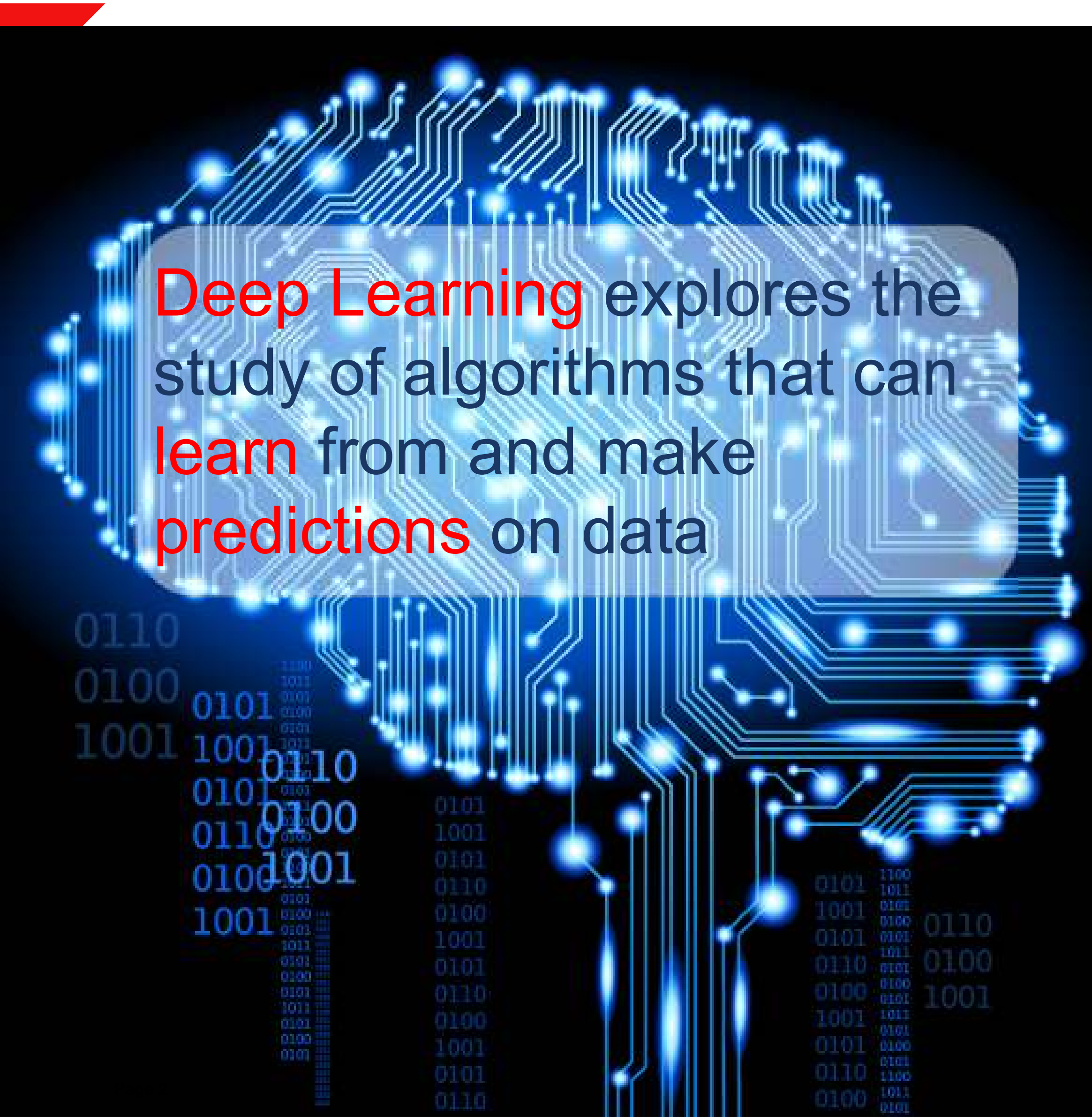
Building the Adaptable, Intelligent World

Machine Learning Suite

Kamran Khan

Sr Product Manager, AI and ML

XILINX



Deep Learning explores the study of algorithms that can learn from and make predictions on data

Deep Learning is Re-Defining many Applications

Cloud Acceleration

Security

Ecommerce Social

Financial

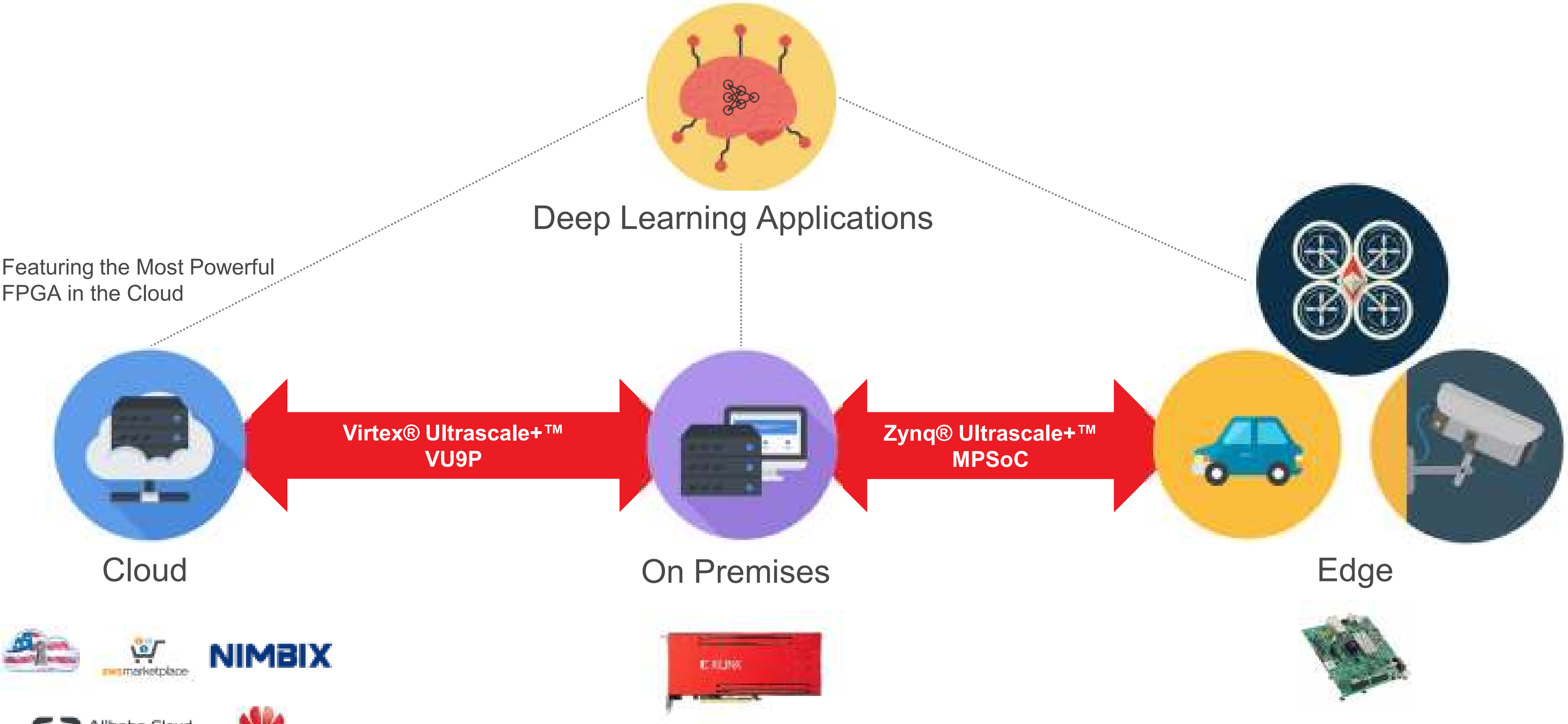
Surveillance

Industrial IOT

Medical Bioinformatics

Autonomous Vehicles

Accelerating AI Inference into Your Cloud Applications



Deep Learning Applications

Featuring the Most Powerful FPGA in the Cloud

Virtex® Ultrascale+™ VU9P

Zynq® Ultrascale+™ MPSoC

Cloud

On Premises

Edge

aws marketplace

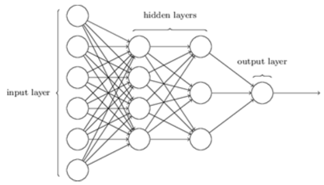
NIMBIX

Alibaba Cloud aliyun.com

HUAWEI

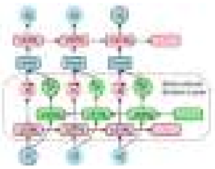
Overlay Architecture
Custom Processors Exploiting Xilinx FPGA Flexibility

- Customized overlays with ISA architecture for optimized implementation
- Easy plug and play with Software Stack

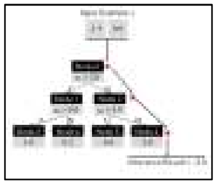


MLP Engine
Scalable sparse and dense implementation

xDNN – CNN Engine for Large 16 nm Xilinx Devices
Deephi DPU – Flexible CNN Engine with Embedded Focus
CHaiDNN – HLS based open source offering

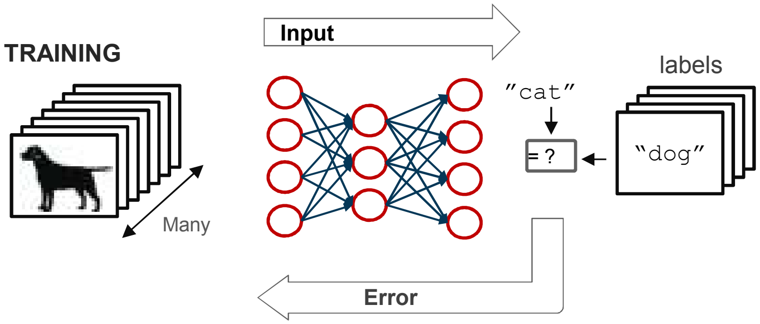


Deephi ESE
LSTM Speech to Text engine



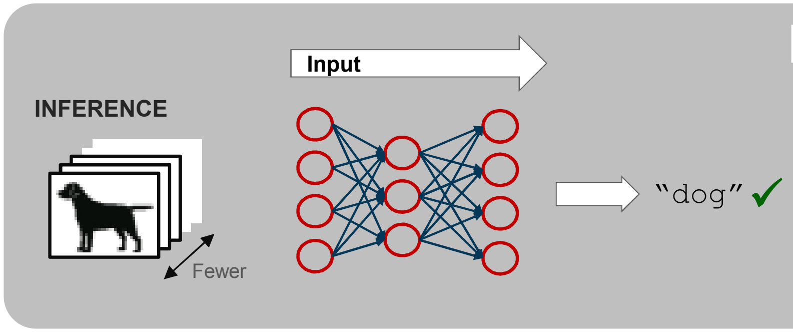
Random Forest
Configurable RF classification

Machine Learning Inference is Xilinx Focus



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

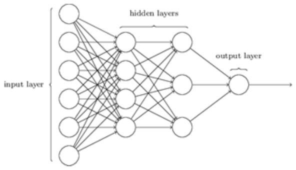
Focus

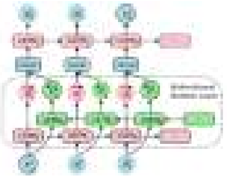


Inference: Using trained models to predict/estimate outcomes from new observations in efficient deployments

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

Deep Learning Models





Multi-Layer Perceptron

- Classification
- Universal Function Approximator
- Autoencoder

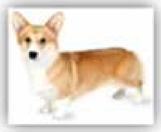
Convolutional Neural Network

- Feature Extraction
- Object Detection
- Image Segmentation

Recurrent Neural Network

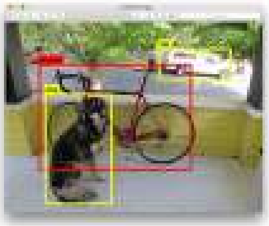
- Sequence and Temporal Data
- Speech to Text
- Language Translation

Classification

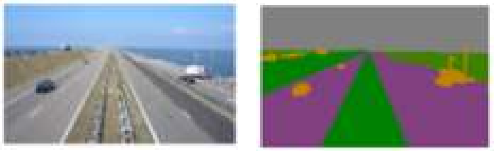


"Dog"

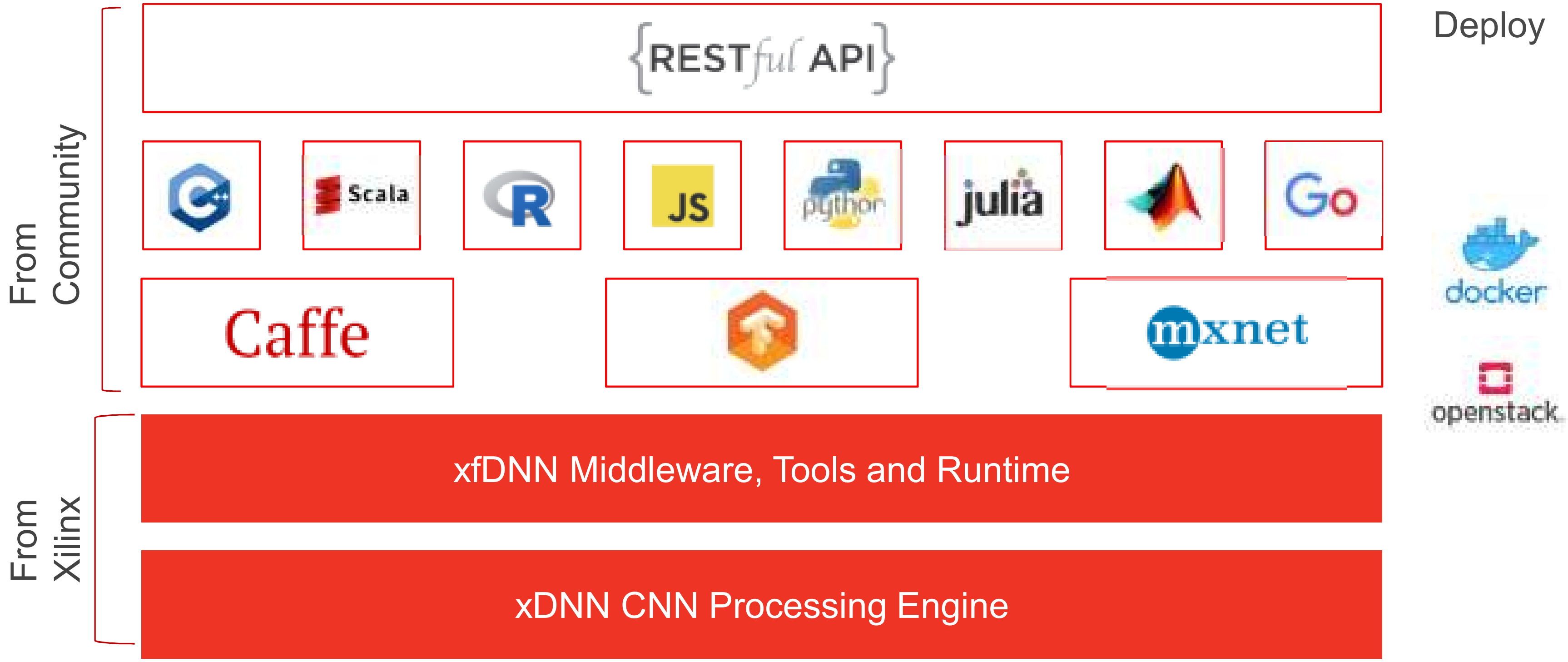
Object Detection



Segmentation



Seamless Deployment with Open Source Software



Deploy

From Community

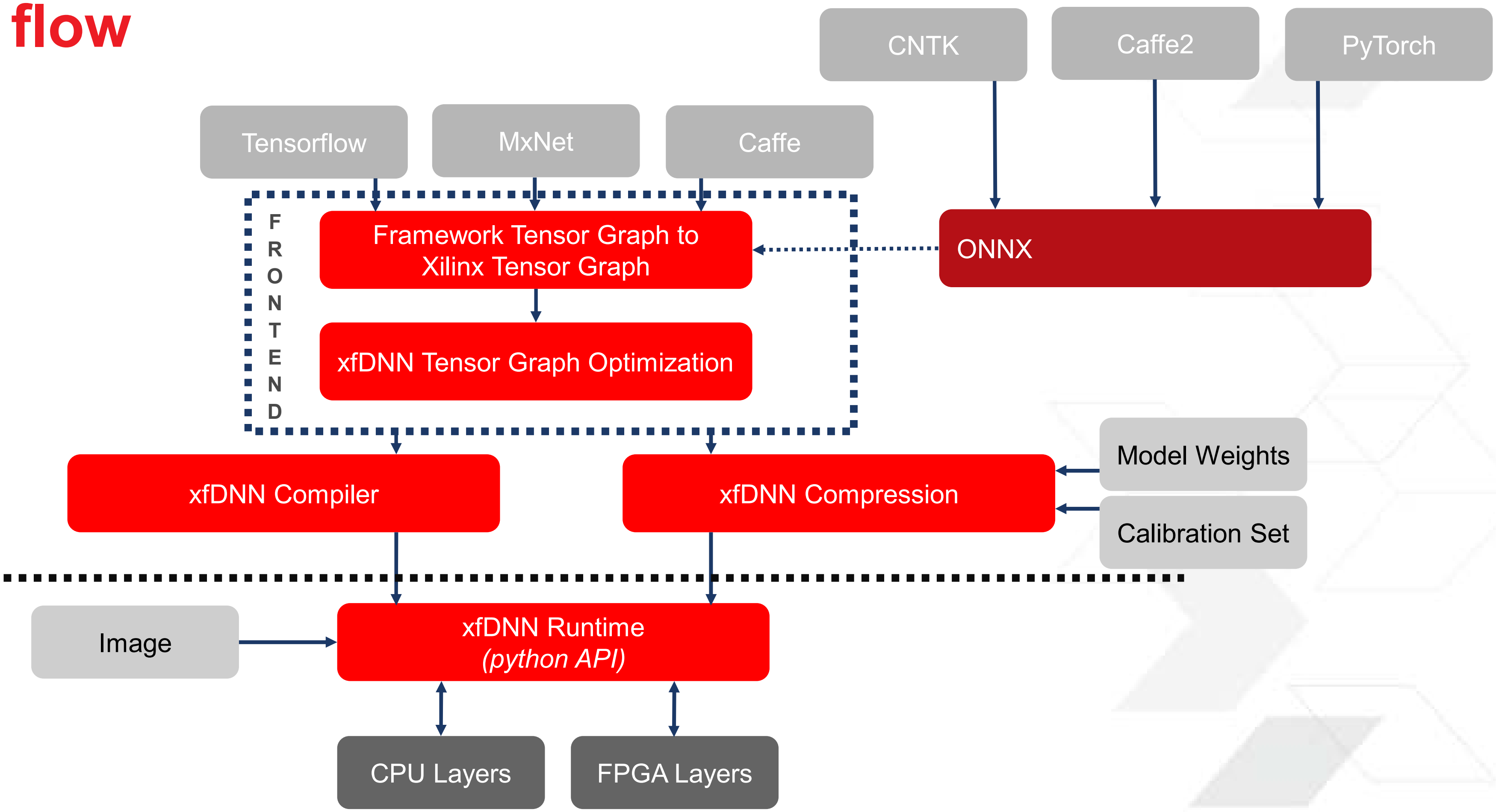
From Xilinx

xfDNN Middleware, Tools and Runtime

xDNN CNN Processing Engine

*TensorFlow Q4 2017

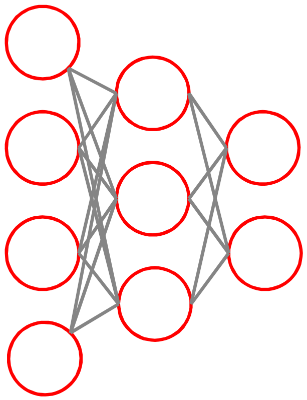
xfDNN flow



https://github.com/Xilinx/ml-suite

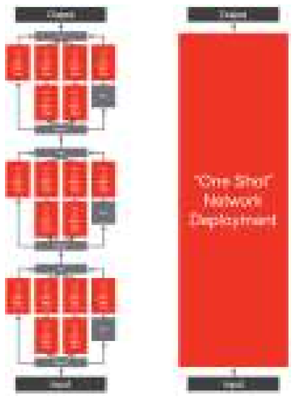
xfDNN Inference Toolbox

Graph Compiler



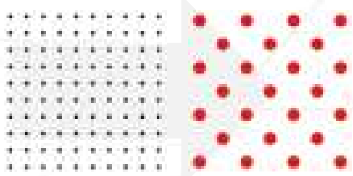
- Python tools to quickly compile networks from common Frameworks – Caffe, MxNet and Tensorflow

Network Optimization



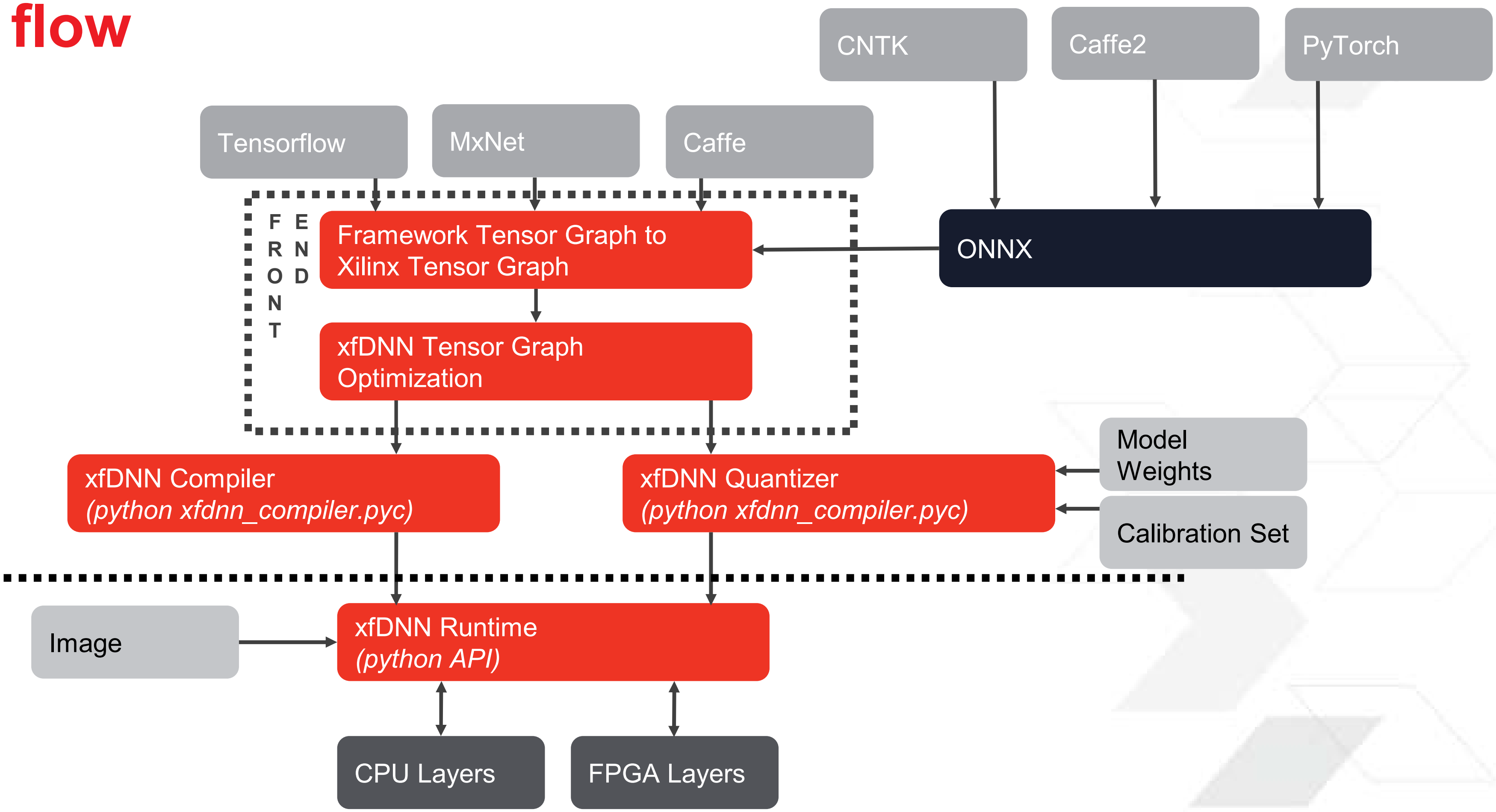
- Automatic network optimizations for lower latency by fusing layers and buffering on-chip memory

xfDNN Quantizer



- Quickly reduce precision of trained models for deployment
- Maintains 32bit accuracy at 8 bit within 2%

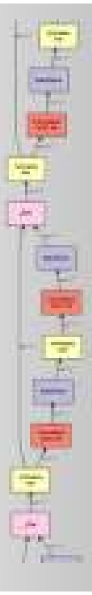
xfDNN flow



https://github.com/Xilinx/ML-Development-Stack-From-Xilinx

xfDNN Graph Compiler

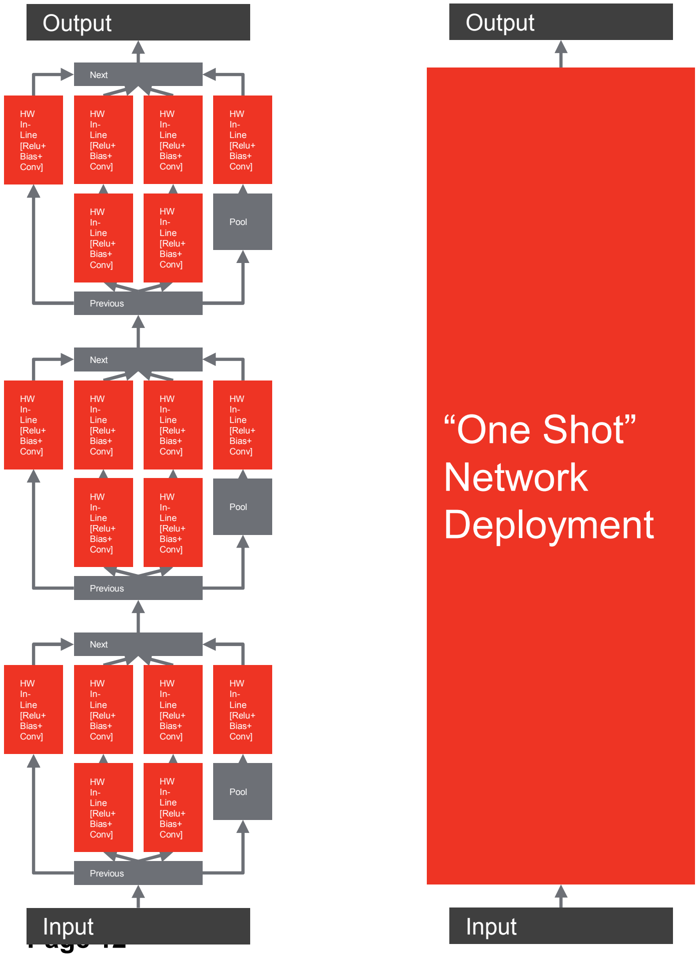
Pass in a Network



xfDNN
Graph Compiler

Microcode for xDNN is Produced

xfDNN Network Deployment



Fused Layer Optimizations

- Compiler can merge nodes
  - (Conv or EltWise)+Relu
  - Conv + Batch Norm
- Compiler can split nodes
  - Conv 1x1 stride 2 -> Maxpool+Conv 1x1 Stride 1

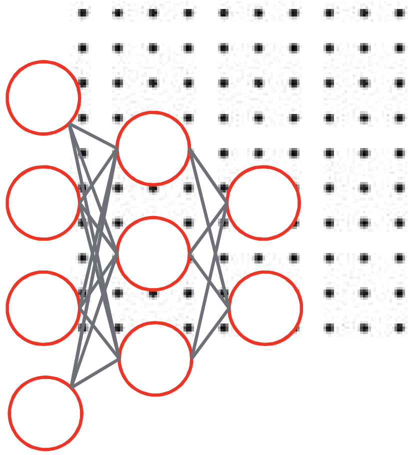
On-Chip buffering reduces latency and increases throughput

- xfDNN analyzes network memory needs and optimizes scheduler
  - For Fused and “One Shot” Deployment

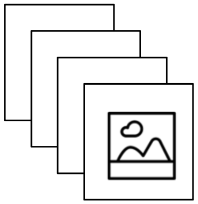
“One Shot” deploys entire network to FPGA

- Optimized for fast, low latency inference
- Entire network, schedule and weights loaded only once to FPGA

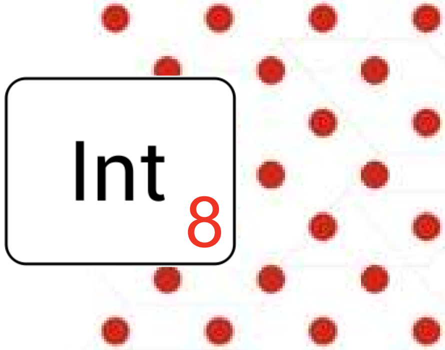
xfDNN Quantizer: Fast and Easy



1) Provide FP32 network and model
   - E.g., prototxt and caffemodel



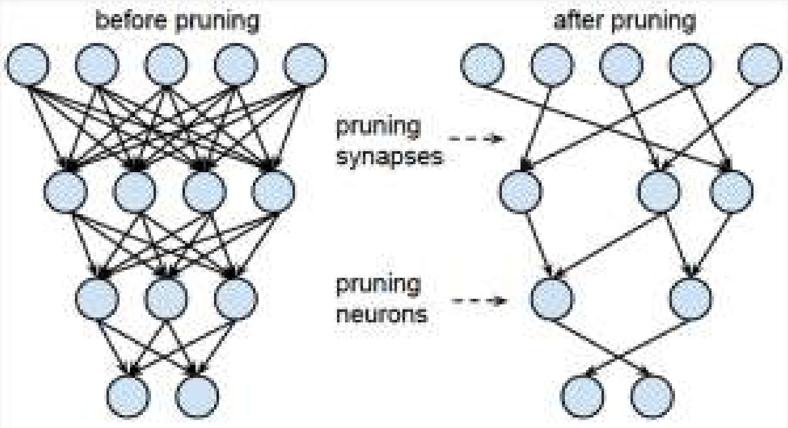
2) Provide a small sample set, no labels required
   - 16 to 512 images



3) Specify desired precision
   - Quantizes to <8 bits to match Xilinx’s DSP

Xilinx Pruning Overview

Deep compression
Makes algorithm smaller and lighter



Highlight

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

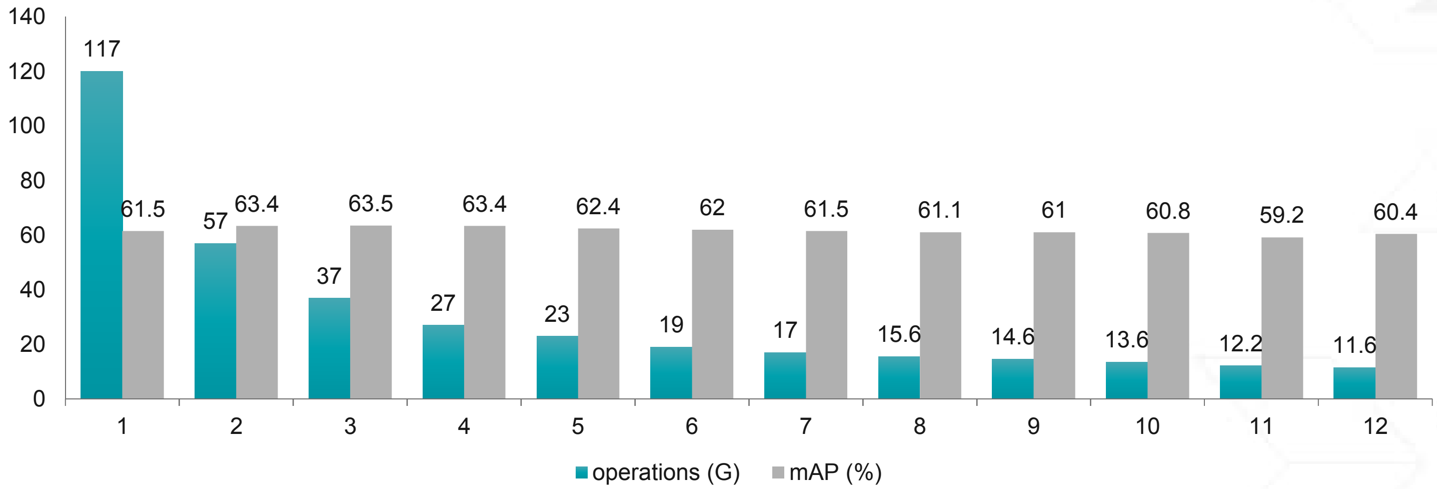
Compression efficiency

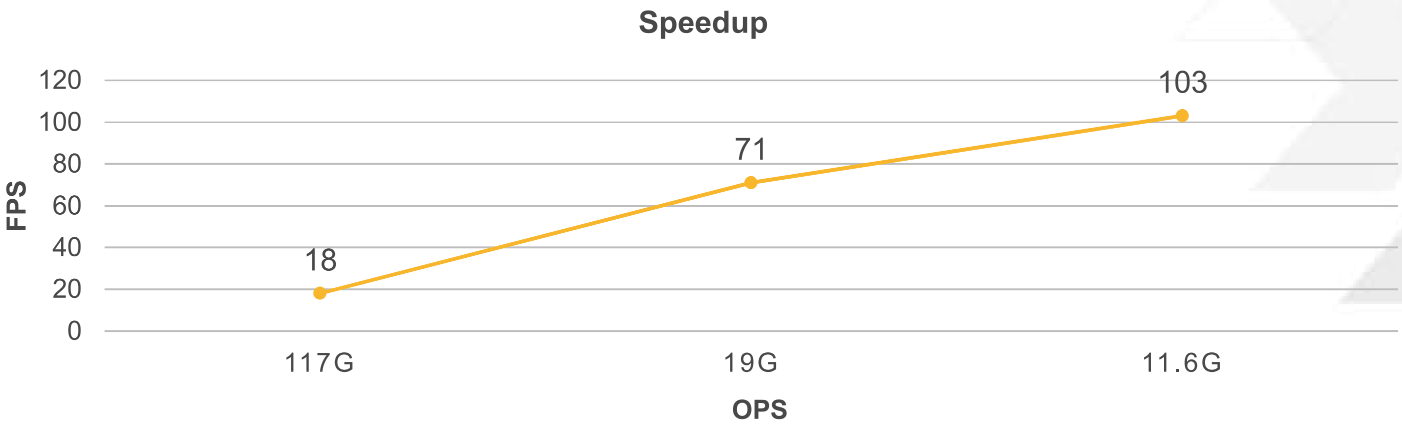
Deep Compression Tool can achieve significant compression on **CNN** and **RNN**

Accuracy

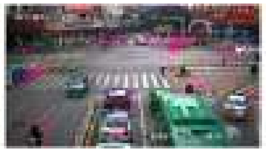
Algorithm can be **compressed 7 times without losing accuracy** under SSD object detection framework

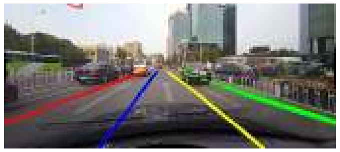
Pruning Example - SSD

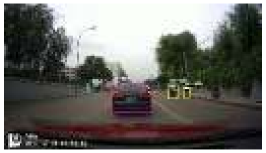




Supported DNN (Deep Neural Network) by Applications





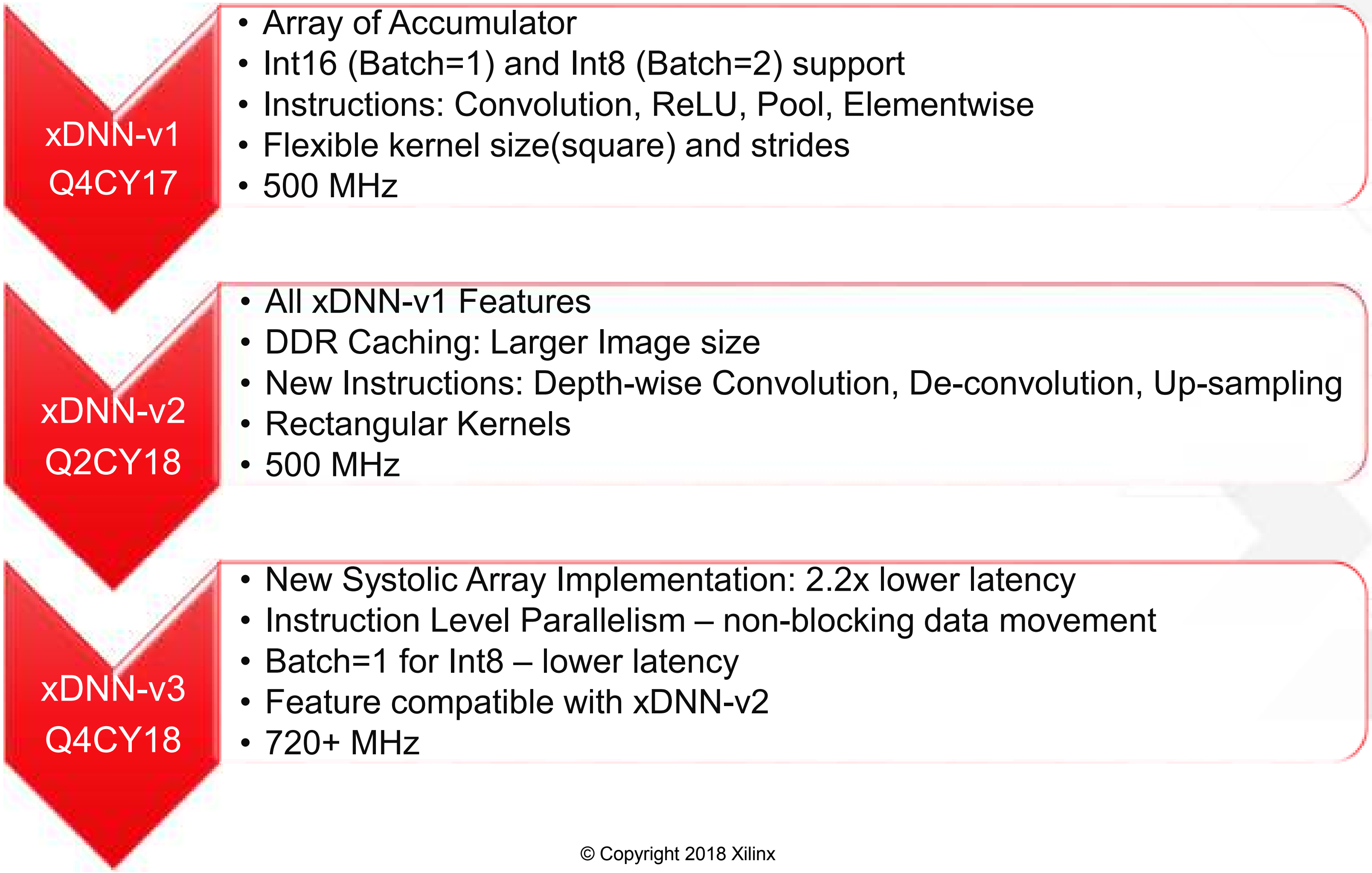


Application	NTT Request	Function	Algorithm
Face		Face detection	SSD, Densebox
		Landmark Localization	Coordinates Regression
		Face recognition	ResNet + Triplet / A-softmax Loss
		Face attributes recognition	Classification and regression
Pedestrian	1	Pedestrian Detection (Crowd Volume)	SSD
		Pose Estimation	Coordinates Regression
		Person Re-identification	ResNet + Loss Fusion
Video Analytics	1	Object detection	SSD, RefineDet
		Pedestrian Attributes Recognition	GoogleNet
		Car Attributes Recognition	GoogleNet
	1	Car Logo Detection	DenseBox
	1	Car Logo Recognition	GoogleNet + Loss Fusion
	1	License Plate Detection	Modified DenseBox
	1	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

xDNN Process Engine

XILINX

Rapid Feature and Performance Improvement



xDNN-v1
Q4CY17

- Array of Accumulator
- Int16 (Batch=1) and Int8 (Batch=2) support
- Instructions: Convolution, ReLU, Pool, Elementwise
- Flexible kernel size(square) and strides
- 500 MHz

xDNN-v2
Q2CY18

- All xDNN-v1 Features
- DDR Caching: Larger Image size
- New Instructions: Depth-wise Convolution, De-convolution, Up-sampling
- Rectangular Kernels
- 500 MHz

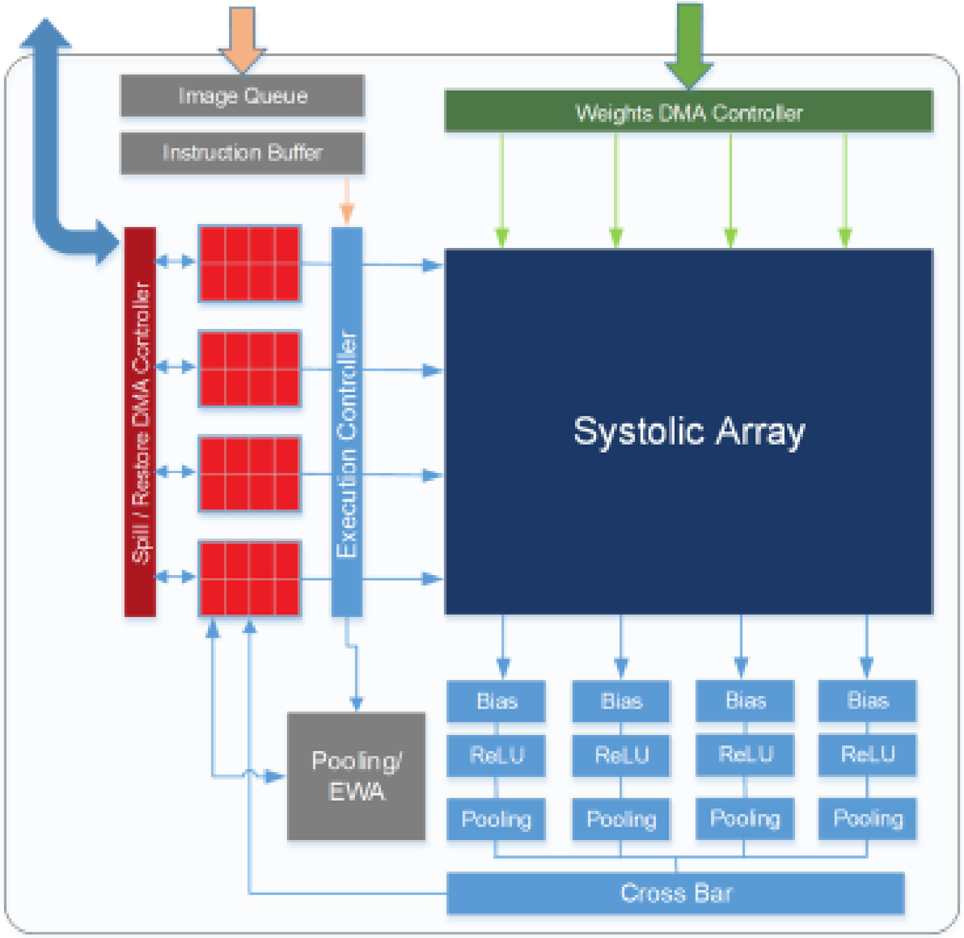
xDNN-v3
Q4CY18

- New Systolic Array Implementation: 2.2x lower latency
- Instruction Level Parallelism – non-blocking data movement
- Batch=1 for Int8 – lower latency
- Feature compatible with xDNN-v2
- 720+ MHz

XDNN v3 Feature Set

Features		Description	
Supported Operations	Convolution / Deconvolution / Convolution Transpose	Kernel Sizes	W: 1-15; H:1-15
		Strides	W: 1,2,4,8; H: 1,2,4,8
		Padding	Same, Valid
		Dilation	Factor: 1,2,4
		Activation	ReLU/pReLU
		Bias	Value Per Channel
		Scaling	Scale & Shift Value Per Channel
	Max Pooling	Kernel Sizes	W: 1-15; H:1-15
		Strides	W: 1,2,4,8; H: 1,2,4,8
		Padding	Same, Valid
	Avg Pooling	Kernel Sizes	W: 1-15; H:1-15
		Strides	W: 1,2,4,8; H: 1,2,4,8
		Padding	Same, Valid
	Element-wise Add	Width & Height must match; Depth can mismatch.	
	Memory Support	On-Chip Buffering, DDR Caching	
	Expanded set of image sizes	Square, Rectangular	
	Upsampling	Strides	Factor: 2,4,8,16
Miscellaneous	Precision	Int16-bit or Int8-bit	

Xilinx DNN Processor (xDNN)



- Configurable Overlay Processor
- DNN Specific Instruction Set
  - Convolution, Max Pool etc.
- Any Network, Any Image Size
- High Frequency & High Compute Efficiency
- Compile and run new networks

ML Suite Overlays with xDNN Processing Engines

Adaptable

> AI algorithms are changing rapidly

> Adjacent acceleration opportunities

Realtime

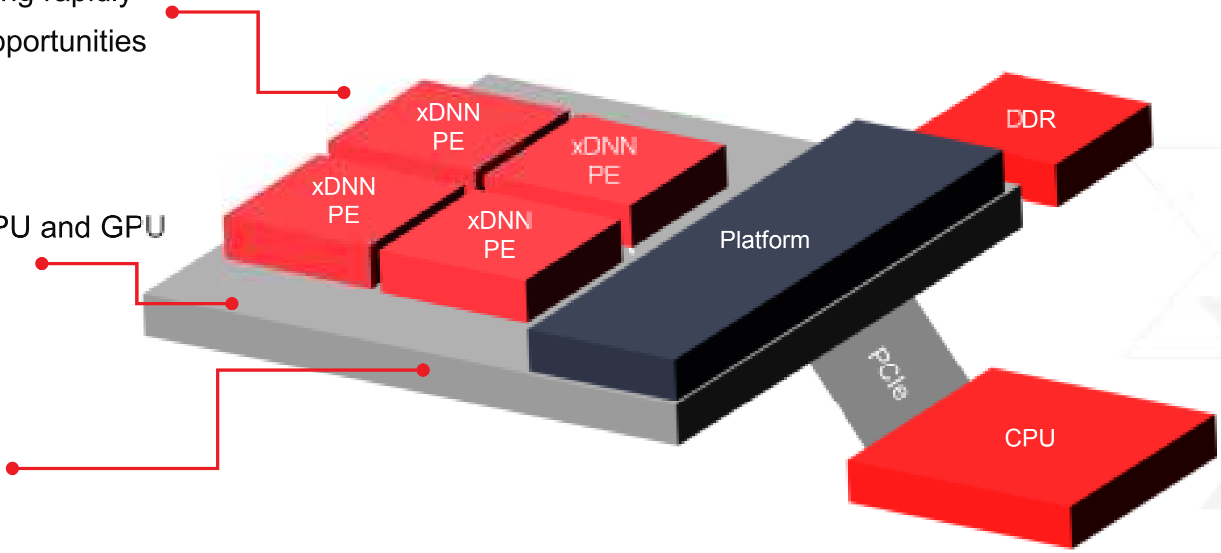
> 10x Low latency than CPU and GPU

> Data flow processing

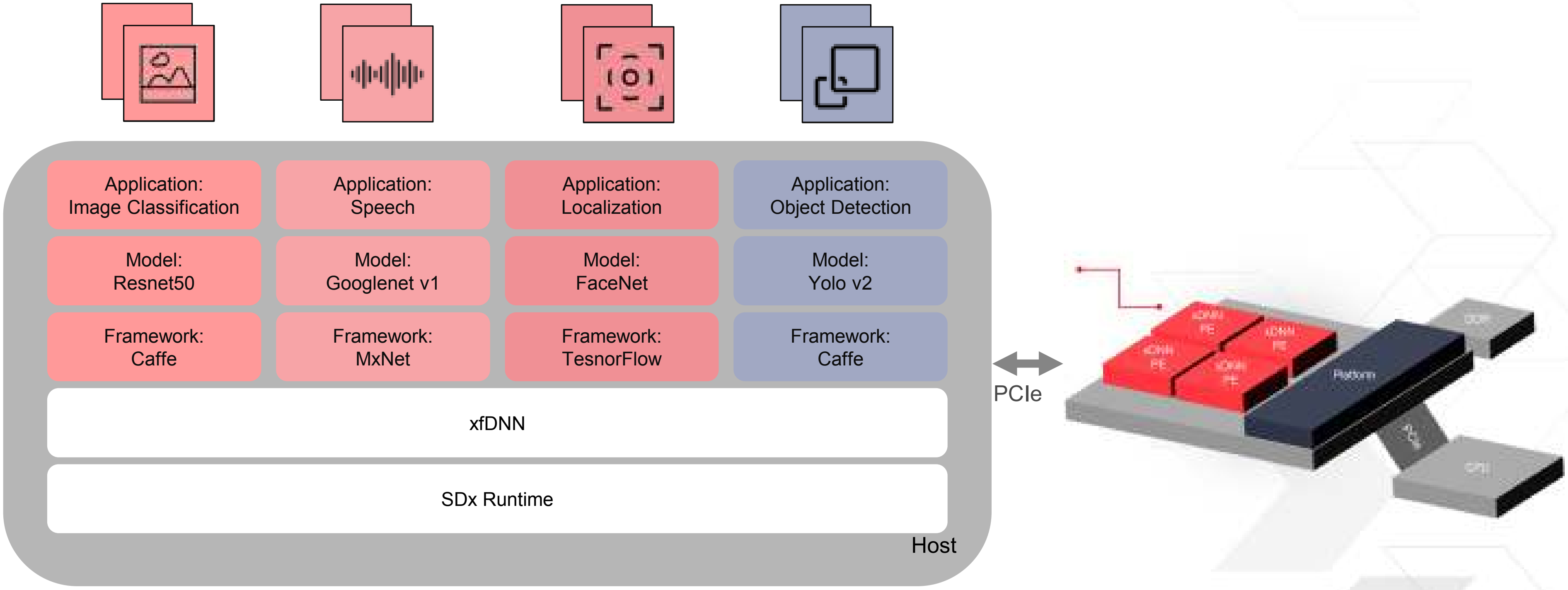
Efficient

> Performance/watt

> Low Power



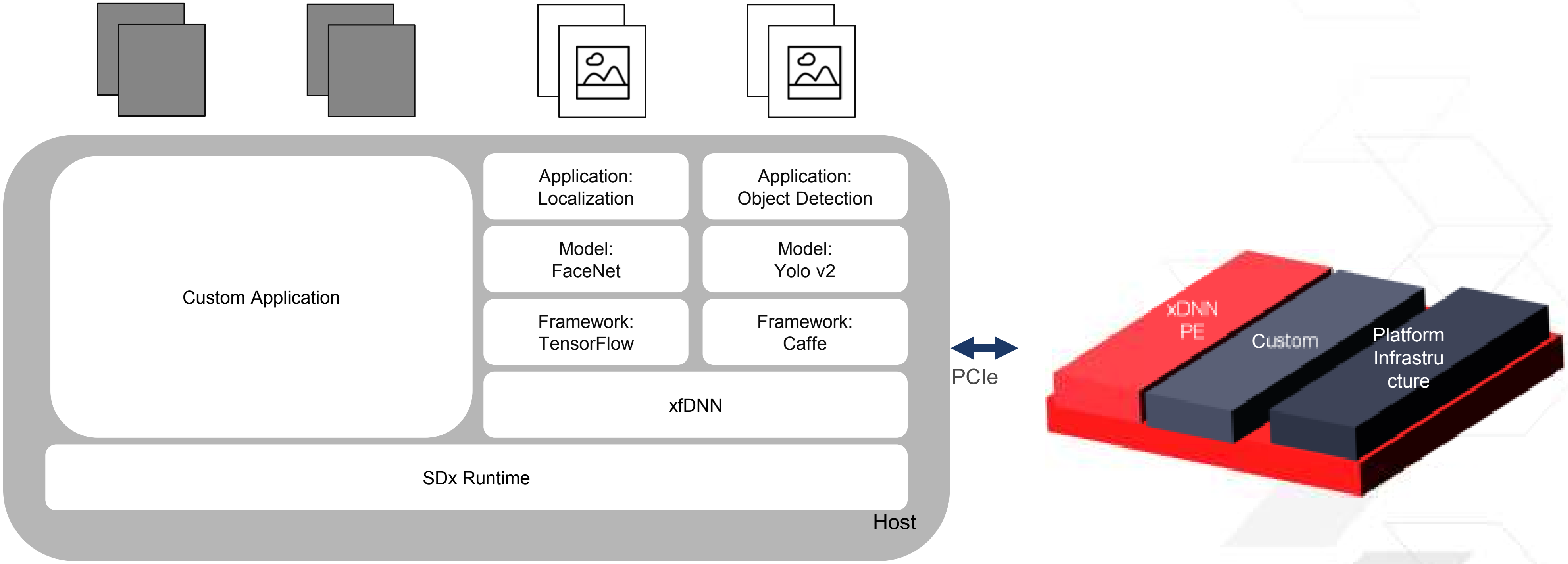
Flexible: Multi-Network Configuration



Application: Image Classification	Application: Speech	Application: Localization	Application: Object Detection
Model: Resnet50	Model: Googlenet v1	Model: FaceNet	Model: Yolo v2
Framework: Caffe	Framework: MxNet	Framework: TesnorFlow	Framework: Caffe

1 FPGA Provides 4 Virtual Accelerators For Real Time Deep Learning

Flexible: Bring Your own IP!



Integrate Custom Applications Directly with xDNN Processing Engines

XILINX
ALVEO
XILINX

Alveo – Breathe New Life into Your Data Center

16nm
UltraScale™ Architecture

Off-Chip Memory Support
- Max Capacity: 64GB
- Max Bandwidth: 77GB/s

Internal SRAM
- Max Capacity: 54MB
- Max Bandwidth: 38TB/s

PCIe Gen3x16
PCIe Gen4x8 w/ CCIX

HBM2 Memory Support
- Max Capacity: 8GB
- Max Bandwidth: 460GB/s

NIMBIX Cloud Deployed

Cloud ↔ On-Premise Mobility

Ecosystem of Applications
- Many available today
- More on the way

Server OEM Support
- Major OEMs in Qualification

SDAccel Environment

Accelerate Any Application
- IDE for compiling, debugging, profiling
- Supports C/C++, RTL, and OpenCL

U200

892K
LUTs

35MB
Internal SRAM Capacity

31TB/s
Internal SRAM Bandwidth

3100img/s
CNN Throughput*

U250

1,341K
LUTs

54MB
Internal SRAM Capacity

38TB/s
Internal SRAM Bandwidth

4100img/s
CNN Throughput*

*Low-latency GoogLeNet v1

It’s All About the Applications

Database

- BLACKLYNX — Elasticsearch
- ALGO-LOGIC — Low Latency KVS
- bigstream — ETL, Streaming, SQL Analytics
- VITESSE DATA — Deepgreen DB
- XELERA — Spark-MLlib
- Titan IC — Hyperion 10G RegEx

Machine Learning

- XILINX — ML Suite for Inference
- Mipsology — Image Classification
- SumUp — Text Analysis

Video

- NGCODEC — ABR Transcoding
- skreens — Experience Engine
- CTACCEL — Image Processing

Financial

- MAXELER Technologies — Real-Time Risk Dashboard
- MAXELER Technologies — SIMM Calculation

HPE & Life Sciences

- Falcon COMPUTING — Accelerated Genomics Pipelines

Application Ecosystem Continues to Grow

Alveo Applications Directory

Solution Stack

Accelerated Solutions

DEVELOPERS

100%

Growth of Published Applications

Hundreds

of Developers Trained

RTL, C, C++, OpenCL

Developer Package

SDAccel Environment

Xilinx ML Suite

GitHub
Open Source Software Tutorial, Apps and Models
xDNN Middleware Tools Runtime
xDNN Middleware
xDNN Compiler and Optimizer
Xilinx xDNN Runtime
xDNN IP

End user

Framework, API, Python/Java/C++ Programmability

HPC &LIFE SCIENCES

FINANCIAL

VIDEO

MACHINE LEARNING

DATABASE

Solutions Xilinx ISVs

Platforms

Cloud

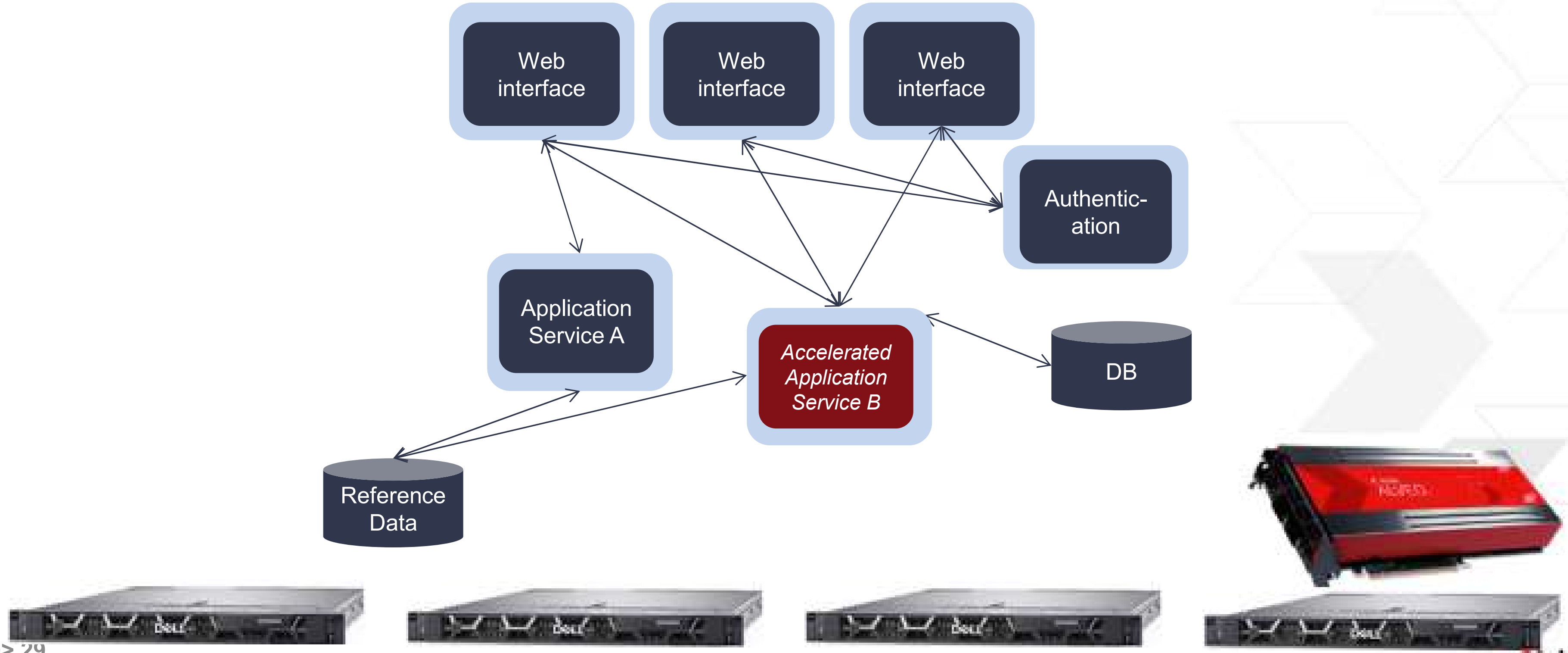
FPGA as a Service (FaaS)

On-premise

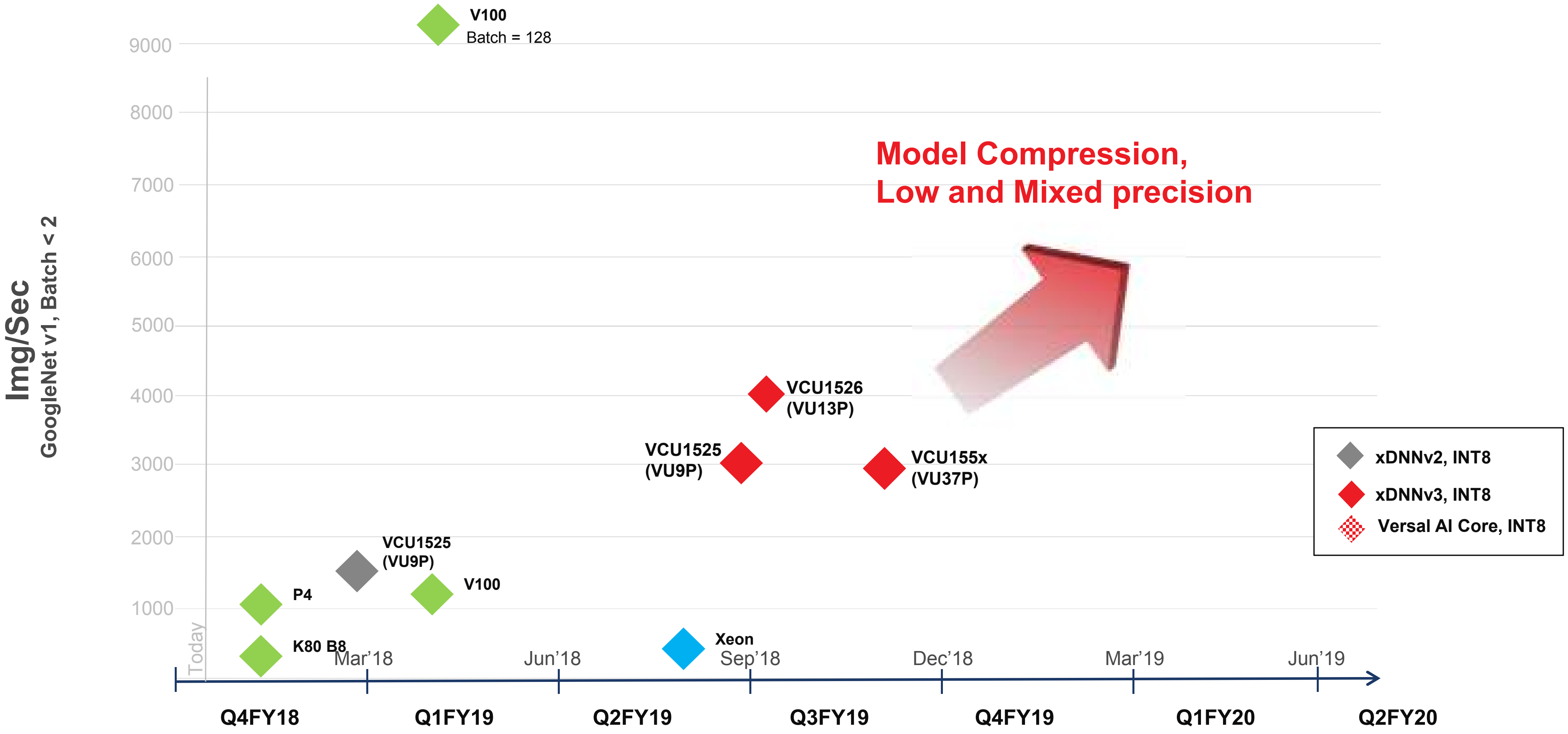
Platform

Orchestration (Kubernetes, OpenShift, Etc)

Migrate the components to the right hardware resource and manage the container communication across the cluster.



ML Suite Performance Roadmap



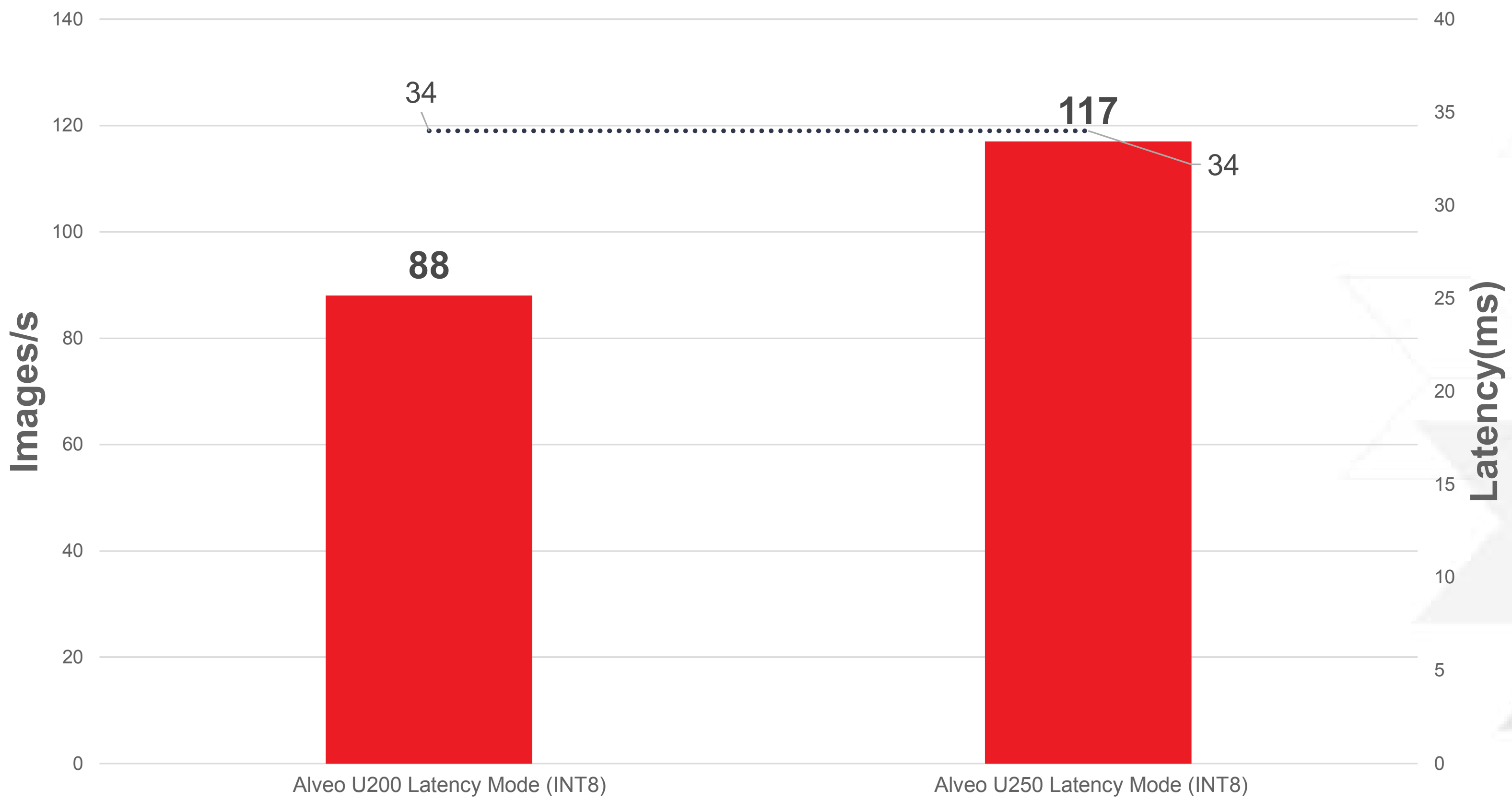
Model Compression,
Low and Mixed precision

▨ ACAP

7x Performance
Improvement

CPU: https://mxnet.incubator.apache.org/faq/perf.html
Nvidia: https://images.nvidia.com/content/pdf/inference-technical-overview.pdf
P4 = int8, v100 = fp16

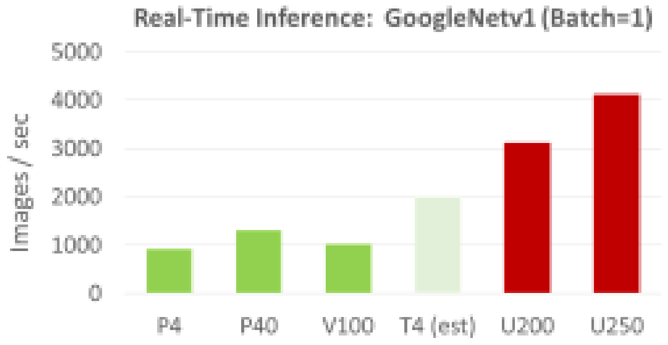
xDNN YOLO v2 Performance – Image Size 608x608



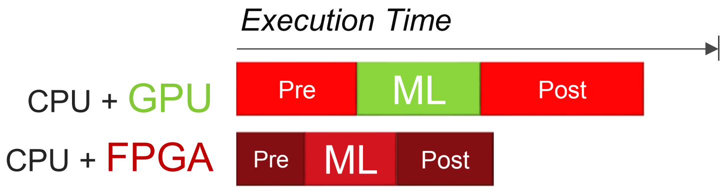
Key Differentiating Value vs. Nvidia

Xilinx Enables . . . .

- **Highest Throughput <u>WITH</u> Low Latency (i.e. low batch)**
  - <u>AND</u> best Perf / Watt



- **Accelerate the <u>Whole Application</u> for Broad Range of Workloads**

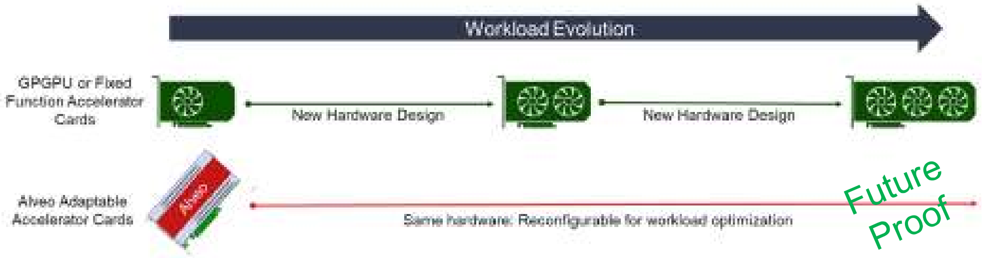


Execution Time

CPU + GPU: Pre | ML | Post

CPU + FPGA: Pre | ML | Post

- **Innovate Faster than the Silicon Cycle with Adaptable Hardware**


Workload Evolution
GPGPU or Fixed Function Accelerator Cards
New Hardware Design
New Hardware Design
Alveo Adaptable Accelerator Cards
Alveo
Same hardware: Reconfigurable for workload optimization
Future Proof

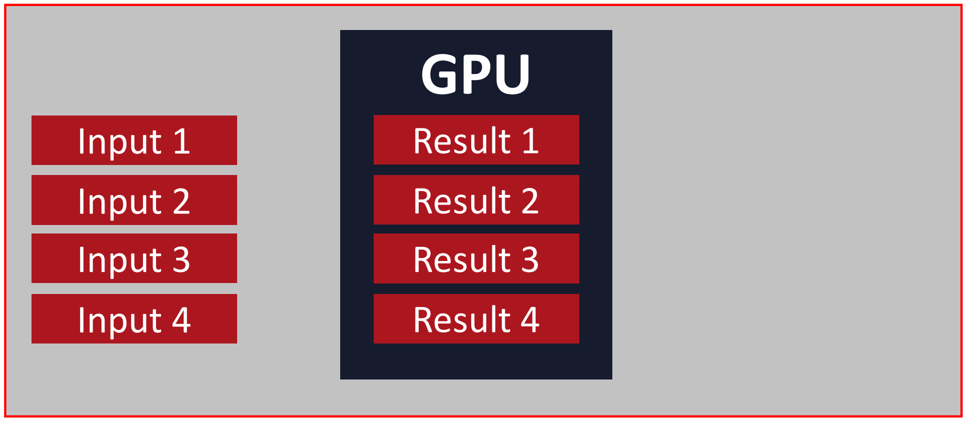
What About Batching?

Fundamental to GPU Architecture

(Software Defined Data Flow)

Batching: Loading up lots of similar Data Sets

- Keep CUDA cores busy
- Hide some memory latency
- Create better SIMT efficiency



GPU

Input 1
Input 2
Input 3
Input 4

Result 1
Result 2
Result 3
Result 4

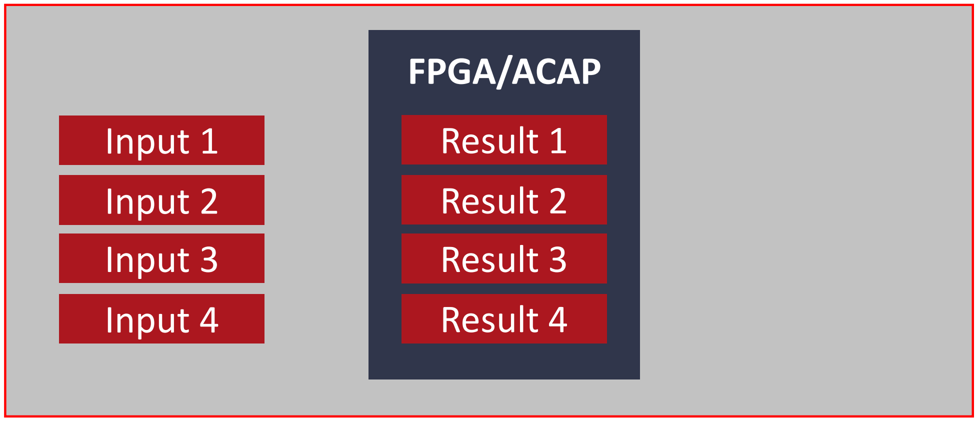
High Throughput OR Low Latency

Not Required for FPGA / ACAP

(Hardware Defined Data Flow)

Independent of Data Set count

- Custom HW kernels
- Custom Memory Hierarchy
- HW pipeline data flow



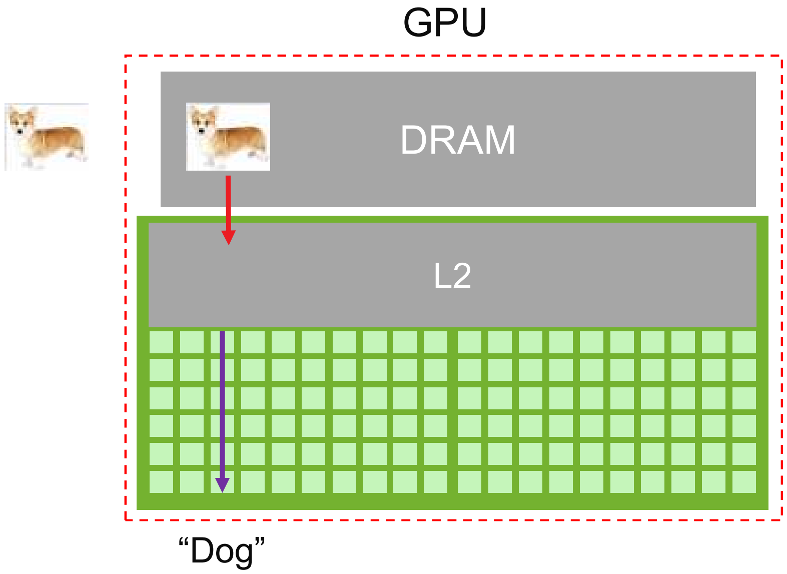
FPGA/ACAP

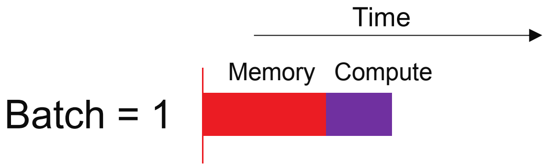
Input 1
Input 2
Input 3
Input 4

Result 1
Result 2
Result 3
Result 4

High Throughput AND Low Latency

A Batching Example: Image Classification

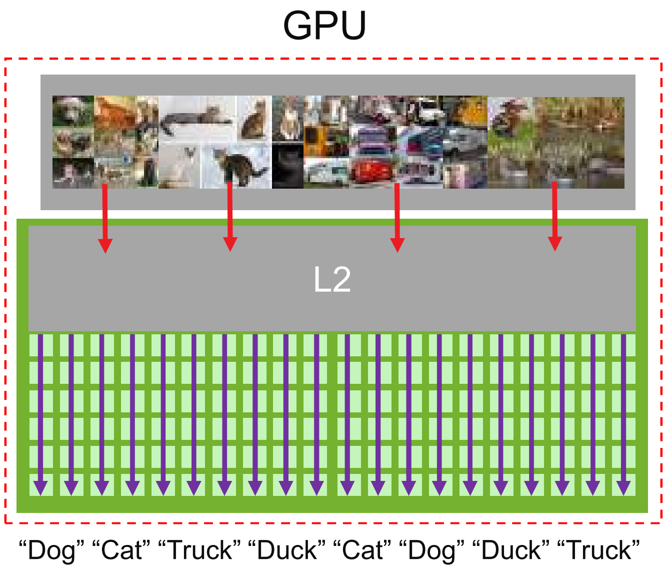


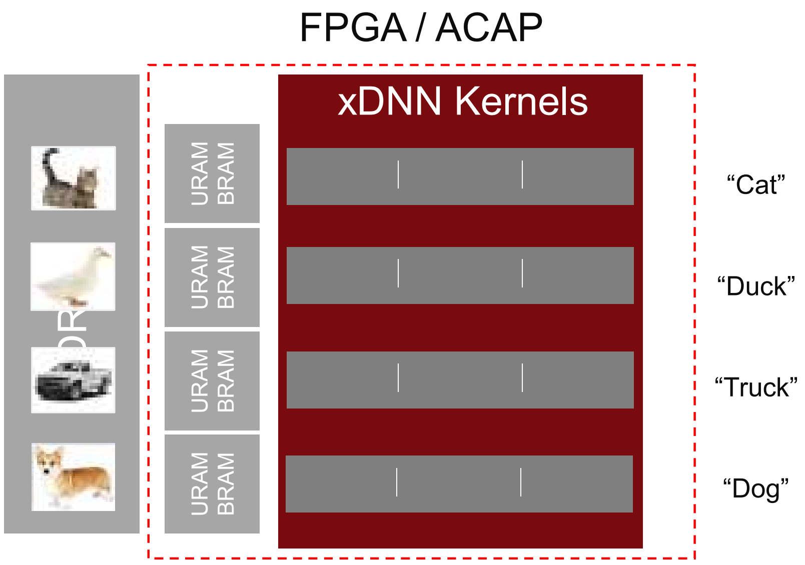


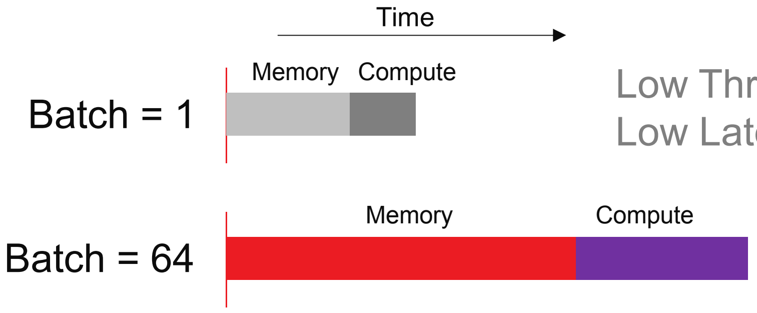
Low Throughput

Low Latency

A Batching Example: Image Classification







Low Throughput
Low Latency

High Throughput
High Latency

High Throughput

AND

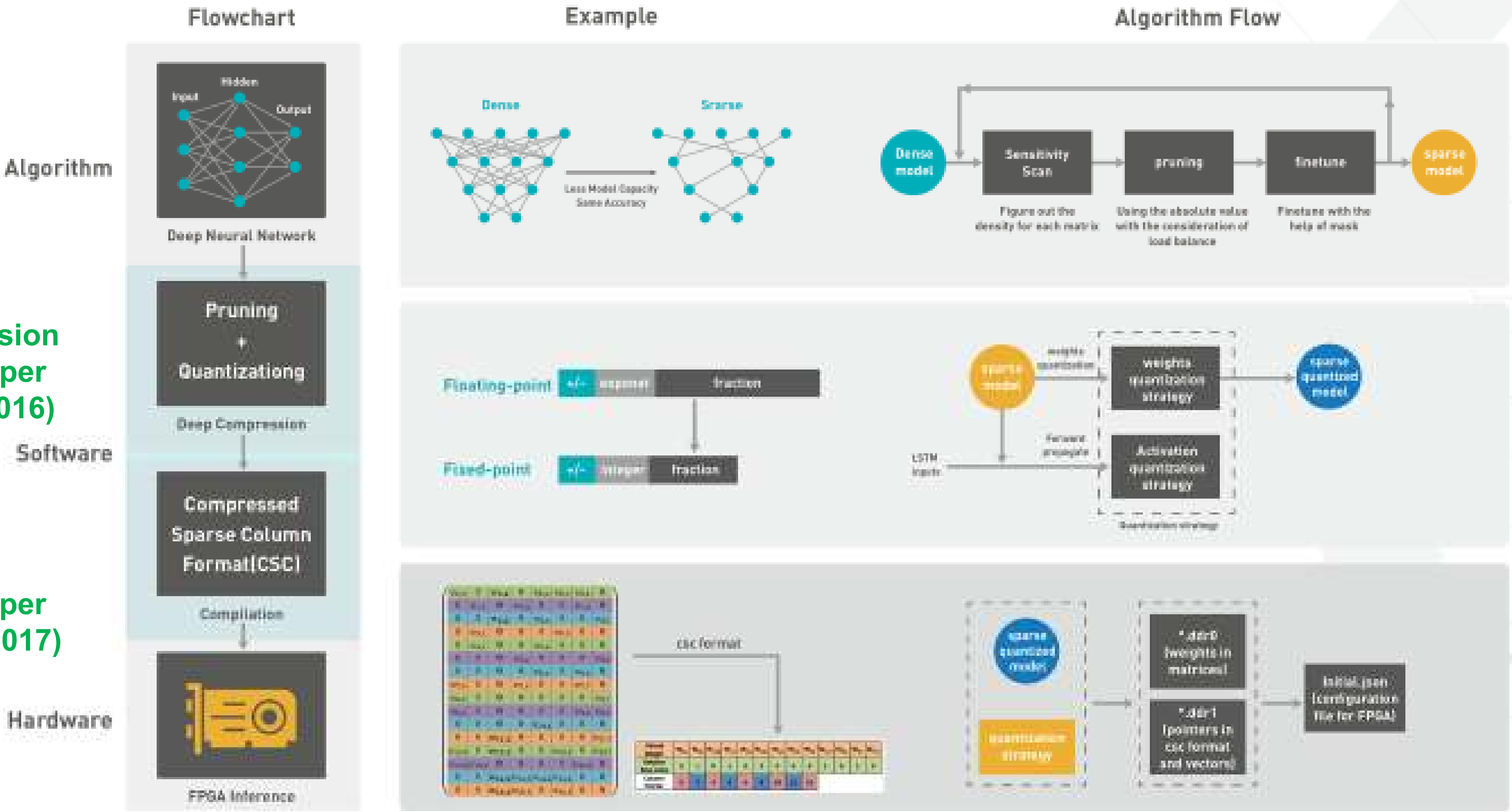
Low Latency

Overview of Compressed LSTM

Deep Compression (Best paper of ICLR2016)

+

ESE (Best paper of FPGA2017)



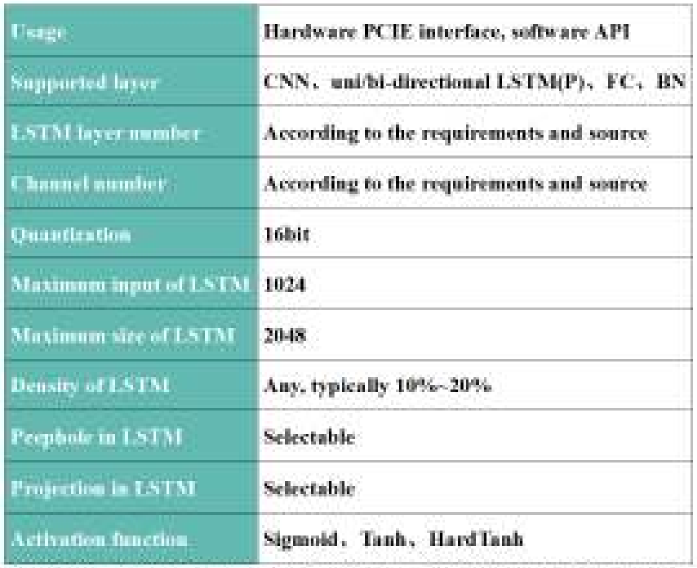
Our algorithm, software and hardware co-design flow for RNN(LSTM) acceleration

LSTM Solution Features

Features

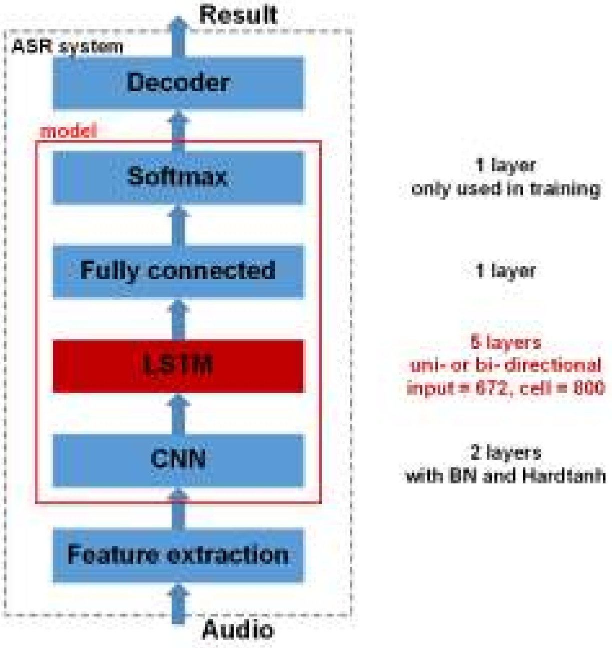
Innovative full-stack acceleration solution for deep learning in acoustic speech recognition (ESE: best paper of FPGA2017)

- Support both unidirectional and bi-directional LSTM acceleration on FPGA for model inference
- Support CNN layers, Fully-Connected (FC) layers, Batch Normalization layers and varieties of activation functions such as Sigmoid, Tanh and HardTanh
- Support testing for both performance comparison of CPU/FPGA and single sentence recognition
- Supporting user's own test audio recognition (English, 16kHz sample rate, no longer than 3 seconds)



Usage	Hardware PCIE interface, software API
Supported layer	CNN、uni/bi-directional LSTM(P)、FC、BN
LSTM layer number	According to the requirements and source
Channel number	According to the requirements and source
Quantization	16bit
Maximum input of LSTM	1024
Maximum size of LSTM	2048
Density of LSTM	Any, typically 10%~20%
Peephole in LSTM	Selectable
Projection in LSTM	Selectable
Activation function	Sigmoid、Tanh、HardTanh

Note: It is available to change the hardware configuration for different requirements such as layer number and model size



Result

ASR system

Decoder

model

Softmax

Fully connected

LSTM

CNN

Feature extraction

Audio

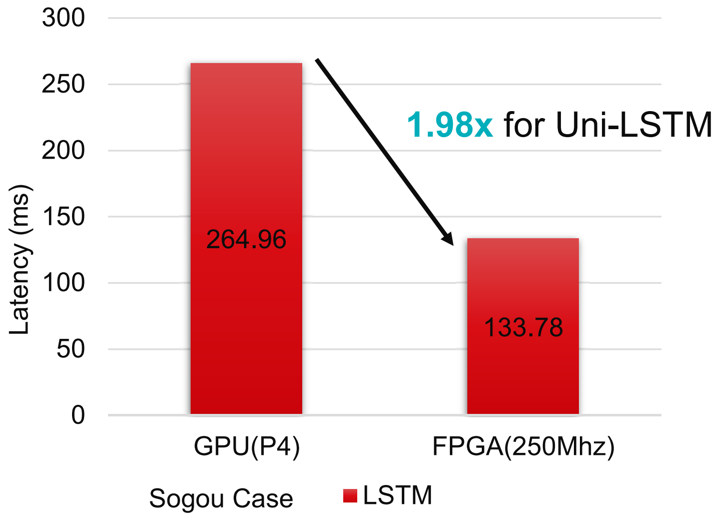
1 layer
only used in training

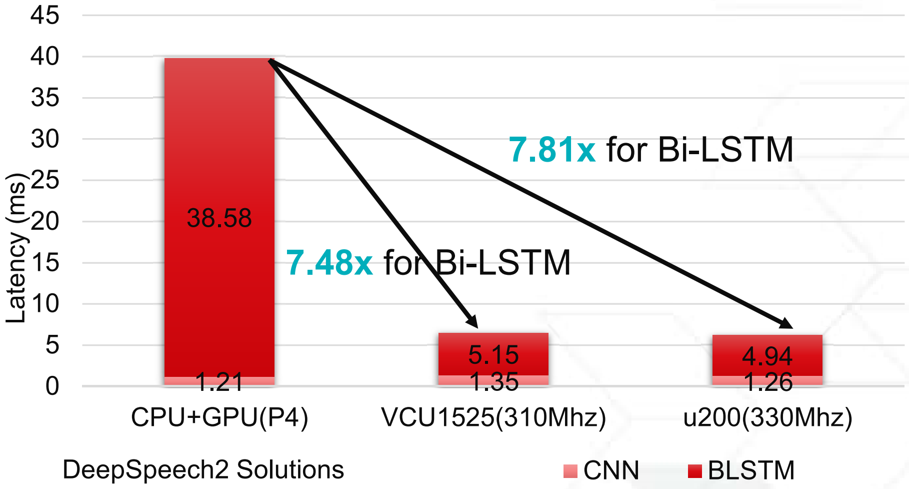
1 layer

5 layers
uni- or bi- directional
input = 672, cell = 800

2 layers
with BN and Hardtanh

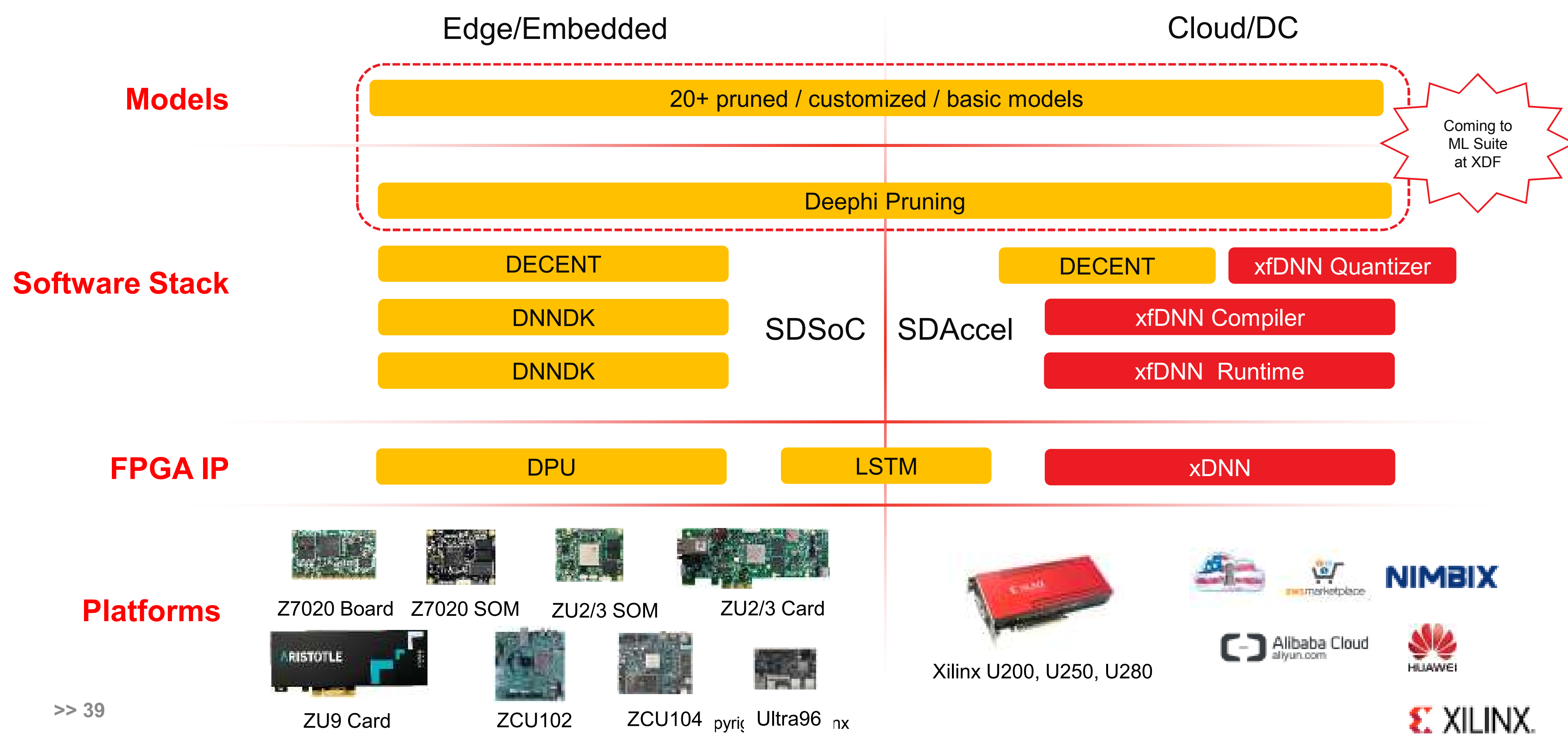
Competitive Performance





	Sogou Case	DeepSpeech2
Model	3-layer LSTM	2-layer CNN + 5-layer bidirectional LSTM
LSTM Density	17.76%	15%
(Bi)LSTM Dimension	Input 153, cell 2048, output 1024	Input 672, cell 800, output 800
CPU Performance	18155.22ms (E5-2690 V4)	118.31ms (E5-2686 V4) [BLSTM Only]
FPGA Implementation	32 channels on ku115 @ 250MHz	1 channel on vcu1525 @ 310MHz (u200@330mhz)
Application	Part of ASR in Sogou	End2End ASR

Integrated Edge – Cloud Roadmap



Edge/Embedded

Cloud/DC

Models

20+ pruned / customized / basic models

Deephi Pruning

Coming to ML Suite at XDF

Software Stack

DECENT

DNNDK

DNNDK

SDSoC

SDAccel

DECENT

xfDNN Quantizer

xfDNN Compiler

xfDNN Runtime

FPGA IP

DPU

LSTM

xDNN

Platforms

Z7020 Board

Z7020 SOM

ZU2/3 SOM

ZU2/3 Card

ZU9 Card

ZCU102

ZCU104

pyri

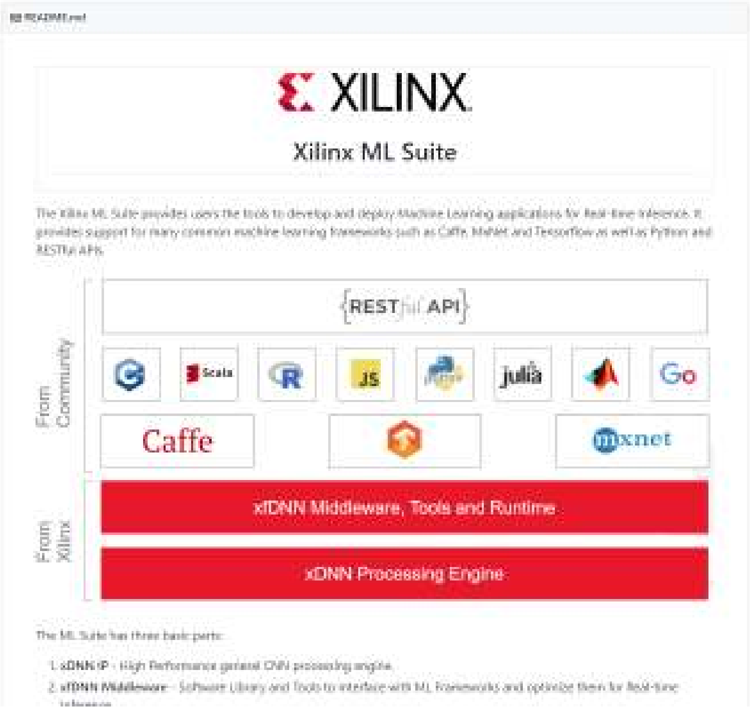
Ultra96

nx

Xilinx U200, U250, U280

Try the Xilinx ML Suite Today

https://github.com/Xilinx/ml-suite



README.md

XILINX

Xilinx ML Suite

The Xilinx ML Suite provides users the tools to develop and deploy Machine Learning applications for Real-time Inference. It provides support for many common machine learning frameworks such as Caffe, MxNet and Tensorflow as well as Python and RESTful APIs.

The ML Suite has three basic parts:

1. **xDNN IP** - High Performance general CNN processing engine.
2. **xfDNN Middleware** - Software Library and Tools to interface with ML Frameworks and optimize them for Real-time Inference.

AWS EC2 Marketplace

Xilinx ML Suite

Overview

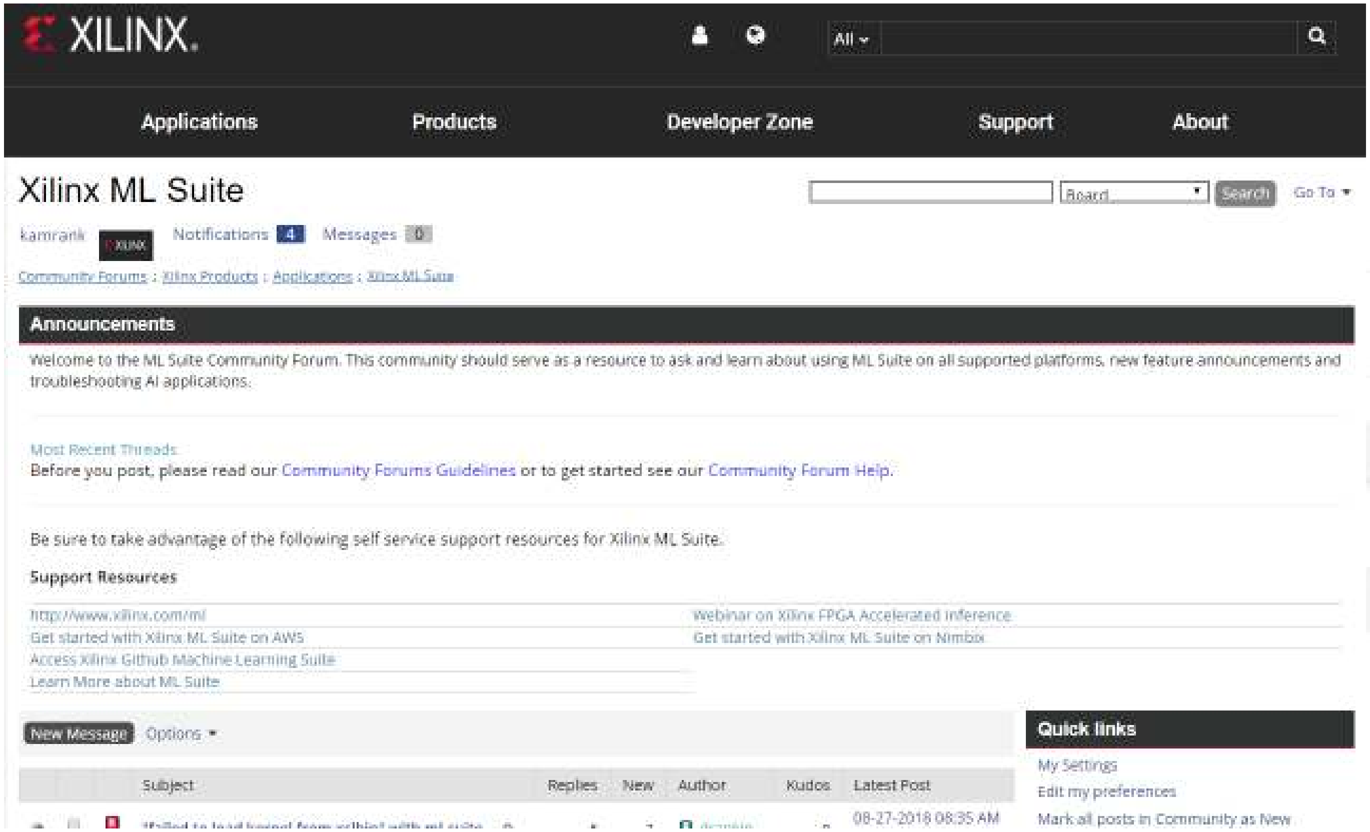
Product Overview

Nimbix NX5

Xilinx ML Suite

© Copyright 2018 Xilinx

Engage with the Community



Xilinx ML Suite

kamrank   Notifications 4   Messages 0

Community Forums : Xilinx Products : Applications : Xilinx ML Suite

Announcements

Welcome to the ML Suite Community Forum. This community should serve as a resource to ask and learn about using ML Suite on all supported platforms, new feature announcements and troubleshooting AI applications.

Most Recent Threads
Before you post, please read our Community Forums Guidelines or to get started see our Community Forum Help.

Be sure to take advantage of the following self service support resources for Xilinx ML Suite.

Support Resources

http://www.xilinx.com/ml	Webinar on Xilinx FPGA Accelerated Inference
Get started with Xilinx ML Suite on AWS	Get started with Xilinx ML Suite on Nimbix
Access Xilinx Github Machine Learning Suite	
Learn More about ML Suite	

Subject	Replies	New	Author	Kudos	Latest Post
'Failed to load kernel from xclbin' with ml suite	[illegible]	[illegible]	[illegible]	[illegible]	08-27-2018 08:35 AM

Quick links

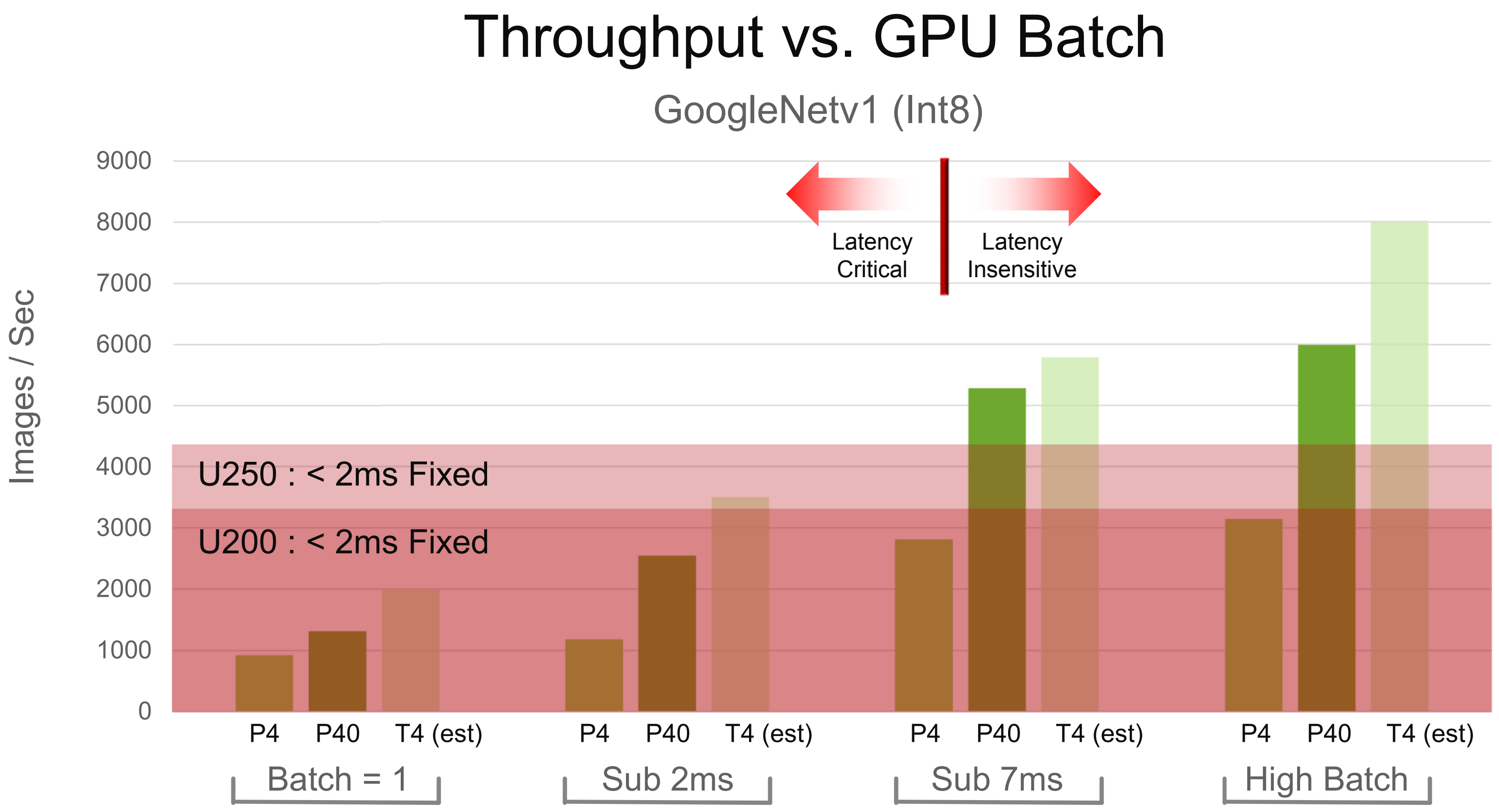
My Settings
Edit my preferences
Mark all posts in Community as New

Building the Adaptable, Intelligent World

Building the Adaptable, Intelligent World

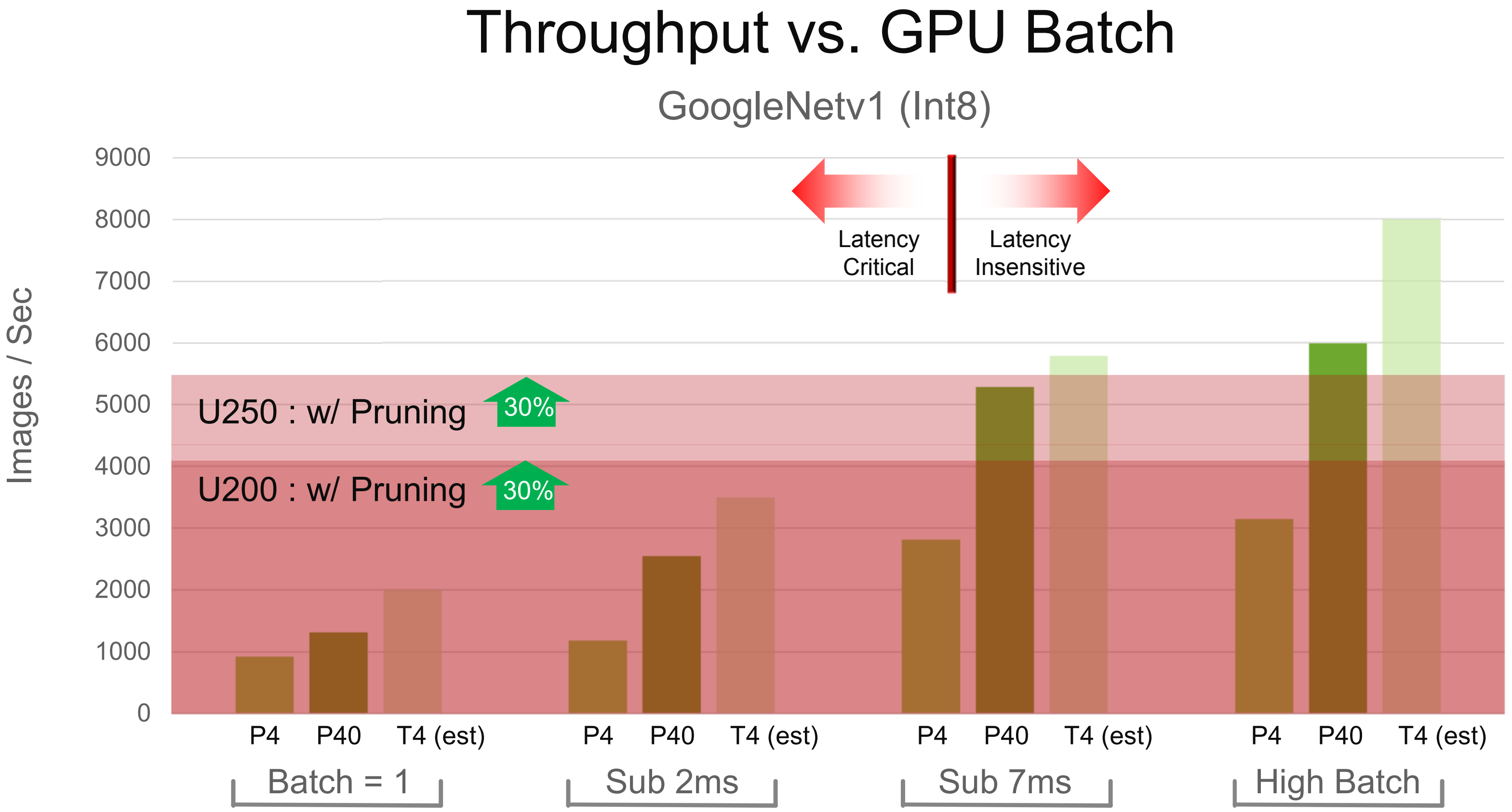
XILINX

A Broader View of ML Benchmarking (Int8)



Alveo Delivers Low, Fixed Latency (< 2ms) in ALL Scenarios

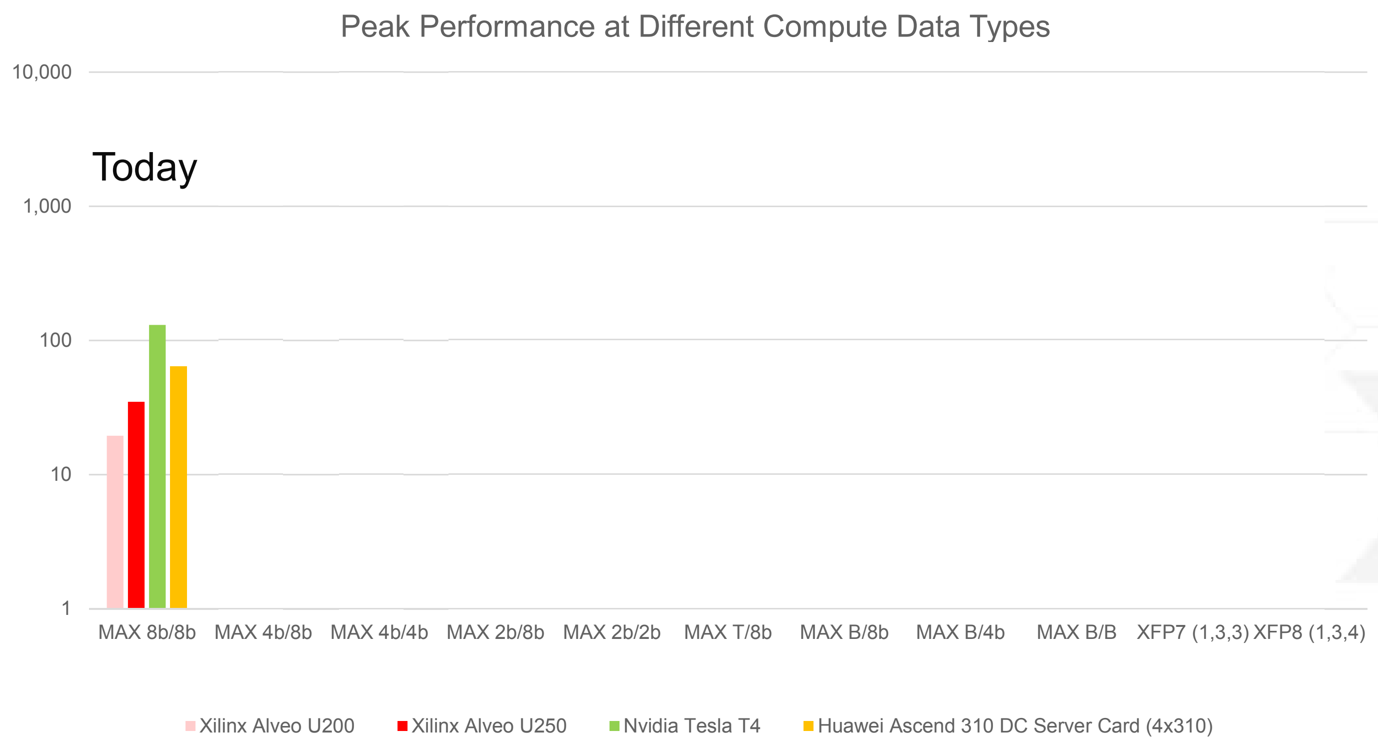
Next Step: Deephi Pruning Techniques



Deephi Proprietary Pruning Increases Performance **30% or More**

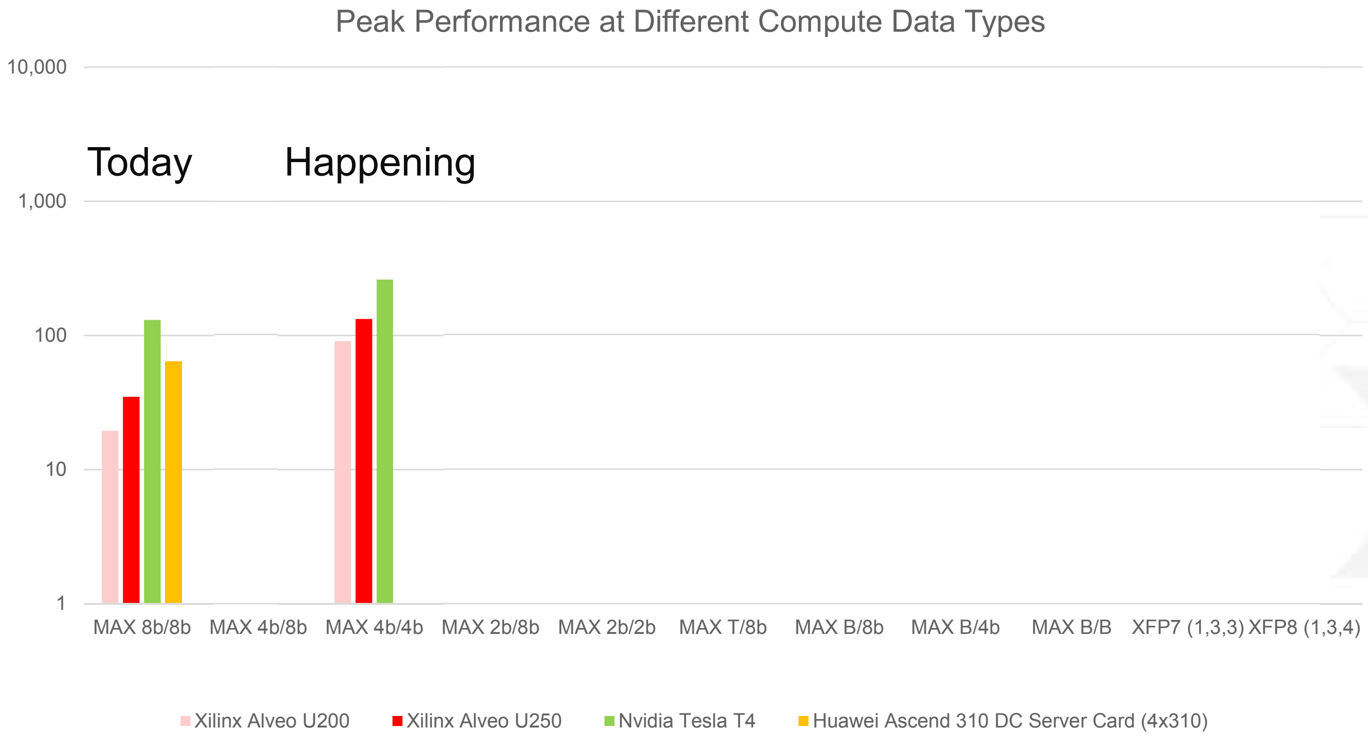
Compute Capability

- INT8 is becoming standard for inference devices



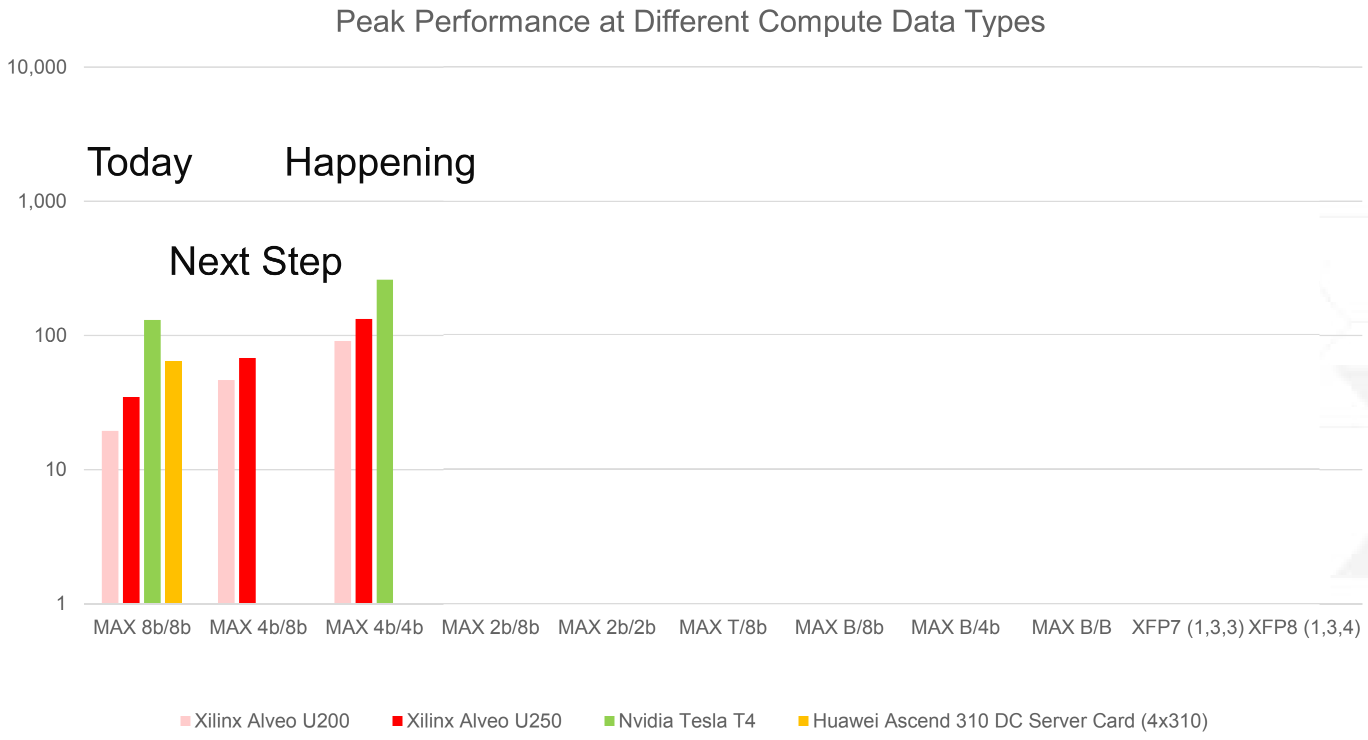
Compute Capability

> **Trends are moving toward lower precision, e.g. INT4 is emerging**



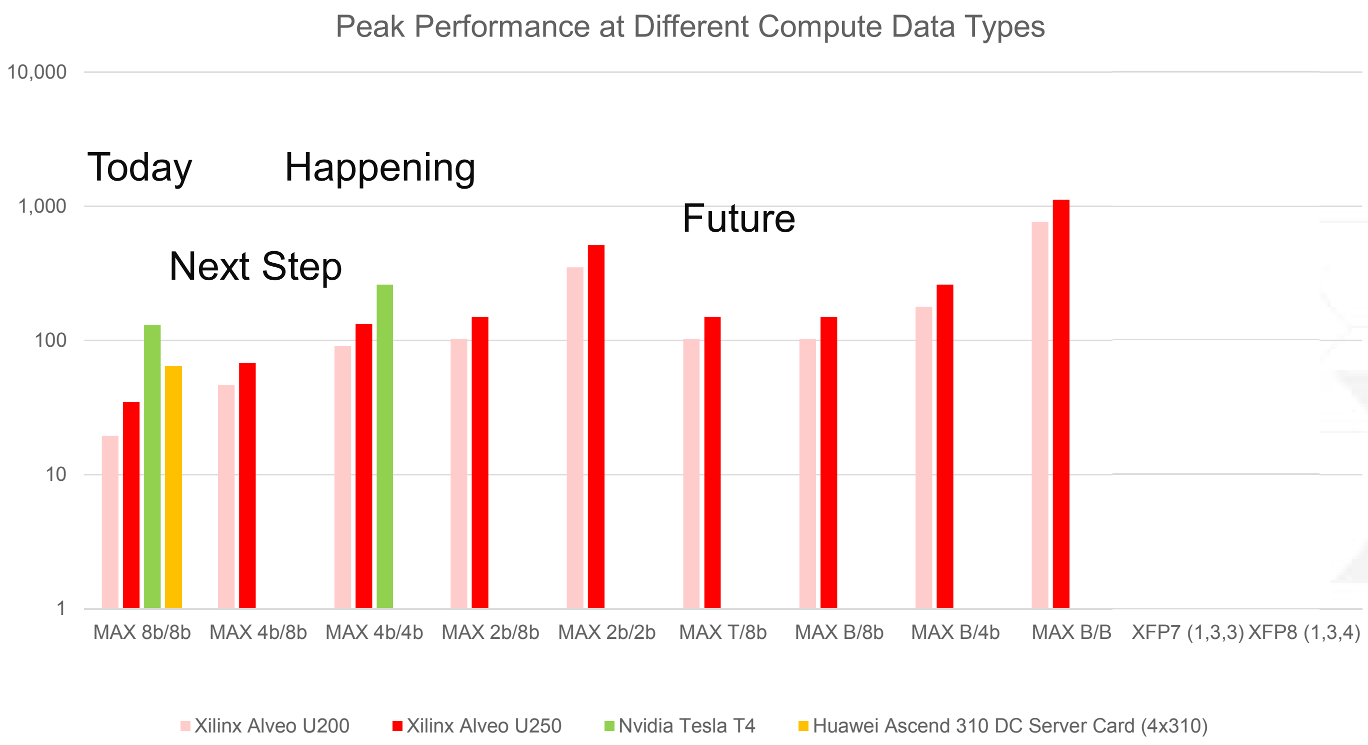
Compute Capability

> **Mixed weight and activation precisions provide additional optimization for inference while keeping high accuracy**



Compute Capability

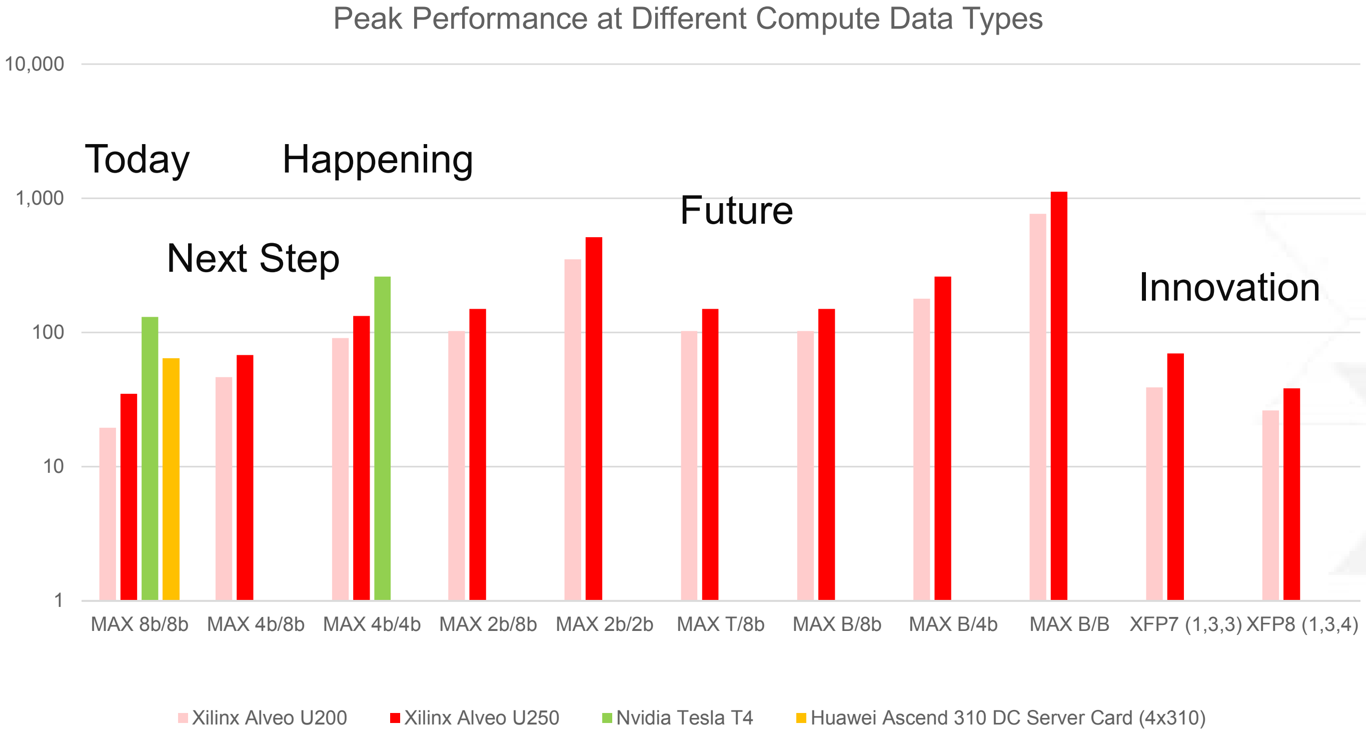
> **The future we see: each model will have its own mix of datatype for optimal efficiency and accuracy**



Compute Capability

> Alveo: Most future proof architecture

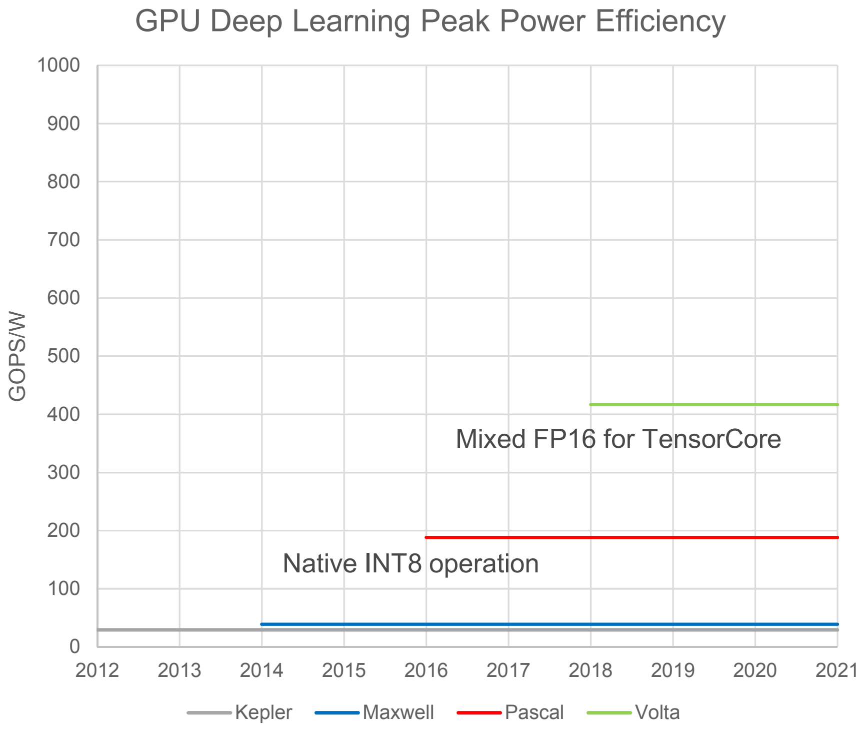
> Scalable performance for any data type

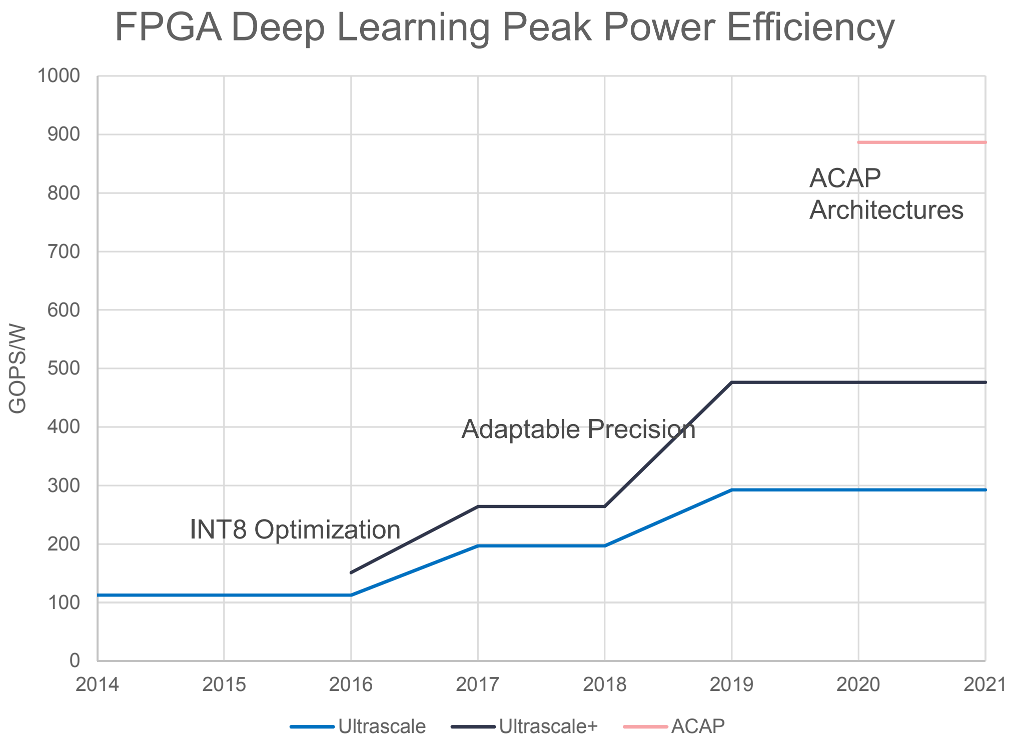


Break Through on Peak Performance

- GPU: Introduce new architectures and silicon

- Xilinx: Adapt the break through of emerging domain knowledge



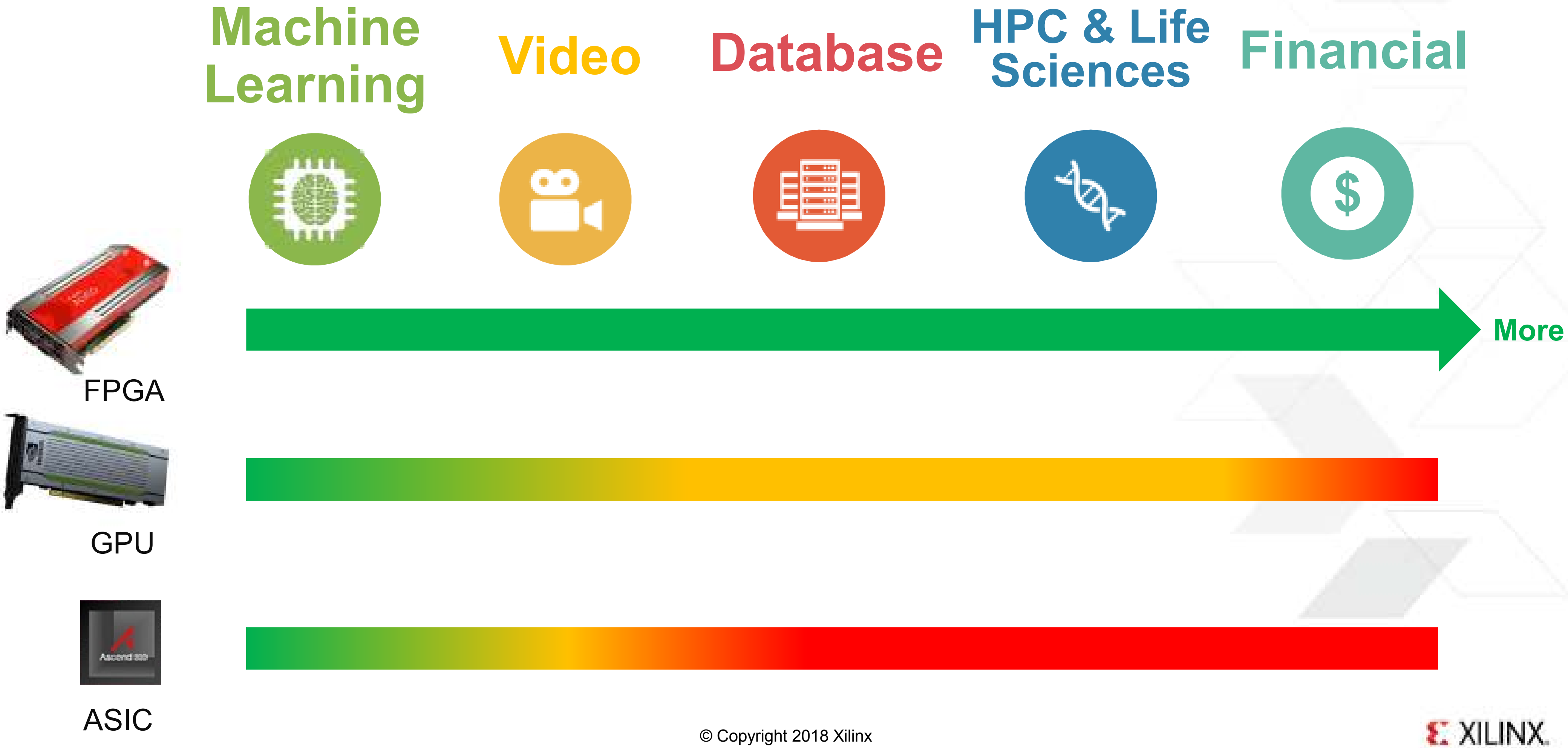


Application Adaptability

🟩 Strong Acceleration

🟨 Some Acceleration

🟥 No Acceleration



Machine Learning

Video

Database

HPC & Life Sciences

Financial

FPGA

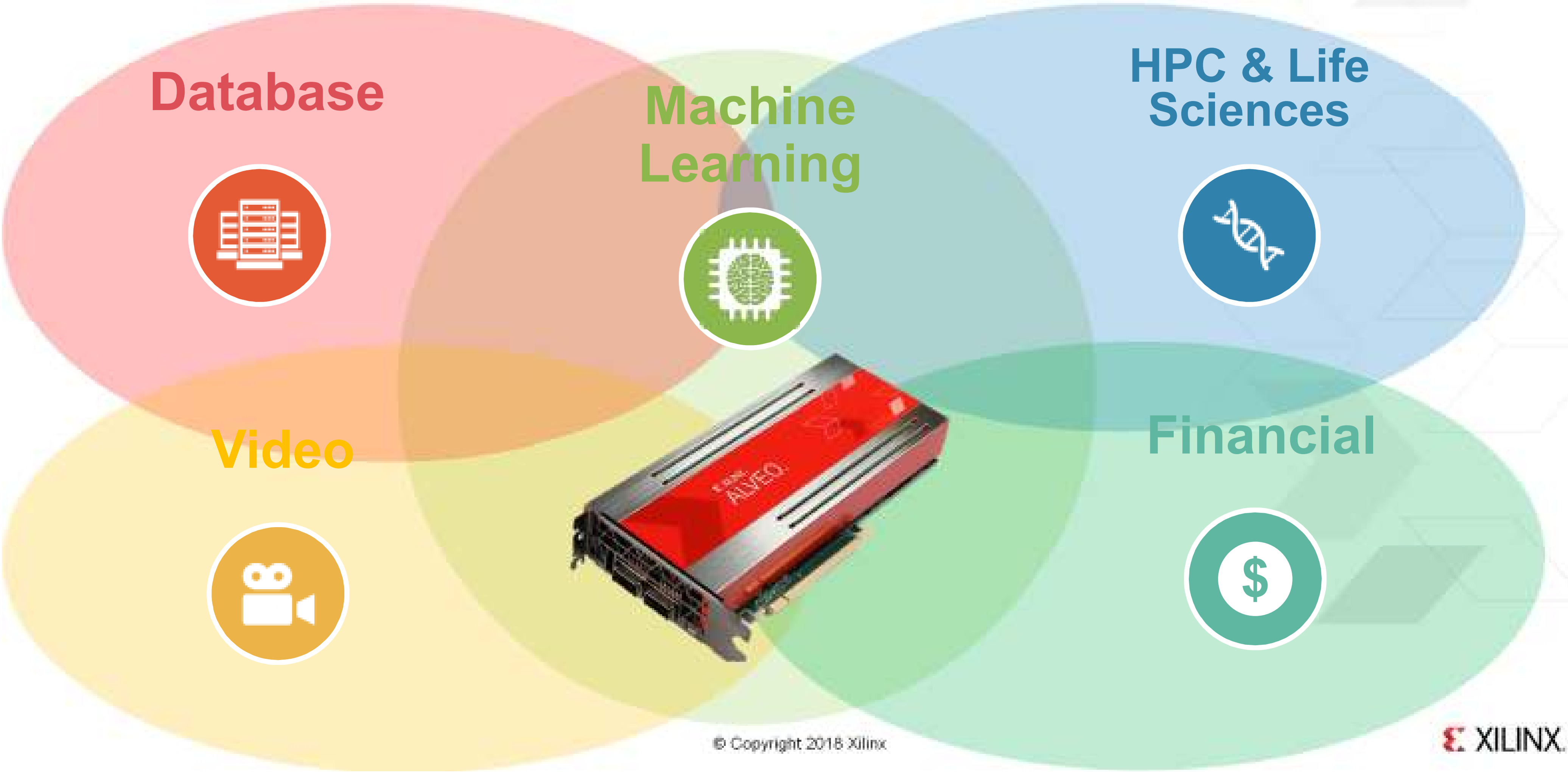
GPU

ASIC

More

XILINX.

Infuse Machine Learning with other accelerations



Database

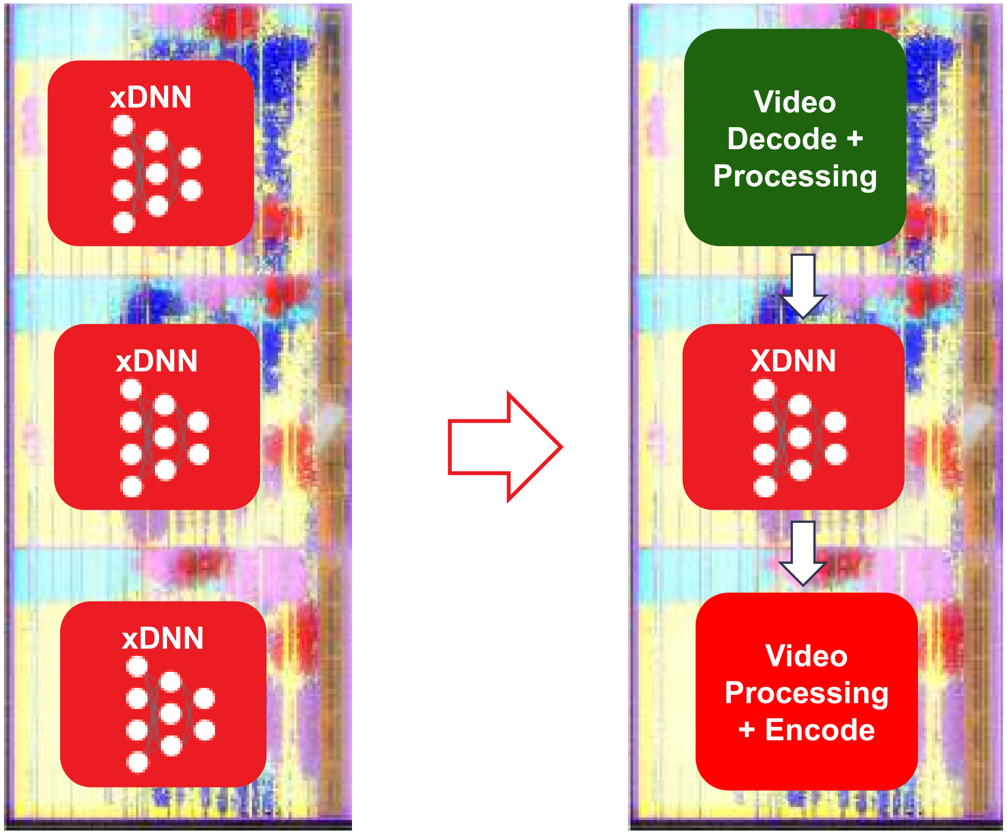
Machine Learning

HPC & Life Sciences

Video

Financial

Custom Deep Learning Pipeline



- Video + ML
- Genomics + ML
- Risk Modelling + ML
- Database + ML
- Network IPS + ML
- Storage + ML

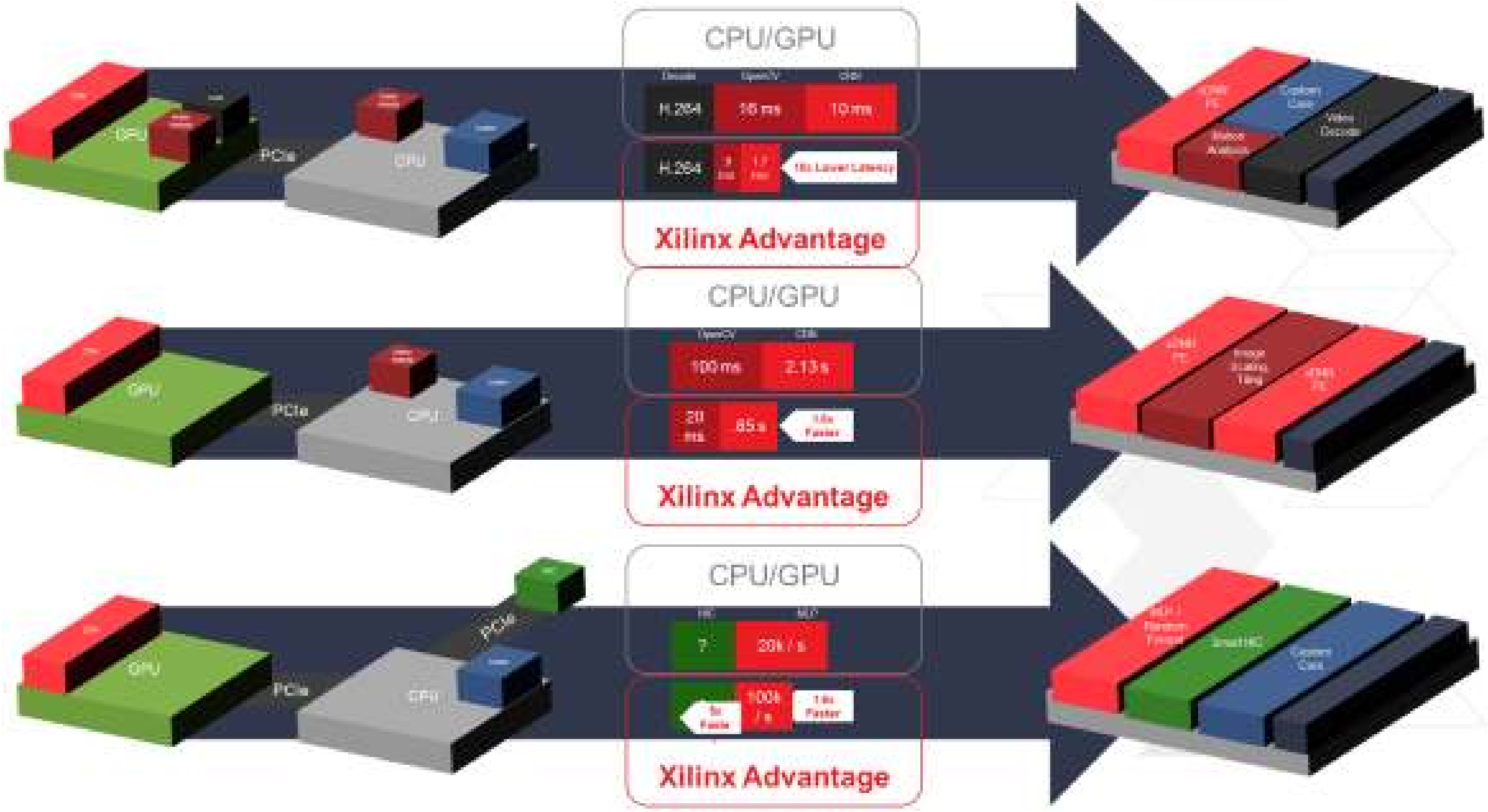
Integrate Custom Applications with xDNN. Lower end-to-end latency

X + ML Focus Applications Summary

Smart City / Cloud Surveillance
- 10x Lower Latency

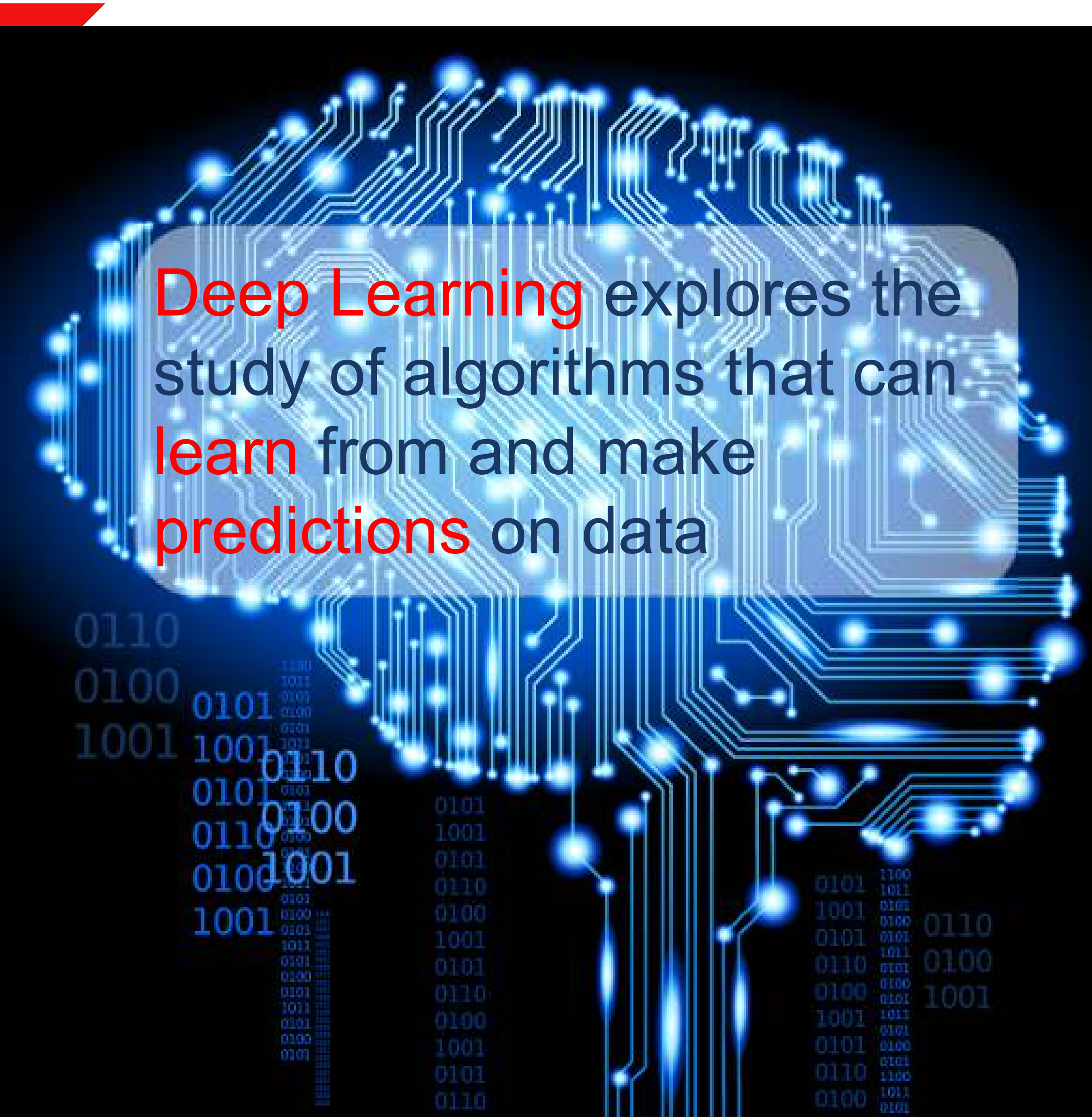
High Resolution Imaging
- 1.6x Faster

Security / Anomaly / Malware
- 5x Faster



ML Suite Solution Stack

XILINX

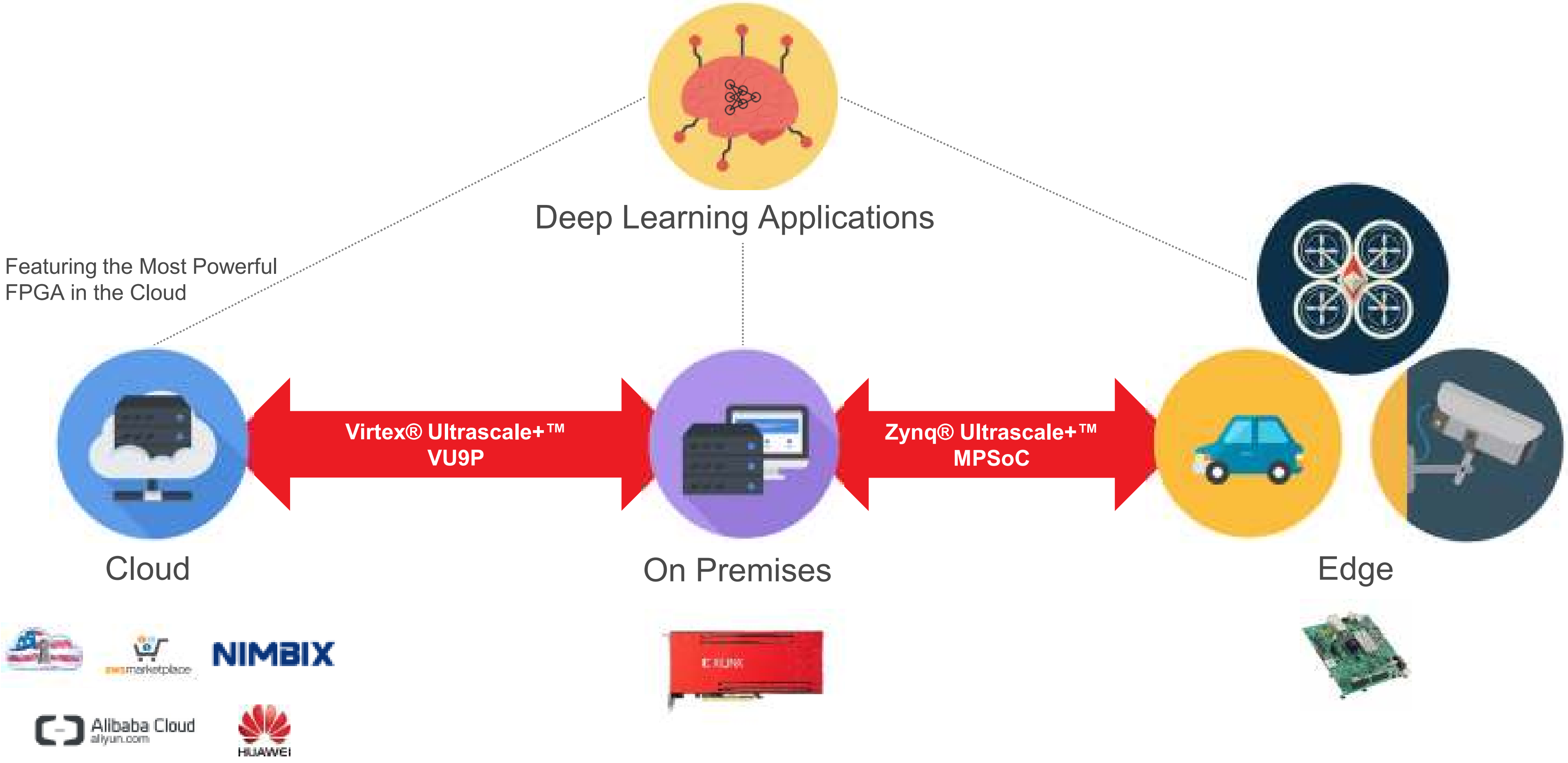


Deep Learning explores the study of algorithms that can learn from and make predictions on data

Deep Learning is Re-Defining many Applications

Cloud Acceleration
Security
Ecommerce Social
Financial
Surveillance
Industrial IOT
Medical Bioinformatics
Autonomous Vehicles

Accelerating AI Inference into Your Cloud Applications



Deep Learning Applications

Featuring the Most Powerful FPGA in the Cloud

Virtex® Ultrascale+™ VU9P

Zynq® Ultrascale+™ MPSoC

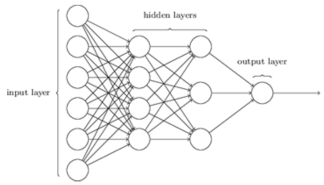
Cloud

On Premises

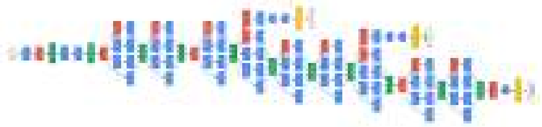
Edge

Overlay Architecture
Custom Processors Exploiting Xilinx FPGA Flexibility

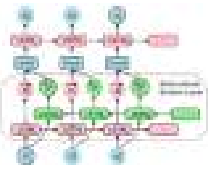
- Customized overlays with ISA architecture for optimized implementation
- Easy plug and play with Software Stack



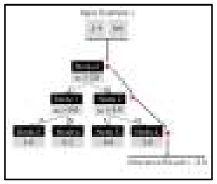
MLP Engine
Scalable sparse and dense implementation



xDNN – CNN Engine for Large 16 nm Xilinx Devices
Deephi DPU – Flexible CNN Engine with Embedded Focus
CHaiDNN – HLS based open source offering

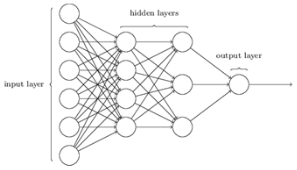


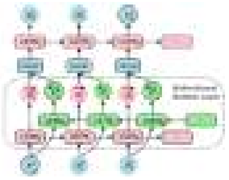
Deephi ESE
LSTM Speech to Text engine



Random Forest
Configurable RF classification

Deep Learning Models





Multi-Layer Perceptron

- Classification
- Universal Function Approximator
- Autoencoder

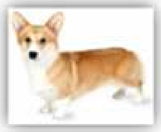
Convolutional Neural Network

- Feature Extraction
- Object Detection
- Image Segmentation

Recurrent Neural Network

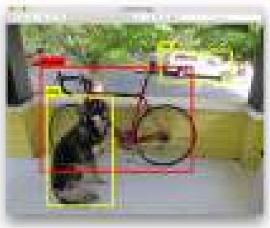
- Sequence and Temporal Data
- Speech to Text
- Language Translation

Classification

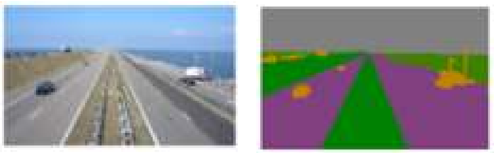


"Dog"

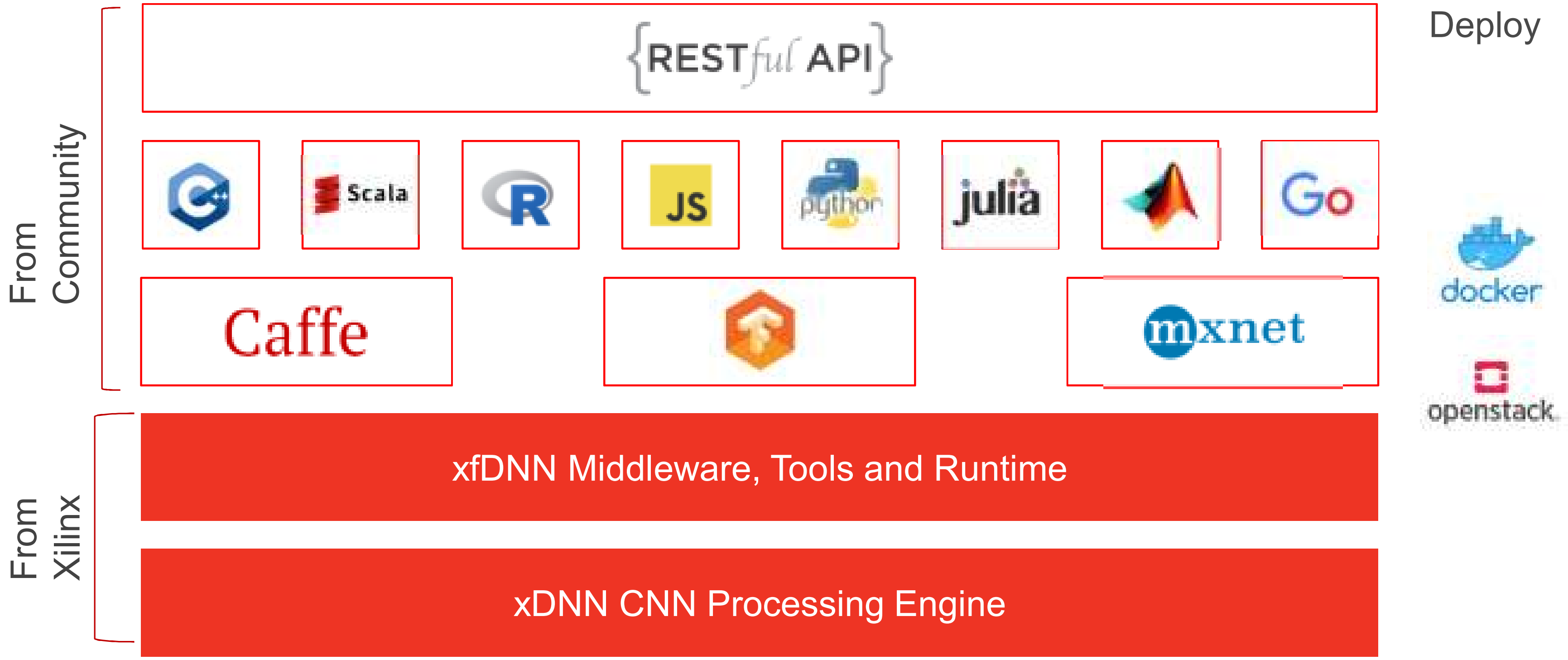
Object Detection



Segmentation



Seamless Deployment with Open Source Software



From Community

From Xilinx

Deploy

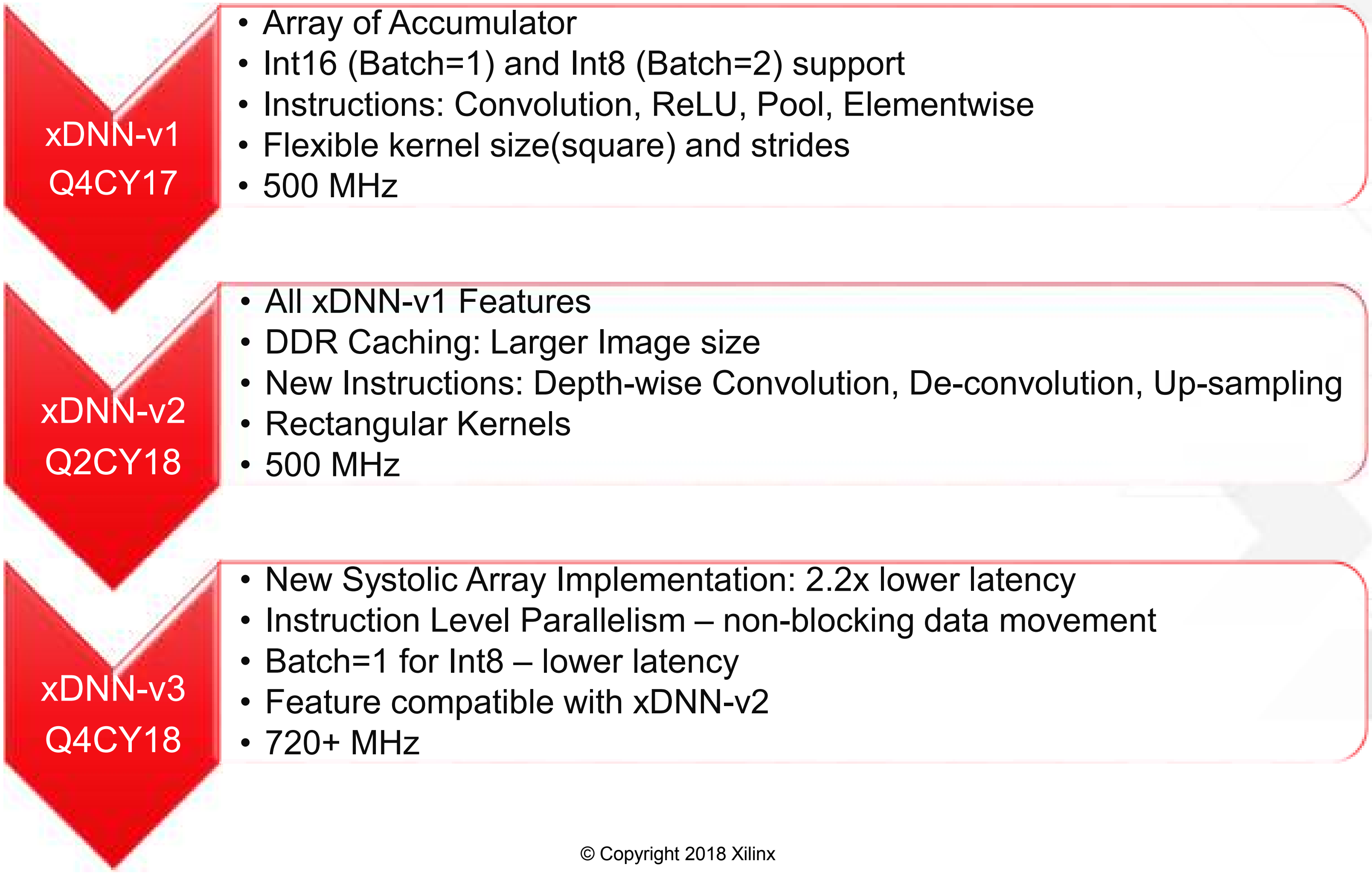
xfDNN Middleware, Tools and Runtime

xDNN CNN Processing Engine

*TensorFlow Q4 2017

xDNN Process Engine

Rapid Feature and Performance Improvement



xDNN-v1
Q4CY17

- Array of Accumulator
- Int16 (Batch=1) and Int8 (Batch=2) support
- Instructions: Convolution, ReLU, Pool, Elementwise
- Flexible kernel size(square) and strides
- 500 MHz

xDNN-v2
Q2CY18

- All xDNN-v1 Features
- DDR Caching: Larger Image size
- New Instructions: Depth-wise Convolution, De-convolution, Up-sampling
- Rectangular Kernels
- 500 MHz

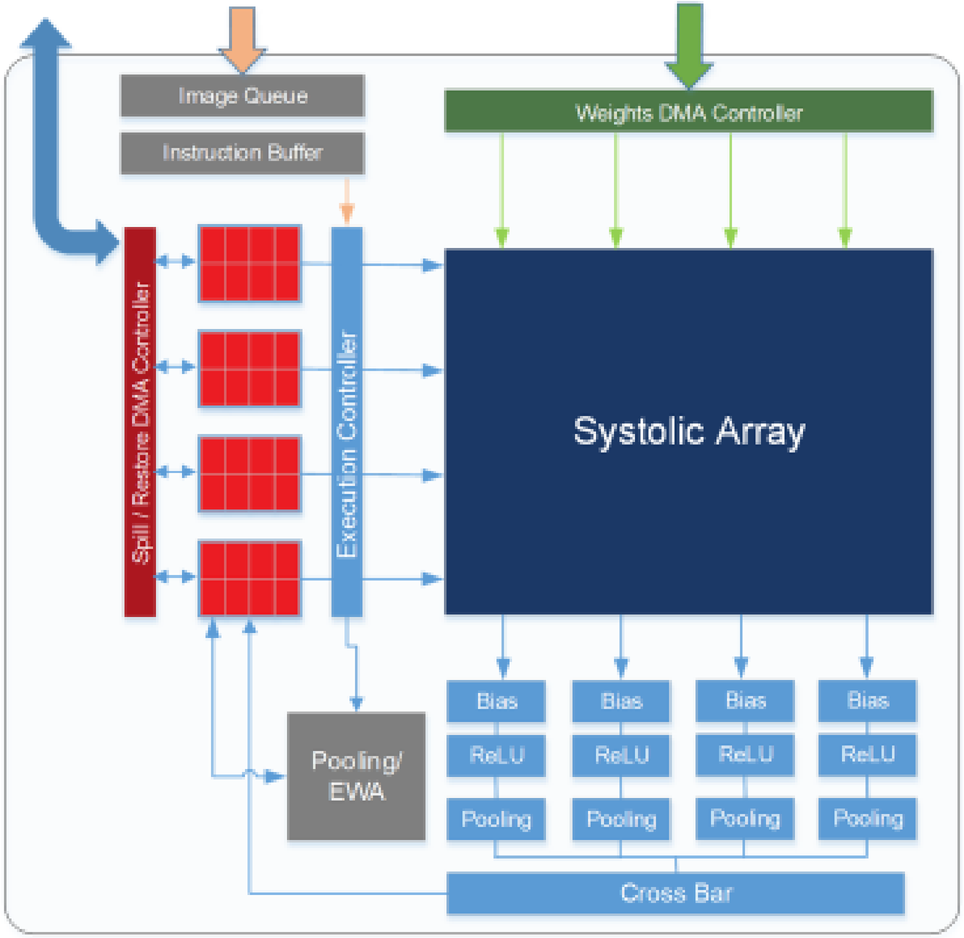
xDNN-v3
Q4CY18

- New Systolic Array Implementation: 2.2x lower latency
- Instruction Level Parallelism – non-blocking data movement
- Batch=1 for Int8 – lower latency
- Feature compatible with xDNN-v2
- 720+ MHz

XDNN v3 Feature Set

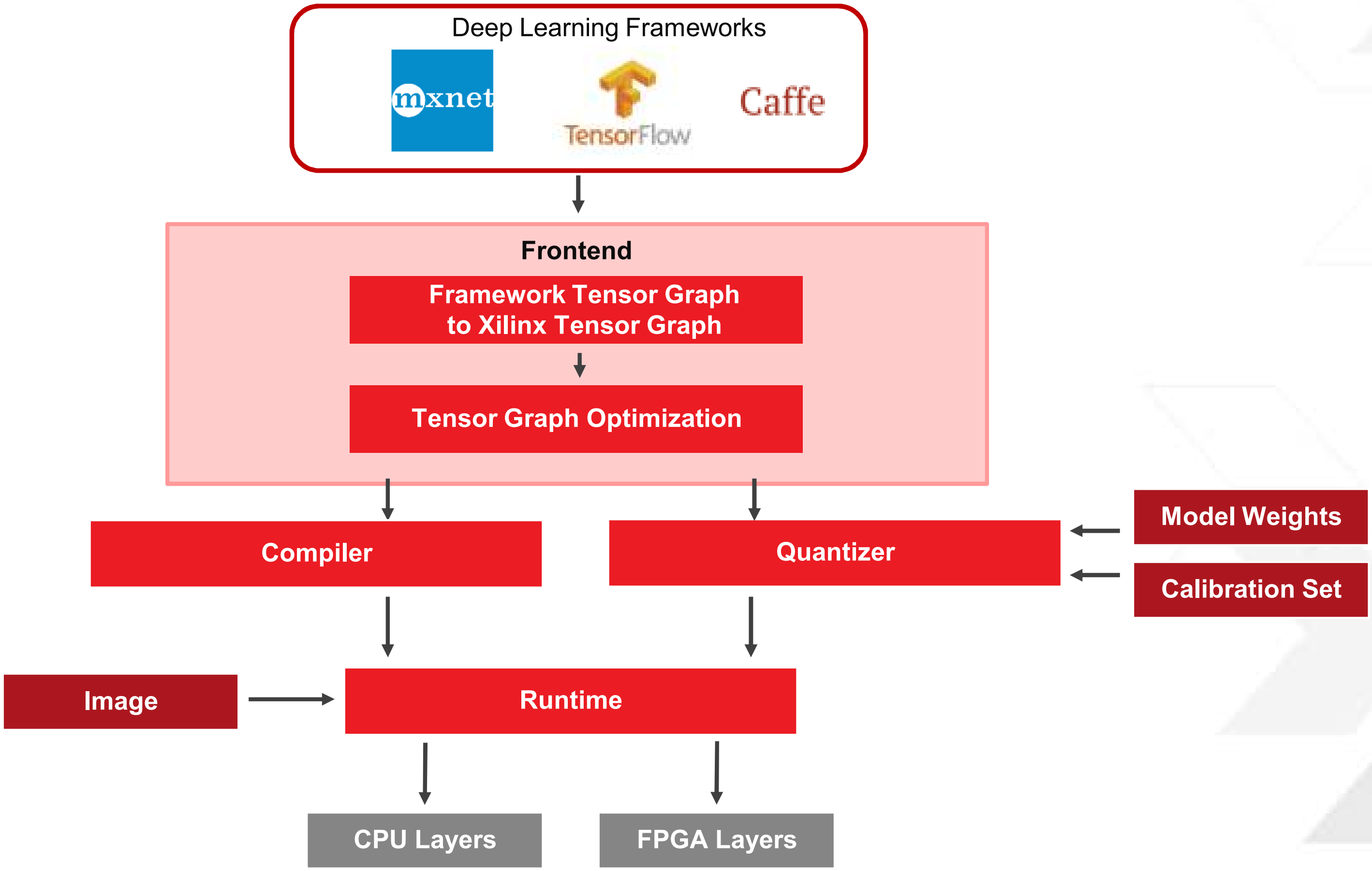
Features		Description	
Supported Operations	Convolution / Deconvolution / Convolution Transpose	Kernel Sizes	W: 1-15; H:1-15
		Strides	W: 1,2,4,8; H: 1,2,4,8
		Padding	Same, Valid
		Dilation	Factor: 1,2,4
		Activation	ReLU/pReLU
		Bias	Value Per Channel
		Scaling	Scale & Shift Value Per Channel
	Max Pooling	Kernel Sizes	W: 1-15; H:1-15
		Strides	W: 1,2,4,8; H: 1,2,4,8
		Padding	Same, Valid
	Avg Pooling	Kernel Sizes	W: 1-15; H:1-15
		Strides	W: 1,2,4,8; H: 1,2,4,8
		Padding	Same, Valid
	Element-wise Add	Width & Height must match; Depth can mismatch.	
	Memory Support	On-Chip Buffering, DDR Caching	
	Expanded set of image sizes	Square, Rectangular	
	Upsampling	Strides	Factor: 2,4,8,16
Miscellaneous	Precision	Int16-bit or Int8-bit	

Xilinx DNN Processor (xDNN)



- Configurable Overlay Processor
- DNN Specific Instruction Set
  - Convolution, Max Pool etc.
- Any Network, Any Image Size
- High Frequency & High Compute Efficiency
- Compile and run new networks

xDNN Compiler + Runtime

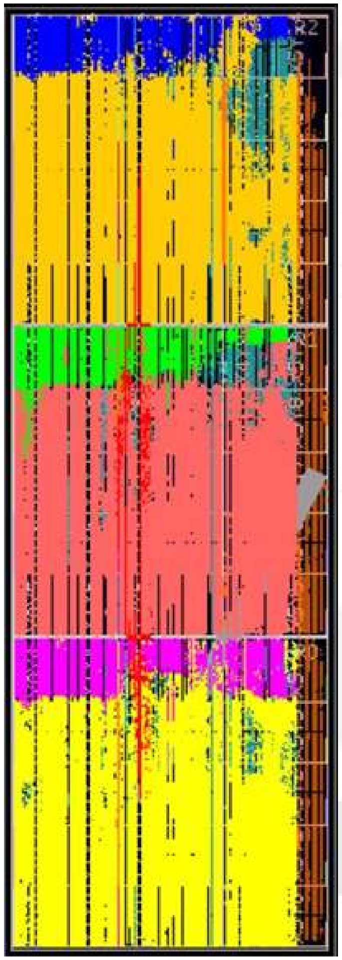


https://github.com/Xilinx/ml-suite

xDNN v3 Implementation on Alveo U200

- 3 Large 96x16 PEs– 1 in each SLR – 5.2 ML Shell
- Kernels @ 720 MHz/360MHz

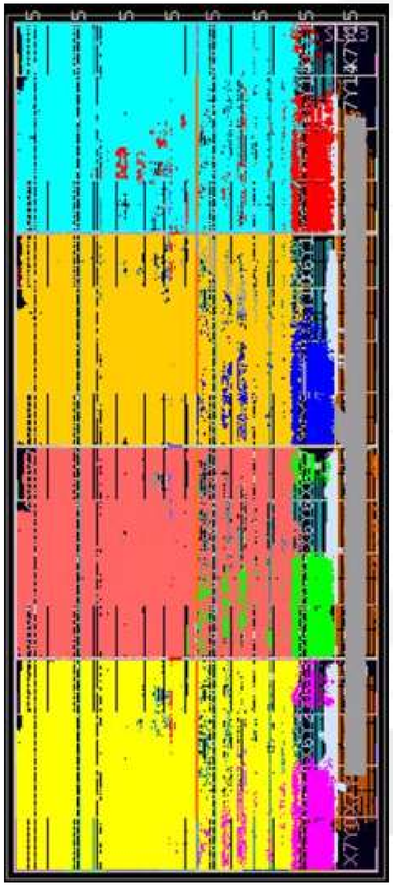
Resource	Count	Utilization
LUTs	658k	52%
DSPs	5661	80%
BRAM	1258	58%
URAM	864	92%



xDNN v3 Implementation on Alveo U250

- 4 Large 96x16 PEs– 1 in each SLR – standard 5.2 Shell
- Kernels at 700 MHz/350 MHz

Resource	Count	Utilization
LUTs	876k	51%
DSPs	7548	62%
BRAM	1632	61%
URAM	1152	90%



ML Suite Overlays with xDNN Processing Engines

Adaptable

> AI algorithms are changing rapidly

> Adjacent acceleration opportunities

Realtime

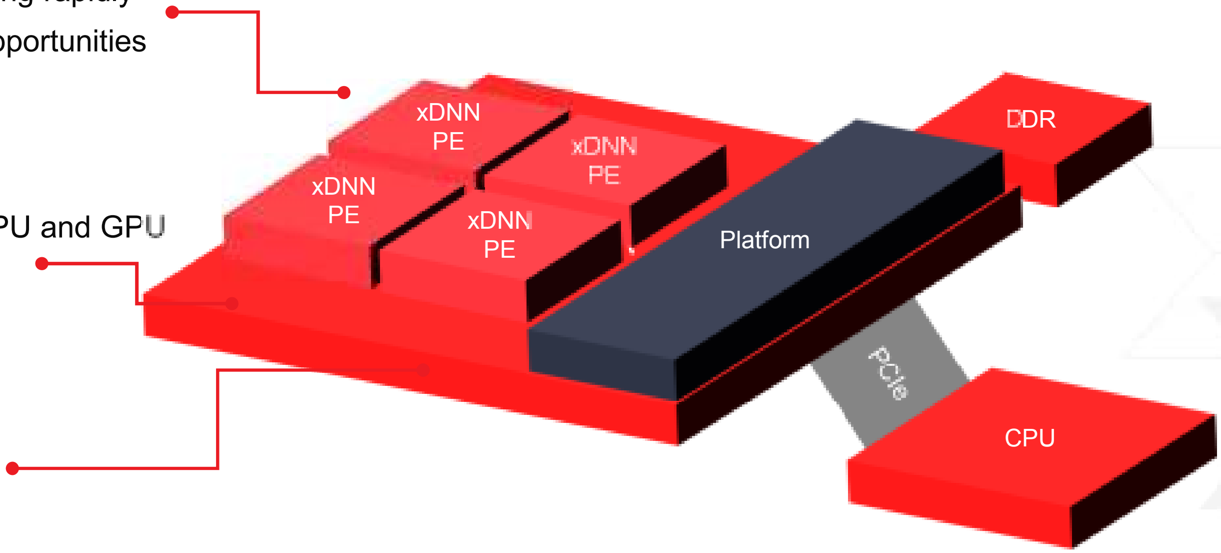
> 10x Low latency than CPU and GPU

> Data flow processing

Efficient

> Performance/watt

> Low Power



Flexible: Multi-Network Configuration

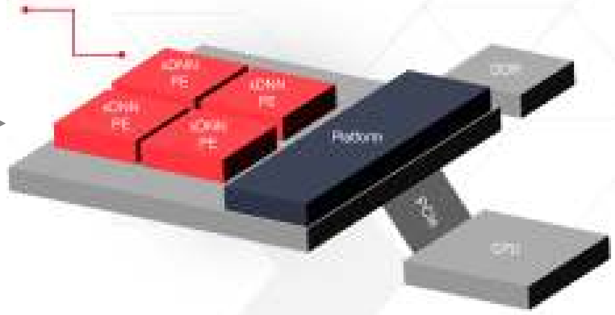
Application: Image Classification	Application: Speech	Application: Localization	Application: Object Detection
Model: Resnet50	Model: Googlenet v1	Model: FaceNet	Model: Yolo v2
Framework: Caffe	Framework: MxNet	Framework: TesnorFlow	Framework: Caffe

xfDNN

SDx Runtime

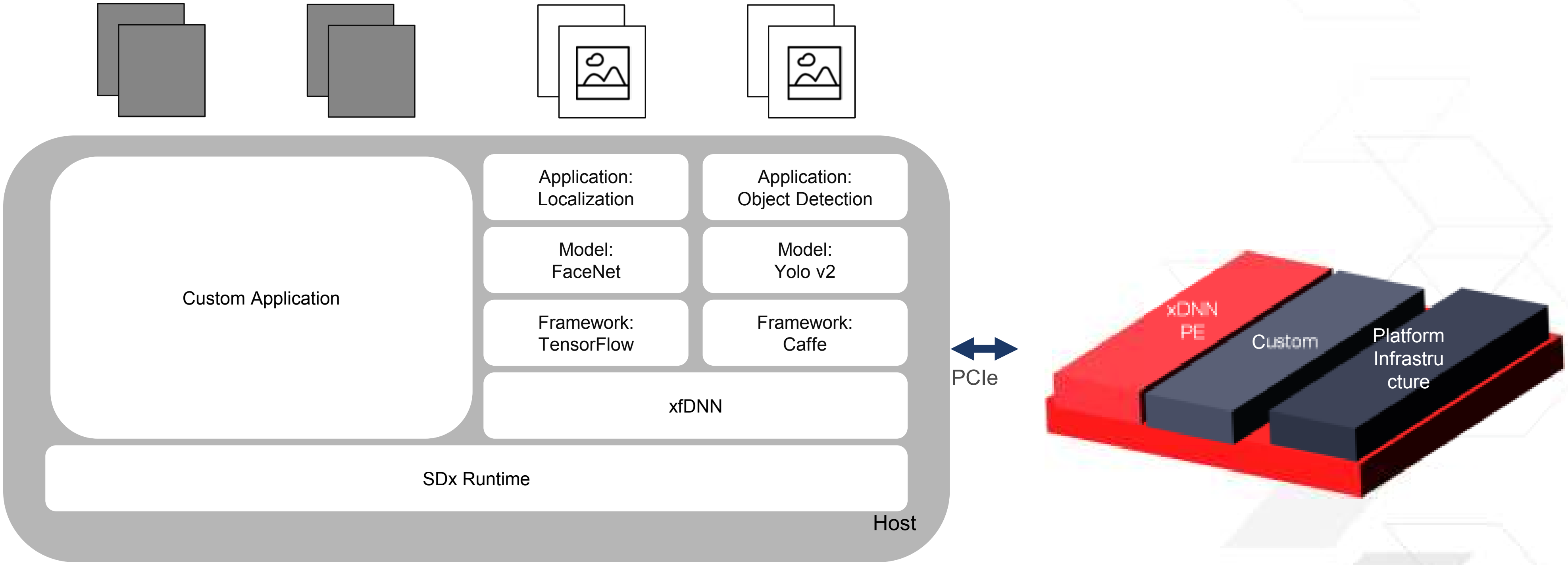
Host

PCIe



1 FPGA Provides 4 Virtual Accelerators
For Real Time Deep Learning

Flexible: Bring Your own IP!



Custom Application

Application:
Localization

Application:
Object Detection

Model:
FaceNet

Model:
Yolo v2

Framework:
TensorFlow

Framework:
Caffe

xfDNN

SDx Runtime

Host

PCIe

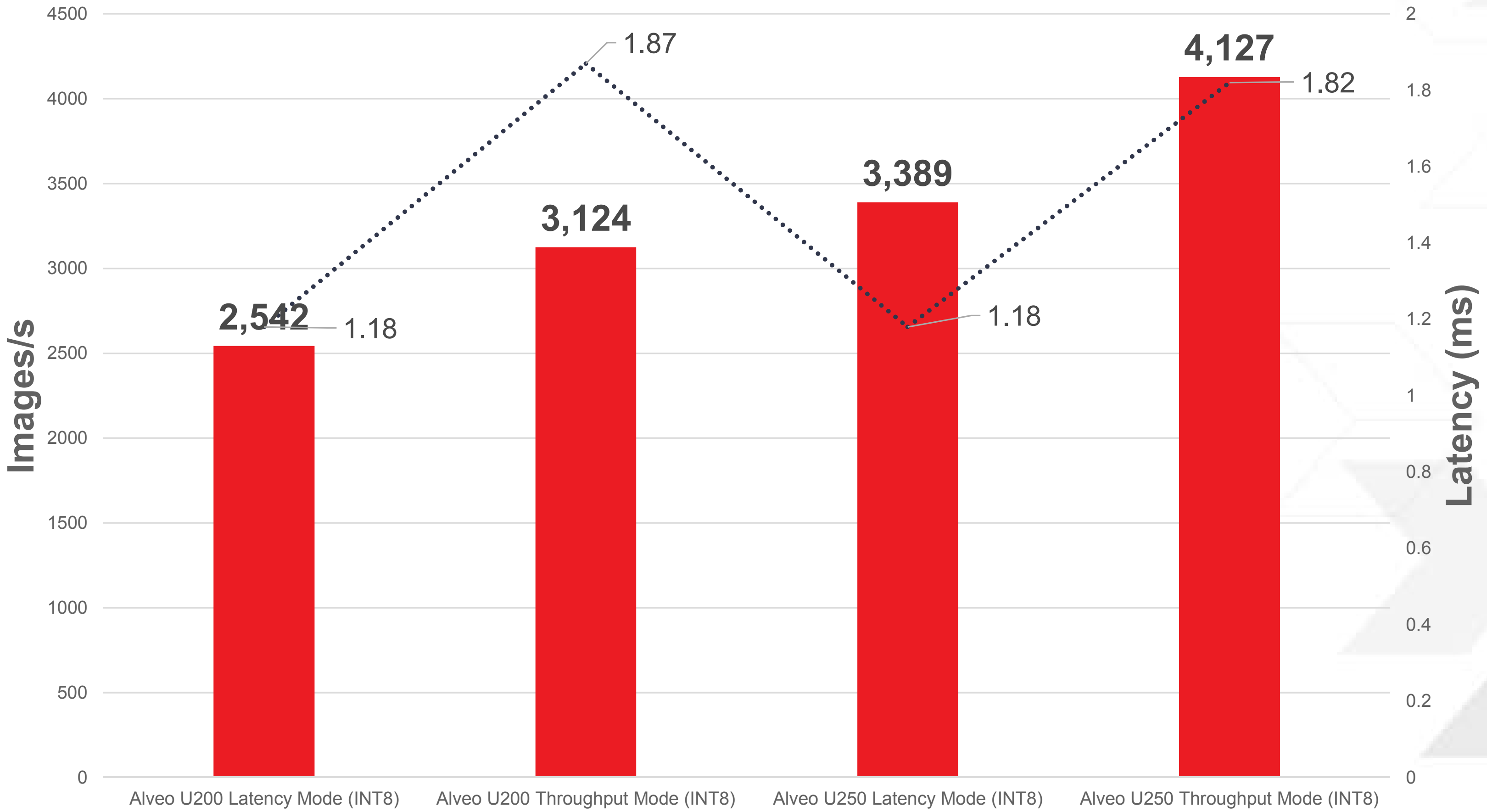
xDNN
PE

Custom

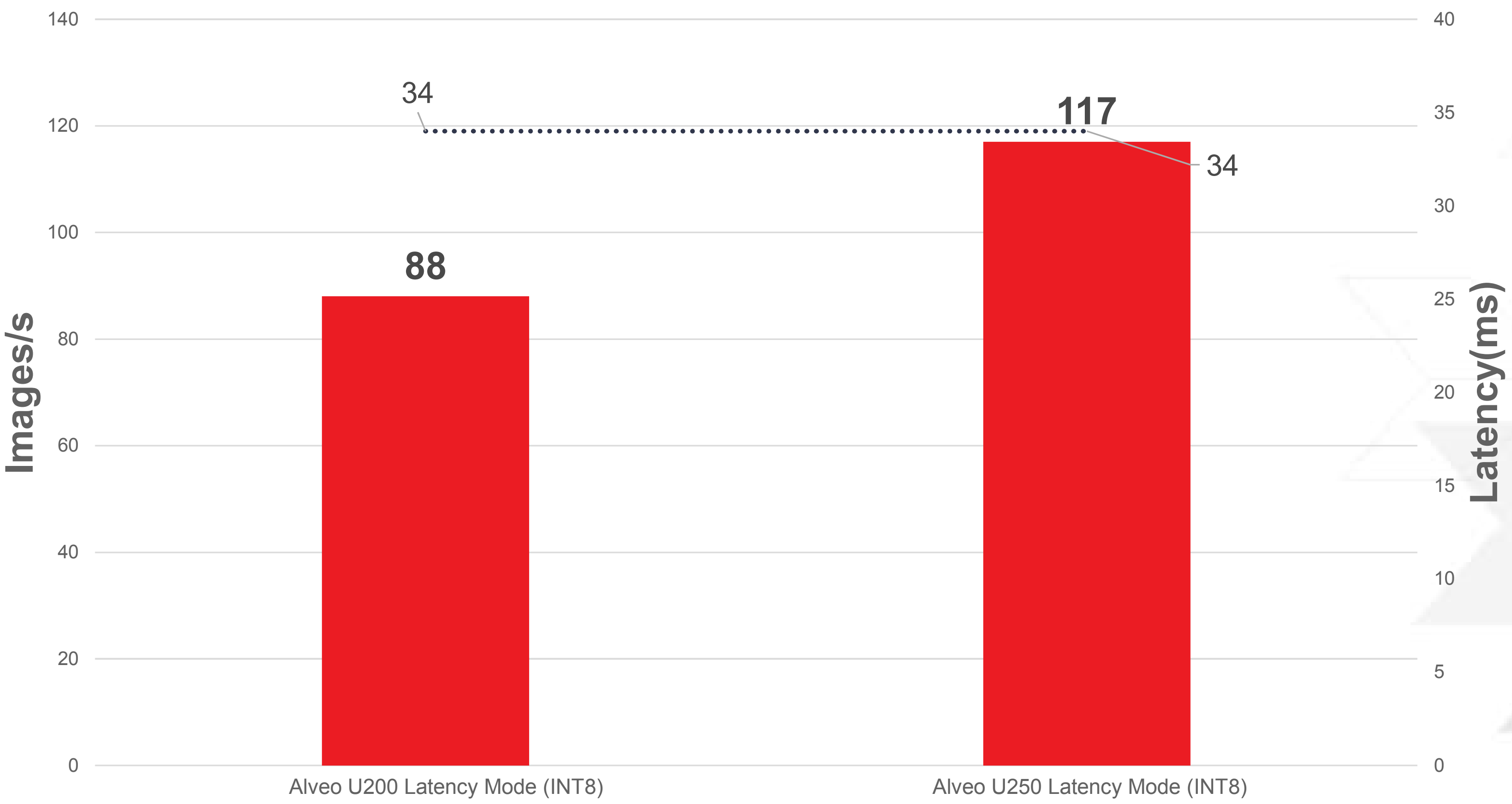
Platform
Infrastru
cture

**Integrate Custom Applications Directly
with xDNN Processing Engines**

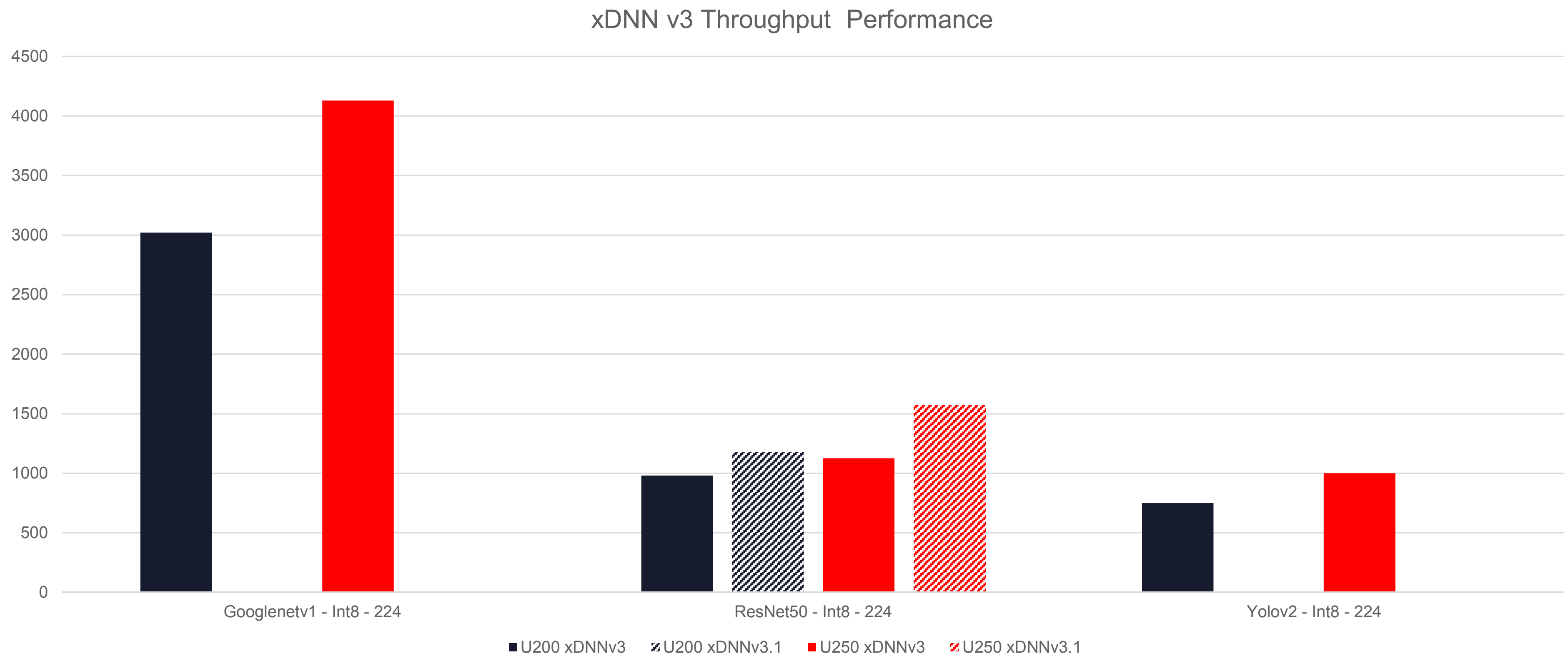
xDNN GoogLeNet v1 Performance – Image Size 224x224



xDNN YOLO v2 Performance – Image Size 608x608

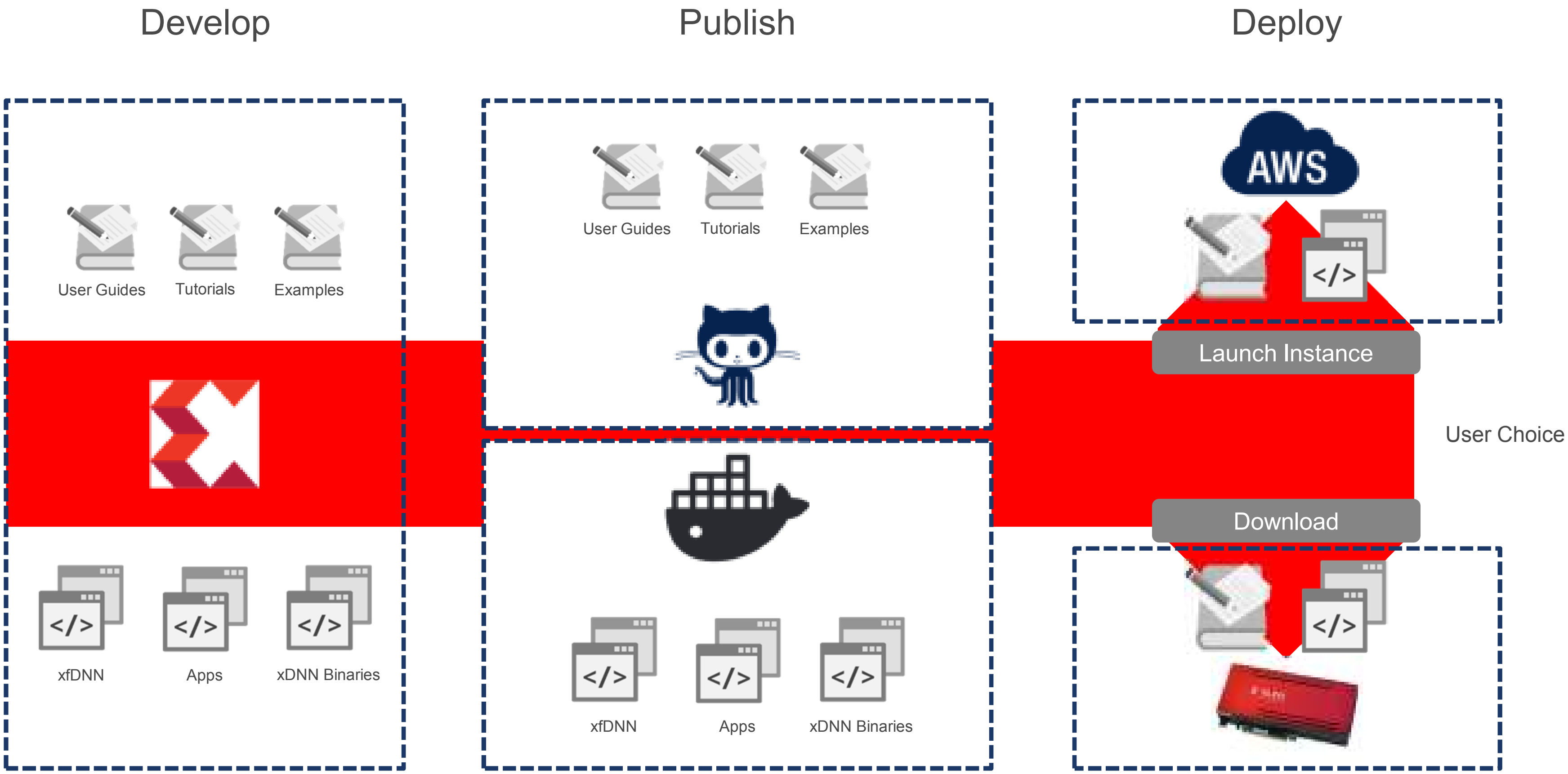


xDNN Real Time Performance



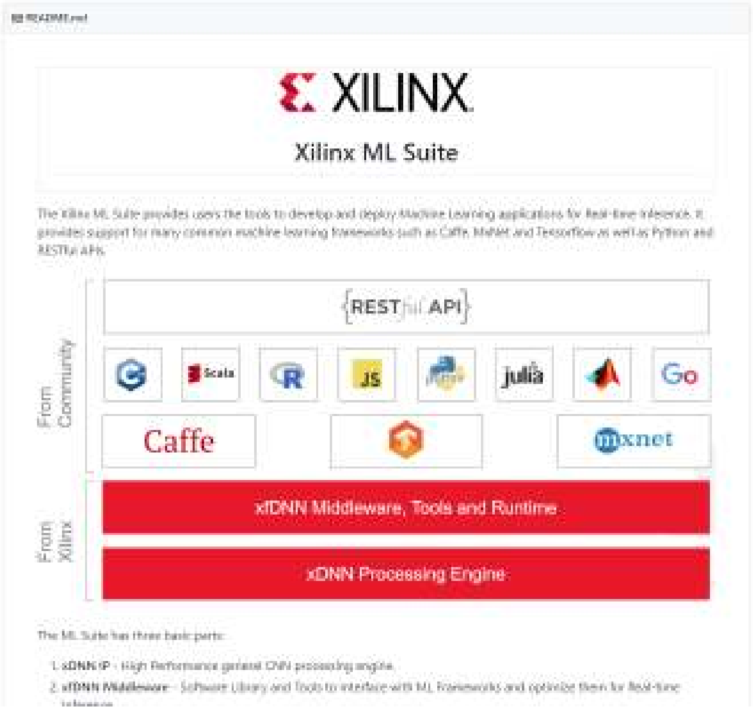
Xilinx Inference Middleware - xfDNN

Unified Simple User Experience from Cloud to XBB



Xilinx ML Suite

- **ML Suite**
  - Supported Frameworks:
    - Caffe
    - MxNet
    - Tensorflow
    - Python Support
    - Darknet
  - Jupyter Notebooks available:
    - Image Classification with Caffe
    - Using the xfDNN Compiler w/ a Caffe Model
    - Using the xfDNN Quantizer w/ a Caffe Model
  - Pre-trained Models
    - Caffe 8/16-bit
      - GoogLeNet v1
      - ResNet50
      - Flowers102
      - Places365
    - Python 8/16-bit
      - Yolov2
    - MxNet 8/16-bit
      - GoogLeNet v1
  - xfDNN Tools
    - Compiler
    - Quantizer



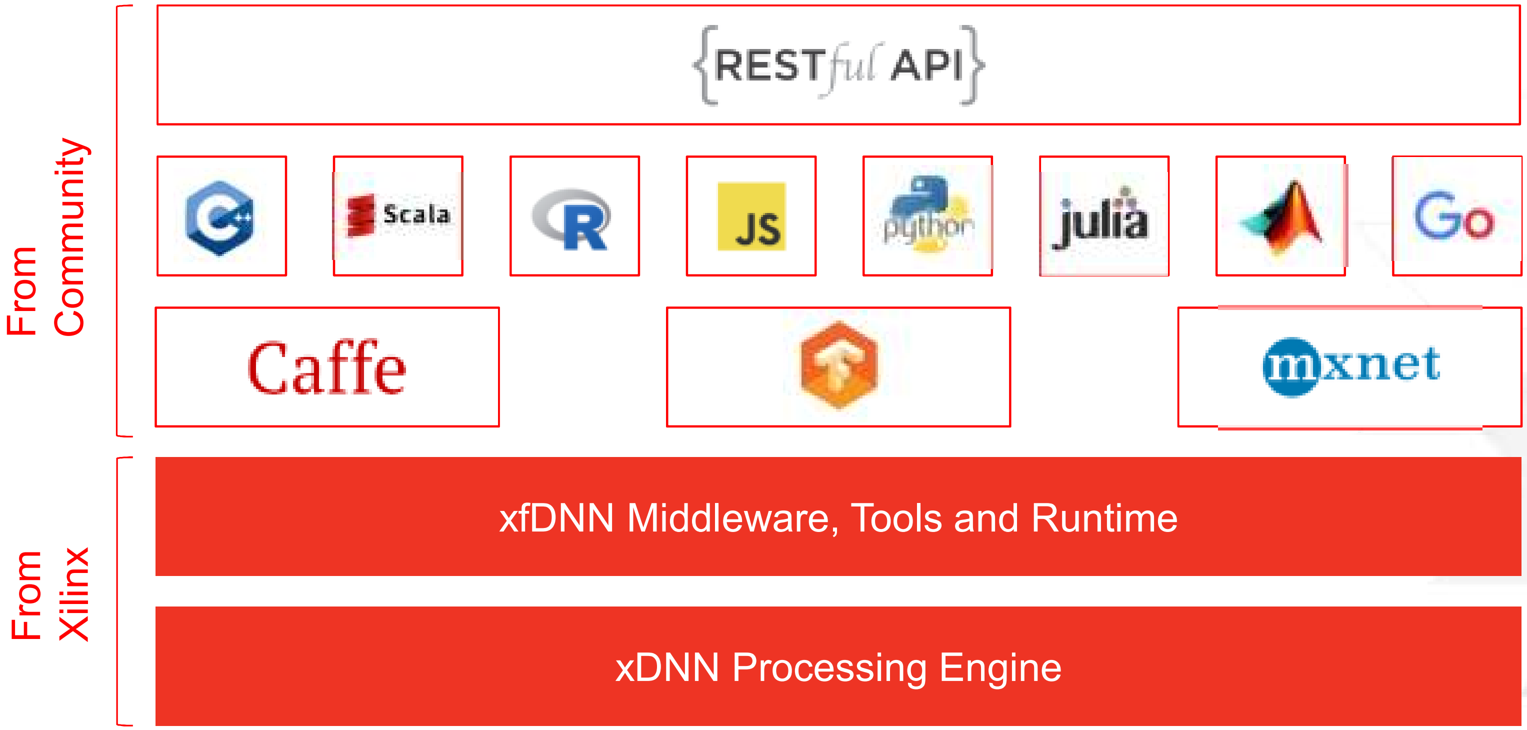
The Xilinx ML Suite provides users the tools to develop and deploy Machine Learning applications for Real-time Inference. It provides support for many common machine learning frameworks such as Caffe, MxNet and Tensorflow as well as Python and RESTful APIs.

The ML Suite has three basic parts:

1. **xDNN IP** - High Performance general CNN processing engine.
2. **xfDNN Middleware** - Software Library and Tools to interface with ML Frameworks and optimize them for Real-time Inference.

https://github.com/Xilinx/ml-suite

Seamless Deployment with Open Source Software



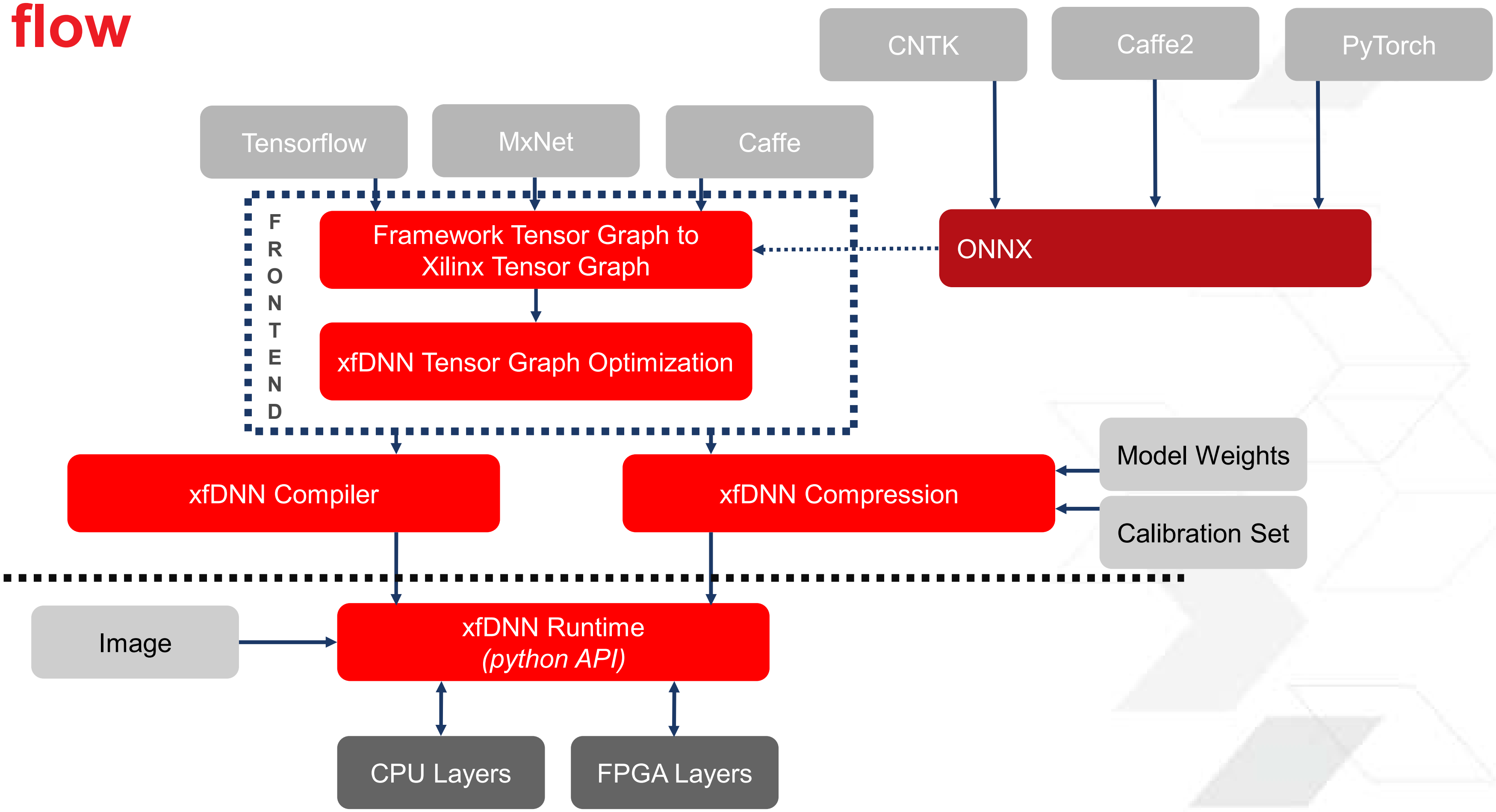
From Community

From Xilinx

xfDNN Middleware, Tools and Runtime

xDNN Processing Engine

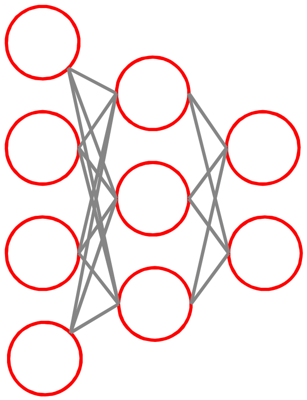
xfDNN flow



https://github.com/Xilinx/ml-suite

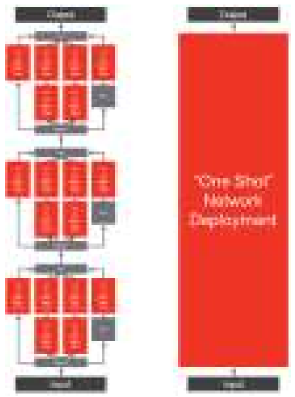
xfDNN Inference Toolbox

Graph Compiler



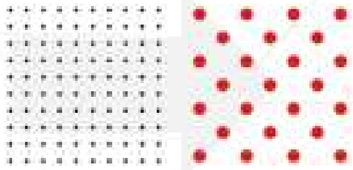
- Python tools to quickly compile networks from common Frameworks – Caffe, MxNet and Tensorflow

Network Optimization



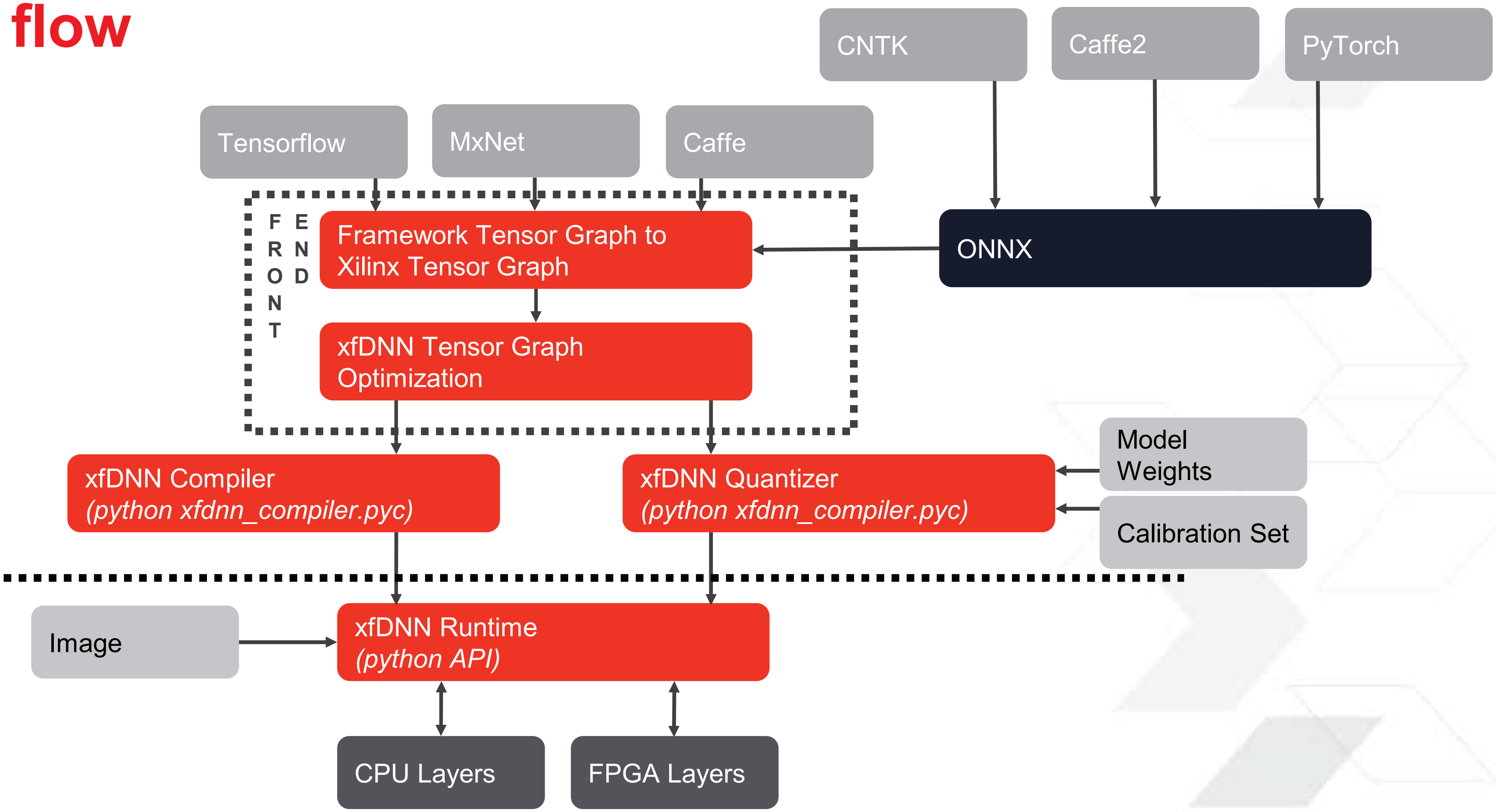
- Automatic network optimizations for lower latency by fusing layers and buffering on-chip memory

xfDNN Quantizer



- Quickly reduce precision of trained models for deployment
- Maintains 32bit accuracy at 8 bit within 2%

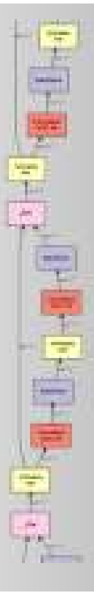
xfDNN flow



https://github.com/Xilinx/ML-Development-Stack-From-Xilinx

xfDNN Graph Compiler

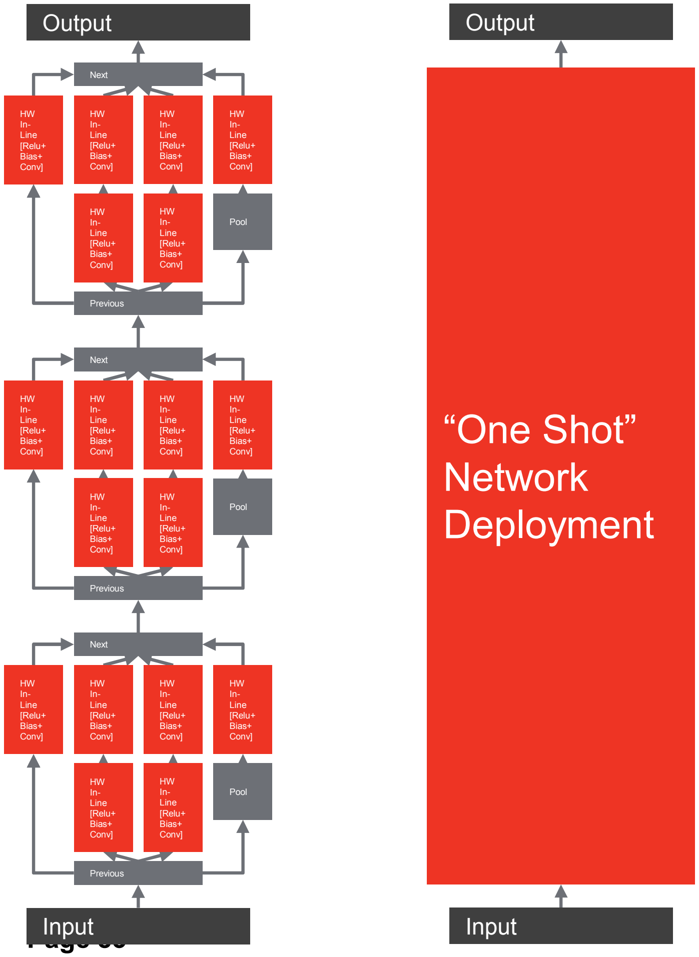
Pass in a Network



xfDNN
Graph Compiler

Microcode for xDNN is Produced

xfDNN Network Deployment



Fused Layer Optimizations

- Compiler can merge nodes
  - (Conv or EltWise)+Relu
  - Conv + Batch Norm
- Compiler can split nodes
  - Conv 1x1 stride 2 -> Maxpool+Conv 1x1 Stride 1

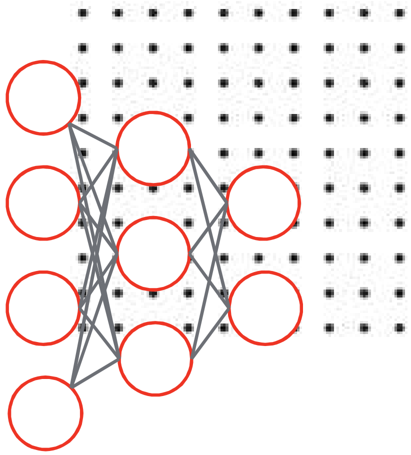
On-Chip buffering reduces latency and increases throughput

- xfDNN analyzes network memory needs and optimizes scheduler
  - For Fused and "One Shot" Deployment

"One Shot" deploys entire network to FPGA

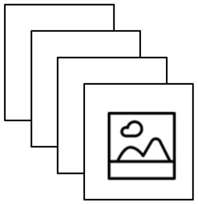
- Optimized for fast, low latency inference
- Entire network, schedule and weights loaded only once to FPGA

xfDNN Quantizer: Fast and Easy



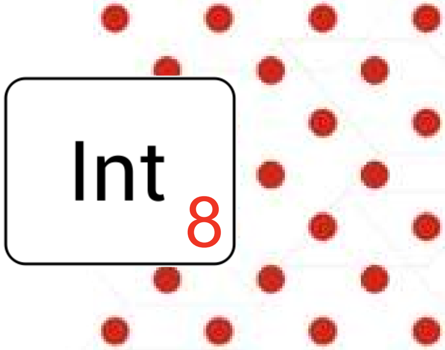
1) Provide FP32 network and model

- E.g., prototxt and caffemodel



2) Provide a small sample set, no labels required

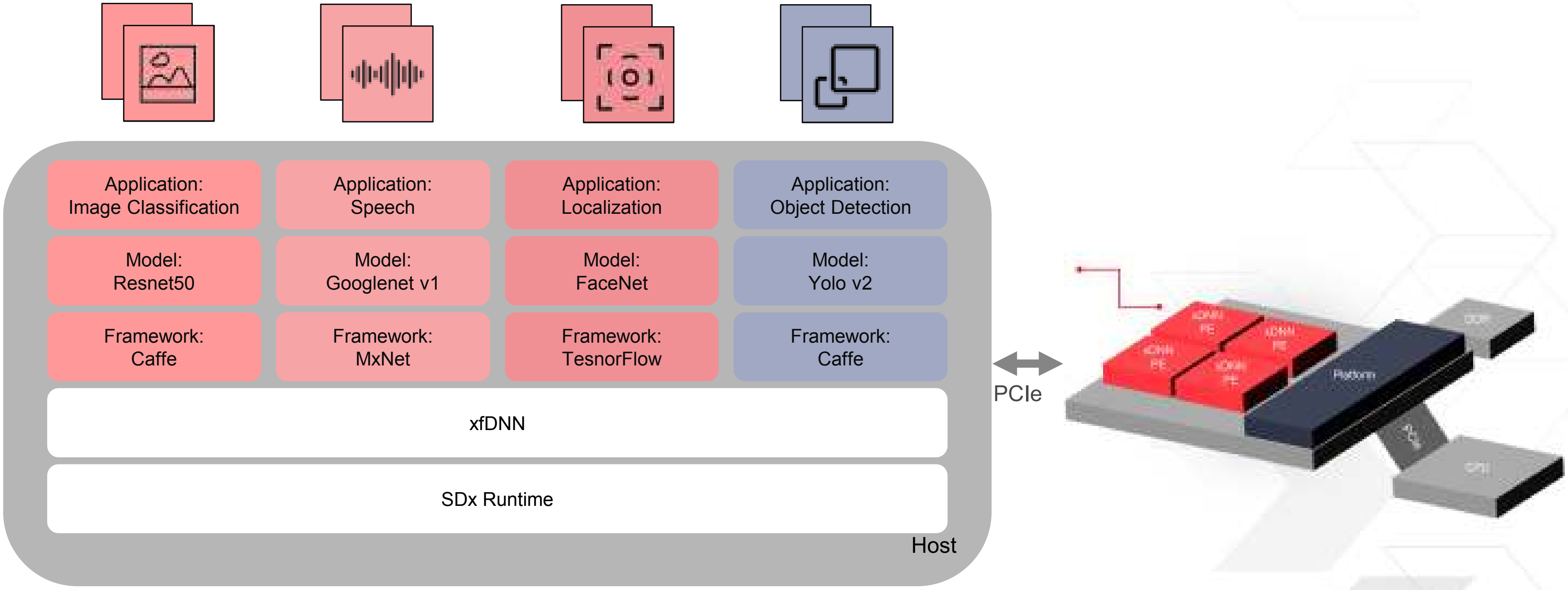
- 16 to 512 images



3) Specify desired precision

- Quantizes to <8 bits to match Xilinx's DSP

Flexible: Multi-Network Configuration



Application: Image Classification	Application: Speech	Application: Localization	Application: Object Detection
Model: Resnet50	Model: Googlenet v1	Model: FaceNet	Model: Yolo v2
Framework: Caffe	Framework: MxNet	Framework: TesnorFlow	Framework: Caffe

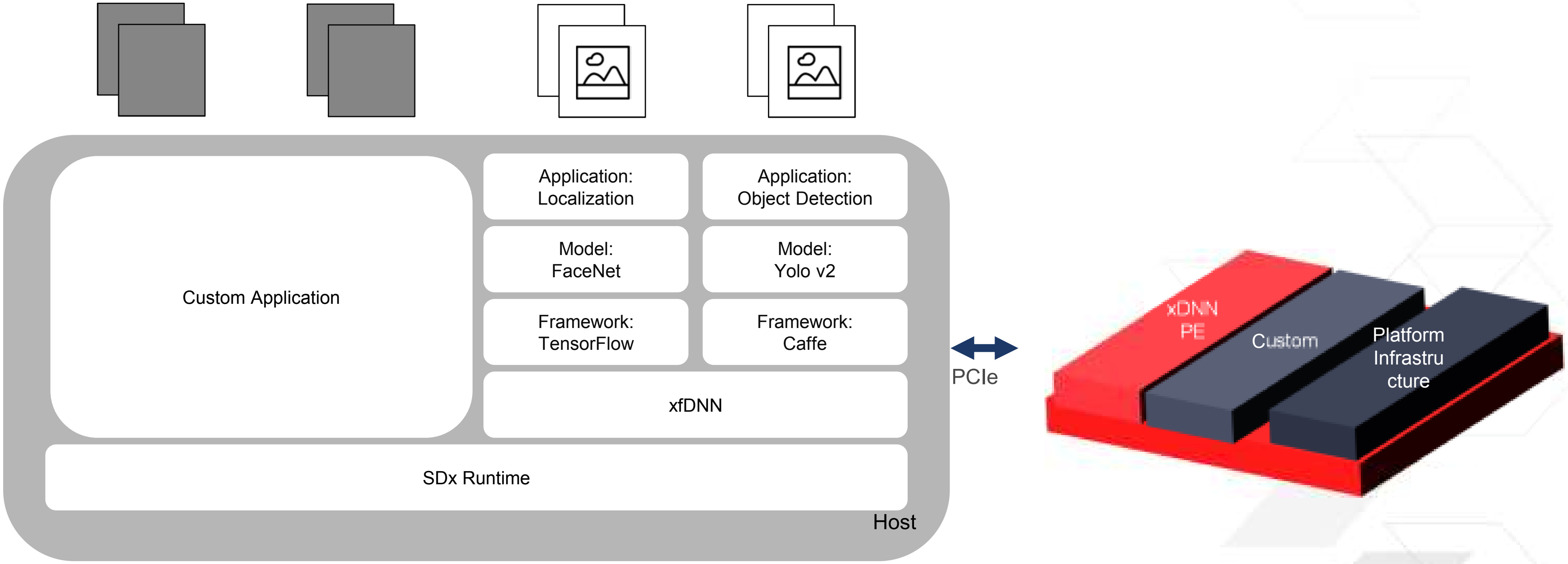
xfDNN

SDx Runtime

Host

1 FPGA Provides 4 Virtual Accelerators For Real Time Deep Learning

Flexible: Bring Your own IP!



Custom Application

Application: Localization

Application: Object Detection

Model: FaceNet

Model: Yolo v2

Framework: TensorFlow

Framework: Caffe

xfDNN

SDx Runtime

Host

PCIe

Integrate Custom Applications Directly with xDNN Processing Engines

X + ML

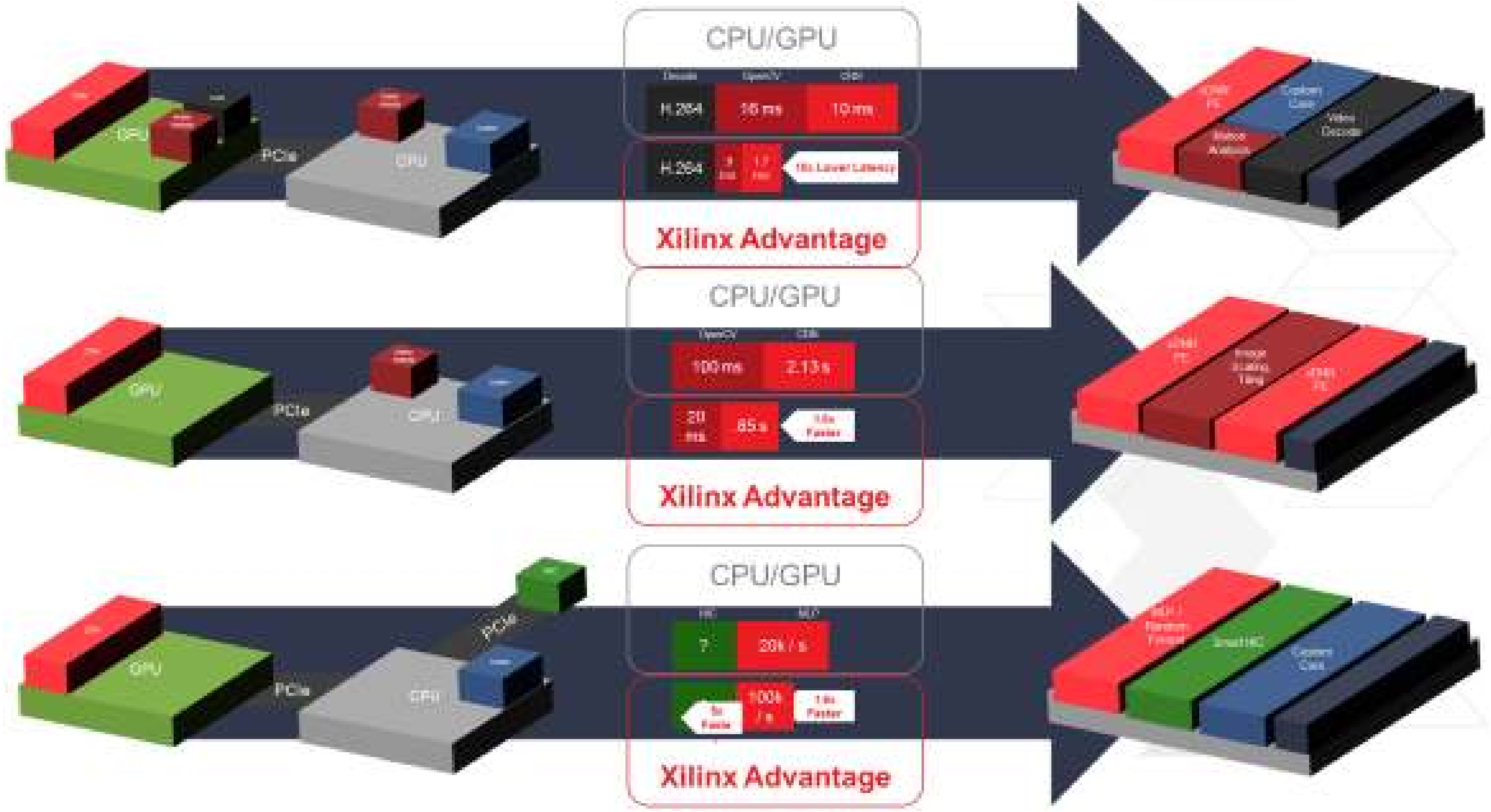
XILINX

X + ML Focus Applications Summary

Smart City / Cloud Surveillance
- 10x Lower Latency

High Resolution Imaging
- 1.6x Faster

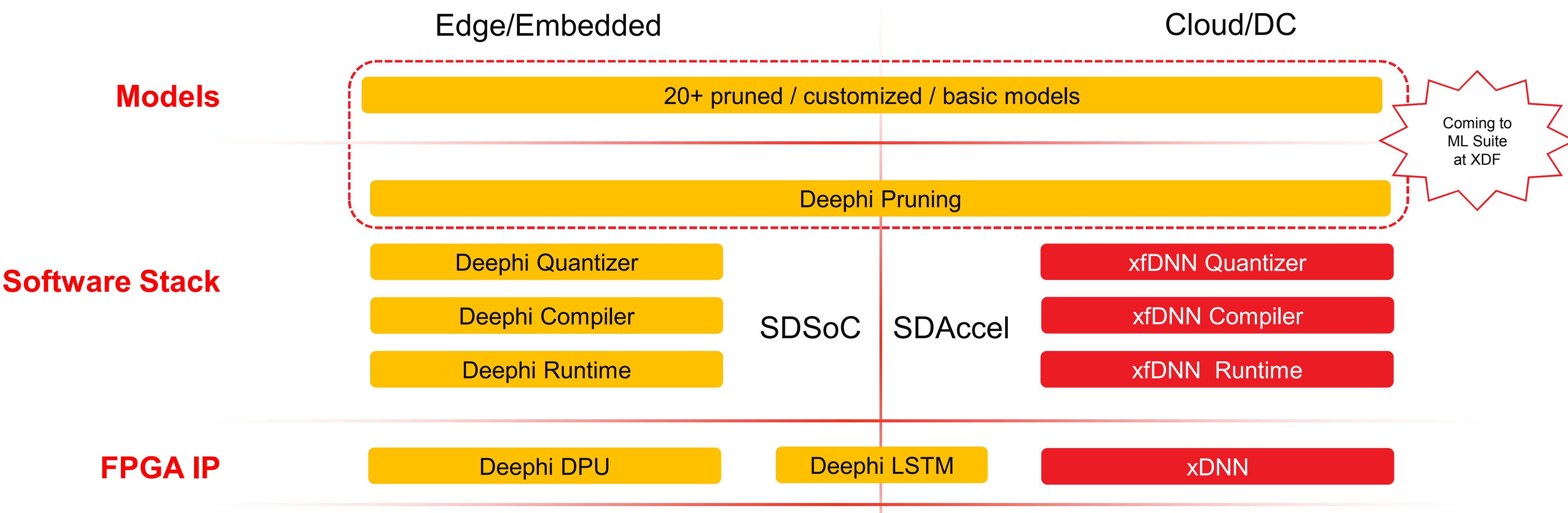
Security / Anomaly / Malware
- 5x Faster



ML Suite + Deephi

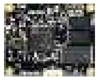
XILINX

Integrated Xilinx-Deephi Roadmap



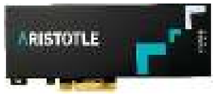
Platforms

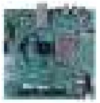
Z7020 Board


Z7020 SOM

ZU2/3 SOM

ZU2/3 Card


ZU9 Card


ZCU102

ZCU104 pyriç

Ultra96 nx

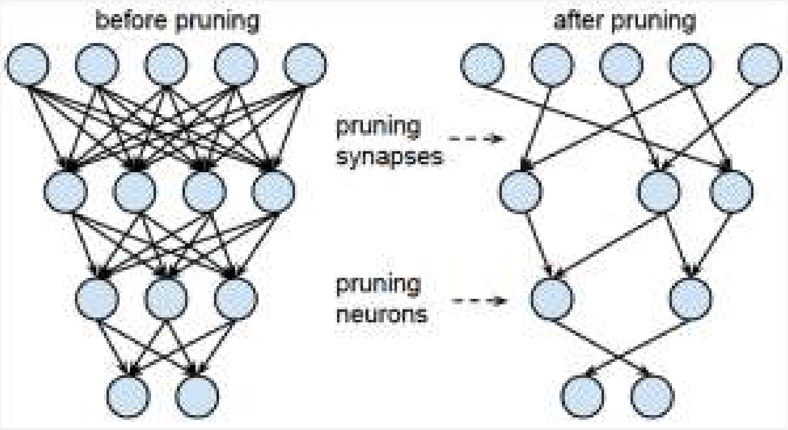
Xilinx U200, U250, U280

marketplace NIMBIX

Alibaba Cloud aliyun.com HUAWEI

Xilinx Pruning Overview

Deep compression
Makes algorithm smaller and lighter



Highlight

- 1/3 Weight number
- 1/10 Bandwidth load
- 1/10 Model size
- 3 X Perfor-mance

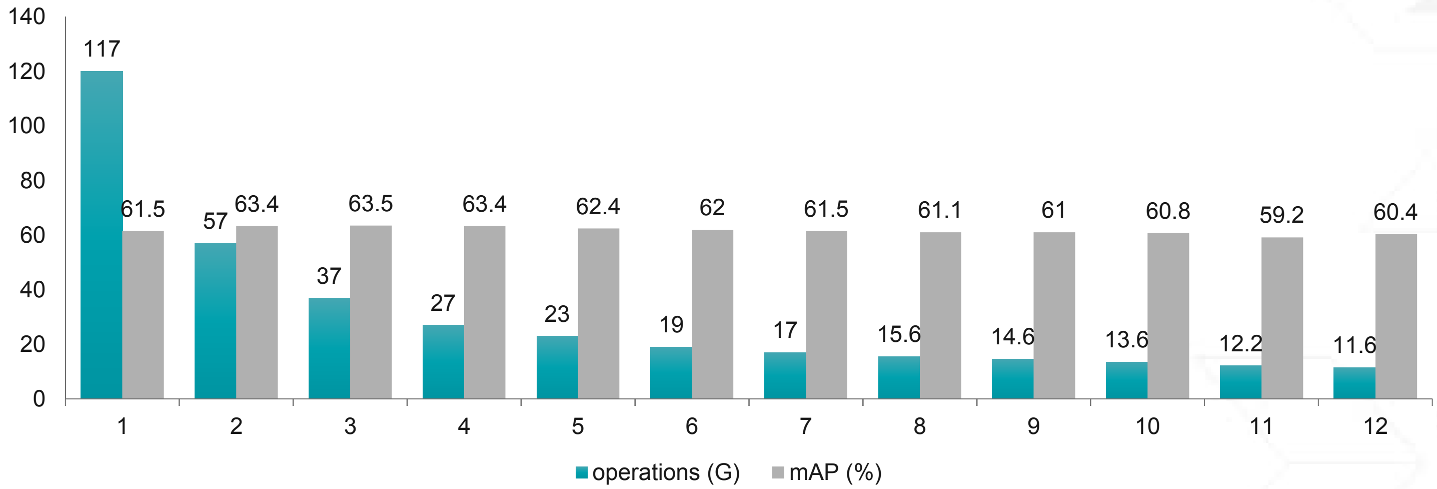
Compression efficiency

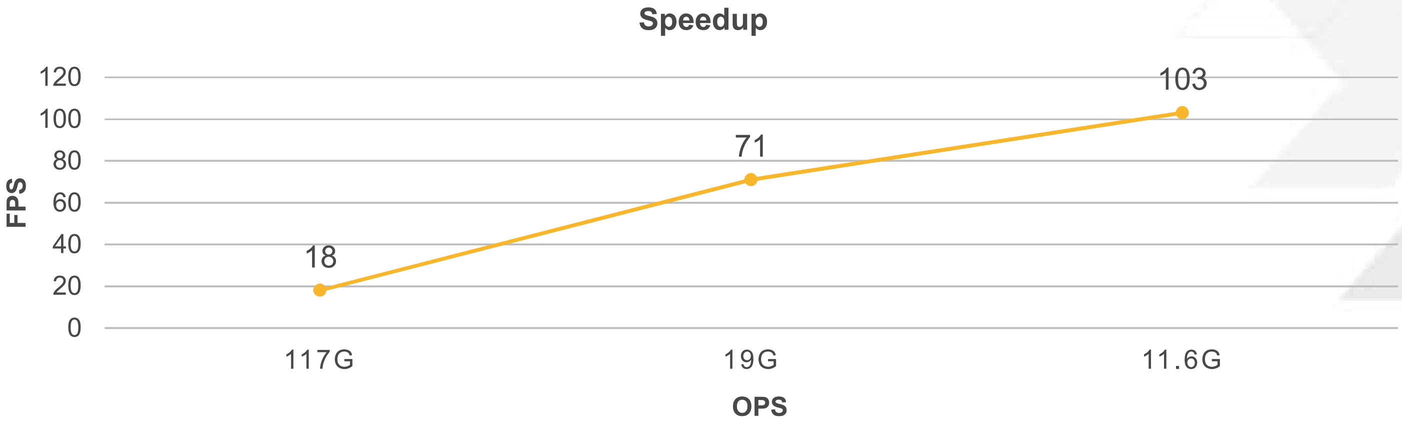
Deep Compression Tool can achieve significant compression on **CNN** and **RNN**

Accuracy

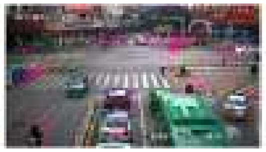
Algorithm can be **compressed 7 times without losing accuracy** under SSD object detection framework

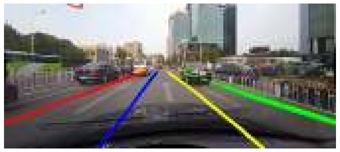
Pruning Example - SSD

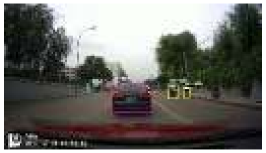




Supported DNN (Deep Neural Network) by Applications





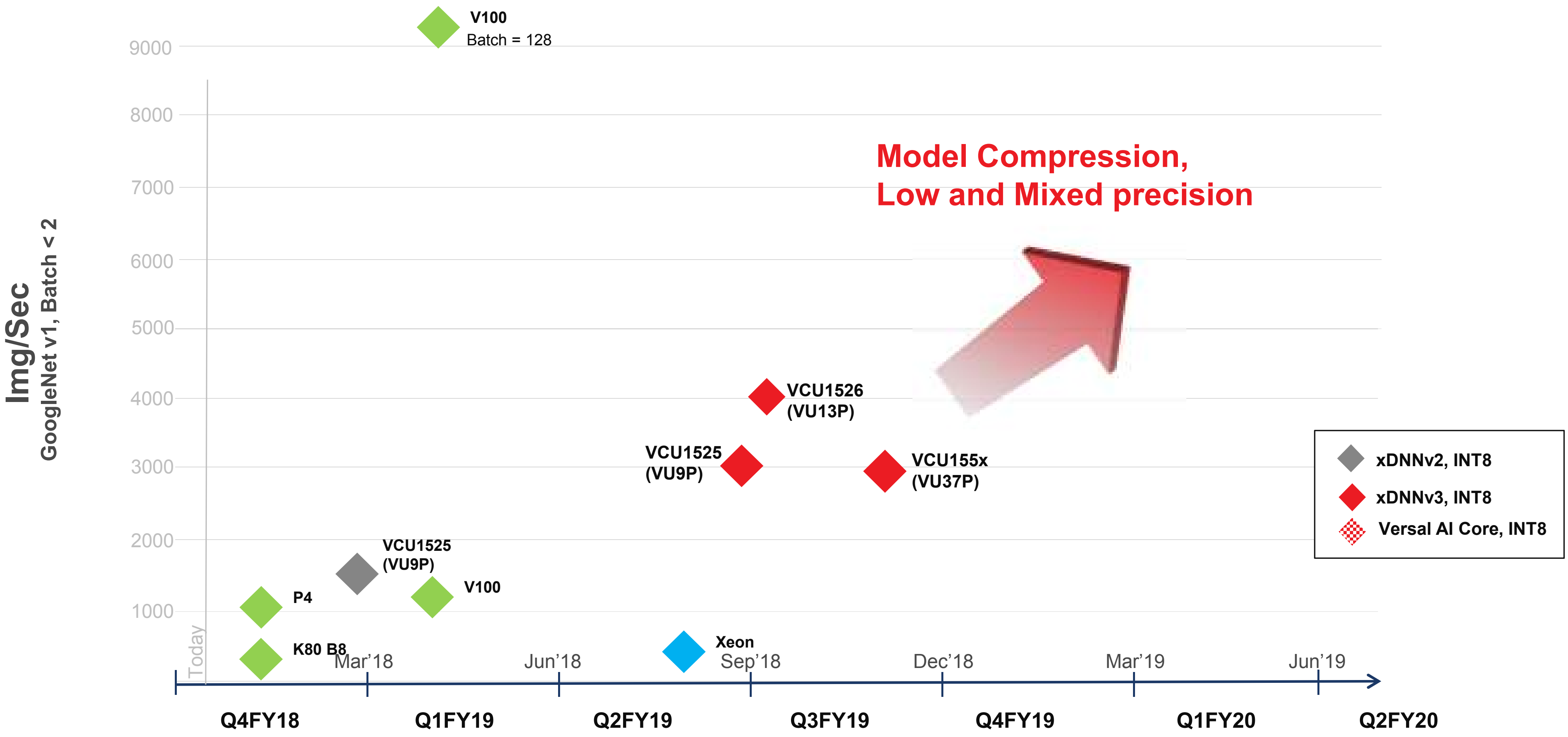


Application	NTT Request	Function	Algorithm
Face		Face detection	SSD, Densebox
		Landmark Localization	Coordinates Regression
		Face recognition	ResNet + Triplet / A-softmax Loss
		Face attributes recognition	Classification and regression
Pedestrian	1	Pedestrian Detection (Crowd Volume)	SSD
		Pose Estimation	Coordinates Regression
		Person Re-identification	ResNet + Loss Fusion
Video Analytics	1	Object detection	SSD, RefineDet
		Pedestrian Attributes Recognition	GoogleNet
		Car Attributes Recognition	GoogleNet
	1	Car Logo Detection	DenseBox
	1	Car Logo Recognition	GoogleNet + Loss Fusion
	1	License Plate Detection	Modified DenseBox
	1	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

ML Suite Performance Roadmap

ACAP

7x Performance Improvement



CPU: https://mxnet.incubator.apache.org/faq/perf.html

Nvidia: https://images.nvidia.com/content/pdf/inference-technical-overview.pdf

P4 = int8, v100 = fp16

Visit Xilinx.com/ML for more information

https://www.xilinx.com/applications/megatrends/machine-learning.html

XILINX

Adaptable.
Intelligent.

XILINX