









Xilinx ML Suite Overview

Yao Fu
System Architect – Data Center Acceleration



Xilinx Accelerated Computing Workloads

	Machine Learning Inference Image classification and object detection	10x	
	Video Streaming Frame rate for HEVC & VP9 encoding	10x	
	Genomics 20 min vs. 33 hours for whole genome analysis	100x	
	Big Data Analytics 40 min vs. 60 hours for logfile query	90x	



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

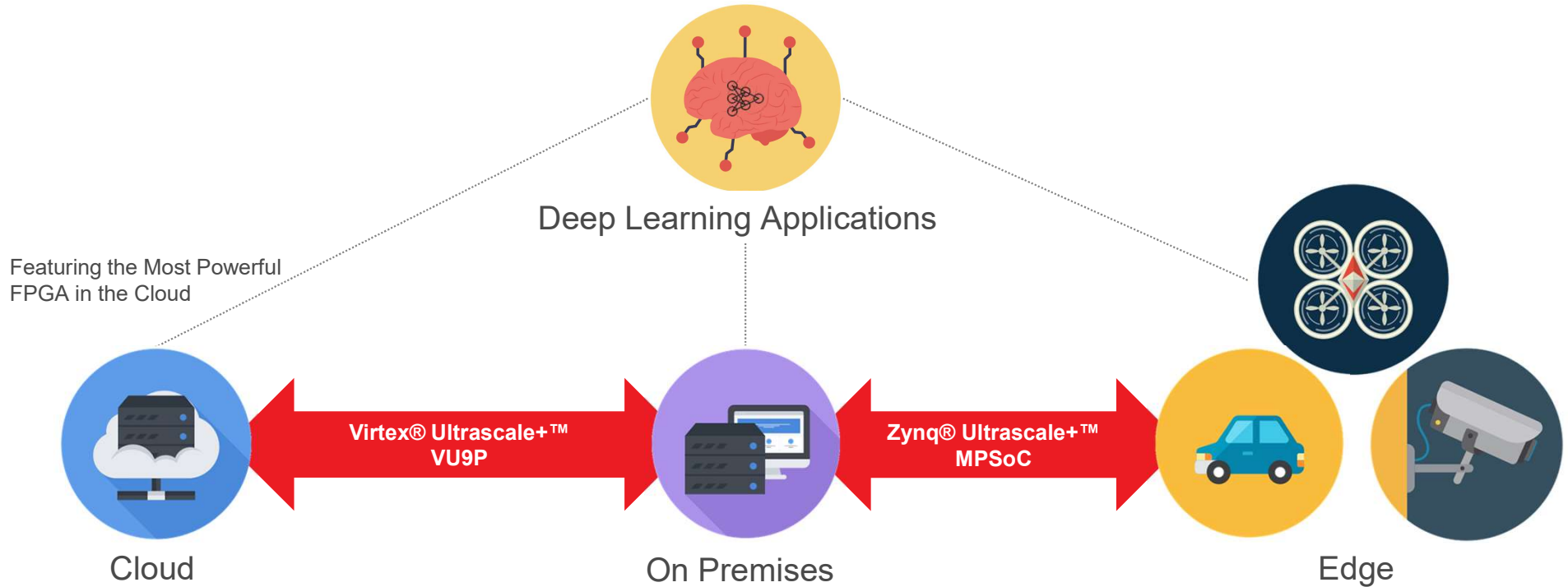


Medical Bioinformatics



Autonomous Vehicles

Accelerating AI Inference into Your Cloud Applications



© Copyright 2018 Xilinx



Xilinx ML Suite - AWS Marketplace



➤ 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

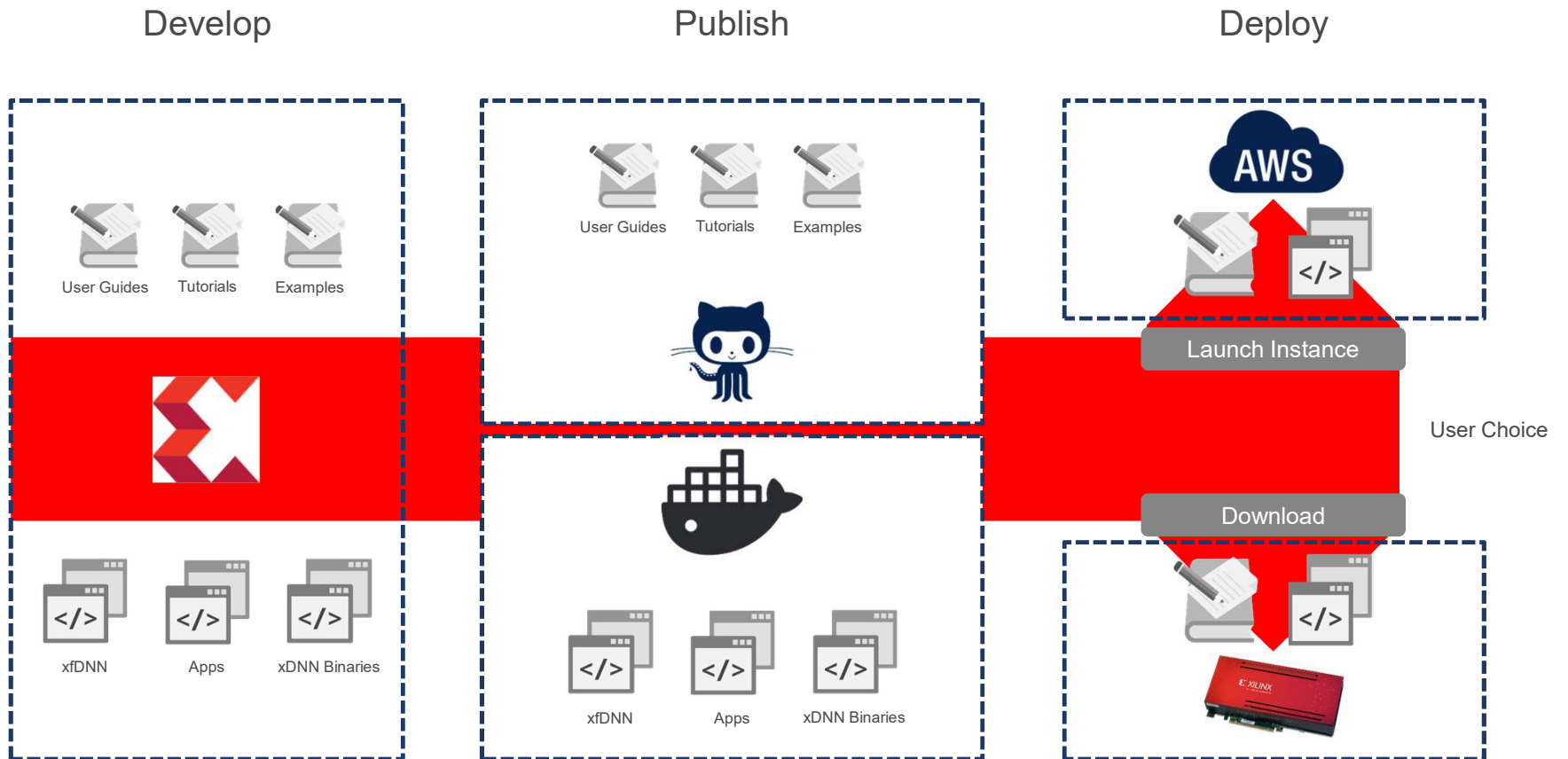
>> xFDNN Tools

- Compiler
- Quantizer

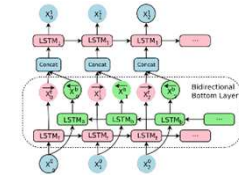
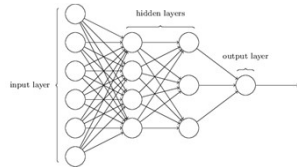
The screenshot displays the AWS Marketplace listing for the Xilinx ML Suite. At the top, the AWS Marketplace logo and navigation links are visible. The main content area features the Xilinx logo and the product name 'Xilinx ML Suite'. Below this, it states 'Sold by: Xilinx' and 'Latest Version: 18.04.02'. A brief description explains that the suite enables users to evaluate, develop, and deploy FPGA-accelerated inference using ready-to-run network models. A 'Show more' link is provided. The pricing section indicates a 'Typical Total Price' of '\$1.650/hr' for services hosted on F1.2xlarge in US East (N. Virginia). Navigation tabs for 'Overview', 'Pricing', 'Usage', 'Support', and 'Reviews' are present. The 'Product Overview' section describes how users can integrate machine learning into their applications and evaluate image classification through Caffe, MxNet, and TensorFlow. A 'Highlights' box lists supported frameworks (Caffe, MxNet, TensorFlow), pre-trained models (GoogLeNet v1, ResNet50, Flowers102, Places365, Yolov2), and DeepDetect support with RESTful and Python APIs. A link to 'www.xilinx.com/ml' is provided for more information.

<https://aws.amazon.com/marketplace/pp/B077FM2JNS>

Unified Simple User Experience from Cloud to XBB



Deep Learning Models



Multi-Layer Perceptron

- Classification
- Universal Function Approximator
- Autoencoder

Convolutional Neural Network

- Feature Extraction
- Object Detection
- Image Segmentation

Recurrent Neural Network

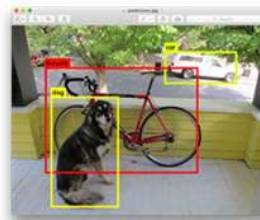
- Sequence and Temporal Data
- Speech to Text
- Language Translation

Classification



“Dog”

Object Detection



Segmentation



Rapid Feature and Performance Improvement

➤ xDNN-v1

- 500 MHz
- URAM for feature maps without caching
- Array of accumulator with
- 16 bit(batch 1), 8 bit(batch 2)
- Instructions: Convolution, Relu, MaxPool, AveragePool, Elementwise
- Flexible kernel size(square) and strides
- Programmable Scaling
- Q4CY17

➤ xDNN-v2

- 500 MHz
- All xDNN-v1 features
- DDR Caching: Larger Image, CNN Networks
- Instructions: Depth wise Convolution, Deconvolution, Convolution, Transpose Upsampling
- Rectangular Kernels
- Q2CY18

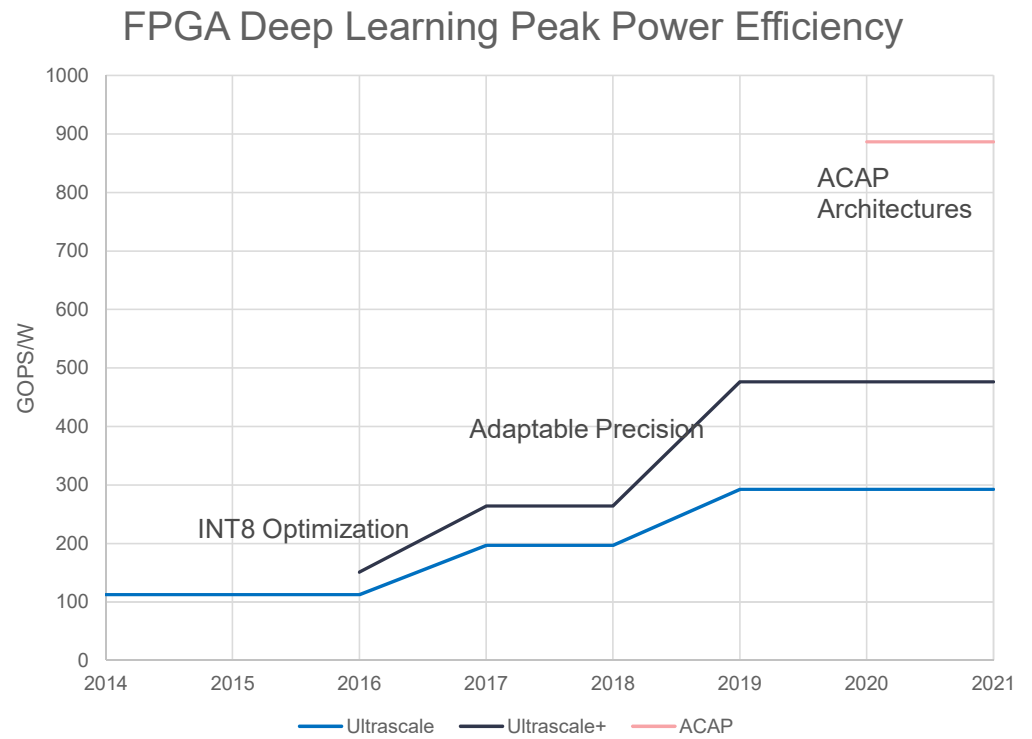
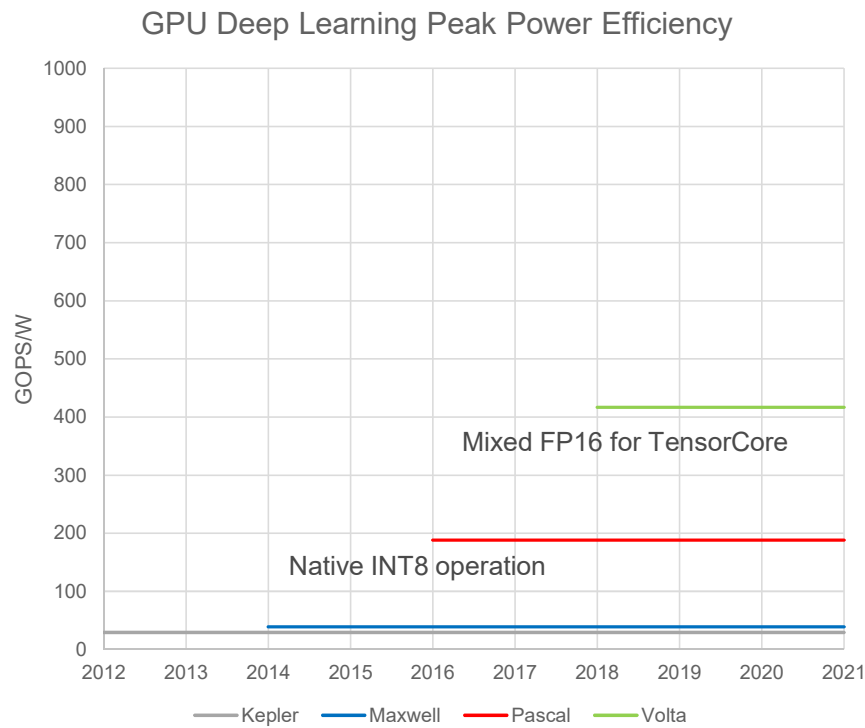
➤ xDNN-v3

- 700 MHz
- Feature compatible with xDNN-v2
- New Systolic Array Implementation: 50% Higher FMAX and 2.2x time lower latency
- Batch of 1 for 8 bit implementation
- Non-blocking Caching and Pooling
- Q4CY18

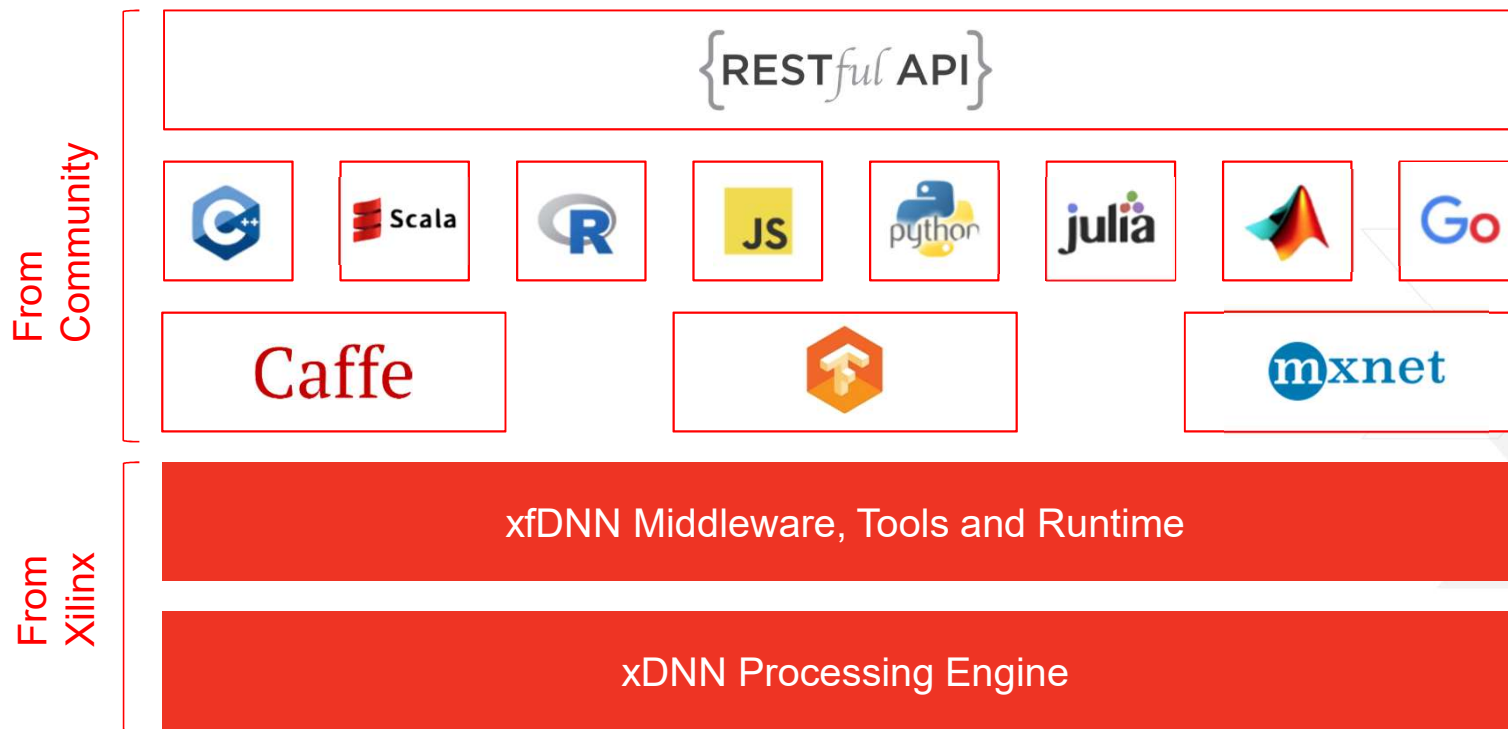
Break Through on Peak Performance

➤ GPU: Introduce new architectures and silicon

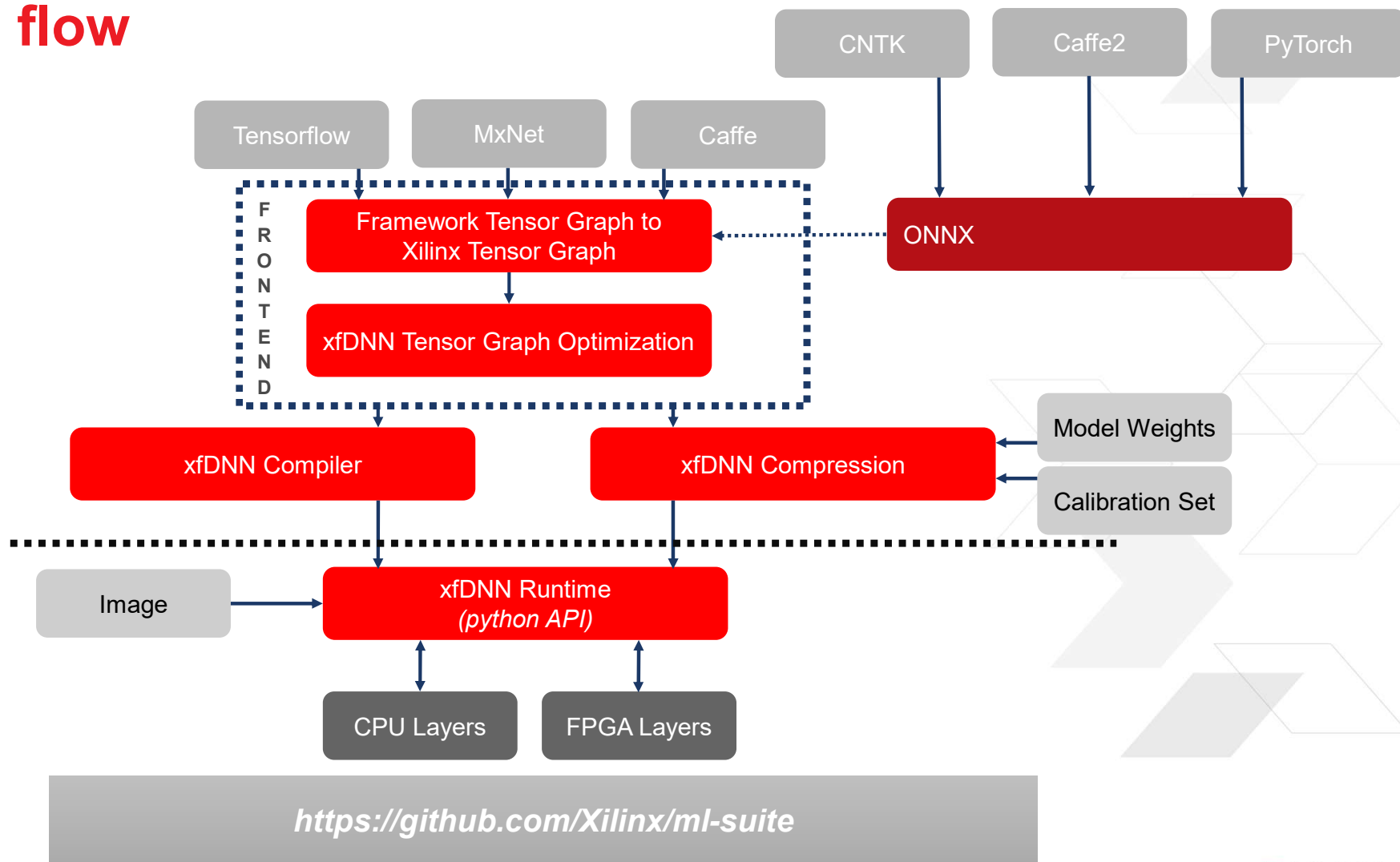
➤ Xilinx: Adapt the break through of emerging domain knowledge



Seamless Deployment with Open Source Software

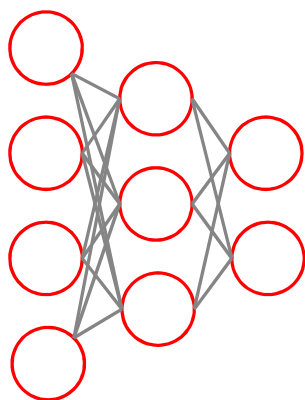


xfDNN flow



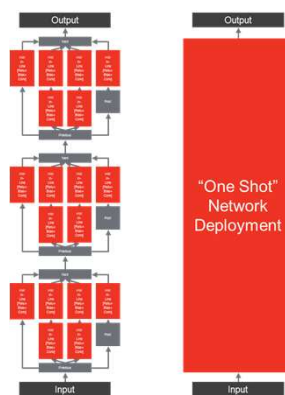
xfDNN Inference Toolbox

Graph Compiler



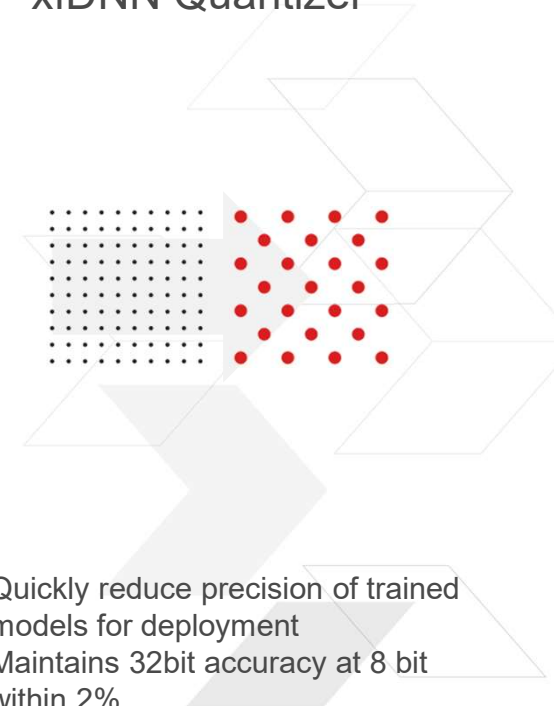
- Python tools to quickly compile networks from common Frameworks – Caffe, MxNet and Tensorflow

Network Optimization



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

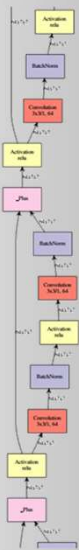
xfDNN Quantizer



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

xfDNN Graph Compiler

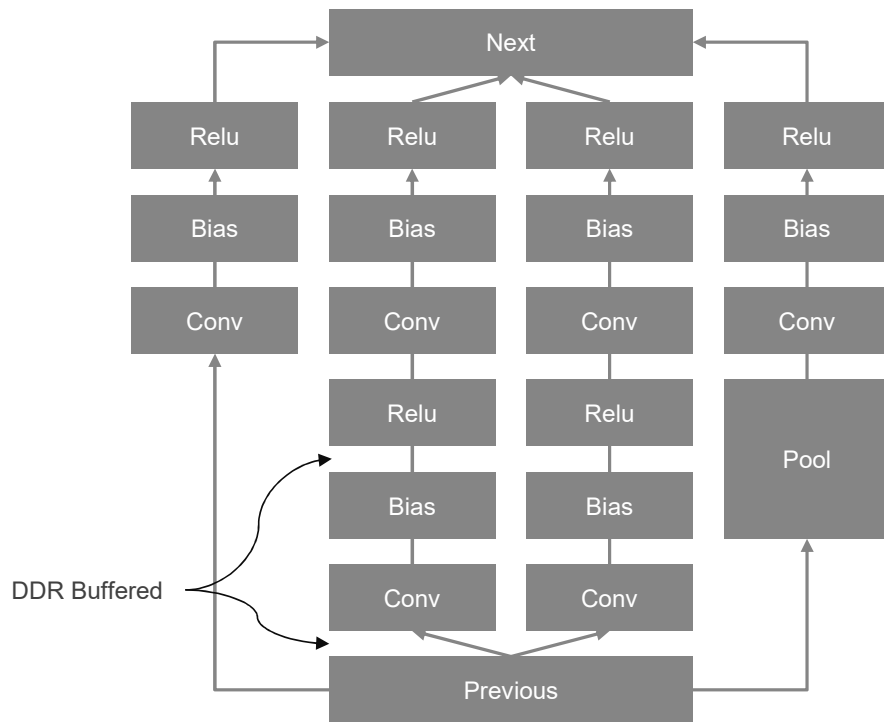
Pass in a Network



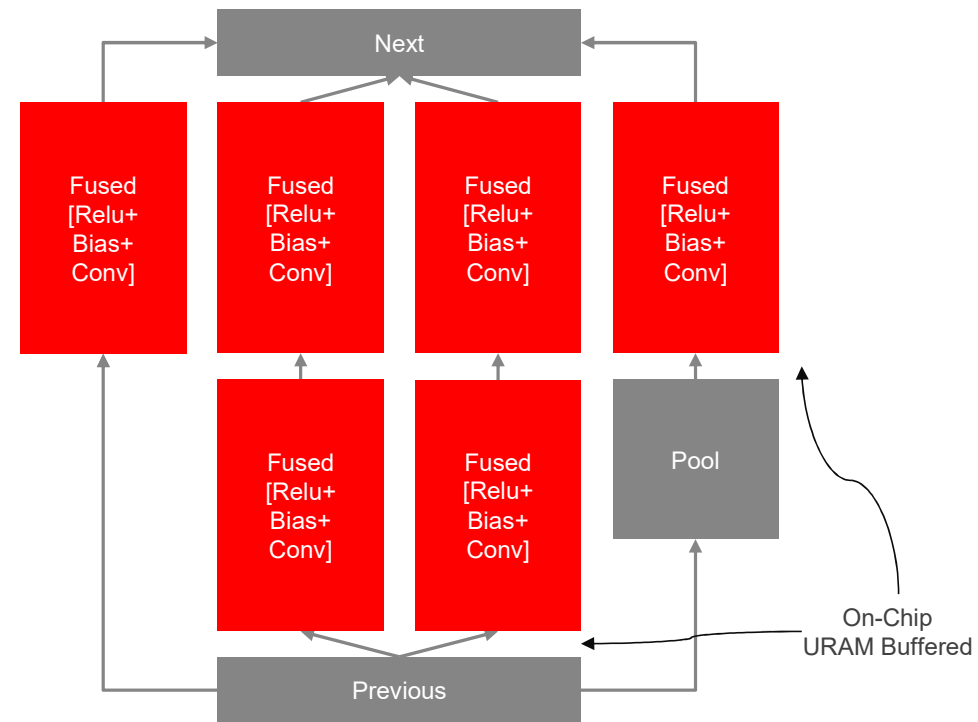
xfDNN
Graph Compiler

Microcode for xDNN is Produced

xfDNN Network Optimization Layer to Layer

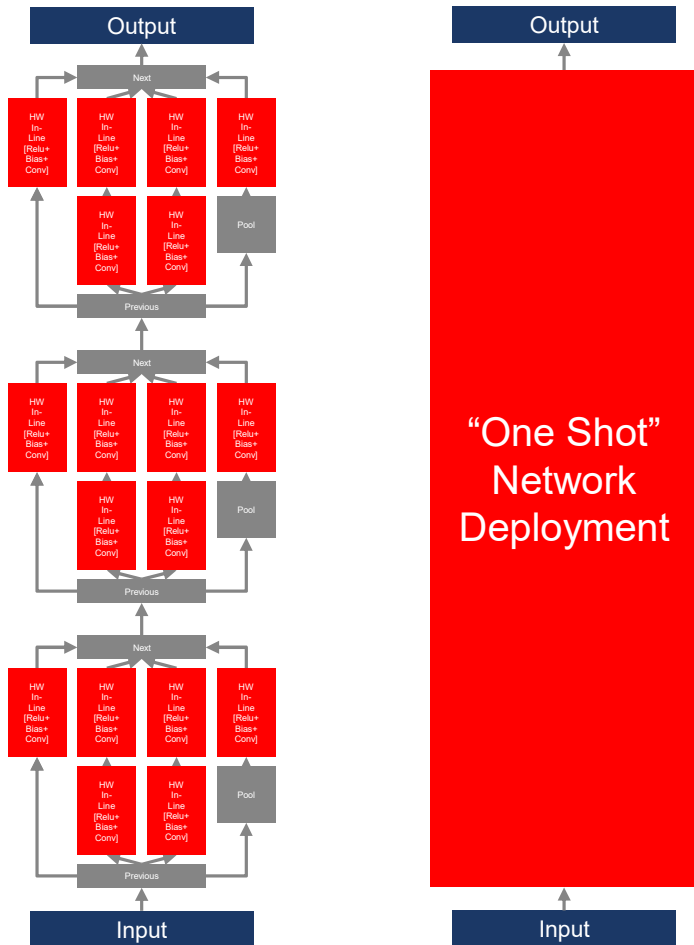


Unoptimized Model



xfDNN Intelligently Fused layers
Streaming optimized for URAM

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: FP to Fixed-Point Quantization

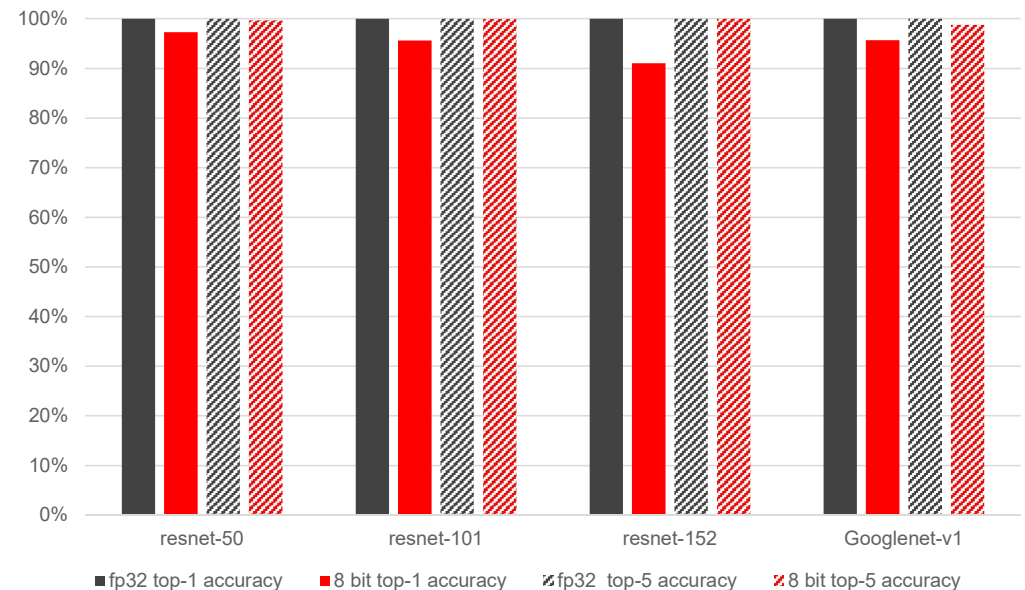
➤ Problem:

- Nearly all trained models are in 32-bit floating-point
- Available Caffe and TensorFlow quantization tools take hours and produce inefficient models
-

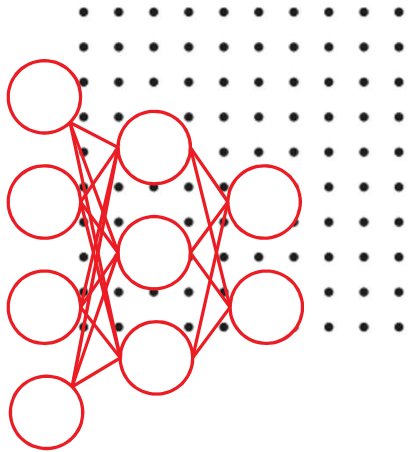
➤ Introducing: xfDNN Quantizer

- A customer friendly toolkit that automatically analyses floating-point ranges layer-by-layer and produces the fixed-point encoding that loses the least amount of information
 - Quantizes GoogleNet in under a minute
 - Quantizes 8-bit fixed-point networks within 1-3% accuracy of 32-bit floating-point networks
 - Extensible toolkit to maximize performance by searching for minimal viable bitwidths and prune sparse networks

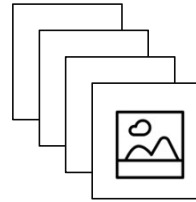
xfDNN Quantized Model Accuracy



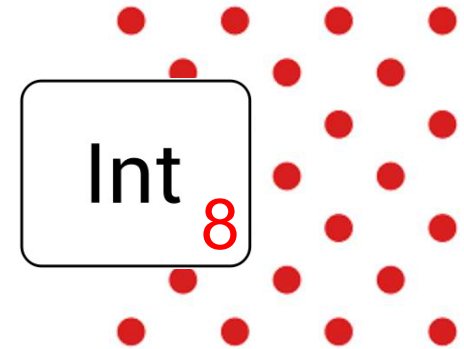
xfDNN Quantizer: Fast and Easy



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

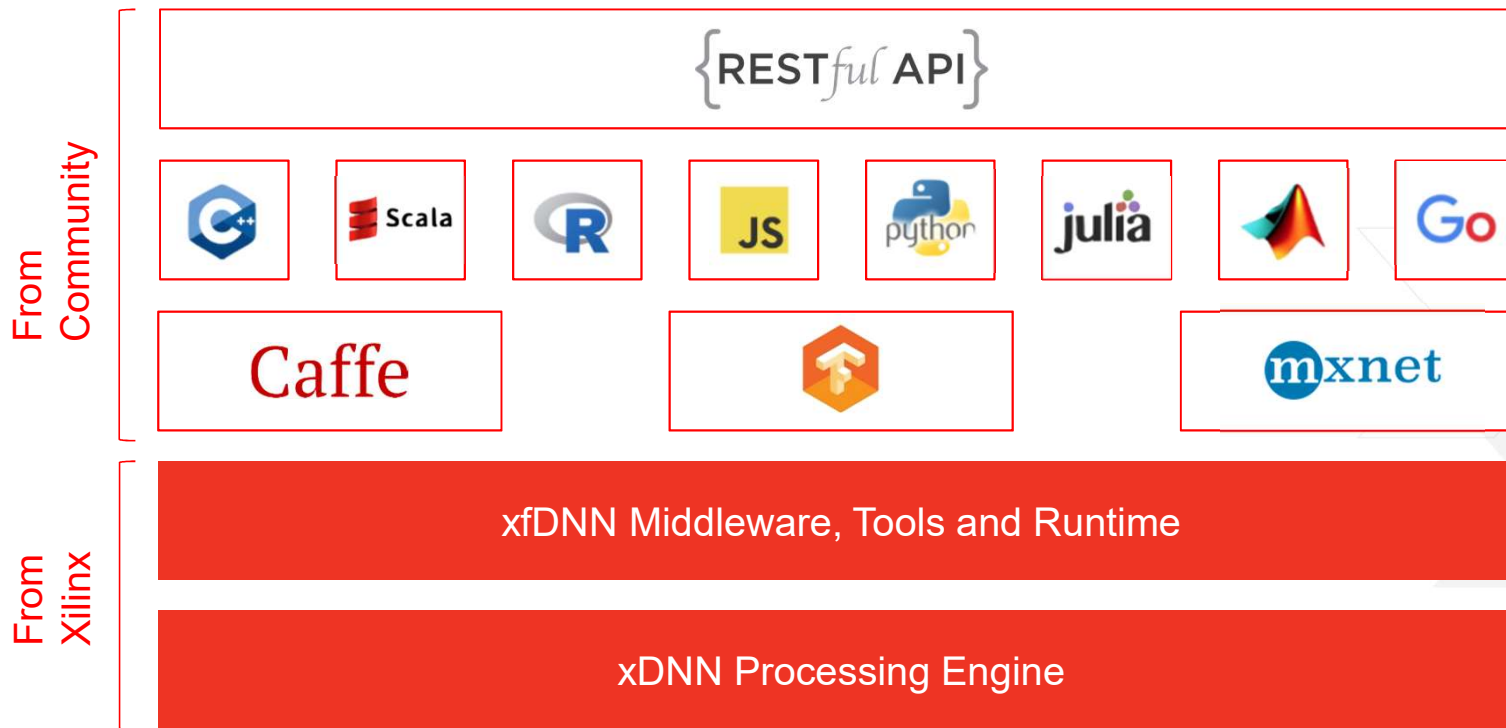


- 2) Provide a small sample set, no labels required
 - 16 to 512 images



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

Seamless Deployment with Open Source Software

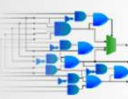


Xilinx ML Processing Engine – xDNN


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
		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	Data width	16-bit or 8-bit	

- Programmable Feature-set
- Tensor Level Instructions
- 500+MHz DSP Freq (VU9P)
- Custom Network Acceleration


Alveo – Breathe New Life into Your Data Center



**16nm
UltraScale™ Architecture**




Cloud Deployed



Off-Chip Memory Support

- Max Capacity: 64GB
- Max Bandwidth: 77GB/s




Cloud ↔ On-Premise Mobility




Internal SRAM

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




Ecosystem of Applications

- Many available today
- More on the way



PCIe Gen3x16



Server OEM Support

- Major OEMs in Qualification



Accelerate Any Application

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

ML Suite Overlays with xDNN Processing Engines

Adaptable

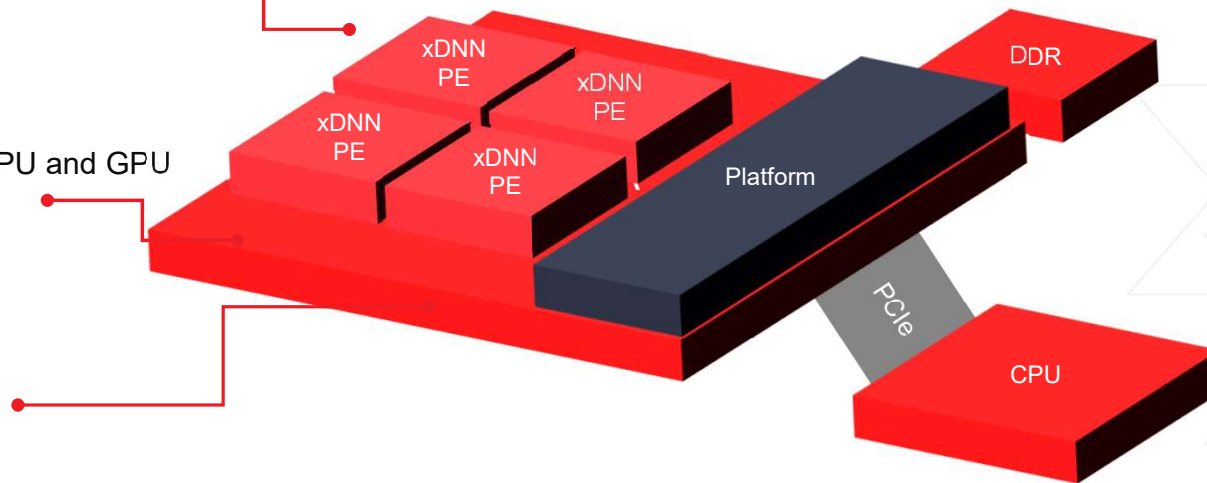
- > AI algorithms are changing rapidly
- > Adjacent acceleration opportunities

Realtime

- > 10x Low latency than CPU and GPU
- > Data flow processing

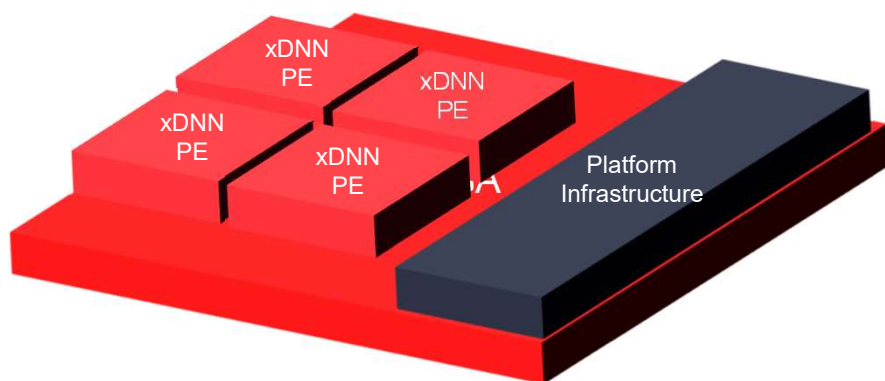
Efficient

- > Performance/watt
- > Low Power

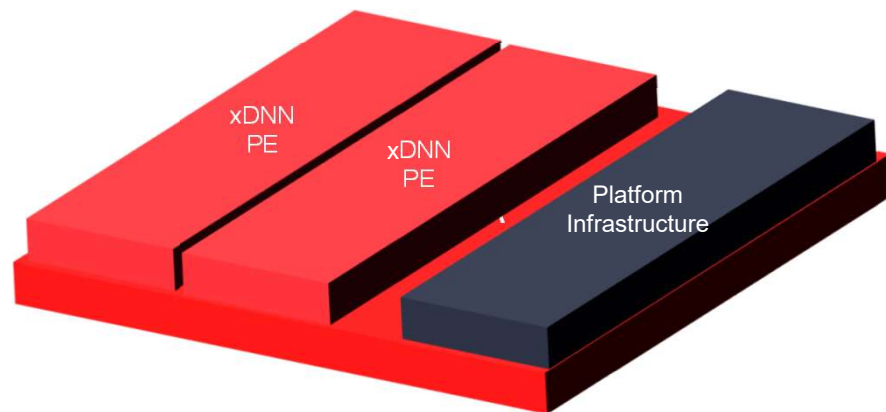


xDNN PEs Optimized for Your Cloud Applications

Throughput, Multi-Network Optimized

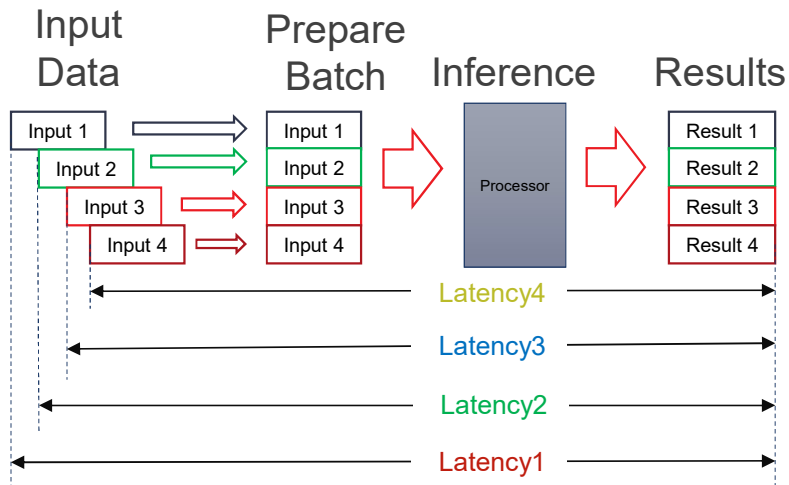


Latency, High Res Optimized



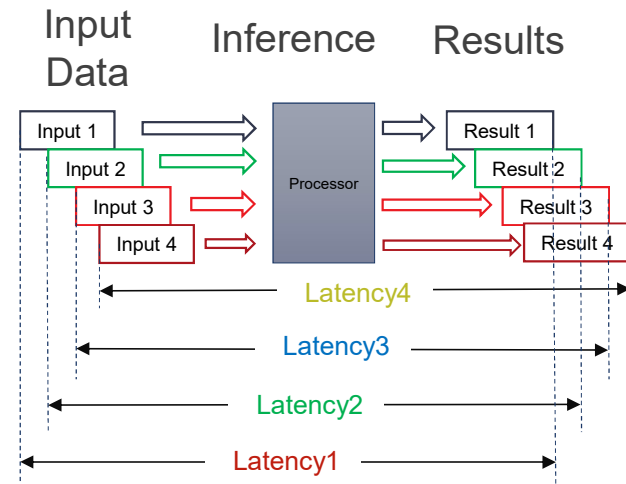
Overlay Name	DSP Array	#PEs	Cache	Precision	GOP/s	Optimized For	Examples Networks
Overlay_0	28x32	4	4 MB	Int16	896	Multi-Network, Maximum Throughput	ResNet50 (224x224)
Overlay_1	28x32	4	4 MB	Int8	1,792	Multi-Network, Maximum Throughput	ResNet50 (224x224)
Overlay_2	56x32	1	5 MB	Int16	1,702	Lowest Latency	Yolov2 (224x224)
Overlay_3	56x32	1	5 MB	Int8	3,405	Lowest Latency	Yolov2 (224x224)

Real-time Inference



➤ Inference with batches

- Require batch of input data to improve data reuse and instruction synchronization
- High throughput depends on high number of batch size
- High and unstable latency
- Low compute efficiency while batch is not fully filled or at lower batch size

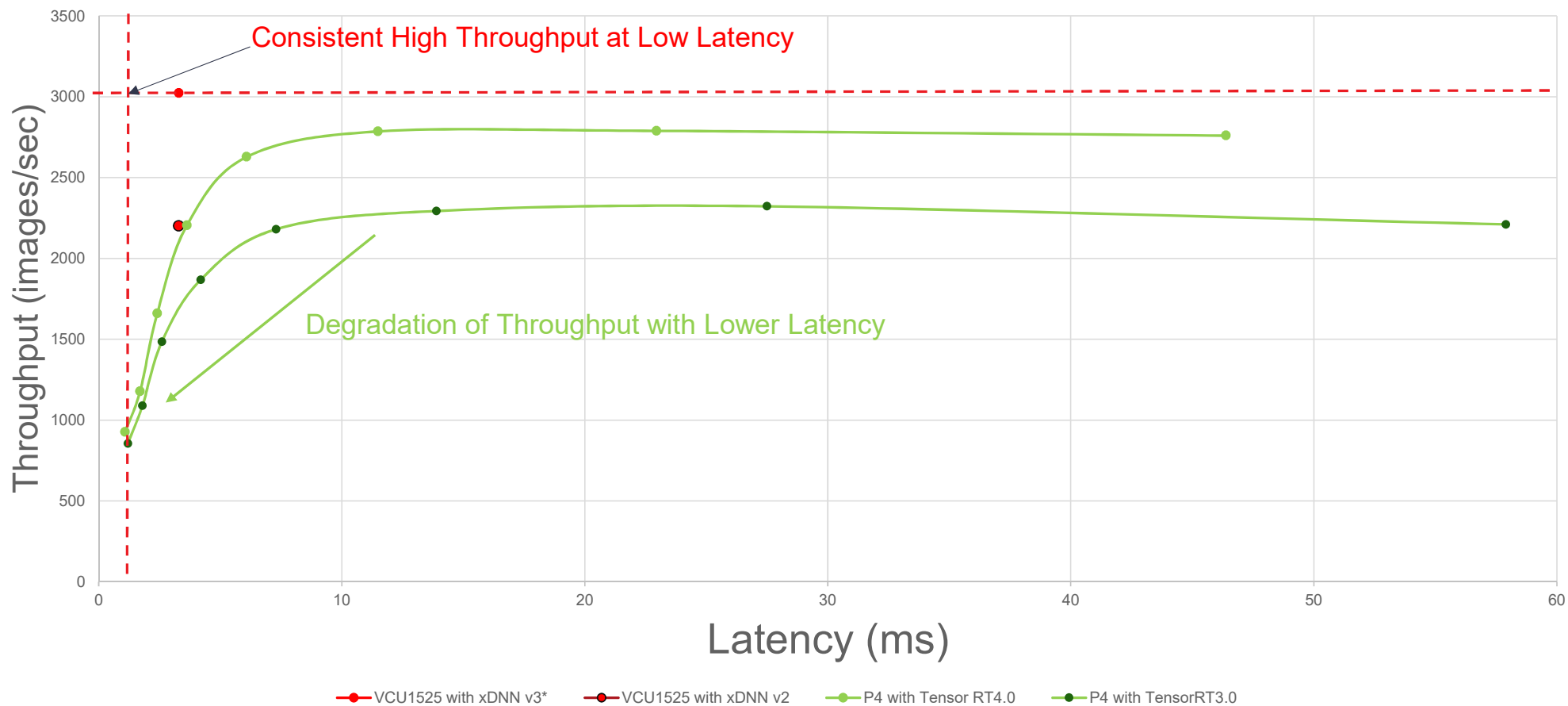


➤ Real Time Inference

- No requirement for batch input data
- Throughput less related to batch size
- Low and deterministic latency
- Consistent compute efficiency

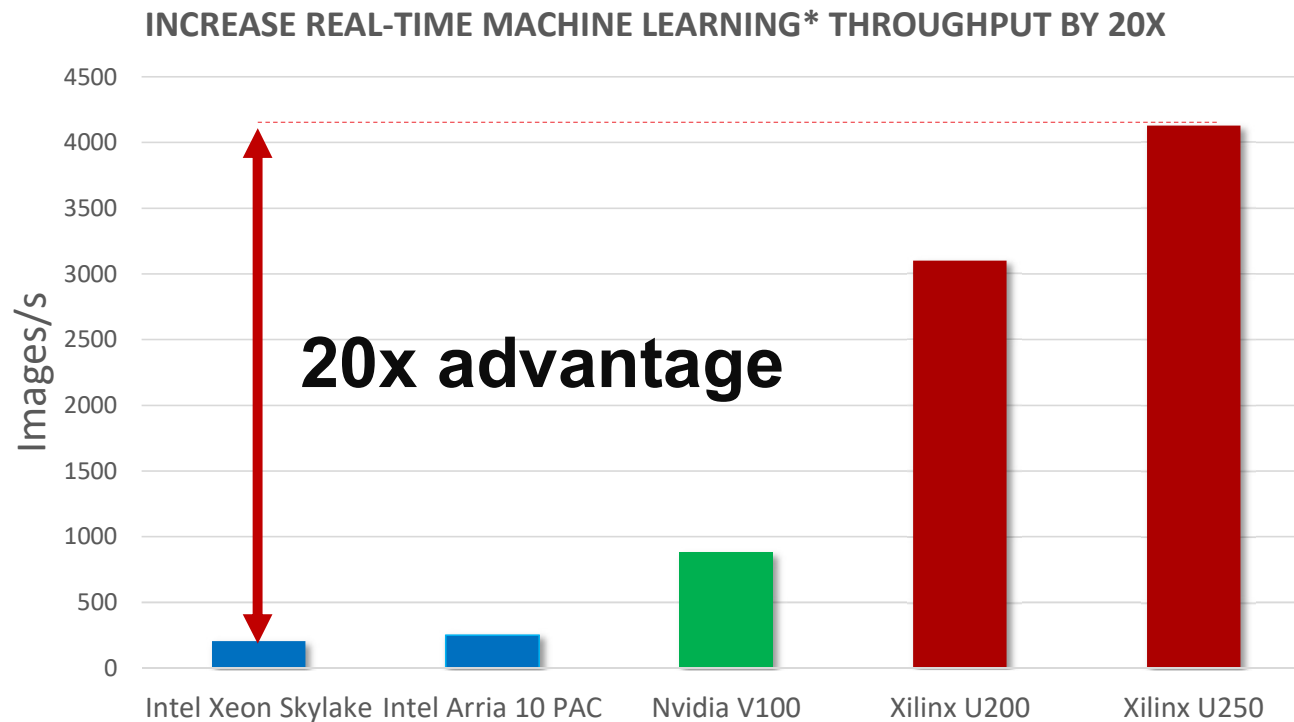
Xilinx - High Throughput at Real-Time

GoogLeNet V1 Performance



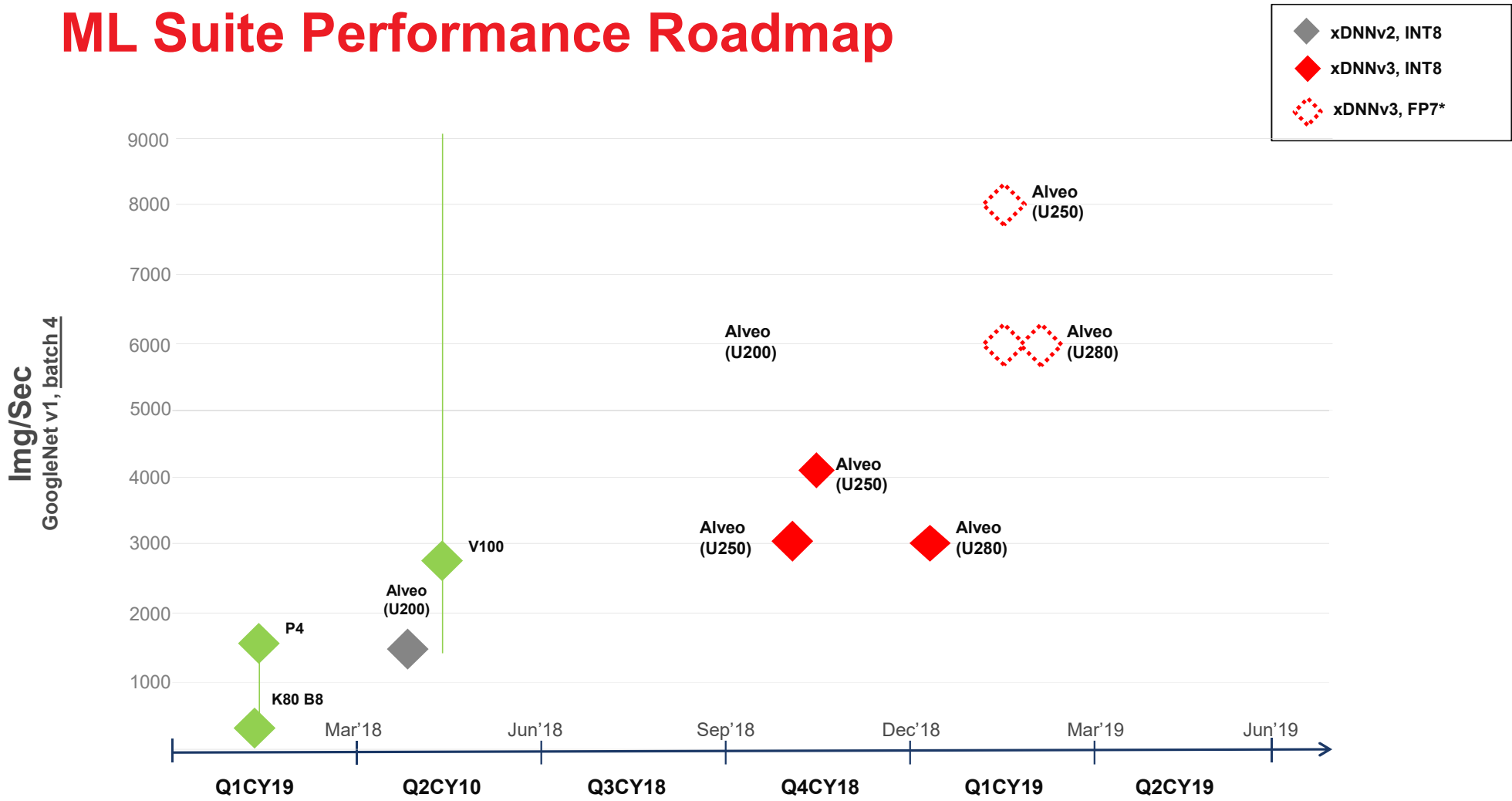
Fast

Advantages in Machine Learning Inference



* Source: [Accelerating DNNs with Xilinx Alveo Accelerator Cards White Paper](#)

ML Suite Performance Roadmap



Visit [Xilinx.com/ML](https://www.xilinx.com/ML) for more information

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



Adaptable.
Intelligent.

