XDF XILINX DEVELOPER FORUM

Xilinx Machine Learning Strategies with Deephi Acquisition

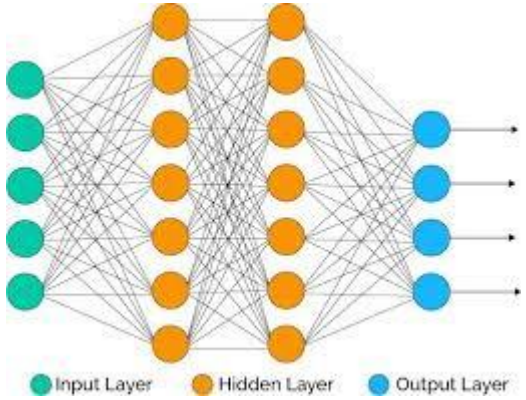
Presented By

Yi Shan, Sr. Director, AI Engineering & Former CTO of DeePhi

XILINX

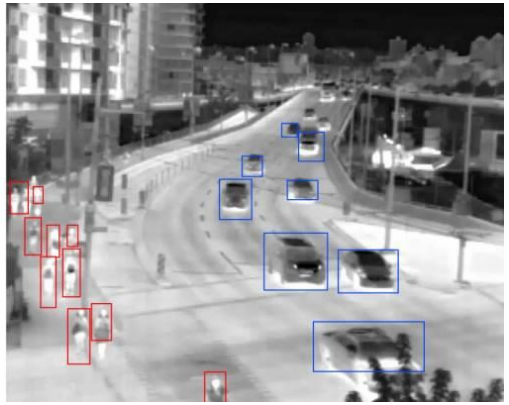
The Hottest Research: AI / Machine Learning

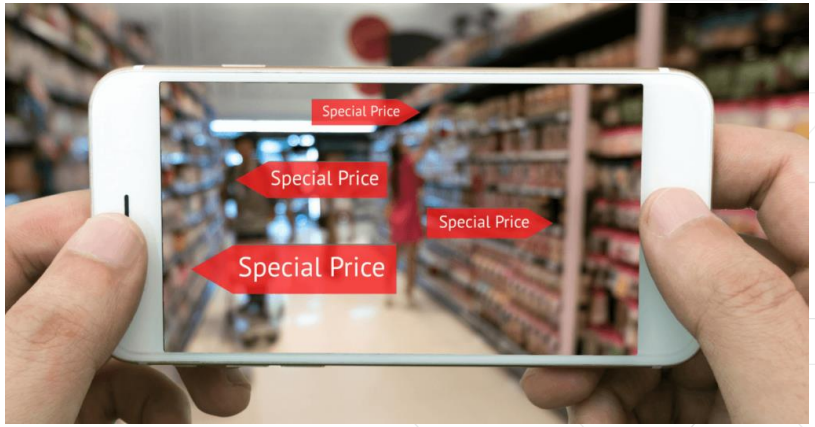
Nick's ML Model
Nick's ML Framework

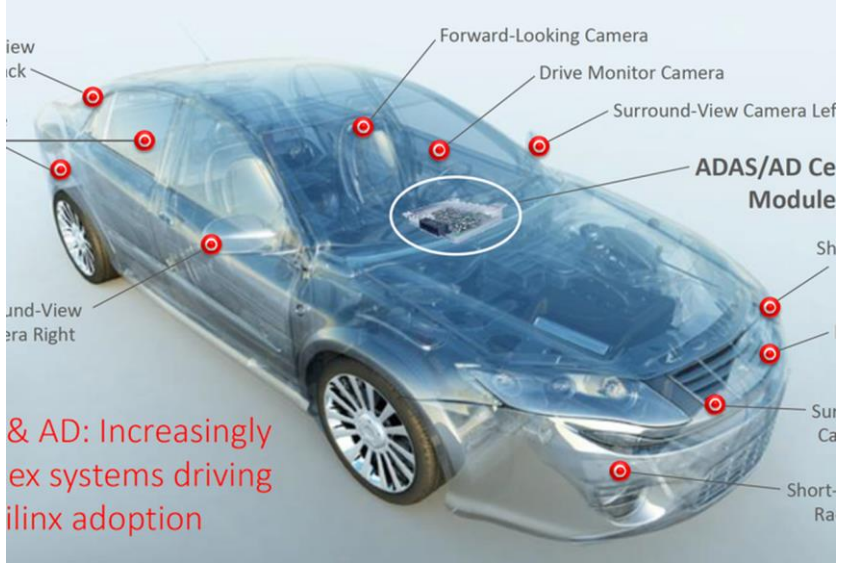


copyright sources: Gospel Coalition

AI/ML Monetization Is Here and Growing

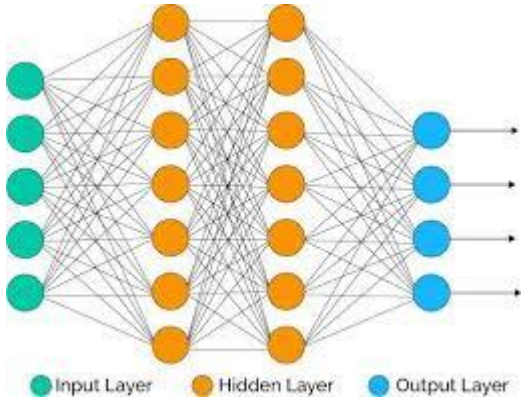






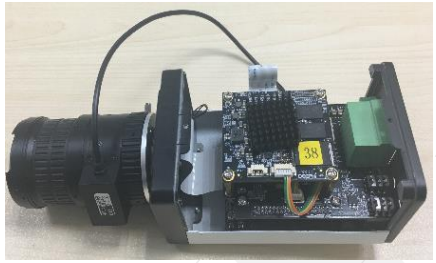
copyright sources: Avigilon, Amazon GO, Daimler, SK Telecom

Challenges in Monetizing AI/ML



= **43 TOPS**

?

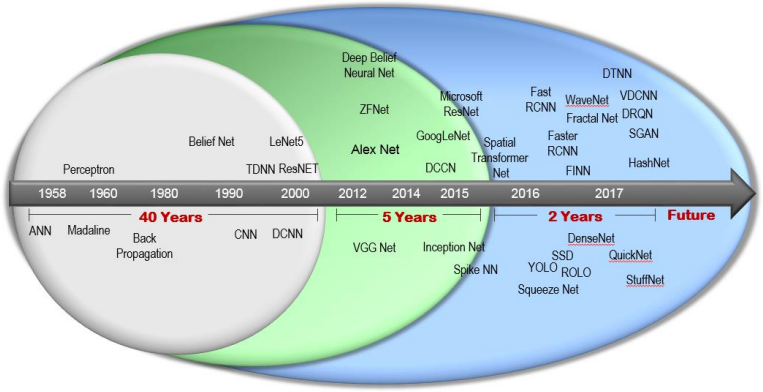


1080p Object Detection (SSD) @ 30 FPS

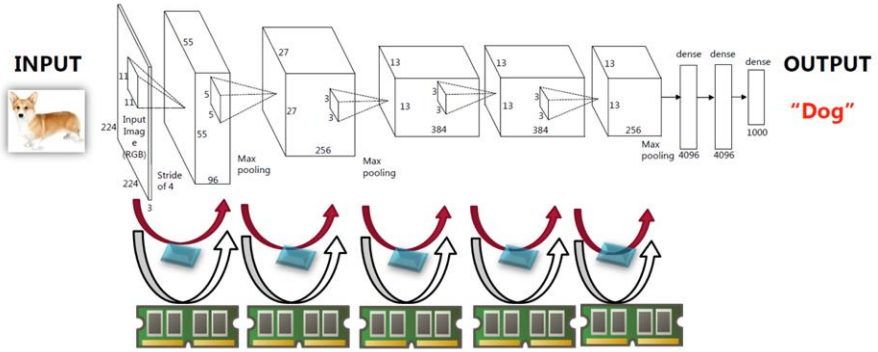
< 10W, < 50 ms latency, <$50

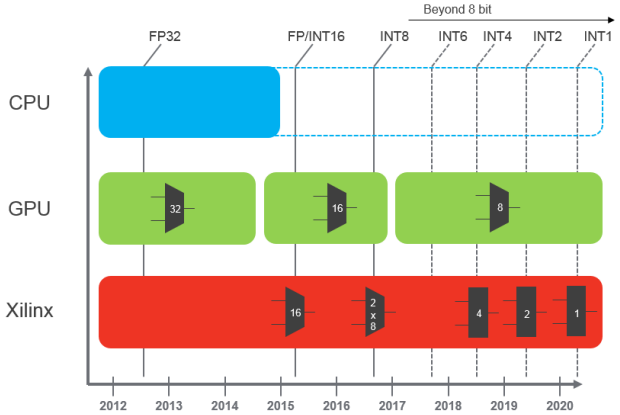
Who is Xilinx? Why Should I Care for ML?

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

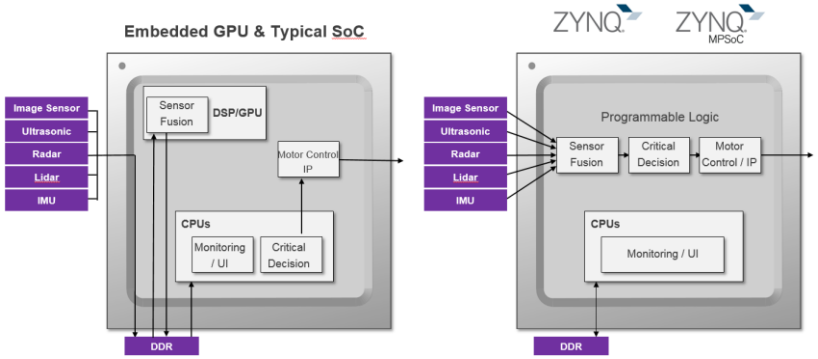


2 **High performance / low power with custom internal memory hierarchy**

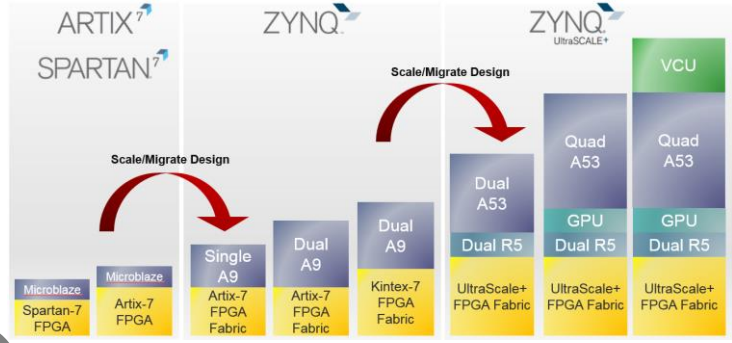




3 **Future proof to lower precisions**



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



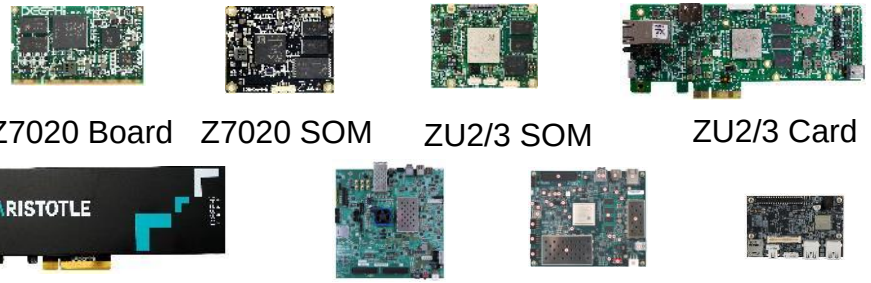
5 **Scalable device family for different applications**

XILINX

Integrated Xilinx-Deephi Roadmap

Xilinx AI Development

	Edge/Embedded	Cloud/DC
Models	20+ pruned / customized / basic models	
Software Stack	Deephi Pruning Deephi Quantizer Deephi Compiler Deephi Runtime SDSoC	SDAccel Deephi Quantizer xfDNN Compiler 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



awsmarketplace NIMBIX Alibaba Cloud aliyun.com HUAWEI

XILINX

Deephi as key part of Embedded Vision Development

Frameworks & Libraries

OpenCV

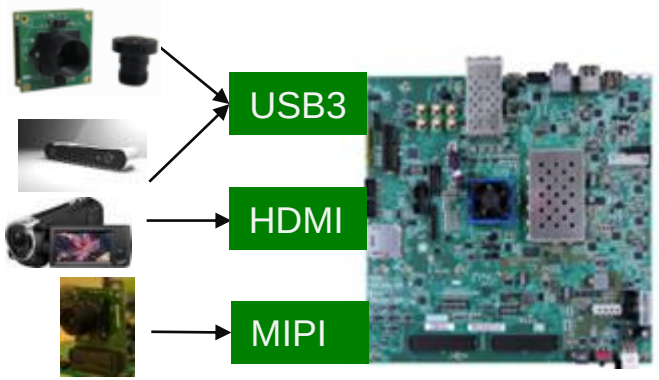
gstreamer

Machine Learning

Development tools

SDSoC™ Environment

Platforms



Xilinx Announces the Acquisition of DeePhi Tech

Deal to Accelerate Data Center and Intelligent Edge Applications

BEIJING and SAN JOSE, Calif., July 17, 2018 – Xilinx, Inc. (NASDAQ: XLNX), the leader in adaptive and intelligent computing, announced today that it has acquired DeePhi Tech, a Beijing-based privately held start-up with industry-leading capabilities in machine learning, specializing in deep compression, pruning, and system-level optimization for neural networks.

DEEPHi 深鉴科技

reVISION™ Stack

XDF XILINX DEVELOPER FORUM

DEEPHi 深鉴科技 Now Part of XILINX®

XILINX®

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

Now Part of Xilinx

DEEPHi

Provide DPU IP + software tools
AI performance level up significantly

Xilinx owns massive industry customers
Provide wide range of applications

Telecom & Data center | Automotive
Industry IoT | Consumers
Aerospace & Defense | Broadcast

Xilinx's main area

XILINX ALL PROGRAMMABLE.

FPGA No. 1
Market and tech leader
Annual revenue of 2.5 billion US dollars

Cloud computing	Intelligent driving	Embedded devices
Cooperation with IT giants	**No. 2 semi-conductor vendor for global ADAS**	**Industry, consumer, etc.**

Typical application scenarios for AI

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

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

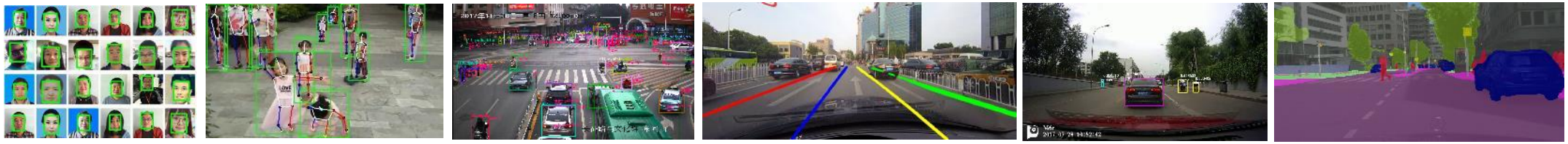
First Practical Case Using Sparse Neural Network Processor

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

Algorithm

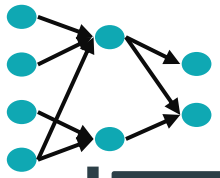
Software

DNNDK

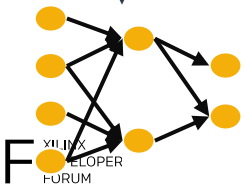
Compilation
- Compiler
- Assembler

Runtime
- Core API
- Loader
- Driver
- Tracer

Pruning



Quantization



Deep Compression

Architecture

FPGA

DPU

LSTM

ZYNQ7020

ZYNQ7020

ZU2

ZU2

ZU9

Designed by DeePhi

B1024/B1152

► Suitable for ZYNQ7020 \ZU2\ZU3\ZU5\ZU7\ZU9

B4096

► Suitable for ZU5\ZU7\ZU9 \KU115\VU9P

Being verified: B512, B2304

amazon web services EC2

Alibaba Cloud aliyun.com

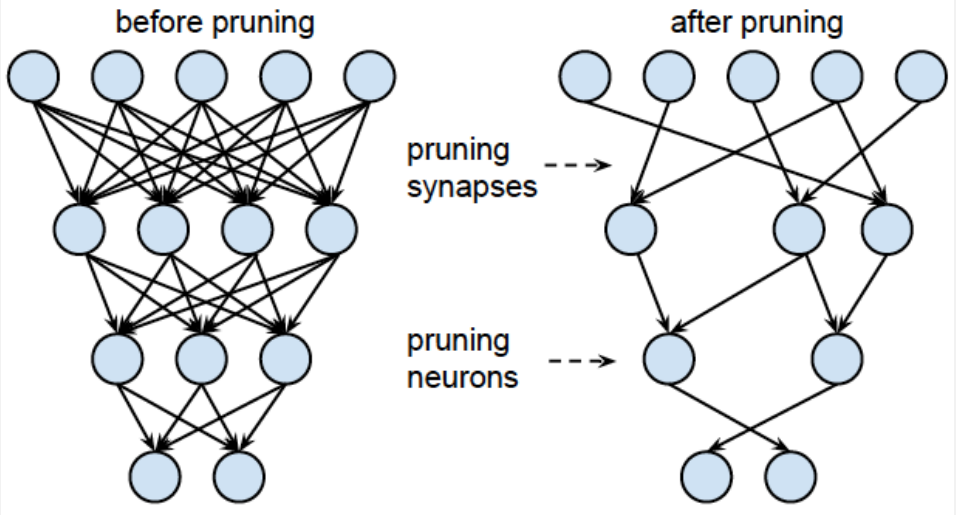
HUAWEI

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

Being designed: ZU2 Arria10, ZU3, ZU5, ZU6ev/ZU7ev, ZU15

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 — Perfor-mance

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

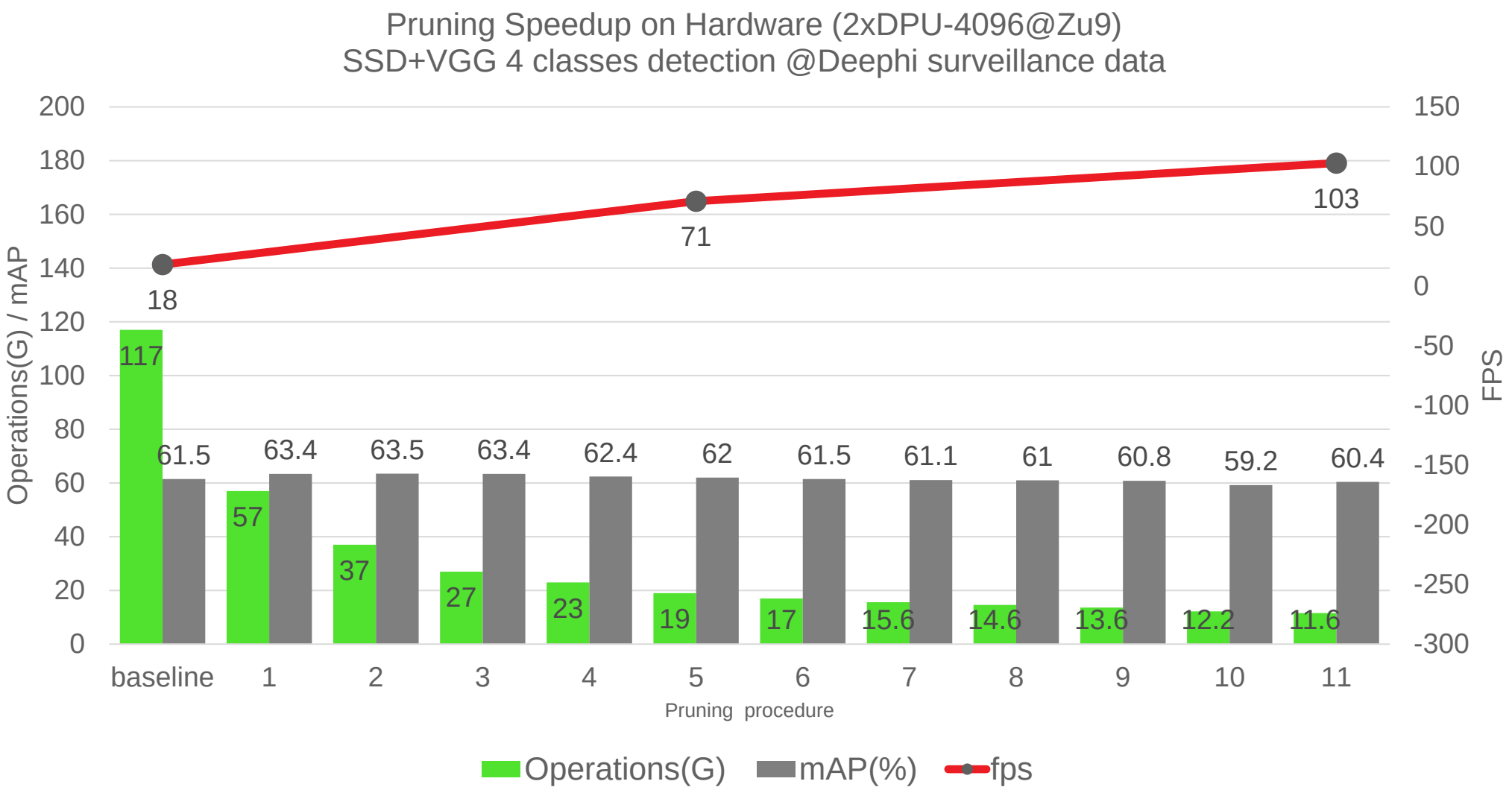
Pruning Results

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%

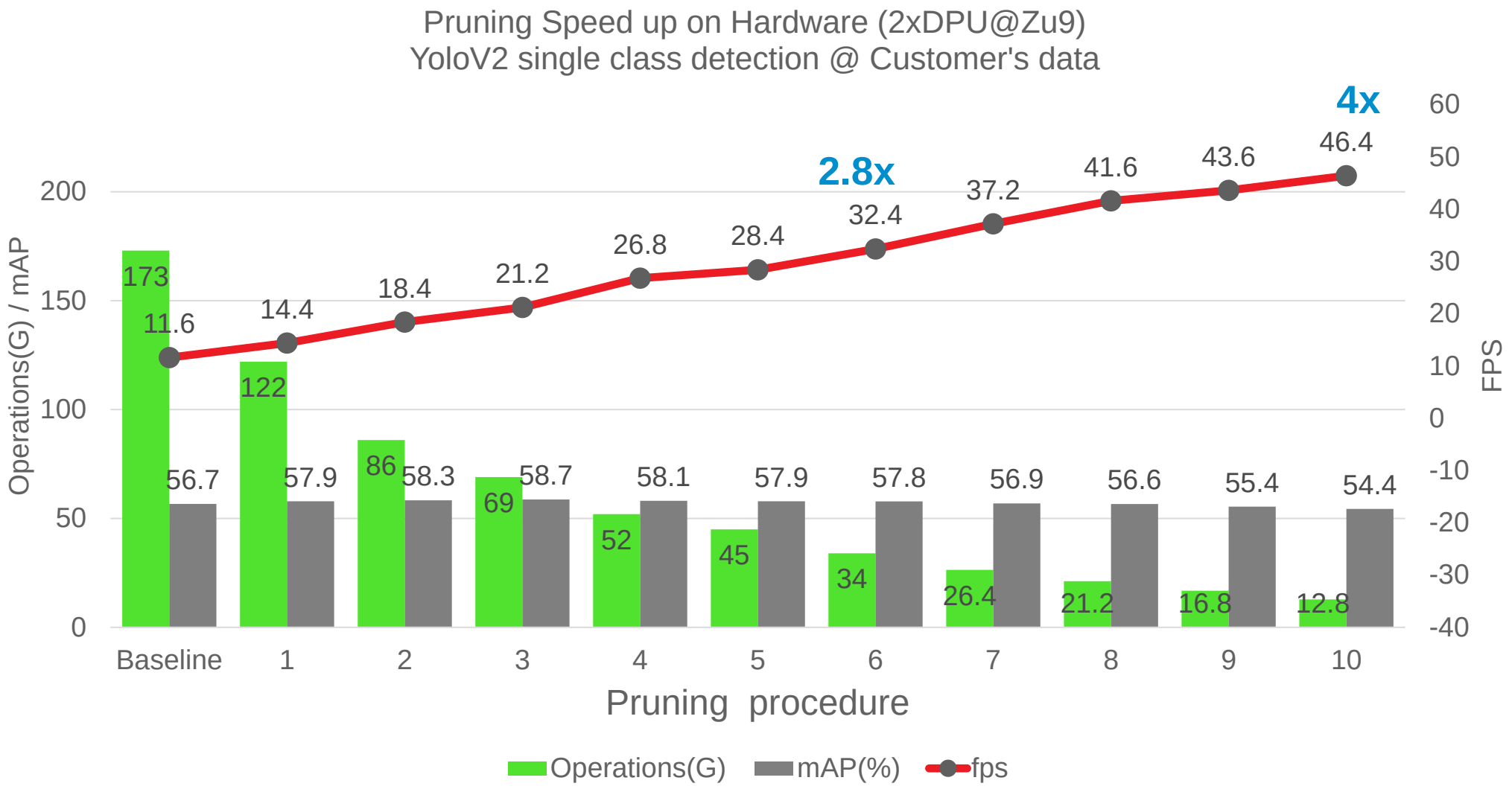
Detection Networks	Baseline mAP	Pruning Result 1			Pruning Result 2		
		mAP	ΔmAP	ratio	mAP	ΔmAP	ratio
DetectNet [17.5G]	44.46	45.7	+1.24	63%	45.12	+0.66	50%
SSD+VGG [117G]	61.5	62.0	+0.5	16%	60.4	-1.1	10%
[A] SSD+VGG [173G]	57.1	58.7	+1.6	40%	56.6	-0.5	12%
[B] Yolov2 [198G]	80.4	81.9	+1.5	28%	79.2	-1.2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

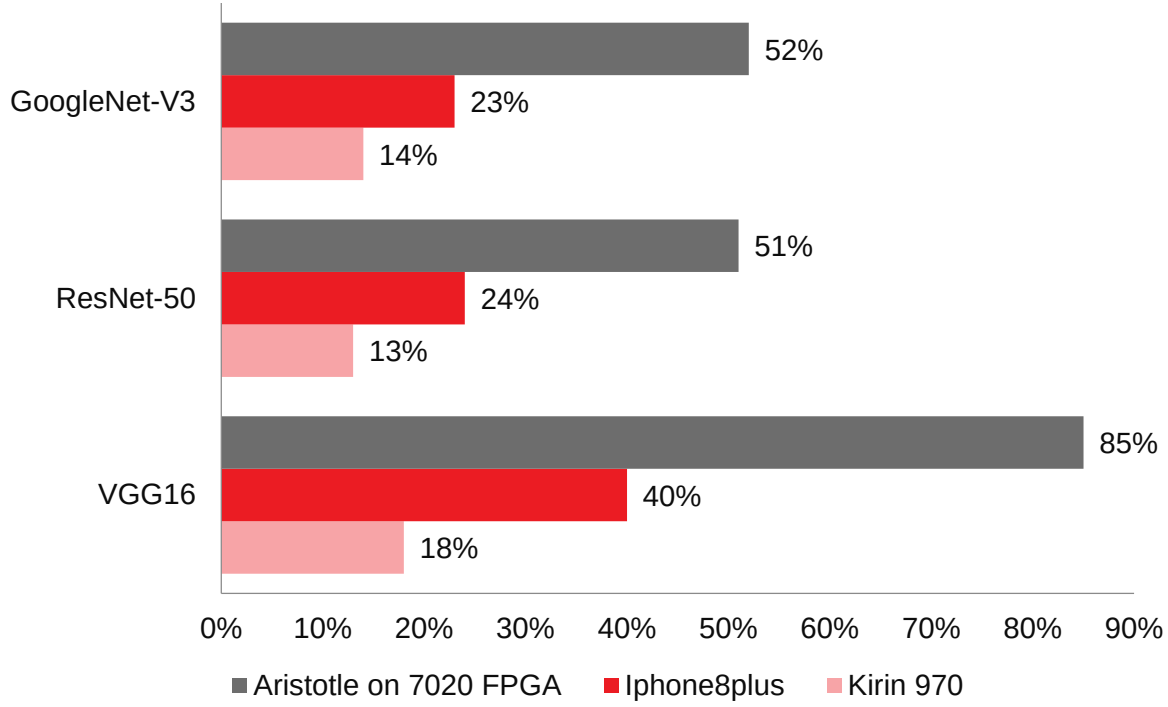
- **Unified compression tool supporting different frameworks**
- **Fully tested tools, ease of use**
- **Improved speed for pruning tool, supporting cluster**

Caffe TensorFlow Pytorch

Core advantage | Instruction set and DPU architecture

DPU Aristotle CNN accelerator

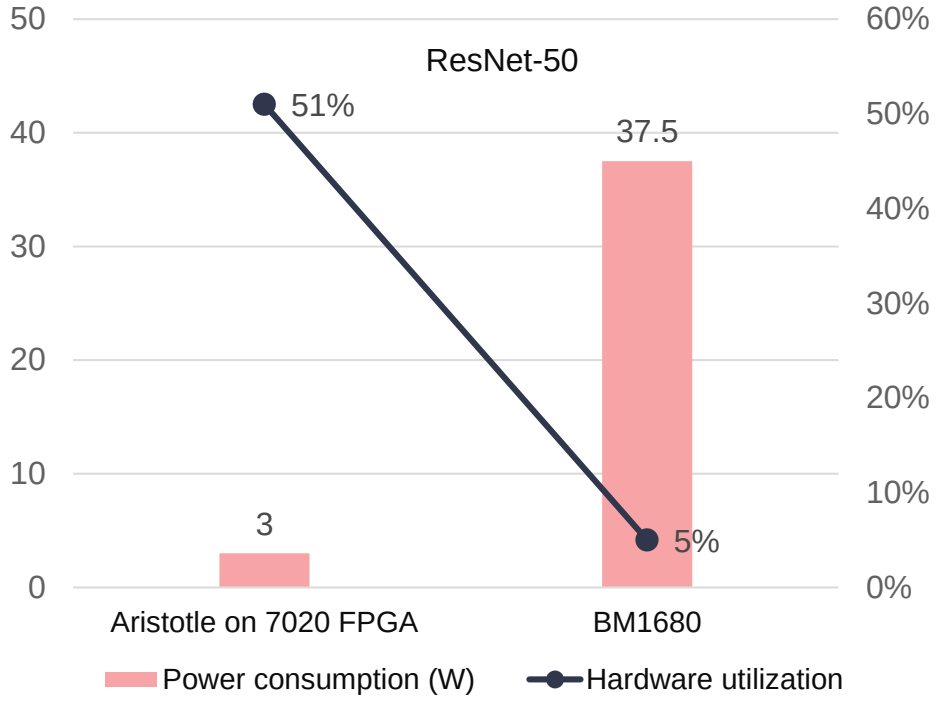
Very high hardware utilization



Source: Published results from Huawei

DPU/FPGA v.s. Sophon BM1680 (ASIC-Bitmain)

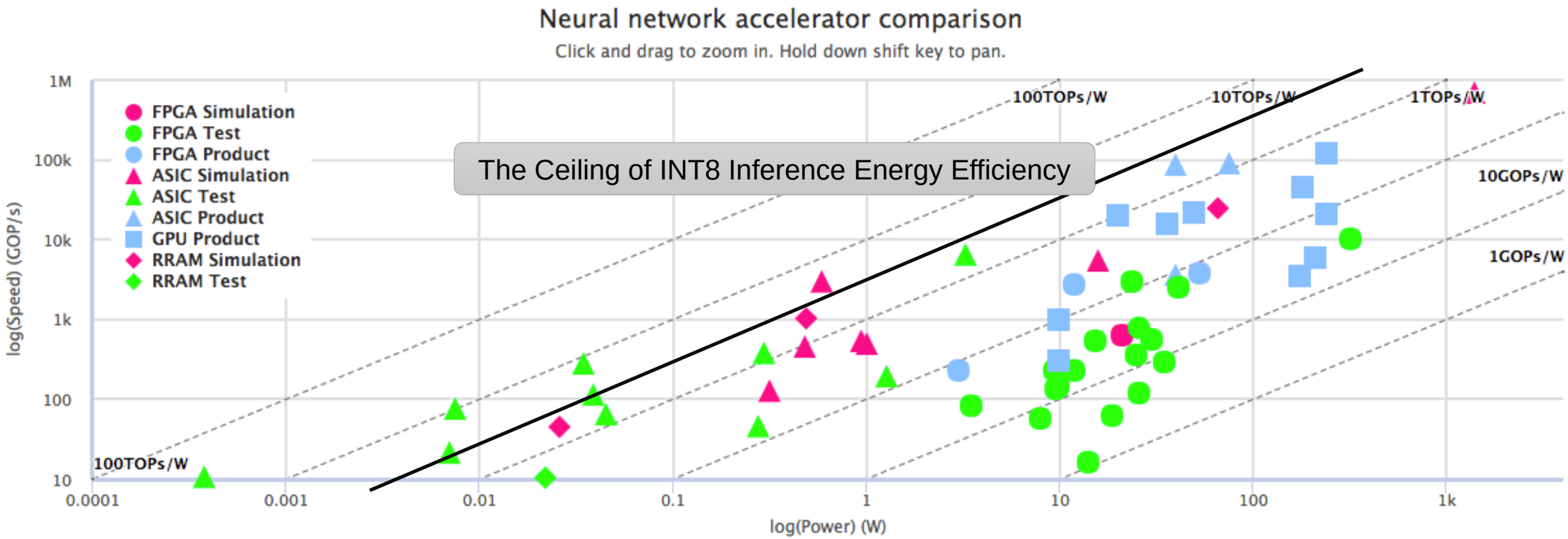
Under the same computing power performance，DeePhi's FPGA lead Sophon significantly both in power consumption and hardware utilization



Source: https://sophon.ai/product/sc1.html

Note: *For ResNet-50, Sophon is 112GOPS with 2TOPS at peak, utilization is 5.5%. Aristotle is 117GOPS with 230GOPS at peak, utilization is 51%

Current Ceiling of CNN Architecture



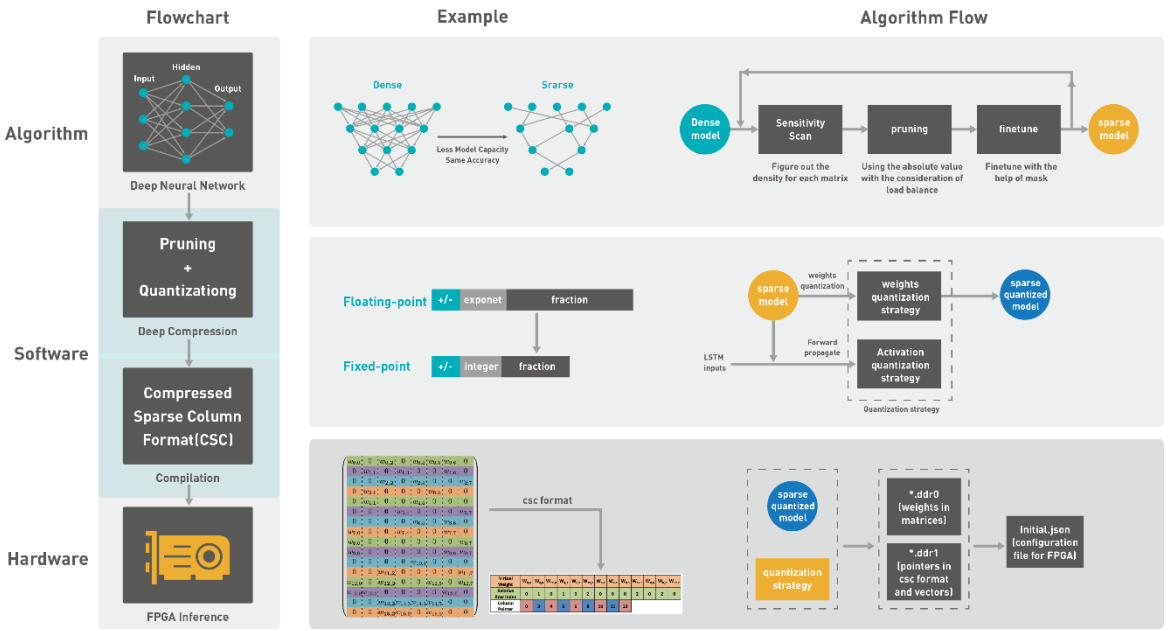
Source:http://nics-efc.org/projects/neural-network-accelerator/

INT8 improvements are slowing down and approaching the ceiling.

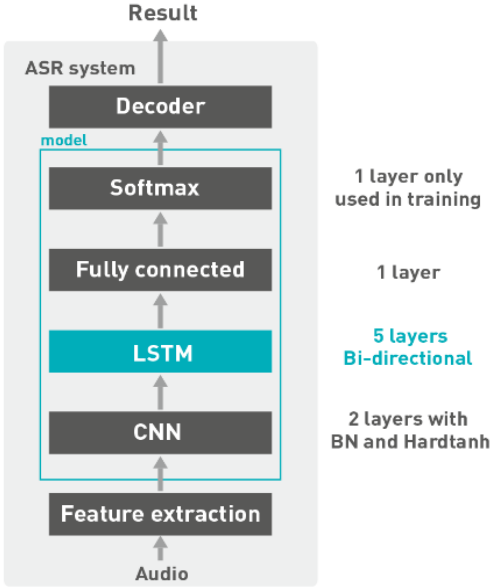
Solutions →

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

Sparsity architecture exploration



for end-to-end speech recognition



Partners

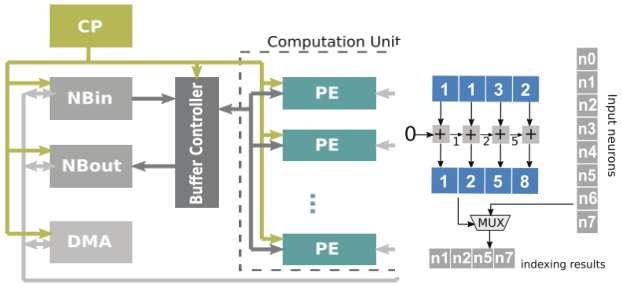
- ✓ On clouds, aiming at customers all over the world
- ✓ Already officially launched in AWS Marketplace and HUAWEI cloud (http://www.deephi.com/ddese.html)
- ✓ Now transplanting to Alibaba cloud

Features

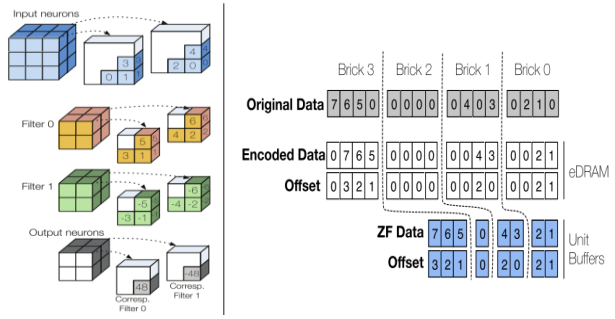
Low storage	Model compressed more than 10X with negligible loss of accuracy
Low latency	More than 2X speedup compared to GPU (P4)
Programmable	Reconfigurable for different requirements

Challenges of Sparse NN Accelerator

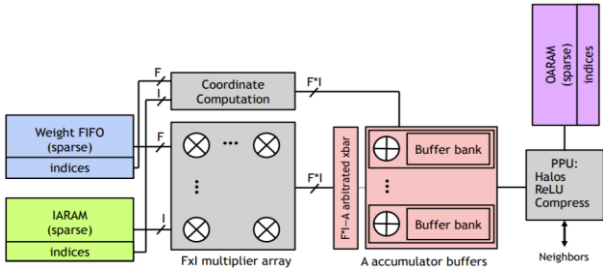
- The conflicts of the irregular pattern of mem access and the regular pattern of calculating
- Difficult to take account of the sparsity of both activation and weights at the same time.
- Additional on-chip memory requirements for indexes



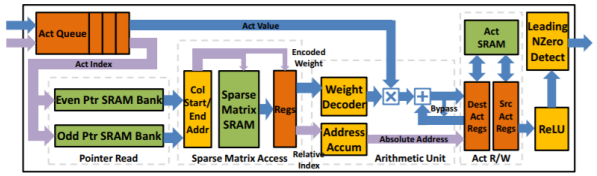
Cambricon-X, MICRO,2016



Cnvlutin, ISCA,2016



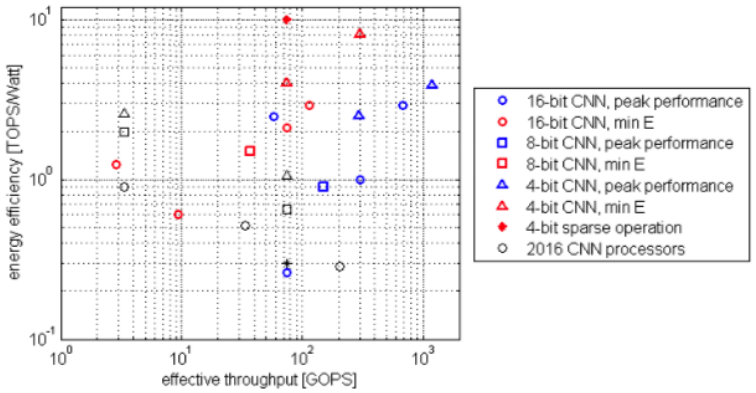
SCNN, ISCA,2017



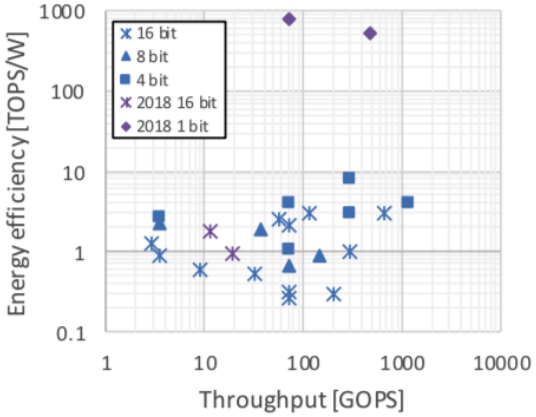
EIE, FPGA,2017

Typical Work of Sparse NN Accelerators

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

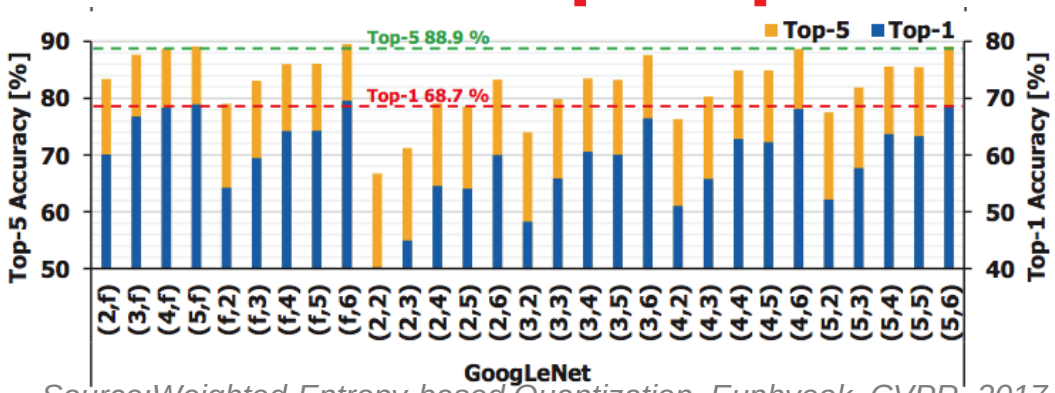
*65nm process,200Mhz,1.2v,25°C

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.

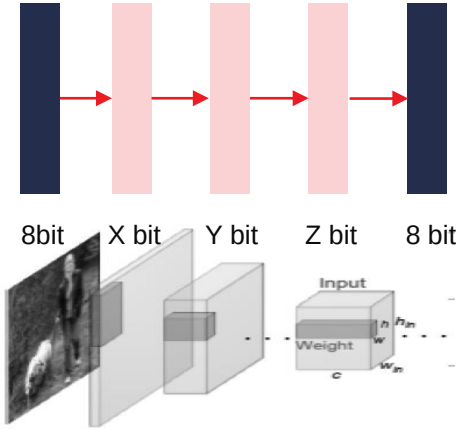
Architecture perspective: Mixed Low-Precision

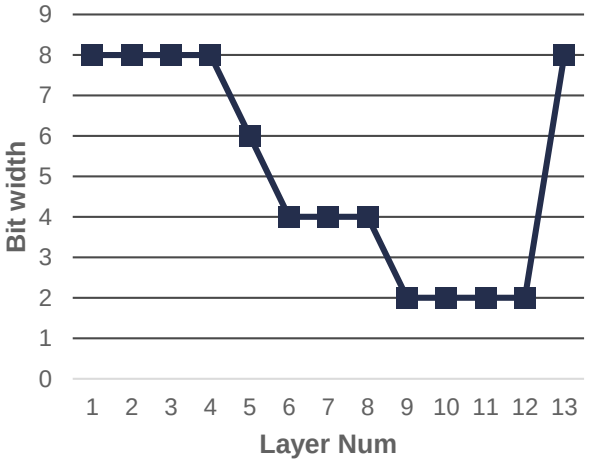


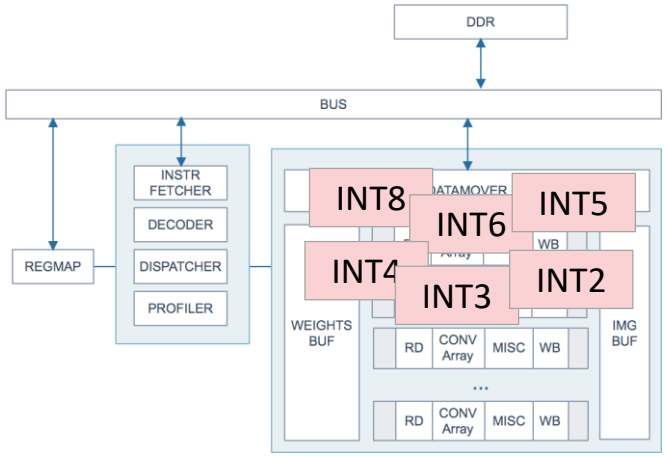
Source:Weighted-Entropy-based Quantization, Eunhyeok, CVPR, 2017-

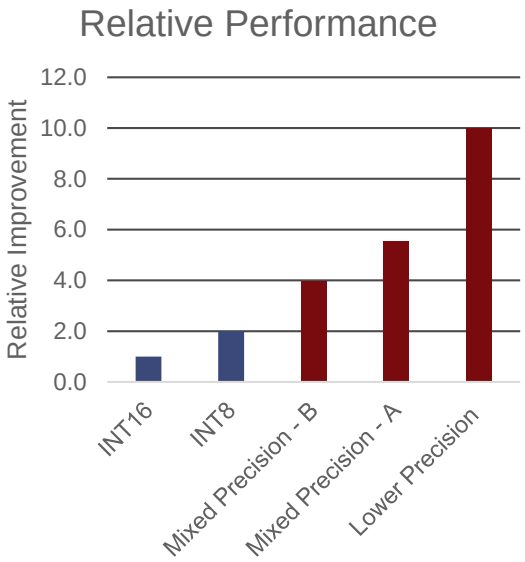
Fixed low-precision quantization already showed competitive results.

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

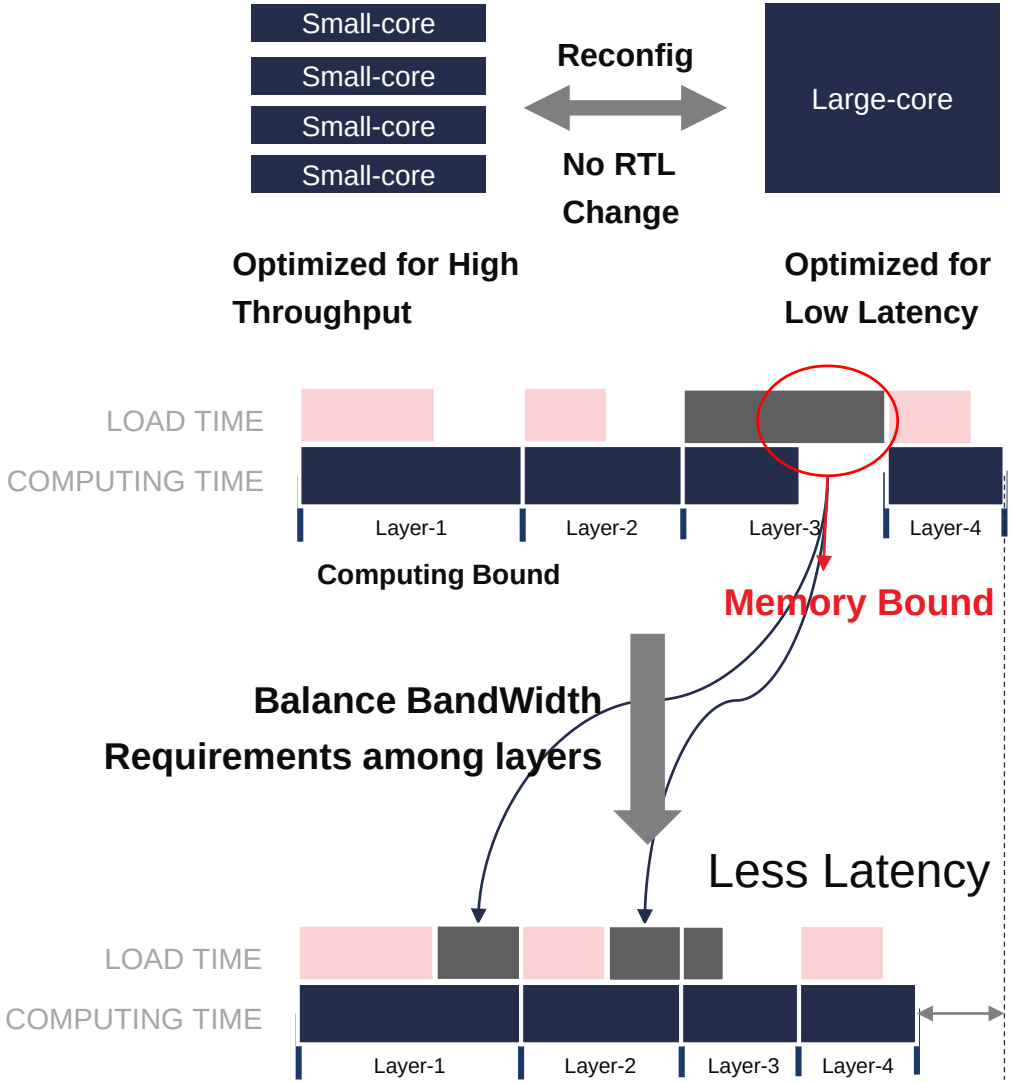
BW	2	3	4	5	6	7	8
wgt	0	0	3	22	17	10	2
act	0	0	0	16	41	13	3

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.



Software perspective

Application

- **Continuous supporting customers for products and solutions**
  - Improving surveillance products and providing more ADAS/AD demonstration to customers
- **System-level optimization for applications**
  - Accelerating time-consuming operations by FPGA and optimizing memory access

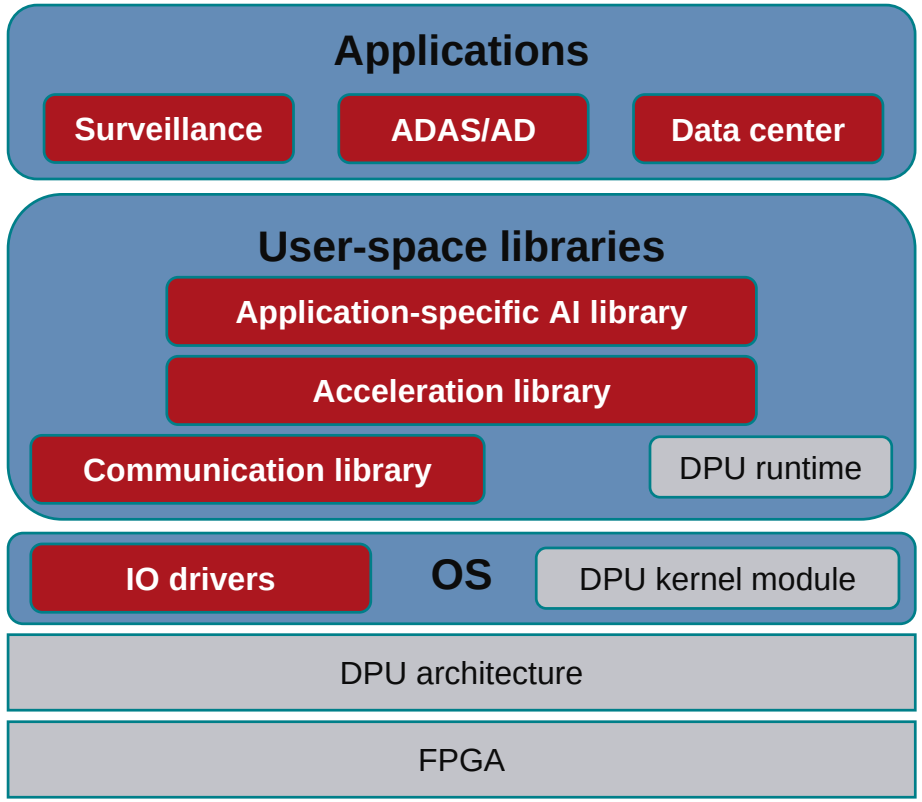
SDK

- **Providing complete SDK for surveillance customers**
  - Such as face and vehicle related SDK
- **Constructing ADAS/AD libraries for internal developers and customers**
  - Such as vehicle detection, segmentation etc.

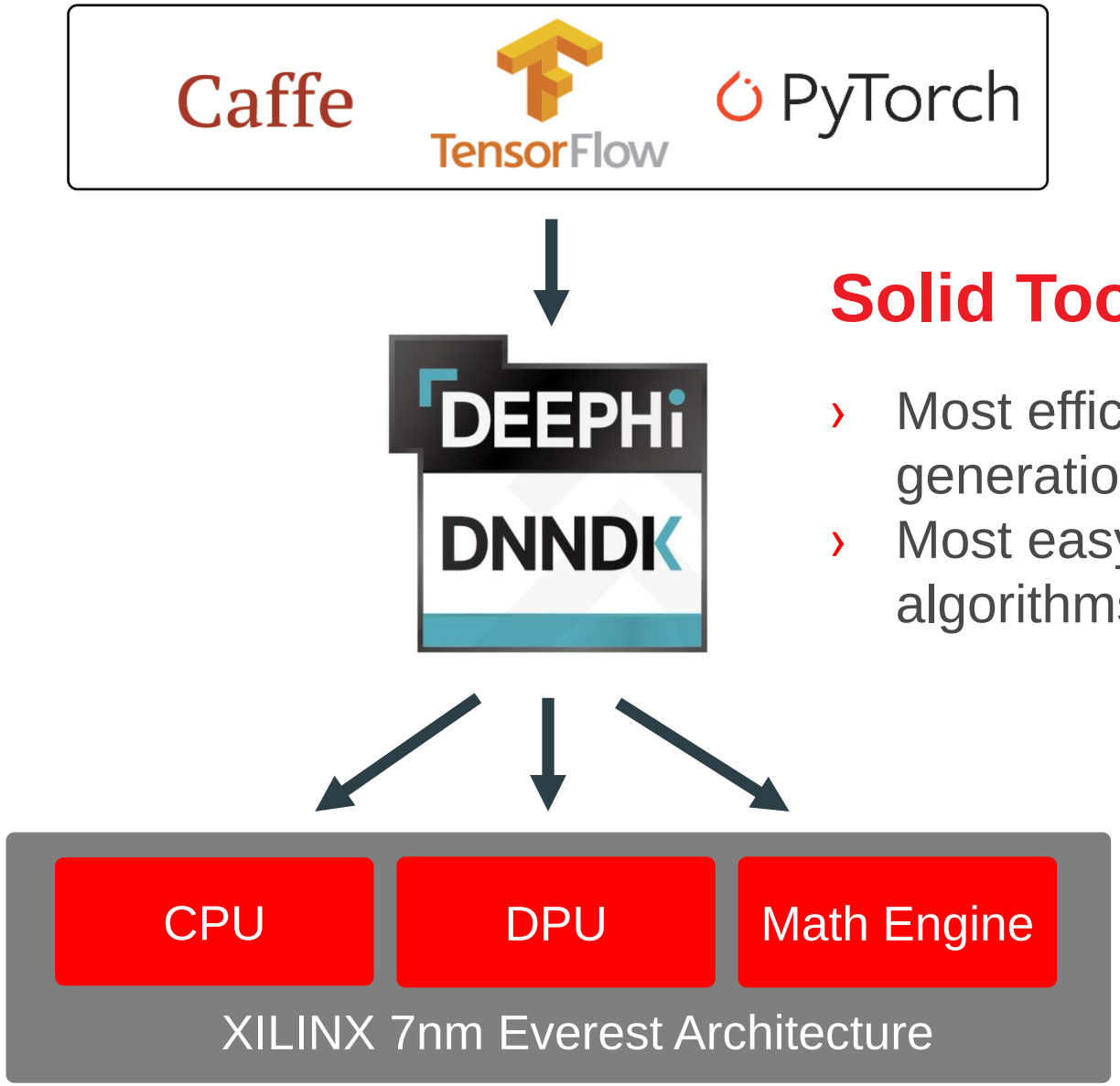
Embedded

- **Providing system for evaluation and product boards**
  - From ZU2 to ZU11
- **Developing more IO drivers**
  - Such as USB 3.0, MIPI etc.
- **Researching other system related with our products**

Software team will provide full stack solutions for AI applications



DNNDK perspective

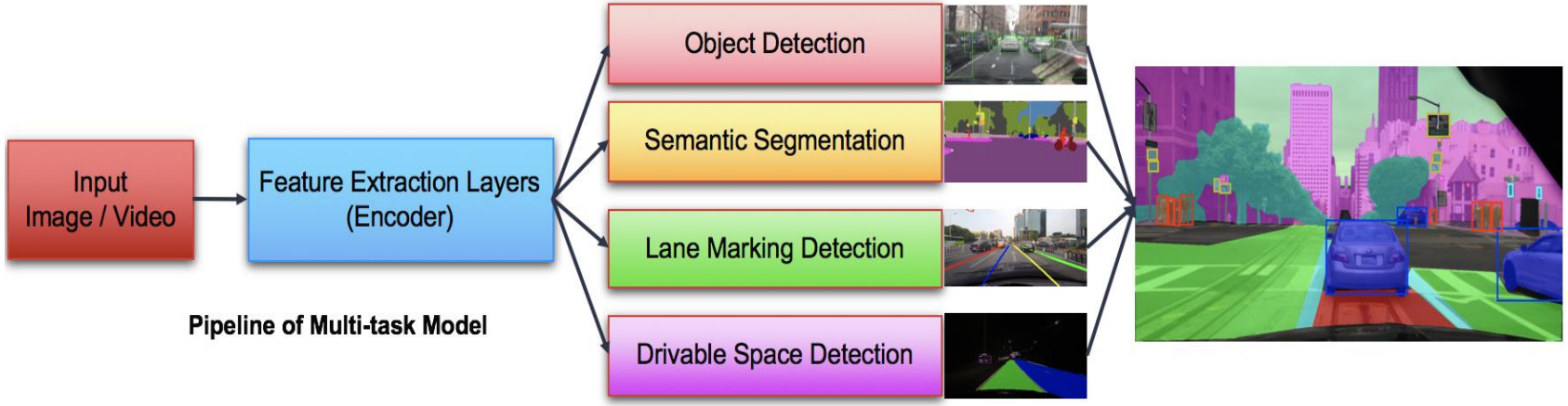


Solid Toolchain Stack for XILINX ACAP

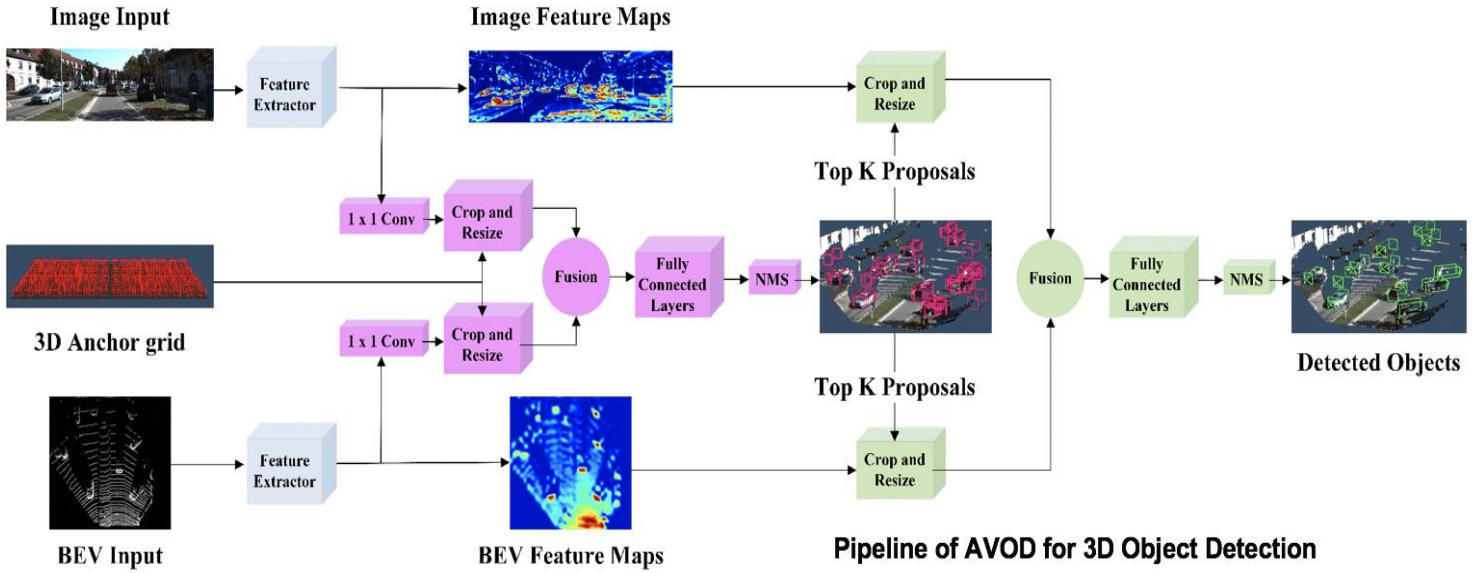
- Most efficiency solution for ML on XILINX next generation computing platform
- Most easy-to-use & productive toolchain for ML algorithms deployment

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 Multi-task Model

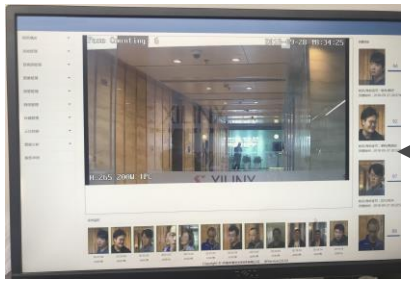


Pipeline of AVOD for 3D Object Detection

System perspective: Video Surveillance in single FPGA

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

ML+X

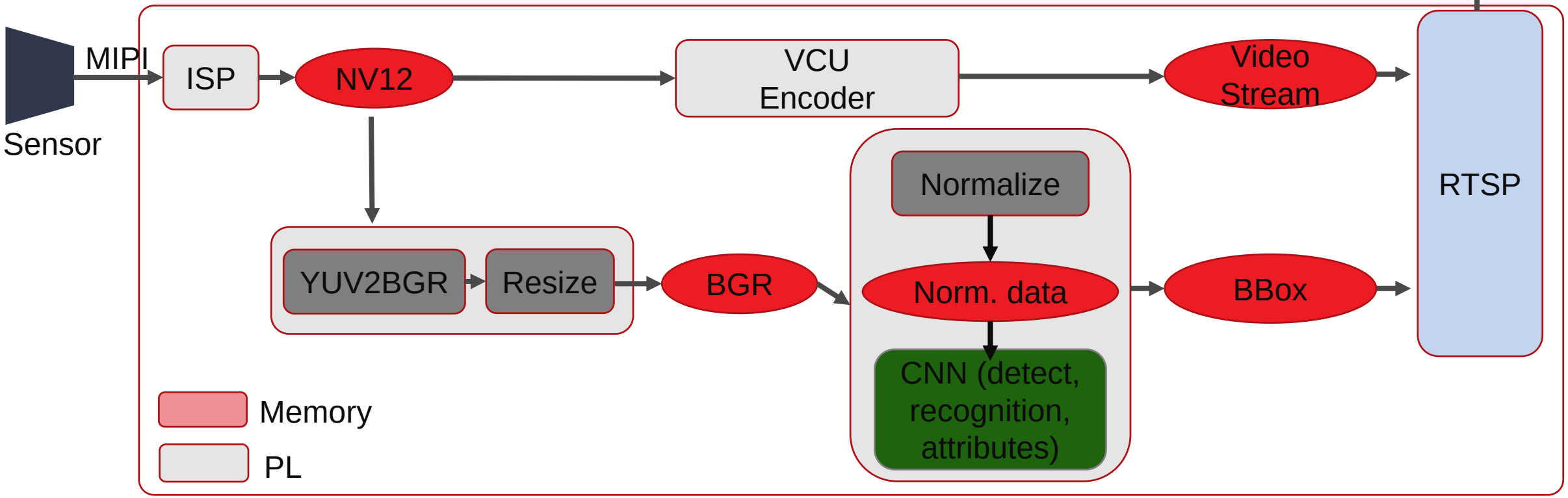


Computer

HUB

Ethernet

Single Chip Solution



This solution needs to further enhance ISP functionality

Attend presentations and workshops for more details

DAY 1

	Crystal	Piedmont
9am-10am		
10am-11am	Machine learning for embedded systems	
11am-12pm		
12pm-1pm	Building ML vision systems with SDSoC	
1pm-2pm		Machine learning with DeePhi
2pm-3pm	Machine learning for embedded systems	Machine learning for embedded
3pm-4pm		Machine Learning with SDSoC for EV
4pm-5pm	Machine learning for embedded systems	

DAY 2

	Regency 1	Sacramento
9am-10am		
10am-11am		
11am-12pm		
12pm-1pm	Machine learning with DeePhi	
1pm-2pm	Ford ADAS	
2pm-3pm	Machine learning for embedded	Building ML vision systems with SDSoC
3pm-4pm	Machine Learning with SDSoC for EV	
4pm-5pm	ML Expert panel	Machine learning for embedded systems
5pm-6pm		

Presentation

Interactive

Lab (own laptop required)

XILINX

XDF XILINX DEVELOPER FORUM

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)
- **Release Plan**
  - First version: 2019Q1