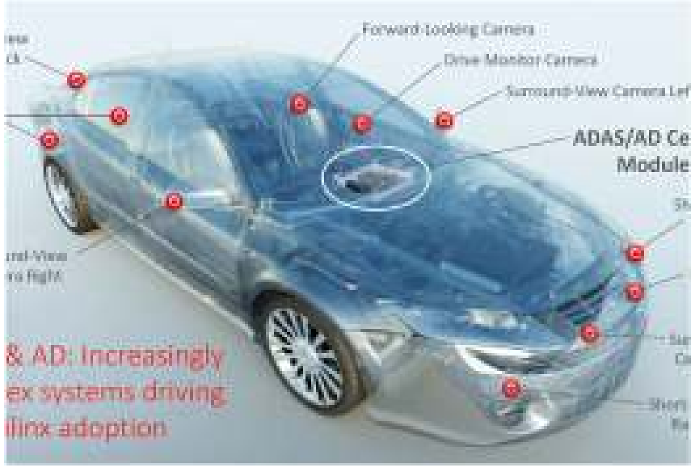
Xilinx Machine Learning Strategies for the Edge

Jon Cory, Embedded Vision Specialist FAE, NA

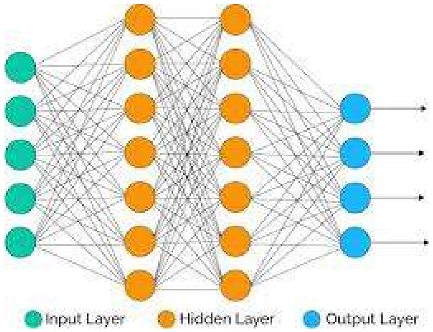
GET READY GET SET GO ADAPT

AI/ML Monetization Is Here and Growing



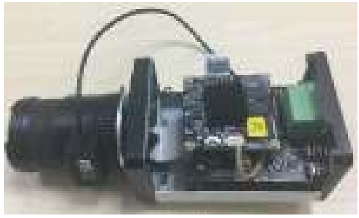
copyright sources: Avigilon, Amazon GO, Daimler, SK Telecom

Challenges in Monetizing AI/ML



= **43 TOPS**

?

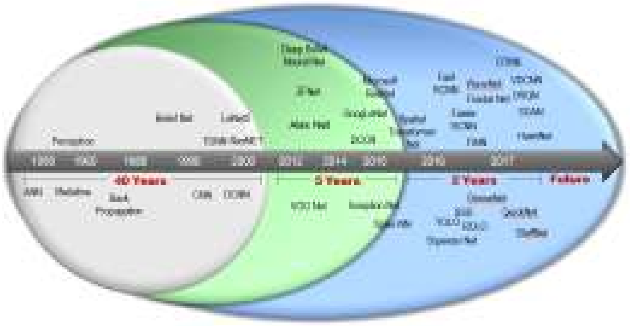


1080p Object Detection (SSD) @ 30 FPS

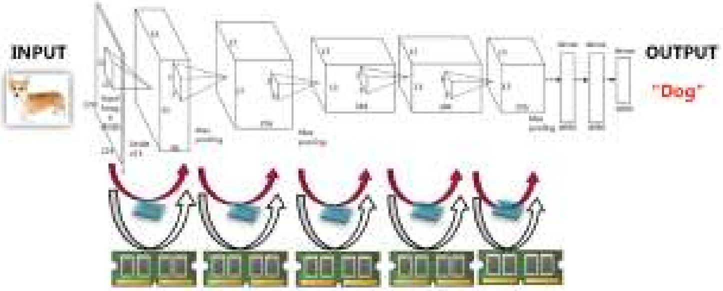
< 10W, < 50 ms latency, <$50

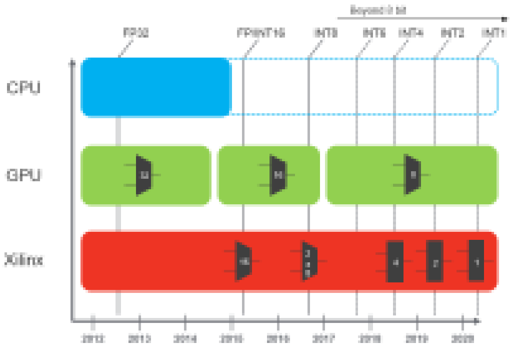
Who is Xilinx? Why Should I Care for ML?

1 **Only HW/SW configurable device for fast changing networks**

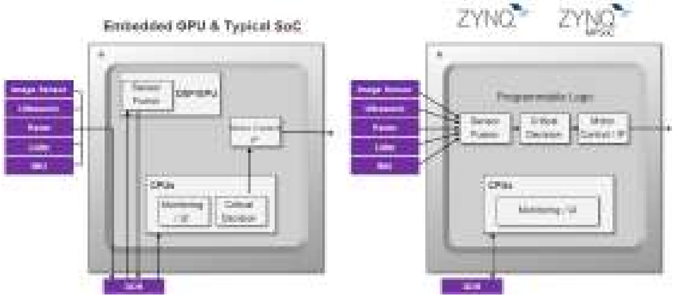


2 **High performance / low power with custom internal memory hierarchy**

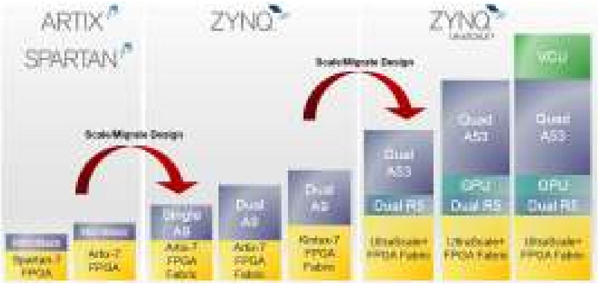




3 **Future proof to lower precisions**



4 **Low latency end-to-end**



5 **Scalable device family for different applications**

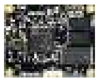
Integrated Xilinx-Deephi Roadmap

Xilinx AI Development

	Edge/Embedded		Cloud/DC
Models	20+ pruned / customized / basic models		
Software Stack	Deephi Pruning		
	Deephi Quantizer		Deephi Quantizer
	Deephi Compiler		xfDNN Compiler
	Deephi Runtime		xfDNN Runtime
FPGA IP	Deephi DPU	Deephi LSTM	xDNN

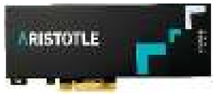
Platforms

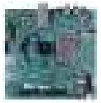
Z7020 Board


Z7020 SOM

ZU2/3 SOM

ZU2/3 Card


ZU9 Card


ZCU102

ZCU104

Ultra96

Xilinx U200, U250, U280

aws marketplace

NIMBIX

Alibaba Cloud
aliyun.com

HUAWEI

XILINX

Xilinx Network Development

Application	Module	Algorithm	Model Development	Compression	Deployment
Face	Face detection	SSD, Densebox	✓	✓	✓
	Landmark Localization	Coordinates Regression	✓	N / A	✓
	Face recognition	ResNet + Triplet / A-softmax Loss	✓	✓	✓
	Face attributes recognition	Classification and regression	✓	N / A	✓
Pedestrian	Pedestrian Detection	SSD	✓	✓	✓
	Pose Estimation	Coordinates Regression	✓	✓	✓
	Person Re-identification	ResNet + Loss Fusion	✓		
Video Analytics	Object detection	SSD, RefineDet	✓	✓	✓
	Pedestrian Attributes Recognition	GoogleNet	✓	✓	✓
	Car Attributes Recognition	GoogleNet	✓	✓	✓
	Car Logo Detection	DenseBox	✓	✓	
	Car Logo Recognition	GoogleNet + Loss Fusion	✓	✓	
	License Plate Detection	Modified DenseBox	✓	✓	✓
	License Plate Recognition	GoogleNet + Multi-task Learning	✓	✓	✓
ADAS/AD	Object Detection	SSD, YOLOv2, YOLOv3	✓	✓	✓
	3D Car Detection	F-PointNet, AVOD-FPN	✓		
	Lane Detection	VPGNet	✓	✓	✓
	Traffic Sign Detection	Modified SSD	✓		
	Semantic Segmentation	FPN	✓	✓	✓
	Drivable Space Detection	MobilenetV2-FPN	✓		
	Multi-task (Detection+Segmentation)	Deephi	✓		

DEEPHi 深鉴科技 Now Part of XILINX

GET READY GET SET GO ADAPT

Long History, Close Collaboration, and Better Future

Collaboration with Xilinx University Program

Deep learning acceleration
Time series analysis
Stereo vision
……

清華大學
Tsinghua University

Development of products on Xilinx FPGA platform since inception of DeePhi

Face recognition
Video analysis
Speech recognition acceleration
……

DEEPHi
深 鉴 科 技

Co-Marketing and Co-Sales with Xilinx Team

Data Center
Automotive
Video surveillance
……

DEEPHi
深 鉴 科 技

DEEPHi
深 鉴 科 技
Now Part of Xilinx

Pioneer in sparse-neural-network-based AI computing, explorer from theory to commercialization

Best Paper Award

Best Paper

NIPS 2015: Top conference in neural information processing

FPGA 2016 & 2017: Top academic conference in FPGA

ICLR 2016 : Top academic conference in machine learning

ISCA 2016 : Top academic conference in computer architecture

Hot Chips 2016 : Top academic conference in semiconductor

First prize of tech innovation China Computer Federation

Registering more than 100 invention patents both in China and US

First Paper in the World on Compressed and Sparse Neural Networks

"Learning both Weights and Connections for Efficient Neural Networks", NIPS 2015

"Deep Compression", ICLR 2016 Best Paper

First Paper in the World on Sparse Neural Network Accelerator

"EIE: Efficient Inference Engine on Compressed Deep Neural Network", ISCA 2016

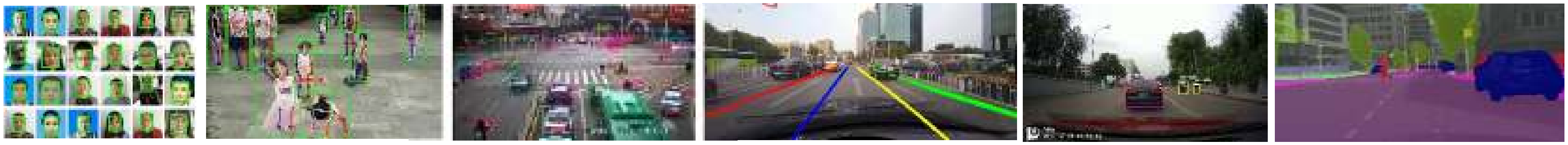
First Practical Case Using Sparse Neural Network Processor

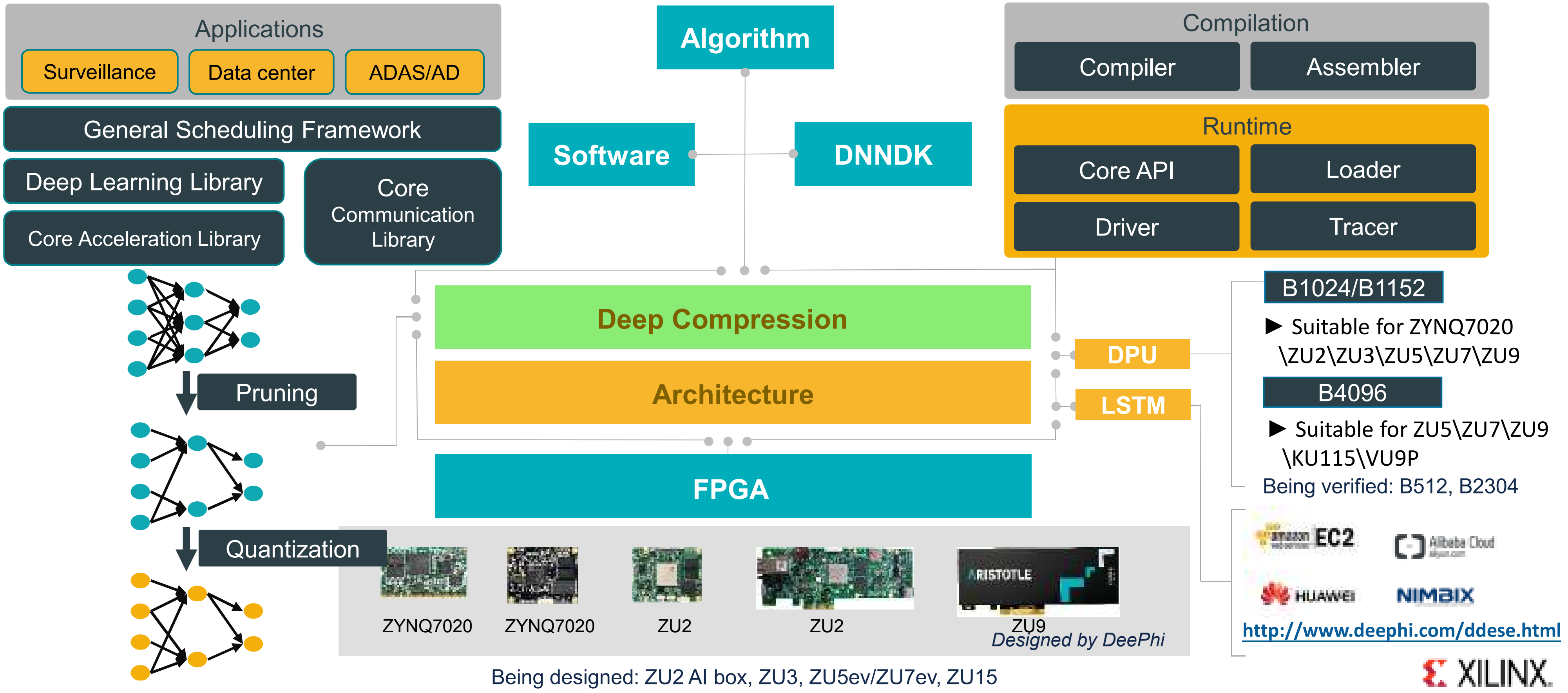
Collaboration with Sogou Inc, partly revealed in :

"ESE: Efficient Speech Recognition Engine with Compressed LSTM on FPGA",

FPGA 2017 Best Paper

Leading Solution for Deep Learning Acceleration





Applications

Surveillance | Data center | ADAS/AD

General Scheduling Framework

Deep Learning Library

Core Acceleration Library

Core Communication Library

Pruning

Quantization

Algorithm

Software

DNNDK

Compilation

Compiler | Assembler

Runtime

Core API | Loader

Driver | Tracer

Deep Compression

Architecture

FPGA

DPU

LSTM

B1024/B1152

- ▶ Suitable for ZYNQ7020 \ZU2\ZU3\ZU5\ZU7\ZU9

B4096

- ▶ Suitable for ZU5\ZU7\ZU9 \KU115\VU9P

Being verified: B512, B2304

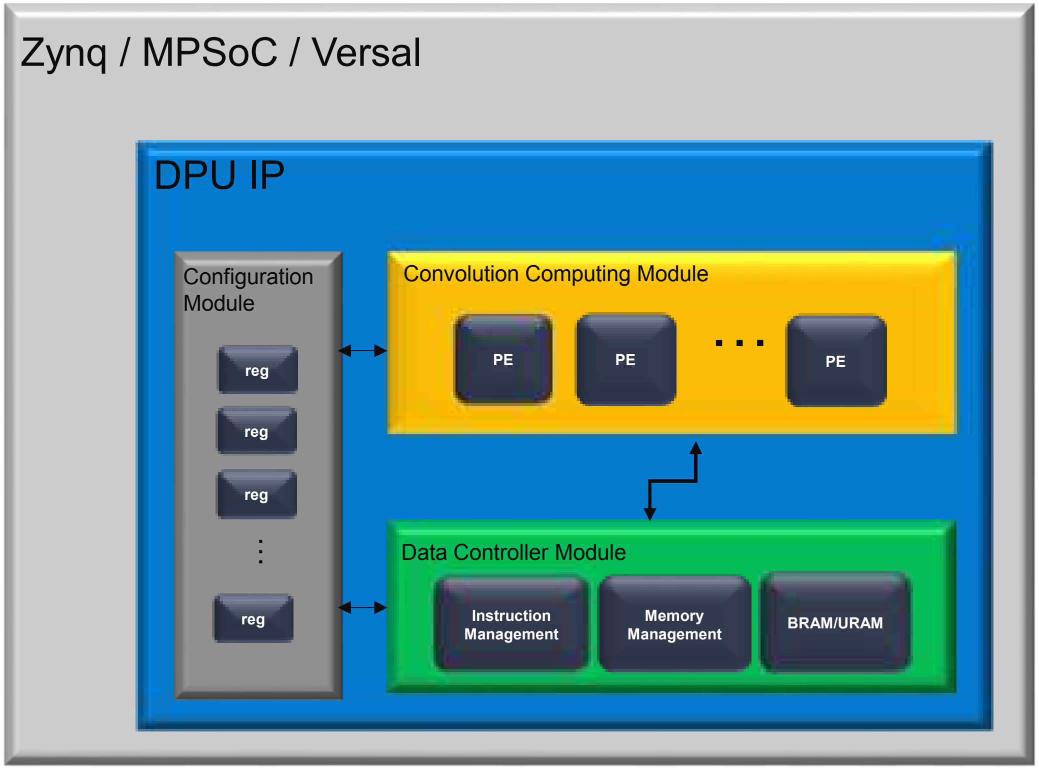
Designed by DeePhi

http://www.deephi.com/ddese.html

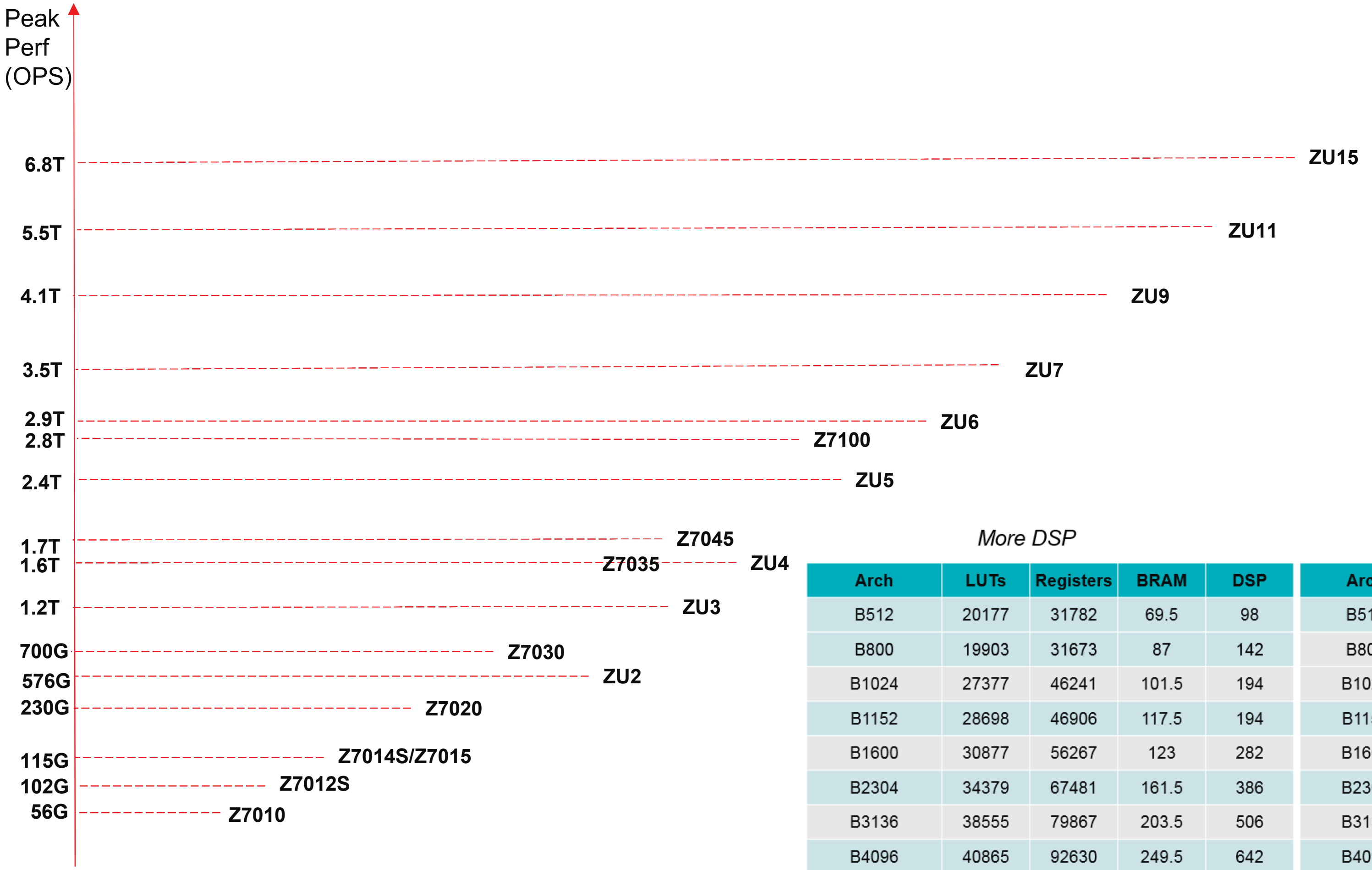
Being designed: ZU2 AI box, ZU3, ZU5ev/ZU7ev, ZU15

How Is Acceleration of CNNs Handled in the PL?

- **DPU Soft IP: Deep learning Processor Unit - Optimized for convolutional neural networks**
- **Consists of 3 Main modules**
  - Configuration module
  - Data Controller module
  - Convolution Computing module
- **Instruction Set**
  - Tensor based instructions
  - Up to 268,435,456 MACs/instruction
- **DPU Targets the Zynq Device Family**
  - APU required:
    - Interrupt handling
    - Data transfers
    - Unsupported operations



DPU Scalability



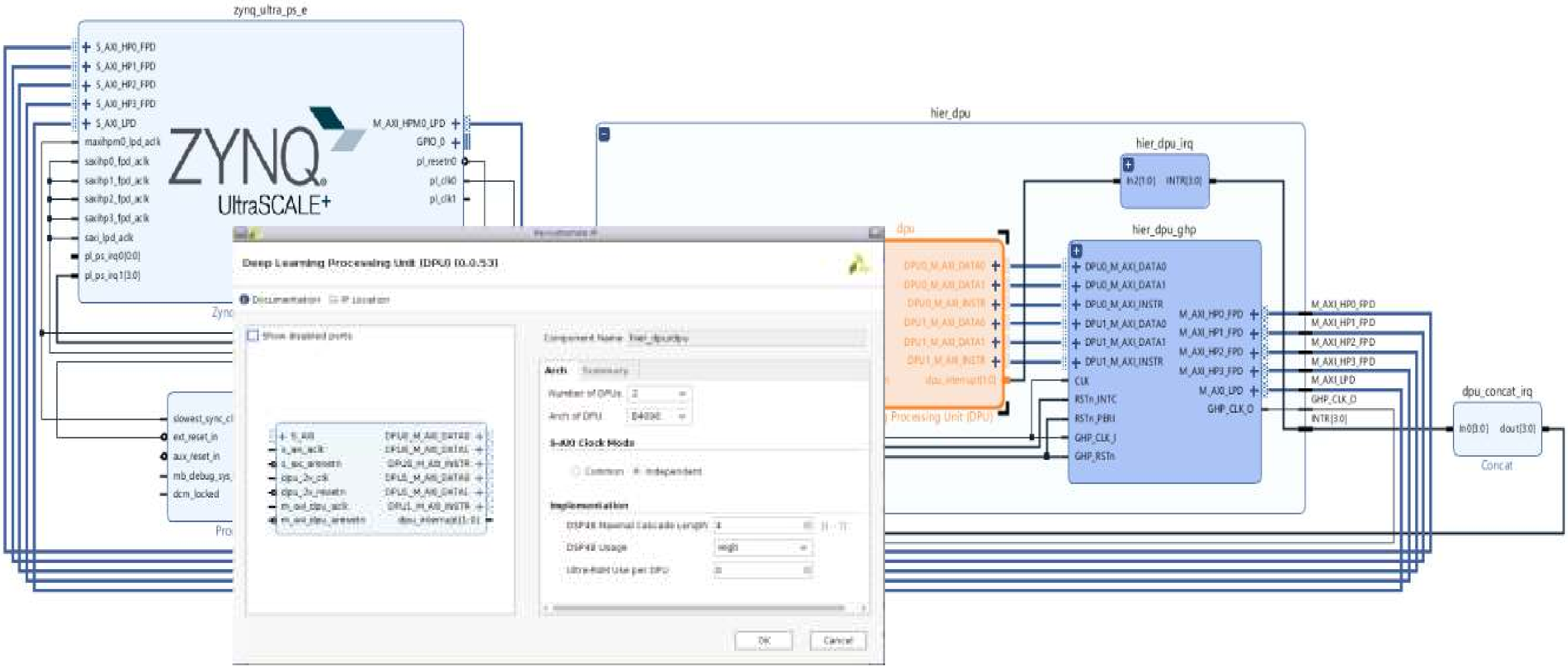
More DSP

Arch	LUTs	Registers	BRAM	DSP
B512	20177	31782	69.5	98
B800	19903	31673	87	142
B1024	27377	46241	101.5	194
B1152	28698	46906	117.5	194
B1600	30877	56267	123	282
B2304	34379	67481	161.5	386
B3136	38555	79867	203.5	506
B4096	40865	92630	249.5	642

More LUT

Arch	LUTs	Registers	BRAM	DSP
B512	20759	33572	69.5	66
B800	21050	33752	87	102
B1024	29155	49823	101.5	130
B1152	30043	49588	117.5	146
B1600	33130	60739	123	202
B2304	37055	72850	161.5	290
B3136	41714	86132	203.5	394
B4096	44583	99791	249.5	514

Zynq UltraScale+ MPSoC DPU TRD



Deep Learning Processing Unit (DPU) (0.0.53)

□ Show disabled ports

Component Name: hier_dpu_dpu

Number of DPUs: 2

Arch of DPU: B4096

S-AXI Clock Mode

○ Common ● Independent

Implementation

DSP48 Maximal Cascade Length: 4 [1 - 7]

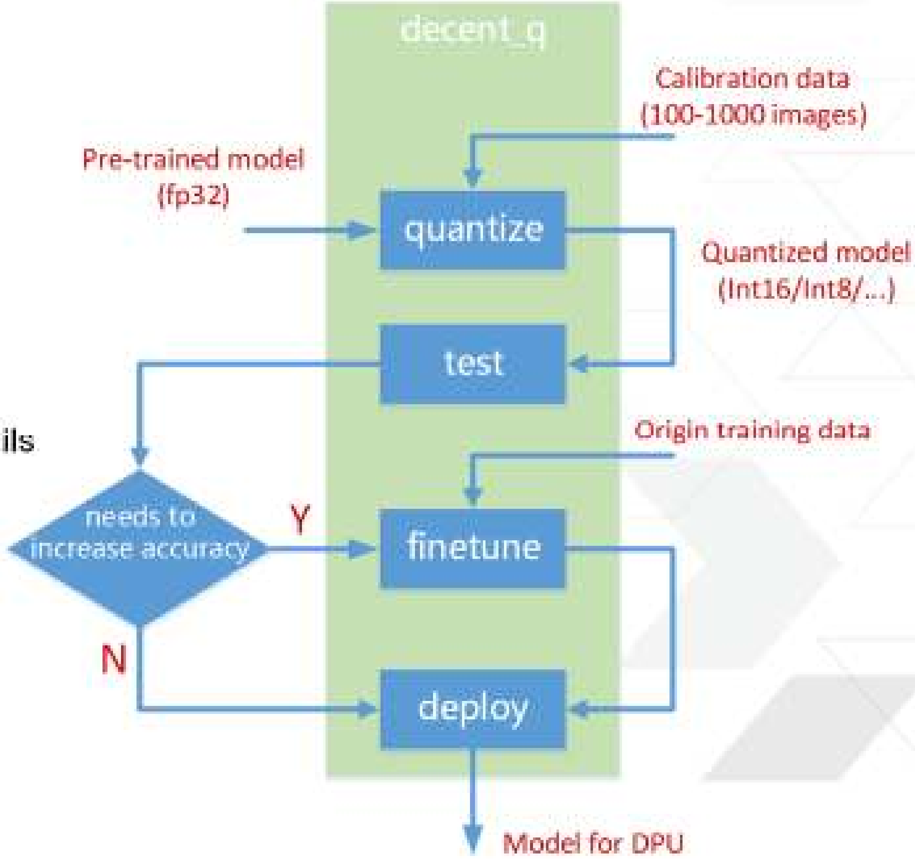
DSP48 Usage: High

UltraRAM Use per DPU: 0

Quantization Tool – decent

- **4 steps in decent quantization**
  - quantize – quantize network
    - Calibration images required
  - test – test network accuracy/mAP
  - finetune – finetune quantized network
    - Usually not needed
    - Requires entire training data set
    - Not documented contact factory for more details
  - deploy – generate model for DPU
    - This is input to the dnnc compiler

- **Data**
  - Calibration data – quantize activation
  - Training data – further increase accuracy

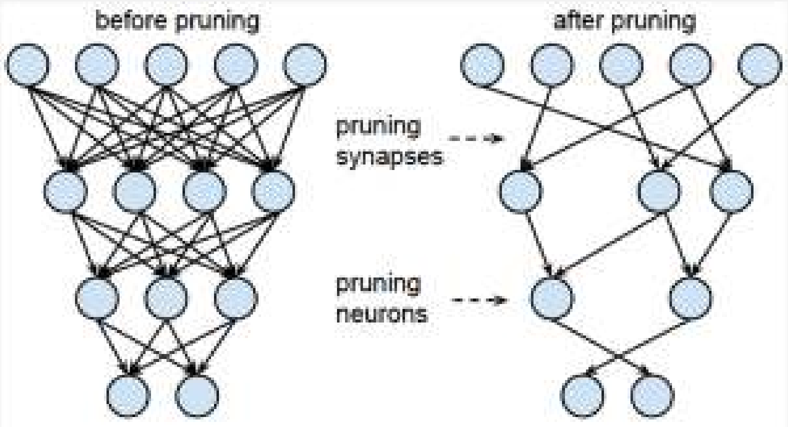


Quantization Results for Popular Networks

Classification	float		8-bit fix	
	Top1	Top5	ΔTop1	ΔTop5
Inception_v1	66.90%	87.68%	-0.28%	-0.10%
Inception_v2	72.78%	91.04%	-0.38%	-0.23%
Inception_v3	77.01%	93.29%	-0.45%	-0.29%
Inception_v4	79.74%	94.80%	-0.32%	-0.16%
ResNet-50	74.76%	92.09%	-0.17%	-0.14%
ResNet-50-v2	75.39%	92.45%	-0.60%	-0.33%
VGG16-3fc-float	70.97%	89.85%	-0.23%	-0.06%
VGG16-1fc-float	70.97%	89.85%	-0.20%	-0.09%
Inception-ResNet-v2	79.95%	95.13%	-0.51%	-0.16%
Detection	Float mAP		8-bit fix mAP	
SSD_VGG	76.47%		-0.20%	

Core advantage | Deep compression algorithm

Deep compression
Makes algorithm smaller and lighter



Highlight

- 1/3 ▼ Weight number
- 1/10 ▼ Bandwidth load
- 1/10 ▼ Model size
- 3 X ▲ Performance

Compression efficiency

Deep Compression Tool can achieve significant compression on **CNN** and **RNN**

Accuracy

Algorithm can be **compressed 7 times without losing accuracy** under SSD object detection framework

Easy to use

Simple software development kit need only **50 lines of code** to run ResNet-50 network

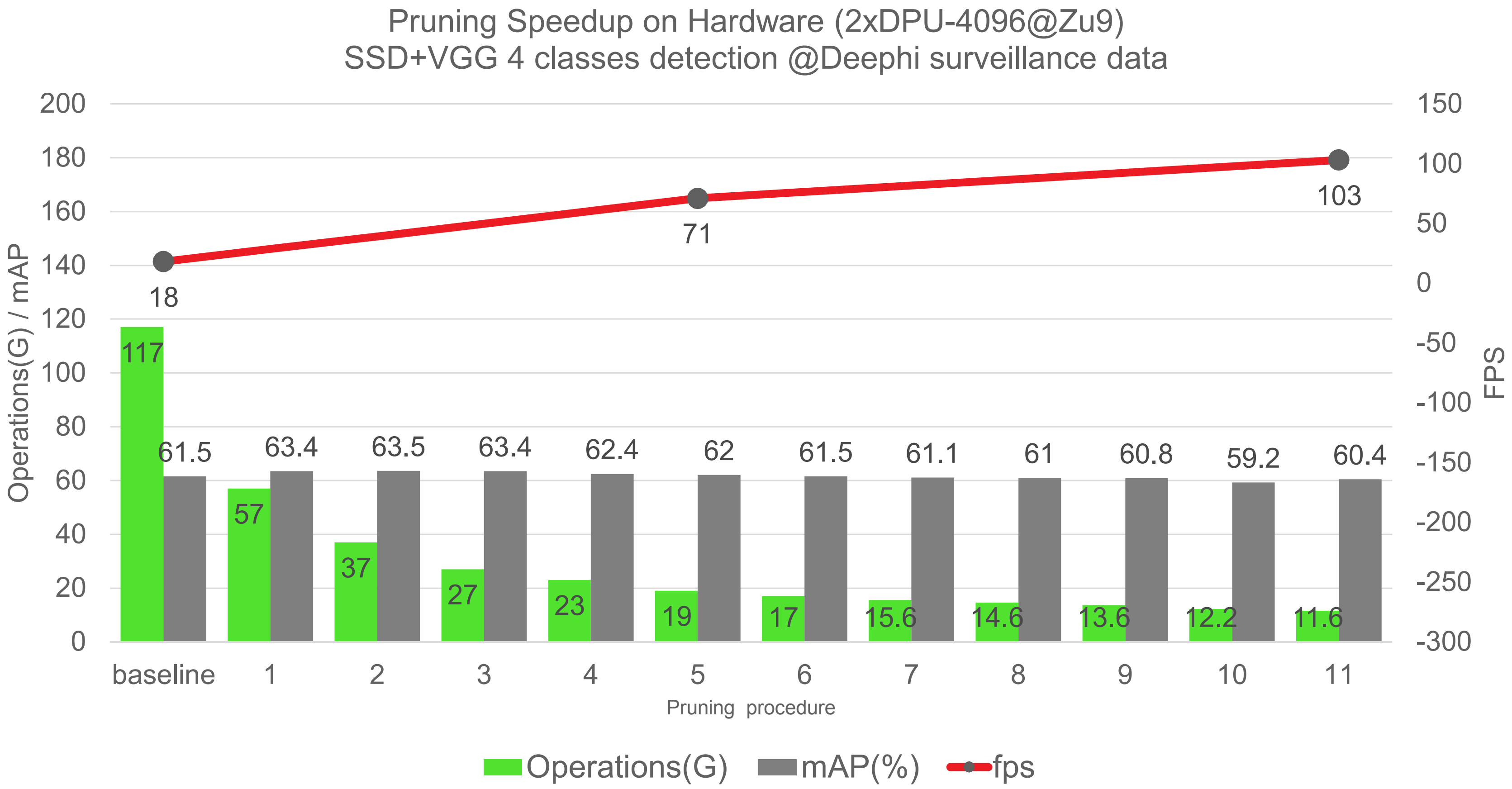
Pruning Results

Classification Networks	Baseline	Pruning Result 1			Pruning Result 2		
	Top-5	Top-5	ΔTop5	ratio	Top-5	ΔTop5	ratio
Resnet50 [7.7G]	91.65%	91.23%	-0.42%	40%	90.79%	-0.86%	32%
Inception_v1 [3.2G]	89.60%	89.02%	-0.58%	80%	88.58%	-1.02%	72%
Inception_v2 [4.0G]	91.07%	90.37%	-0.70%	60%	90.07%	-1.00%	55%
SqueezeNet [778M]	83.19%	82.46%	-0.73%	89%	81.57%	-1.62%	75%

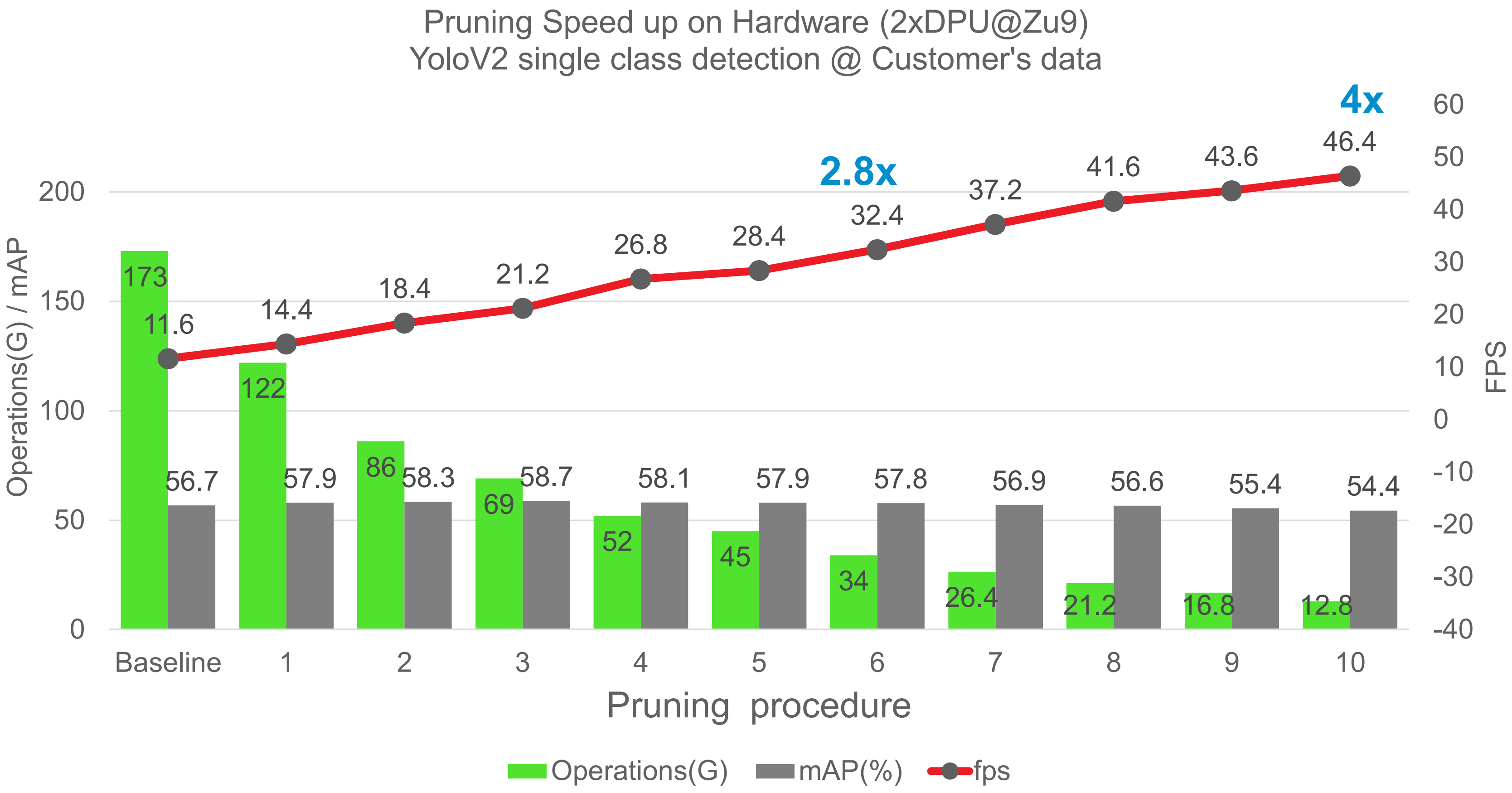
Detection Networks	Baseline mAP	Pruning Result 1			Pruning Result 2		
		mAP	ΔmAP	ratio	mAP	ΔmAP	ratio
DetectNet [17.5G]	44.46	45.7	+1.24	63%	45.12	+0.66	50%
SSD+VGG [117G]	61.5	62.0	+0.5	16%	60.4	-1.1	10%
[A] SSD+VGG [173G]	57.1	58.7	+1.6	40%	56.6	-0.5	12%
[B] Yolov2 [198G]	80.4	81.9	+1.5	28%	79.2	-1.2	7%

Segmentation Networks	Baseline	Pruning Result 1			Pruning Result 2		
	mIoU	mIoU	ΔmIoU	ratio	mIoU	ΔmIoU	ratio
FPN [163G]	65.69%	65.21%	-0.48%	80%	64.07%	-1.62%	60%

Pruning Speedup Example – SSD



Pruning Speedup Example – Yolo_v2



Compression perspective

Research

Quantization

- **Low-bit and hybrid low-bit quantization**
  - Some simple hybrid low-bit experiments [Compared to 8bit results, without finetune]
    - 20% model size reduce, <1% accuracy drop
    - 10% model size reduce, <1% accuracy drop (hardware-friendly low-bit patterns)
- **7nm FPGA with AI Engines**
  - Some fp32/fp16 resources -> Relax some restrictions for quantization -> Better performance
    - For low-bit quantization, non-uniform quantization with lookup tables is possible
    - Some layers can run without quantization
- **AutoML for quantization**
  - Automated quantization for hybrid low-bit quantization

Pruning

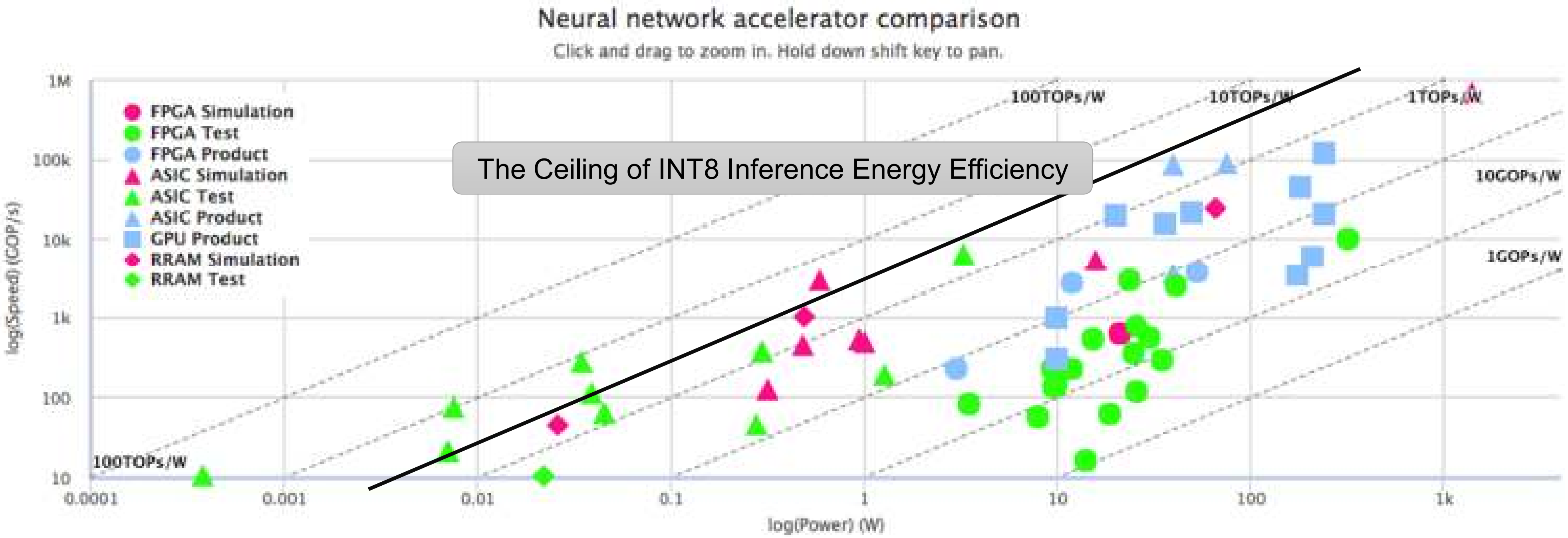
- **AutoML for pruning**
  - Automated pruning by reinforcement learning

Tools

- **Unified compression tool supporting different frameworks**
- **Fully tested tools, ease of use**
- **Improved speed for pruning tool, supporting cluster**

Caffe TensorFlow Pytorch

Current Ceiling of CNN Architecture



The Ceiling of INT8 Inference Energy Efficiency

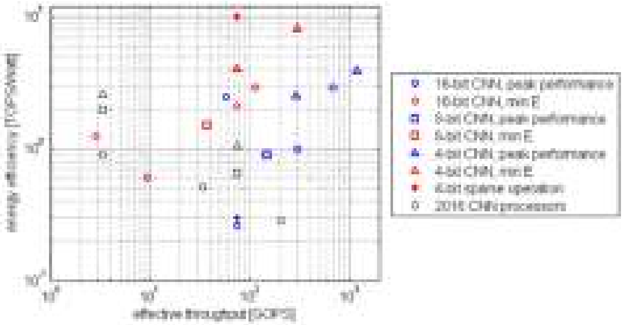
Source:http://nics-efc.org/projects/neural-network-accelerator/

INT8 improvements are slowing down and approaching the ceiling.

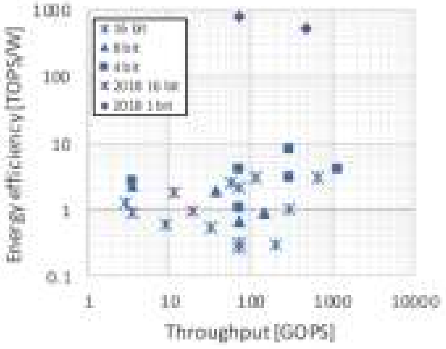
Solutions

- **Sparsity**
- **Low Precision**

Potentials of low precision

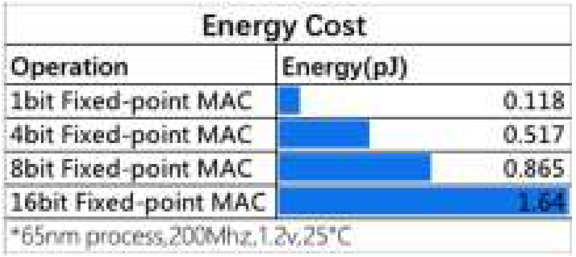


ISSCC,2017



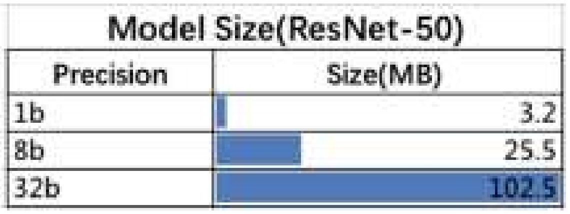
ISSCC,2018

Low Precision Becomes Popular



Energy Cost	
Operation	Energy(pJ)
1bit Fixed-point MAC	0.118
4bit Fixed-point MAC	0.517
8bit Fixed-point MAC	0.865
16bit Fixed-point MAC	1.64

*65nm process,200Mhz,1.2v,25°C

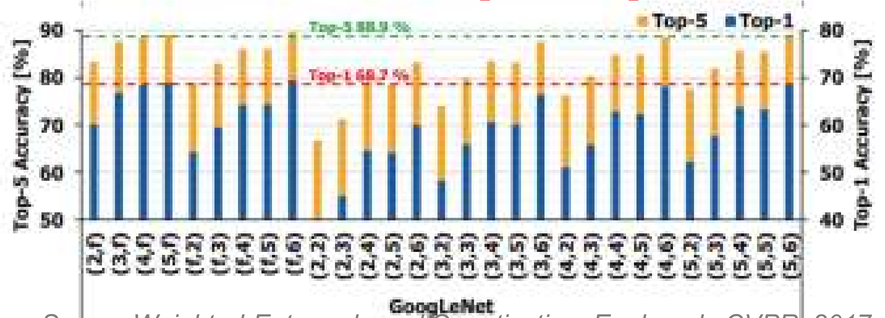


Model Size(ResNet-50)	
Precision	Size(MB)
1b	3.2
8b	25.5
32b	102.5

- Scales performance
- Reduces hardware resources
- Less bandwidth/on-chip memory requirement
- Regular memory access pattern and calculating pattern

FPGA benefits a lot from low-precision.

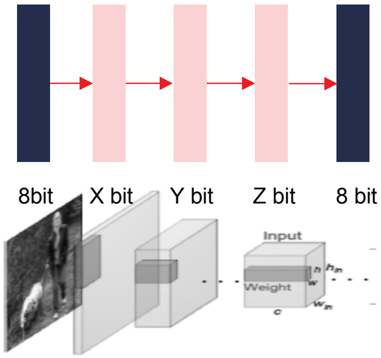
Architecture perspective: Mixed Low-Precision

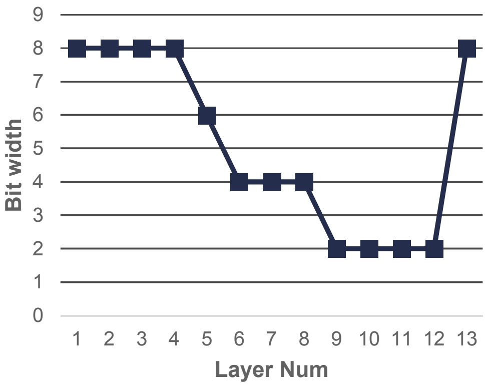


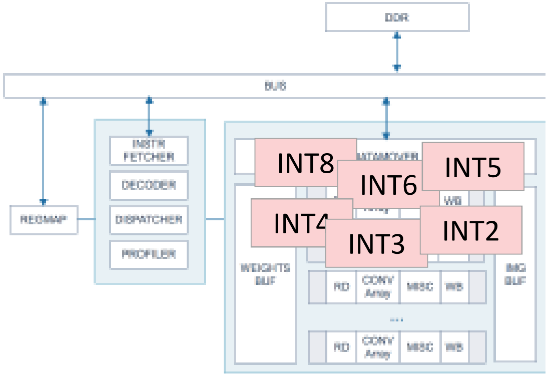
Source:Weighted-Entropy-based Quantization, Eunhyeok, CVPR, 2017-

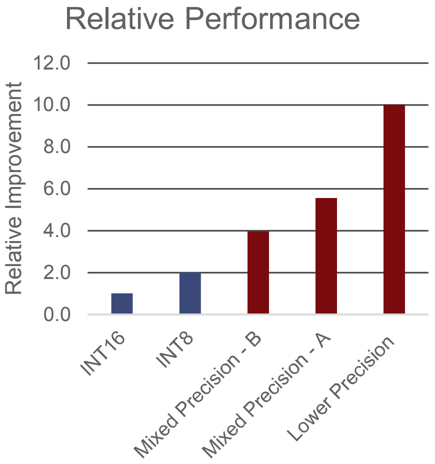
Fixed low-precision quantization already showed competitive results.

Next generation: **Variable** precision of activation/weights among layers









*accuracy drop less than 1%

BW	2	3	4	5	6	7	8
wgt	0	3	4	6	0	0	3
act	0	0	0	2	5	10	5

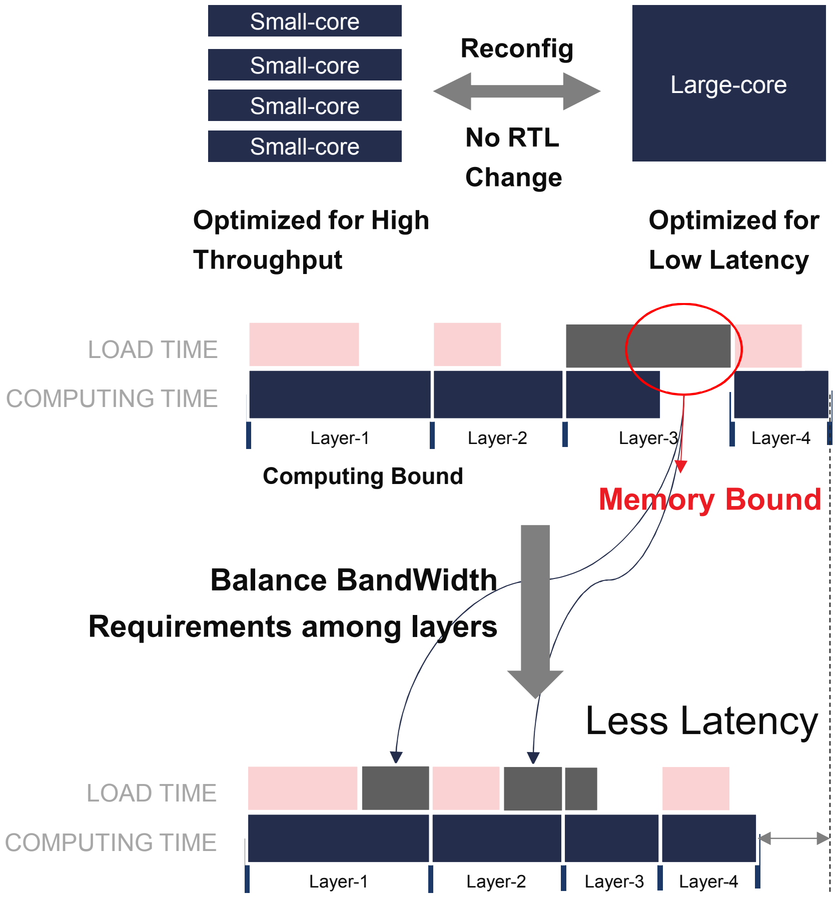
BW	2	3	4	5	6	7	8
wgt	0	0	3	22	17	10	2
act	0	0	0	16	41	13	3

BW	2	3	4	5	6	7	8
wgt	0	0	0	15	84	38	13
act	0	0	0	0	6	84	99

Preliminary experiments on popular networks. (vgg-16,resNet-50,inception-v4)

Architecture perspective: Mixed Low-Precision CNN

- **Mixed Precision Support**
  - INT8/6/5/4/3/2
- **Flexible Between Throughput and Latency**
  - Switch between Throughput-Opt-Mode and Latency-Opt-Mode without RTL change
- **Enhanced Dataflow Techniques**
  - Make the balance among different layers. Do NOT require the model can be fully placed on chip, but load the data at the right time.
  - Physical-aware data flow design to meet higher frequency.
  - Supports high-resolution images at high utilization.
- **Performance Target** (googlenet_v1)
  - 3103 FPS (INT8)
  - 5320 FPS (INT8/4/2 mixed)
  - 12412 FPS (INT2 only)



Small-core
Small-core
Small-core
Small-core

Reconfig

No RTL Change

Large-core

Optimized for High Throughput

Optimized for Low Latency

LOAD TIME
COMPUTING TIME
Layer-1
Layer-2
Layer-3
Layer-4

Computing Bound

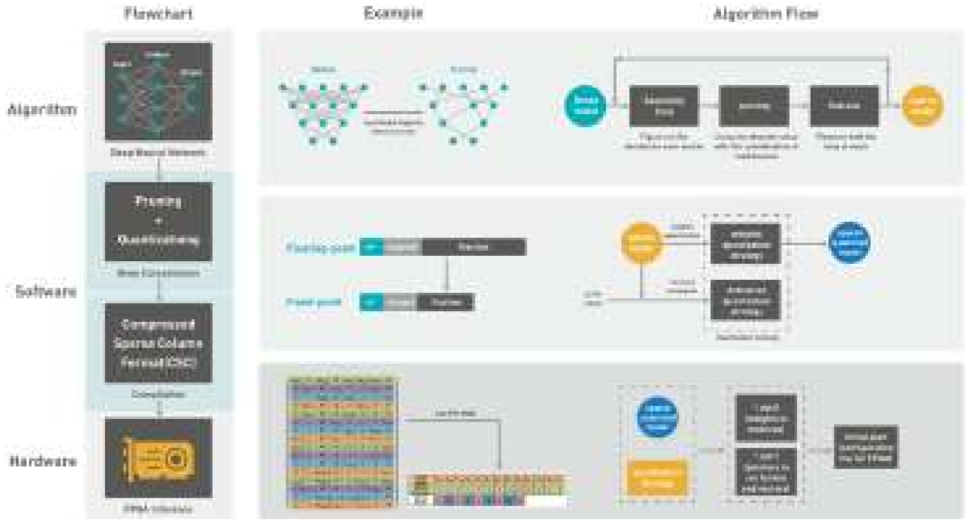
Memory Bound

Balance BandWidth Requirements among layers

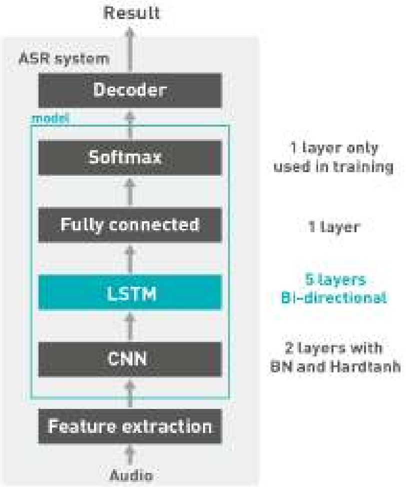
Less Latency

LOAD TIME
COMPUTING TIME
Layer-1
Layer-2
Layer-3
Layer-4

Sparsity architecture exploration



for end-to-end speech recognition



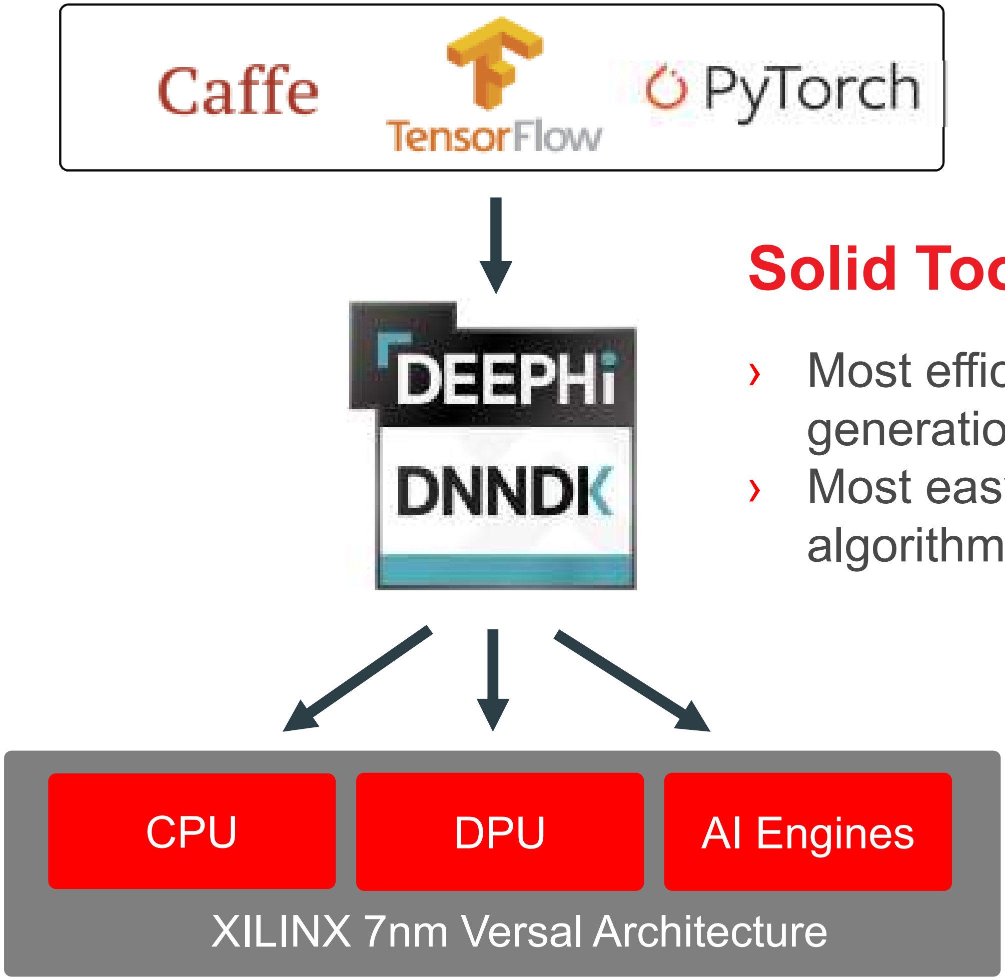
Partners

- ✓ On cloud, aiming at customers all over the world
- ✓ Already officially launched in AWS Marketplace and HUAWEI cloud (http://www.deephi.com/ddese.html)
- ✓ Now transplanting to Alibaba cloud

Features

Low storage	Model compressed more than 10X with negligible loss of accuracy
Low latency	More than 2X speedup compared to GPU (P4)
Programmable	Reconfigurable for different requirements

DNNDK perspective



Solid Toolchain Stack for XILINX Versal

- Most efficiency solution for ML on XILINX next generation computing platform
- Most easy-to-use & productive toolchain for ML algorithms deployment

System perspective: schedule ADAS tasks in single FPGA

Multi-task Models

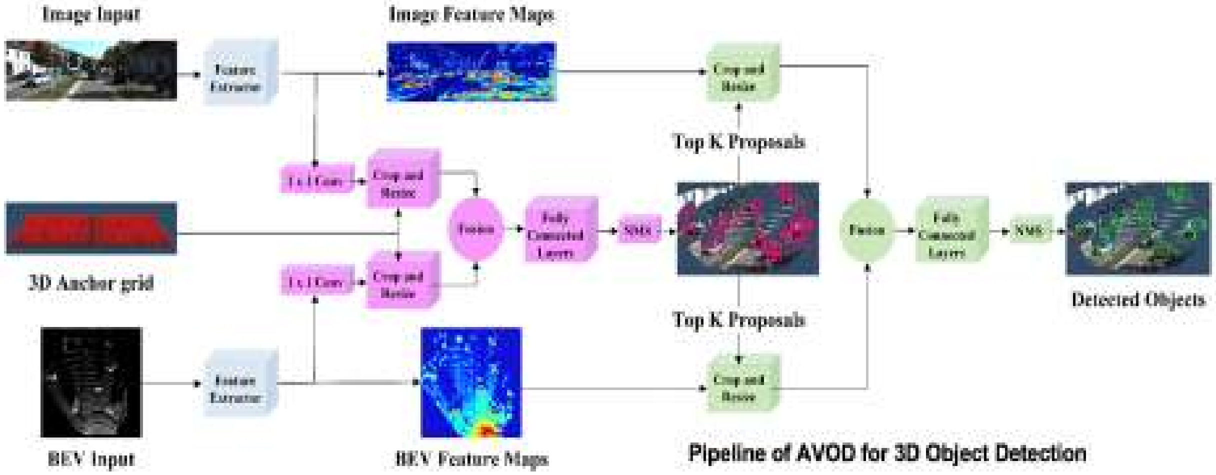
- Training:
  - Knowledge sharing
  - Reduce computation cost
- Pruning:
  - Balance different objective functions

Sensor Fusion

- Sensor alignment & Data Fusion

Task scheduling

- Resource constrained scheduling: Serialization & Parallelization
- Task scheduling and memory management framework with low context-switching cost
- Support new operations with runtime variable parameter by software and hardware co-design

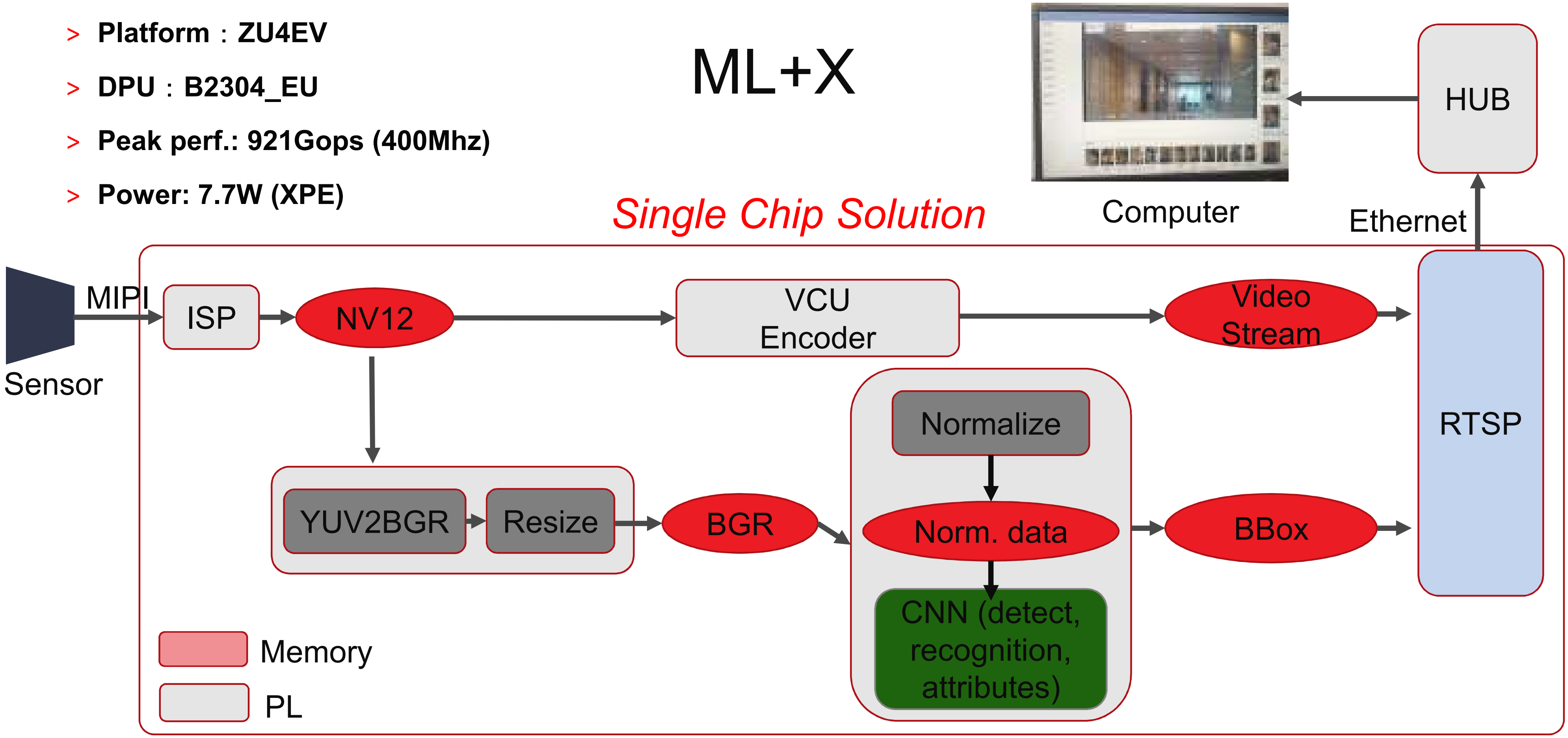


Pipeline of AVOD for 3D Object Detection

System perspective: Video Surveillance in single FPGA

- Platform：ZU4EV
- DPU：B2304_EU
- Peak perf.: 921Gops (400Mhz)
- Power: 7.7W (XPE)

ML+X



This solution needs to further enhance ISP functionality

Adaptable.
Intelligent.

XDF XILINX DEVELOPER FORUM

XILINX