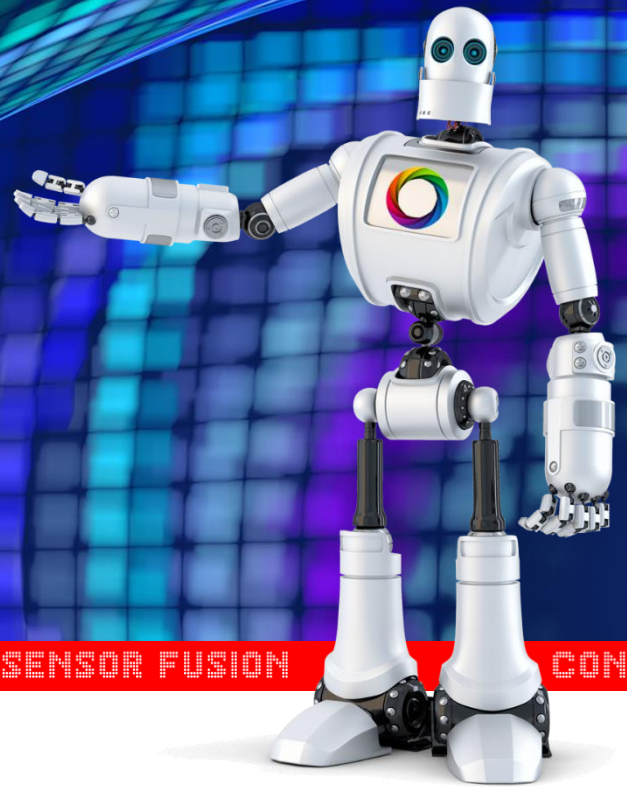
reVISION™

Responsive and Reconfigurable Vision Systems



MACHINE LEARNING | COMPUTER VISION | SENSOR FUSION | CONNECTIVITY

XILINX
ALL PROGRAMMABLE™

Caffe to Zynq: State-of-the-Art Machine Learning Inference Performance in Less Than 5 Watts

Vinod Kathail, Distinguished Engineer

May 24, 2017

XILINX
ALL PROGRAMMABLE™

Agenda

- **Why Zynq SoCs for Deep Learning Inference**
- **Caffe to Zynq SoC in Seconds**
- **A Full System Example**

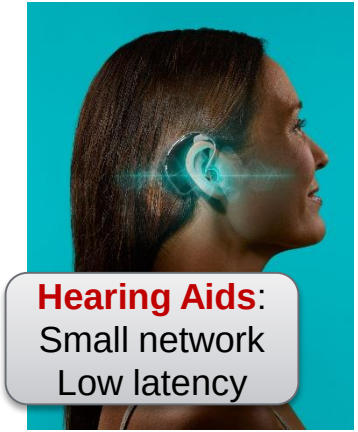
Diverse Applications with Diverse Design Targets

Translate & AlphaGo:
Huge networks

AlphaGo

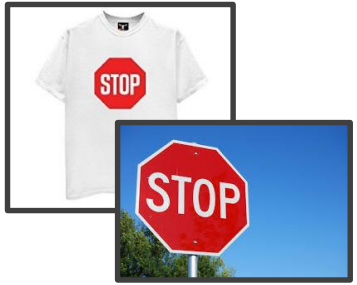
Medical Diagnosis:
Small networks

Robotics:
Real-time

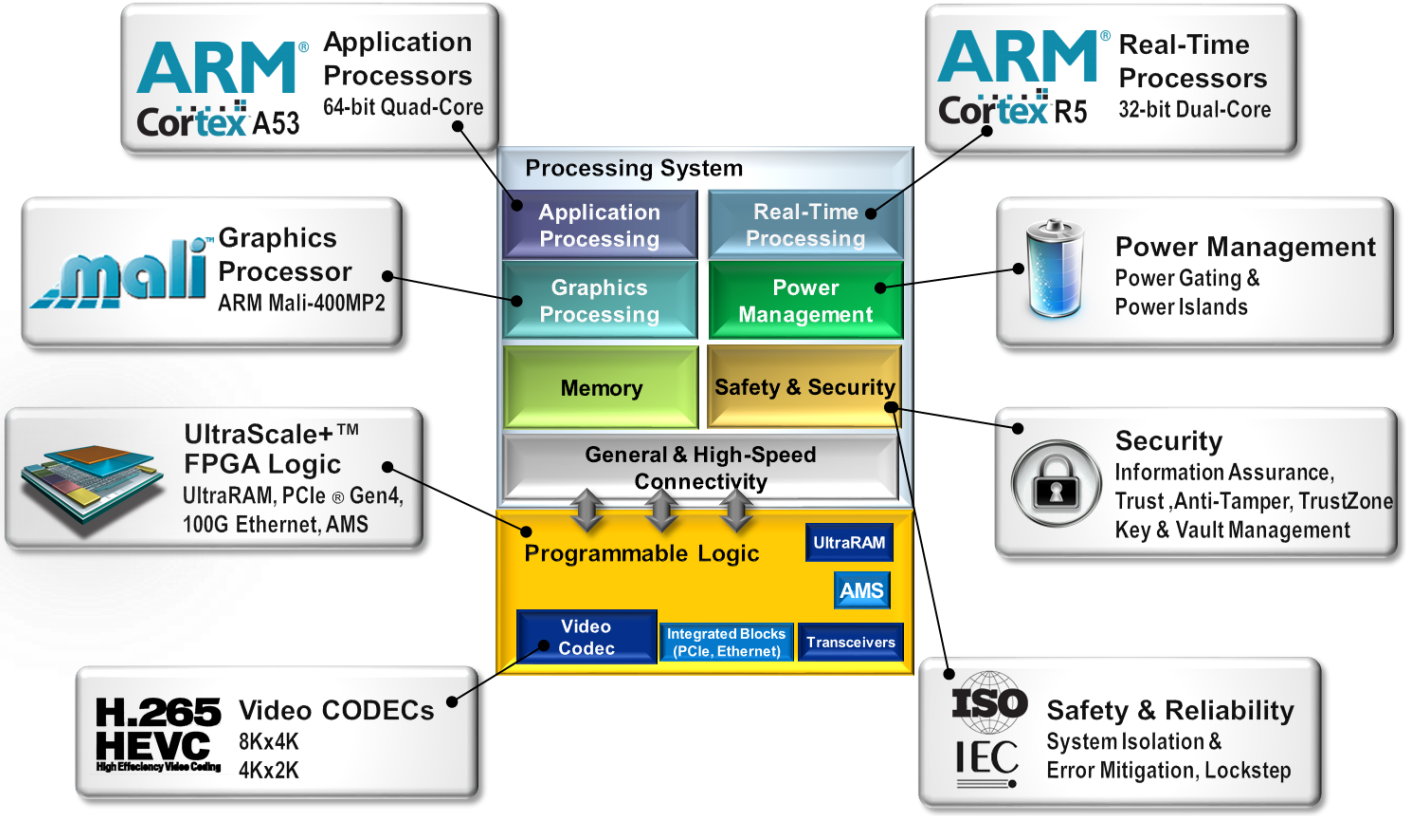


Hearing Aids:
Small network
Low latency

ADAS
High accuracy
Low latency



Zynq Offers the Most Efficient Deep Learning Inference



Zynq SoCs Offer Superior Throughput, Latency

6x
Images/sec/watt

Machine Learning
Inference

Xilinx Benchmark

1/5
Latency (ms)

Real Time Applications
Latency

Xilinx Benchmark

		Xilinx ZU9	Xilinx ZU5	eGPU*
GoogLeNet @ batch = 1	Images/s	**370.0**	**155.0**	70
	Power (W)	**7.0**	**4.5**	7.9
	Images/s/watt	**53.0**	**34.5**	8.9

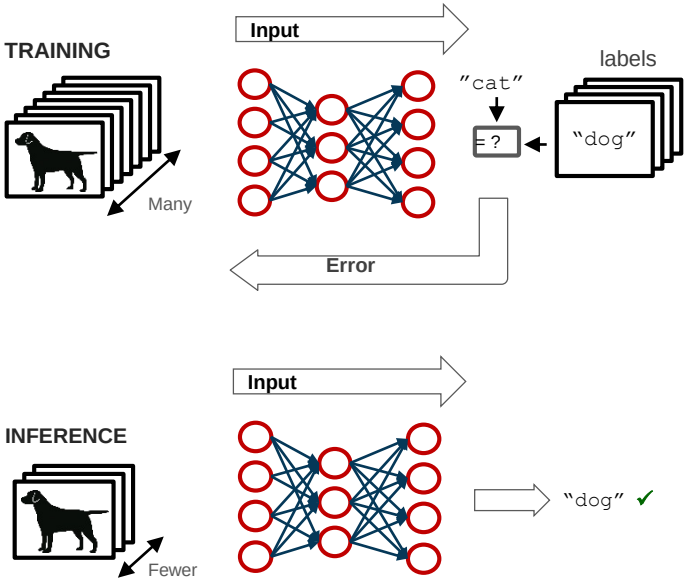
		Xilinx ZU9	Xilinx ZU5	eGPU*
GoogLeNet @ batch = 1	Images/s	**370.0**	155.0	70
	Latency (ms)	**2.7**	6.4	14.2

		Xilinx ZU9	Xilinx ZU5	eGPU*
GoogLeNet @ batch = 8	Images/s	**370.0**	155.0	163
	Latency (ms)	**2.7**	6.4	49.0

For large batch, CPU/GPU/DSPs latency increases significantly

- eGPU = nVidia TX1: nVidia GoogLeNet performance: https://devblogs.nvidia.com/parallelforall/jetpack-doubles-jetson-tx1-deep-learning-inference/

The Divergence of Training and Inference



Training: Process for machine to "learn" and optimize model from data

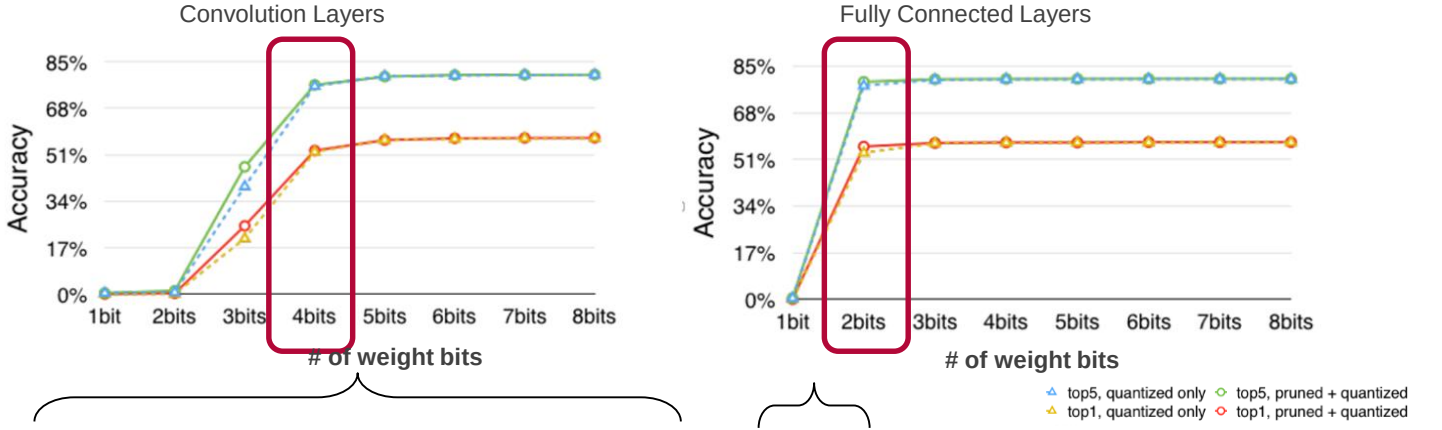
Inference: Using trained models to predict/estimate outcomes from new observations in efficient deployments

Inference now 8 bit and below for maximum efficiency

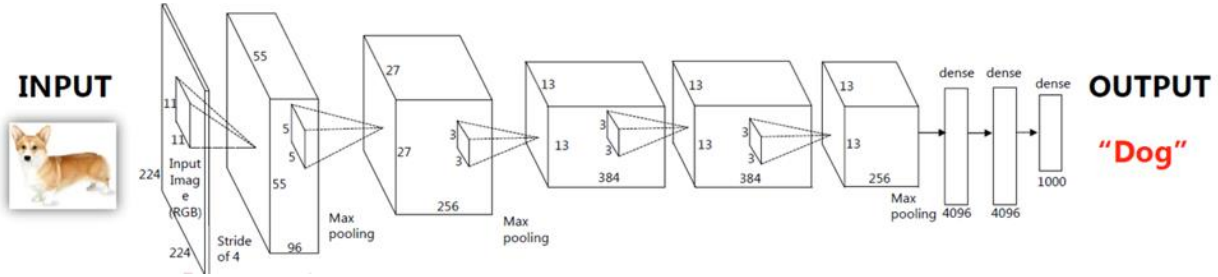
Top-5 Accuracy	FP-32	FIXED-16 (INT16)	FIXED-8 (INT8)	Difference vs FP32
VGG-16	86.6%	86.6%	86.4%	(0.2%)
GoogLeNet	88.6%	88.5%	85.7%	(2.9%)
SqueezeNet	81.4%	81.4%	80.3%	(1.1%)

https://arxiv.org/pdf/1510.00149v5.pdf

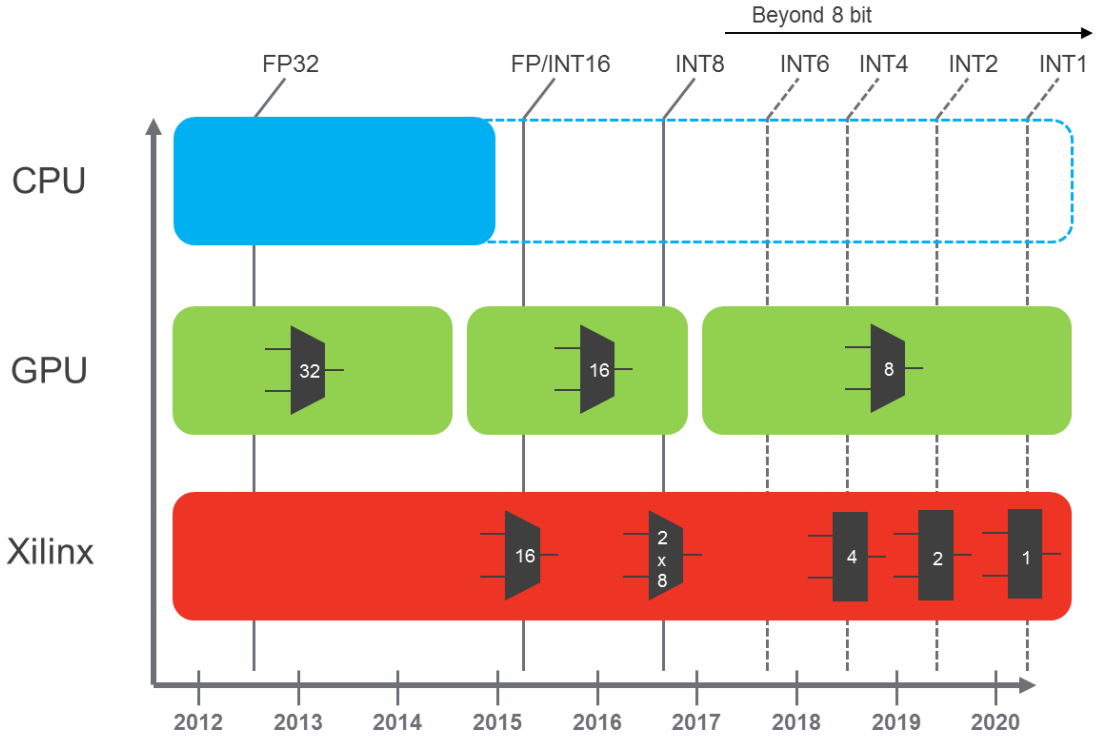
Inference Precisions Moving to Lower and Variable Precision



Citation: https://arxiv.org/pdf/1510.00149.pdf



Future Proof Architecture for Any Precisions



Limited to 32 bit operations
SIMD operations at INT8/16

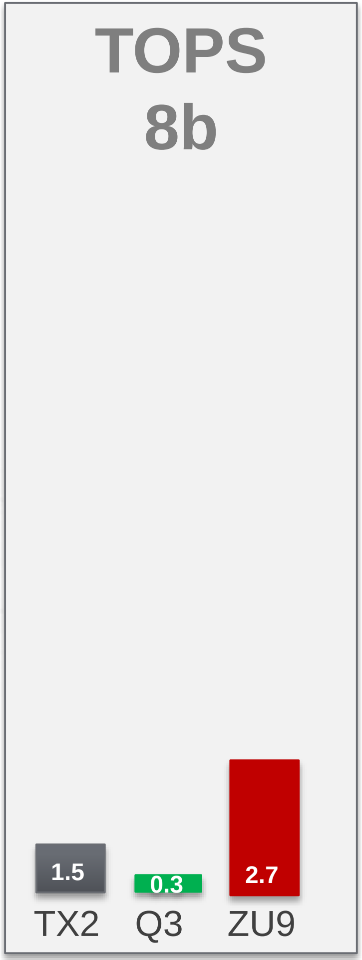
New devices required to
support change in precision efficiently

Reconfigurable to scale and
optimize for different precisions

BNN: Unparalleled Performance

- **Reducing precision from 8b to 1b shrinks LUT cost by 40x**
- **Potential to scale CNN performance to above 23TOPS (ZU9)**

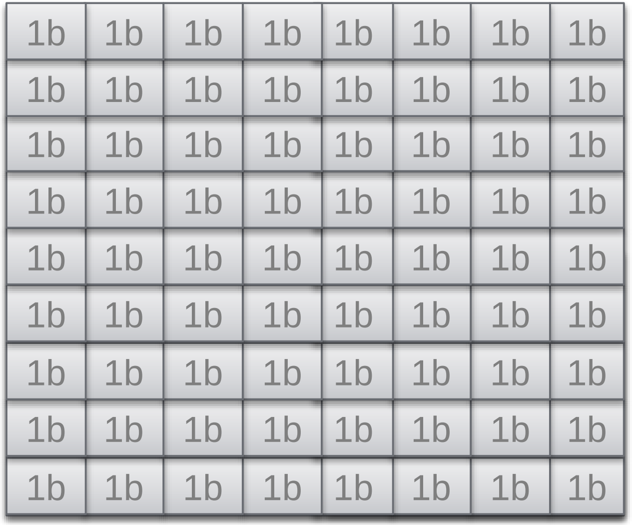
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b

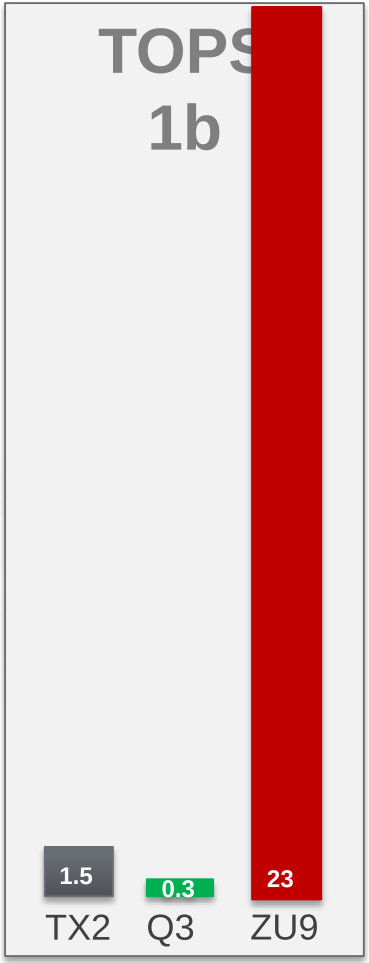


- Assuming 300 MHz with 90%/70% DSP/LUT utilizations
- Resource consumption assumption: 2.5 LUTs/op (INT1), 16 LUTs/op (INT4), 0.25 DSP/op (INT8)

BNN: Unparalleled Performance

- Reducing precision from 8b to 1b shrinks LUT cost by 40x
- Potential to scale CNN performance to above **23TOPS (ZU9)**



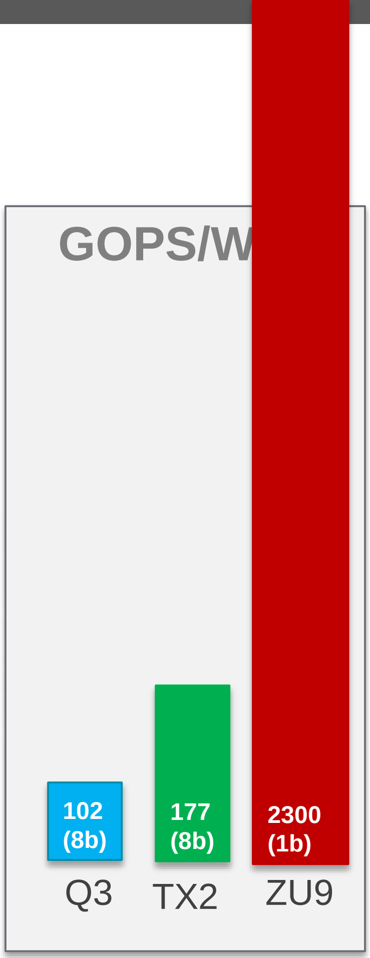


- Assuming 300 MHz with 90%/70% DSP/LUT utilizations
- Resource consumption assumption: 2.5 LUTs/op (INT1), 16 LUTs/op (INT4), 0.25 DSP/op (INT8)

BNN: Unparalleled Performance

- **Reducing precision from 8b to 1b shrinks LUT cost by 40x**
- **Potential to scale CNN performance to above 23TOPS (ZU9)**

1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b
1b	1b	1b	1b	1b	1b	1b	1b

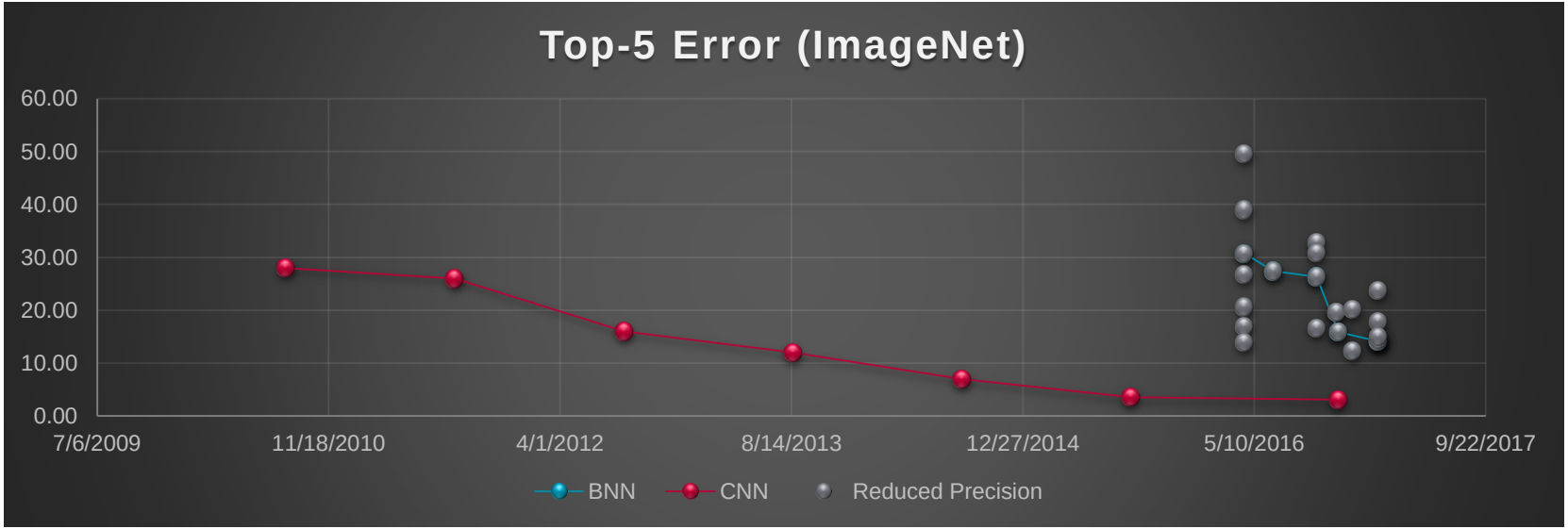


Embedded

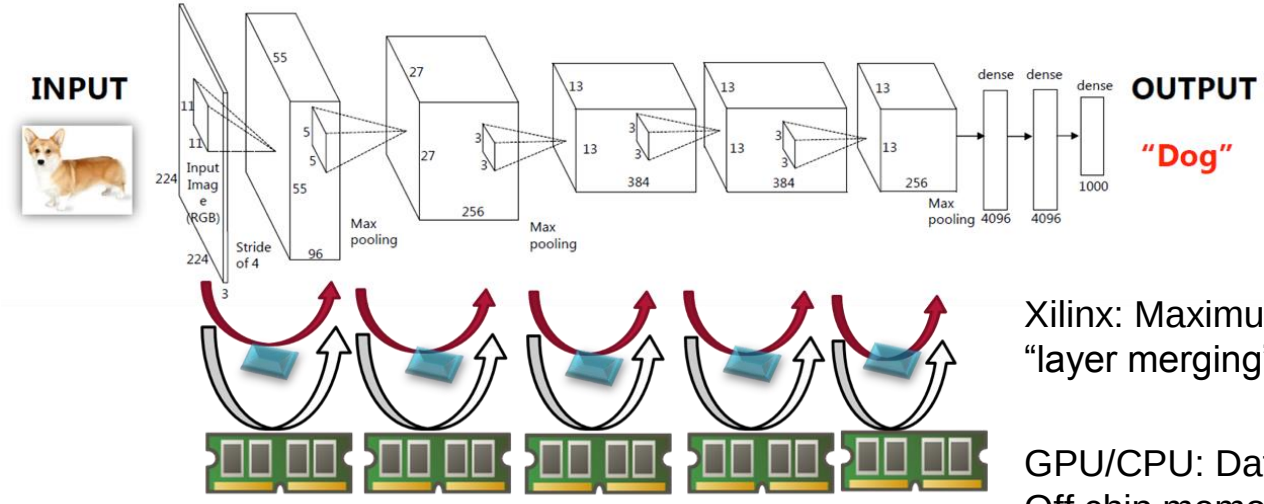
- Assuming 300 MHz with 90%/70% DSP/LUT utilizations
- Resource consumption assumption: 2.5 LUTs/op (INT1), 16 LUTs/op (INT4), 0.25 DSP/op (INT8)
- 10W power assumption on ZU9

8bits to 1bit: What is the Challenge?

- **Small degradation in accuracy but fast improving**



Low Latency Inference by Layer to Layer Dataflow On Chip



Xilinx: Maximum local data reuse
"layer merging" between layers

GPU/CPU: Dataflow using
Off chip memory

On-chip Memory	Nvidia Tegra X1 (GPU Regfile + L2)	Xilinx ZU7 (BRAM + URAM)
	6 Mb	38 Mb

Up to 6x More On-chip Memory than SoCs and eGPUs

Nvidia TX1 spec: http://wccftech.com/nvidia-tegra-x1-super-chip-announced-ces-2015-features-maxwell-core-architecture-256-cuda-cores/

reVISION Stack

OpenVX integration

Caffe

Frameworks

OpenCV

SDSoC Environment

DNN

CNN

GoogLeNet
SSD
FCN …

Libraries and Tools

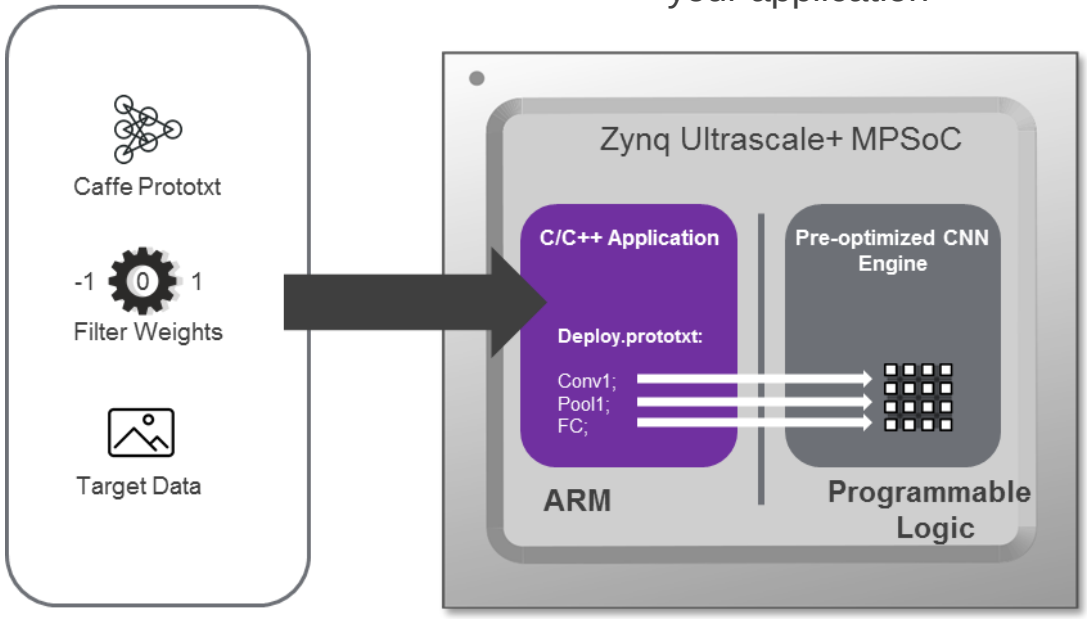
Development Kits

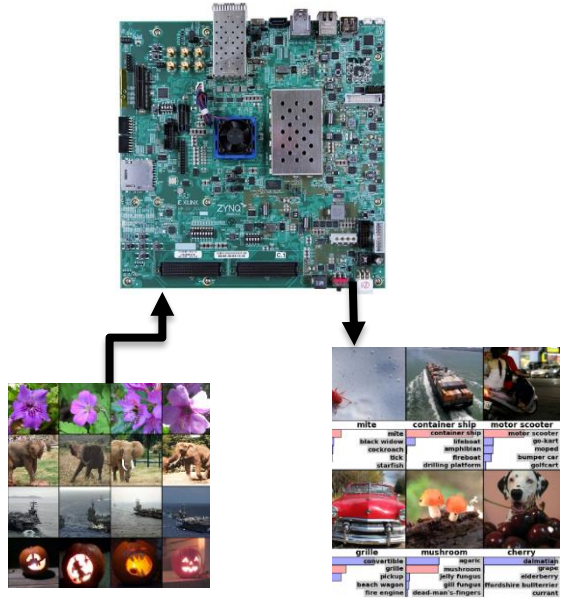
xFdnn: Direct Deep Learning Inference from Caffe

1 Import .prototxt and trained weights

2 Call prototxt runtime API in your application

3 Cross-compile for Cortex-A53 and run on a board

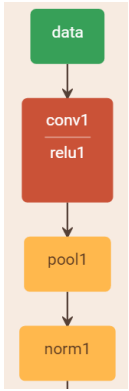




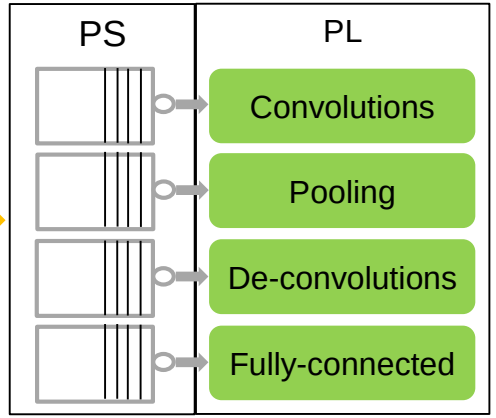
Compiles only ARM software code in minutes. No hardware compilation.

Caffe Prototxt to Zynq

```
name: "CaffeNet"
layer {
  name: "data"
  type: "Input"
  top: "data"
  input_param { shape: {
dim: 10 dim: 3 dim: 227
dim: 227 } }
}
layer {
  name: "conv1"
  type: "Convolution"
  bottom: "data"
  top: "conv1"
  convolution_param {
    num_output: 96
    kernel_size: 11
    stride: 4
  }
}
layer {
  name: "relu1"
  type: "ReLU"
  bottom: "conv1"
  top: "conv1"
}
```

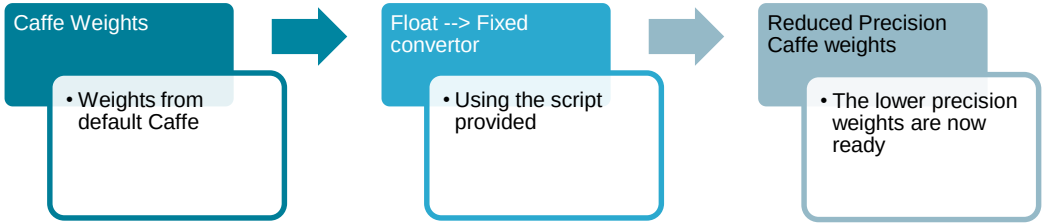


Network Scheduler (ARM)



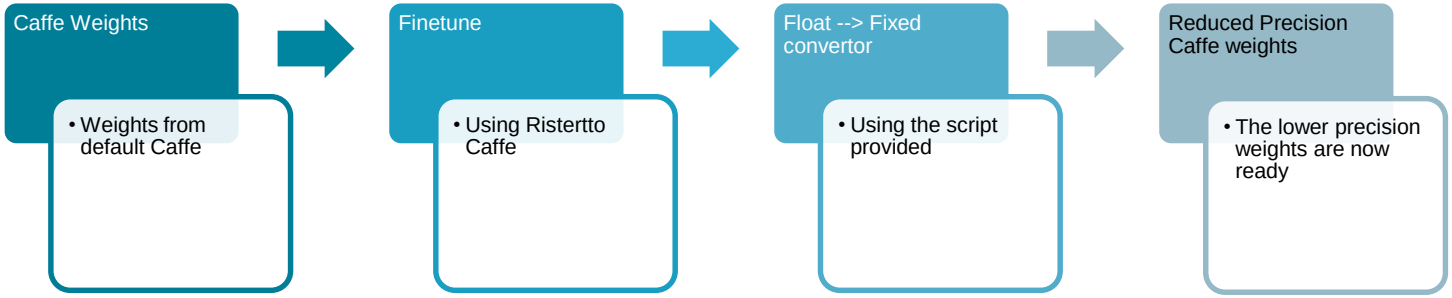
32 Bit Training to 8 Bit Inference

- Approach 1: Quick evaluation



Network	Default Caffe	Approach# 1	Approach# 2
GoogleNet	67.35	65.33	66.15
Alexnet	55.66	53.55	54.01

- Approach 2: Use Ristertto Caffe to retrain and fine-tune the weights for better accuracy

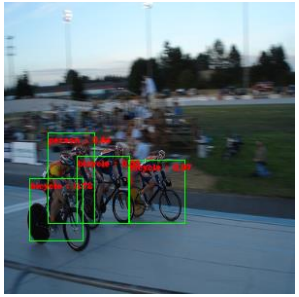


Deep Learning Design Examples

		May 2017	Roadmap
GoogLeNet @ batch = 1 3.2 Gops/img	Images/s	**114**	**370**
	Power (W)	6.0	7.0
	Images/s/watt	**19.0**	**52.9**

		May 2017	Roadmap
SSD @ batch = 1 62.4 Gops/img	Images/s	**6.3**	
	Power (W)	6.0	
	Images/s/watt	**1.1**	

		May 2017	Roadmap
FCN-AlexNet @ batch = 1 42.0 Gops/img	Images/s	**7.0**	
	Power (W)	6.0	
	Images/s/watt	**1.2**	



Deep Learning IP Export Flow

SDSoC Generated

Platform

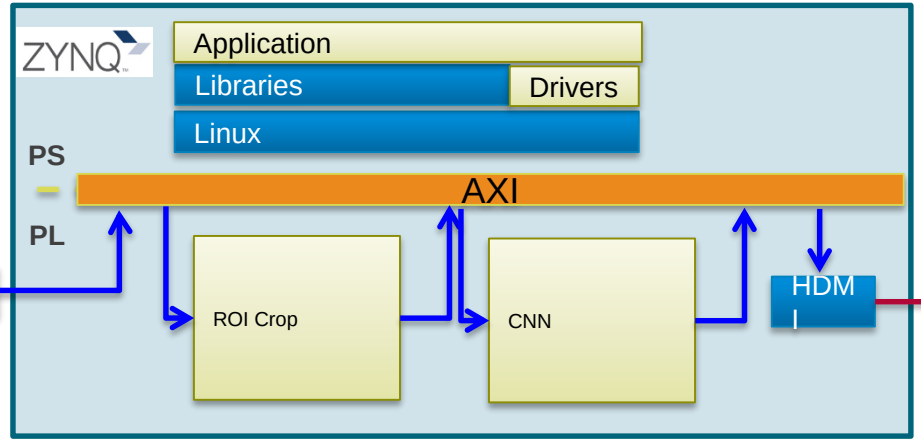
DMA

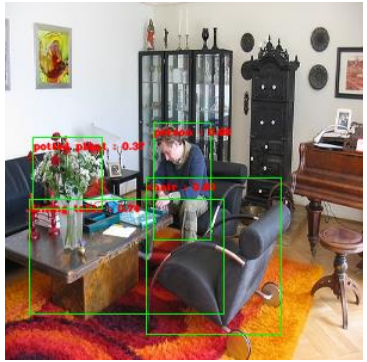
AXI-S

- Export DNN IP and ARM scheduler to integrate into real system
- Compile-time configuration of DNN IP (e.g. DSP, BRAM, buffer size …)

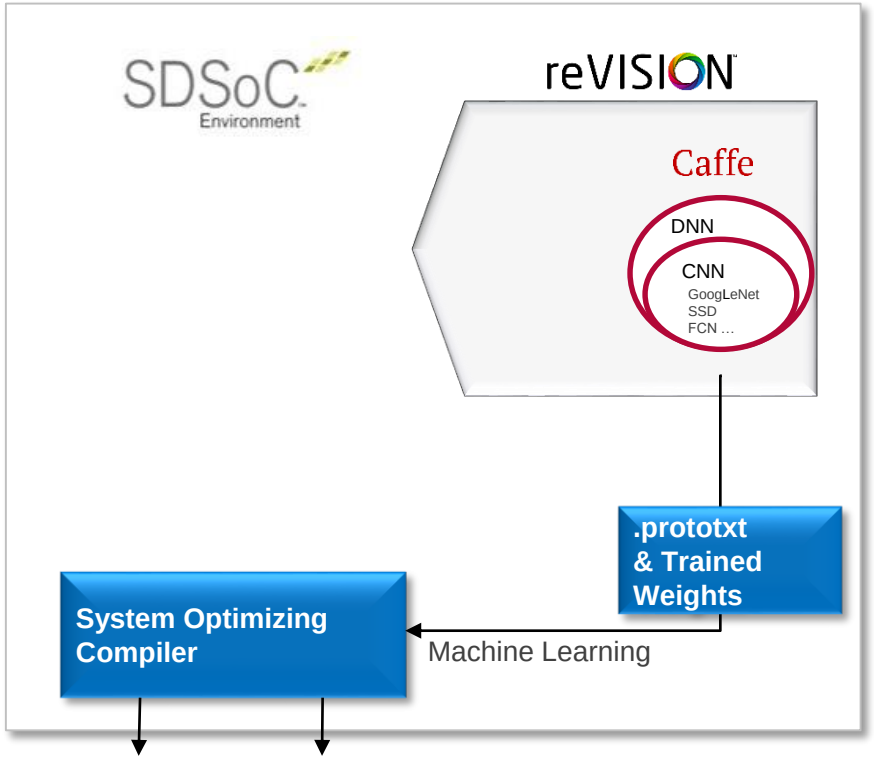
```
main(){
 imread(A);
 imread(B);
 roi_crop(A,img)
 xFdnn <DSP,BRAM,BUF,…>(img,prototxt,weights,out)
 imshow(out);}
```

SDSoC Environment

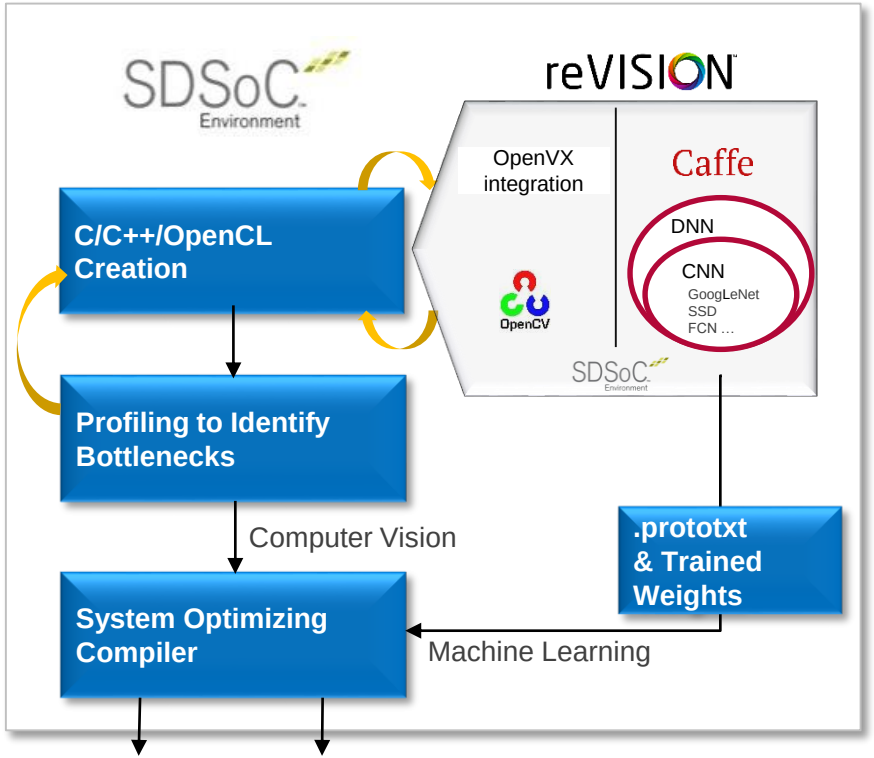




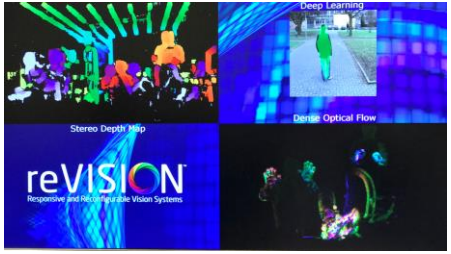
Building a Full Embedded Vision System



Building a Full Embedded Vision System



Putting It All Together: CV and CNN with Multiple Sensors



Summary

- Zynq SoCs offer superior performance and lower latency compared to other SoC offerings
- reVISION stack provides seamless inference of custom deep learning networks from Caffe to Zynq SoCs
- Visit Xilinx.com/reVISION for more information

Empowering Product Creators to Harness Embedded Vision

The Embedded Vision Alliance (www.Embedded-Vision.com) is a partnership of 60+ leading embedded vision technology and services suppliers

Mission: Inspire and empower product creators to incorporate visual intelligence into their products

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.Embedded-Vision.com

The Alliance enables vision technology providers to grow their businesses through leads, ecosystem partnerships, and insights

For membership, email us: membership@Embedded-Vision.com

Q&A

For more information and resources visit www.xilinx.com/reVISION