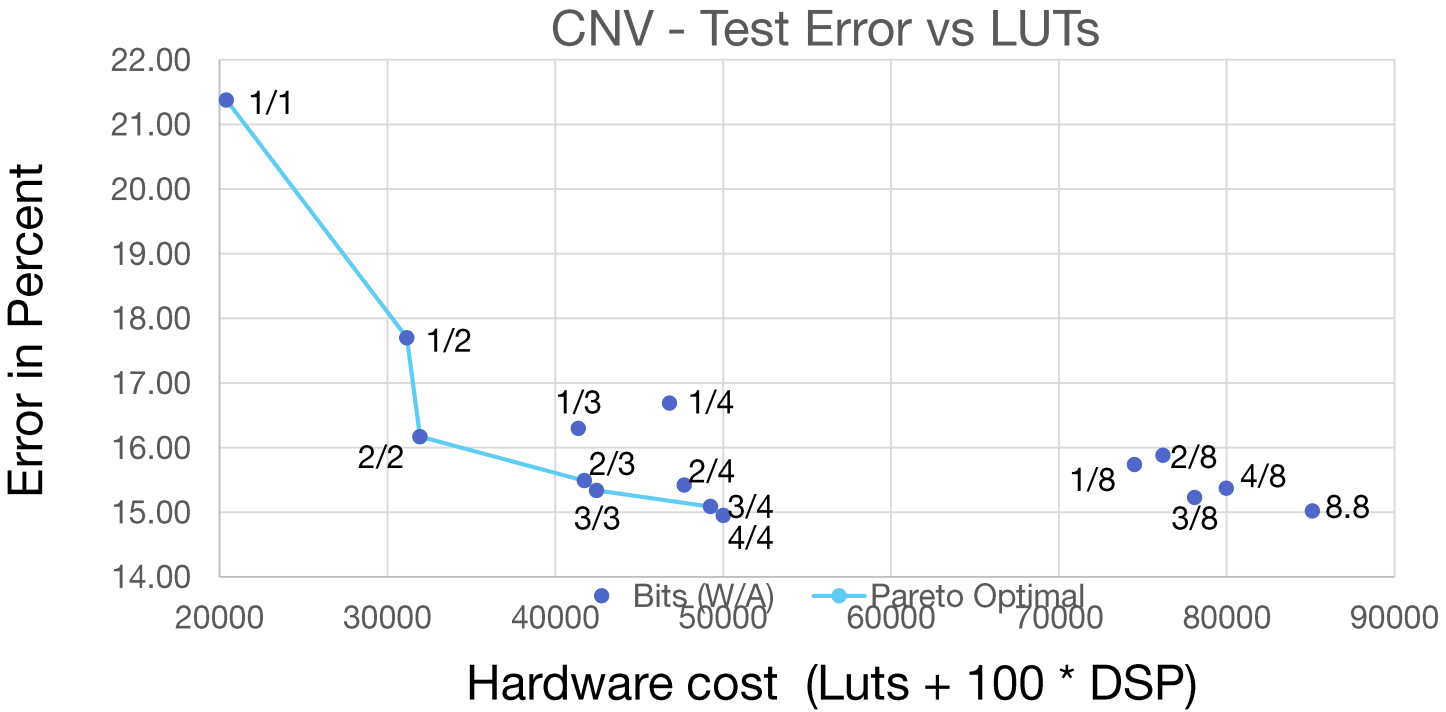


CNV on CIFAR-10 – Error vs Memory Blocks Used



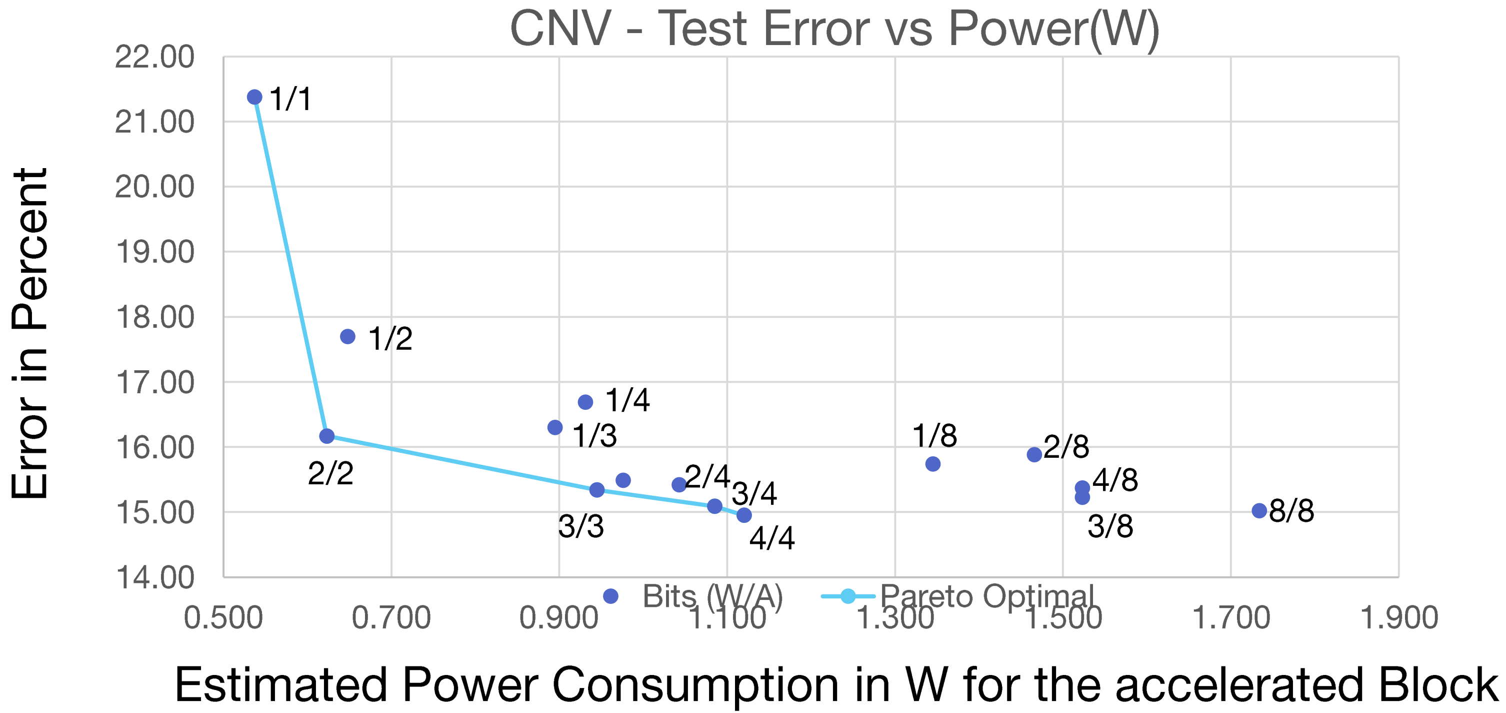
*Target Device ZU7EV ● Vivado 2017.3 tool suite ● 200 MHz target frequency ● Post-placed utilization ● #Blocks considered as BRAM36 + 4*URAM*

CNV on CIFAR-10 – Error vs LUTs



Target Device ZU7EV ● Vivado 2017.3 tool suite ● Post-placed LUT utilization ● 200 MHz target frequency ● Flow_PerfOptimized_high strategy for synthesis ● Performance_ExtraTimingOpt strategy for implementation

CNV on CIFAR-10 – Error vs Estimated Power

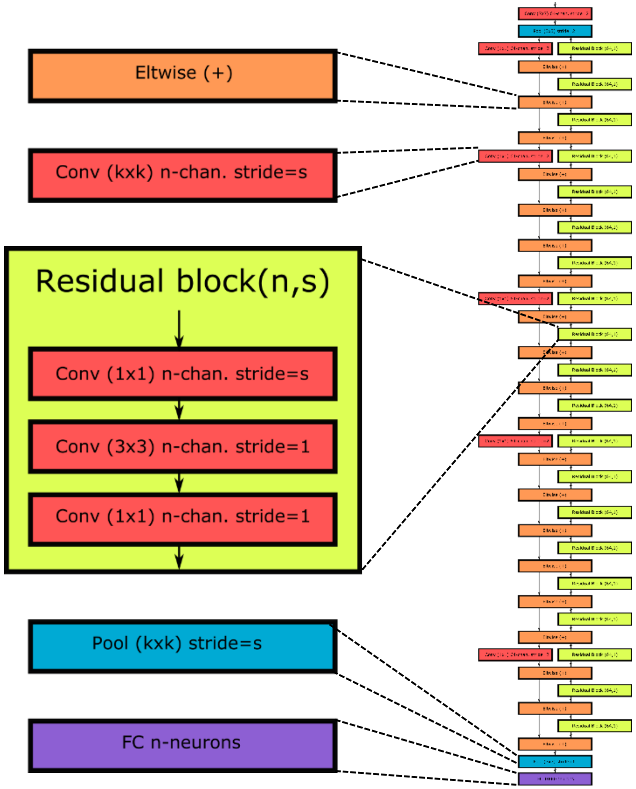


Target Device ZU7EV ● Ambient temperature: 25 °C ● 12.5% of toggle rate ● 0.5 of Static Probability ● Power reported for PL accelerated block only

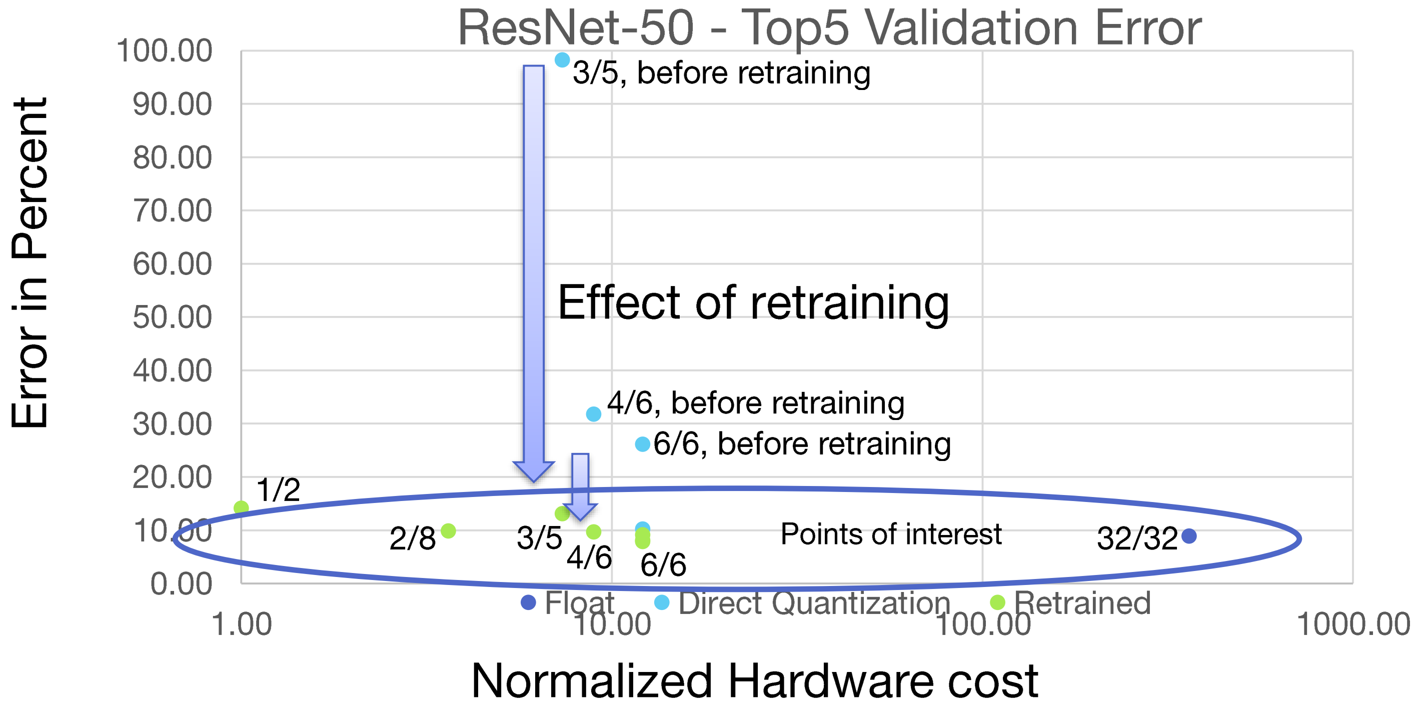
Resnet50 with ImageNet accuracy study

ResNet-50 on ImageNet

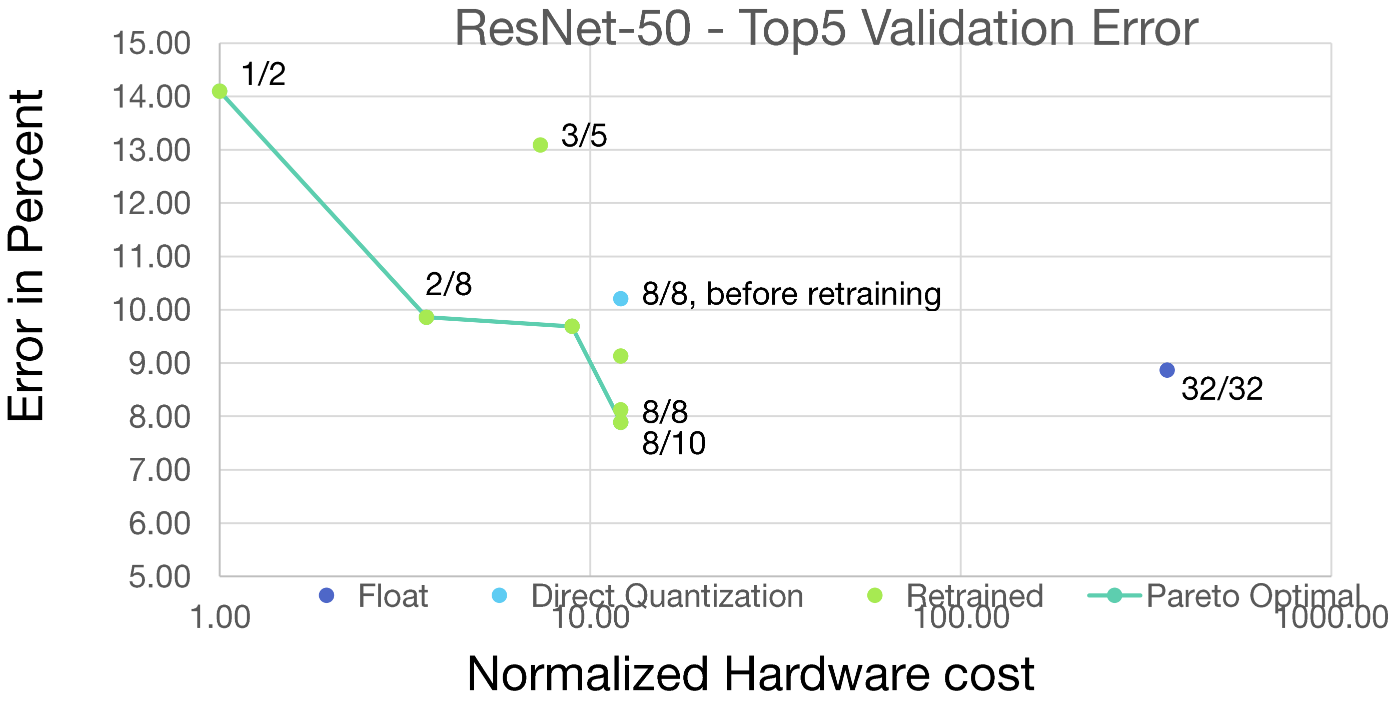
- Topology;
  - Number of layers: 53 Conv + 2 Pool + 1 FC
  - Compute requirement: 7.6 GOPS/Frame
  - # Parameters: 25.5 M
- Comparison between direct quantization and retraining
- Performance Model;
  - Combination of LUTs and DSPs used
  - Hardware cost:
    - a weighted sum of LUTs and DSPs required per operation (LUTs + 100*DSPs)


Eltwise (+)
Conv (kxk) n-chan. stride=s
Residual block(n,s)
Conv (1x1) n-chan. stride=s
Conv (3x3) n-chan. stride=1
Conv (1x1) n-chan. stride=1
Pool (kxk) stride=s
FC n-neurons

ResNet-50 on ImageNet – Error vs Hardware Cost



ResNet-50 on ImageNet – Error vs Hardware Cost



Conclusions

Reduced Precision Conclusions on FPGAs

- **Pareto optimal solutions** show best implementations for a certain error
- Precisions **well below 8-bit** are very promising, benefits are:
  - Lower power
  - Less Hardware
  - At acceptable error rates
- **Re-training** is essential to exploit reduced precision implementations
- Xilinx FPGAs are an excellent implementation platform for reduced precision Neural Networks

Resources

Xilinx products:

- www.xilinx.com
- https://www.xilinx.com/video/application/revision.html
- https://www.xilinx.com/products/design-tools/embedded-vision-zone.html

University support and open source:

- https://www.xilinx.com/support/university/.html
- http://www.pynq.io/home.html
- https://github.com/Xilinx
- Embedded Vision Alliance:
- https://www.embedded-vision.com/summit