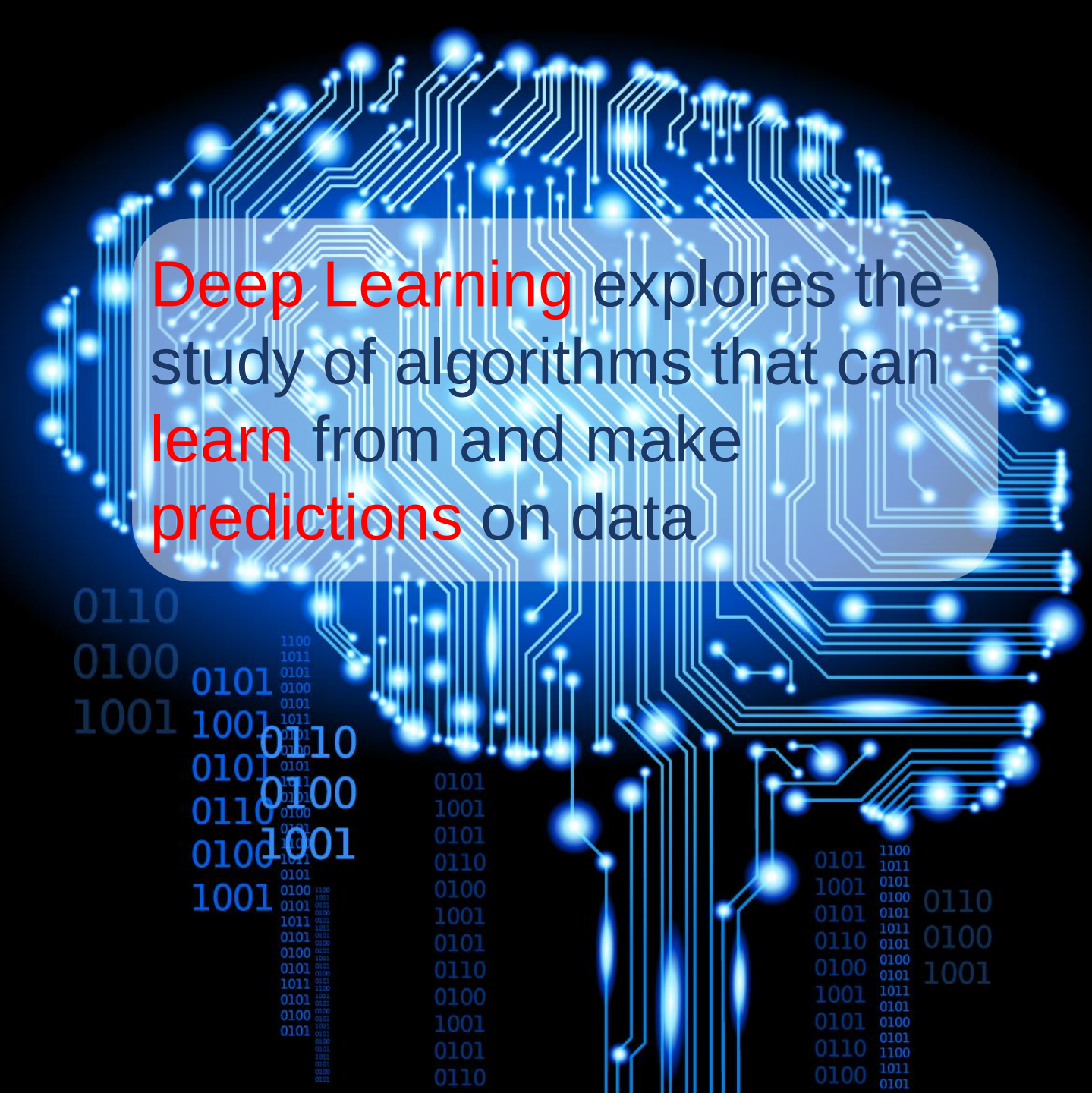
Xilinx ML Suite Overview

Jim Heaton
Sr. FAE

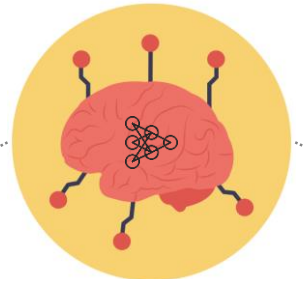
XILINX.



Deep Learning is Re-defining Many Applications

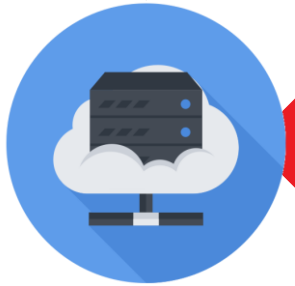
- Cloud Acceleration
- Security
- Ecommerce Social
- Financial
- Surveillance
- Industrial IOT
- Medical Bioinformatics
- Autonomous Vehicles

Accelerating AI Inference into Your Cloud Applications

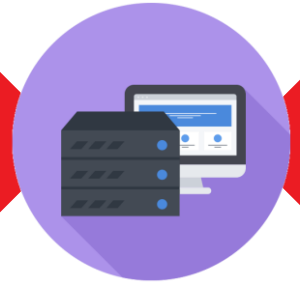


Deep Learning Applications

Featuring the Most Powerful
FPGA in the Cloud



Virtex® Ultrascale+™
VU9P



Zynq® Ultrascale+™
MPSoC

Cloud

On Premises

Edge

awsmarketplace

NIMBIX

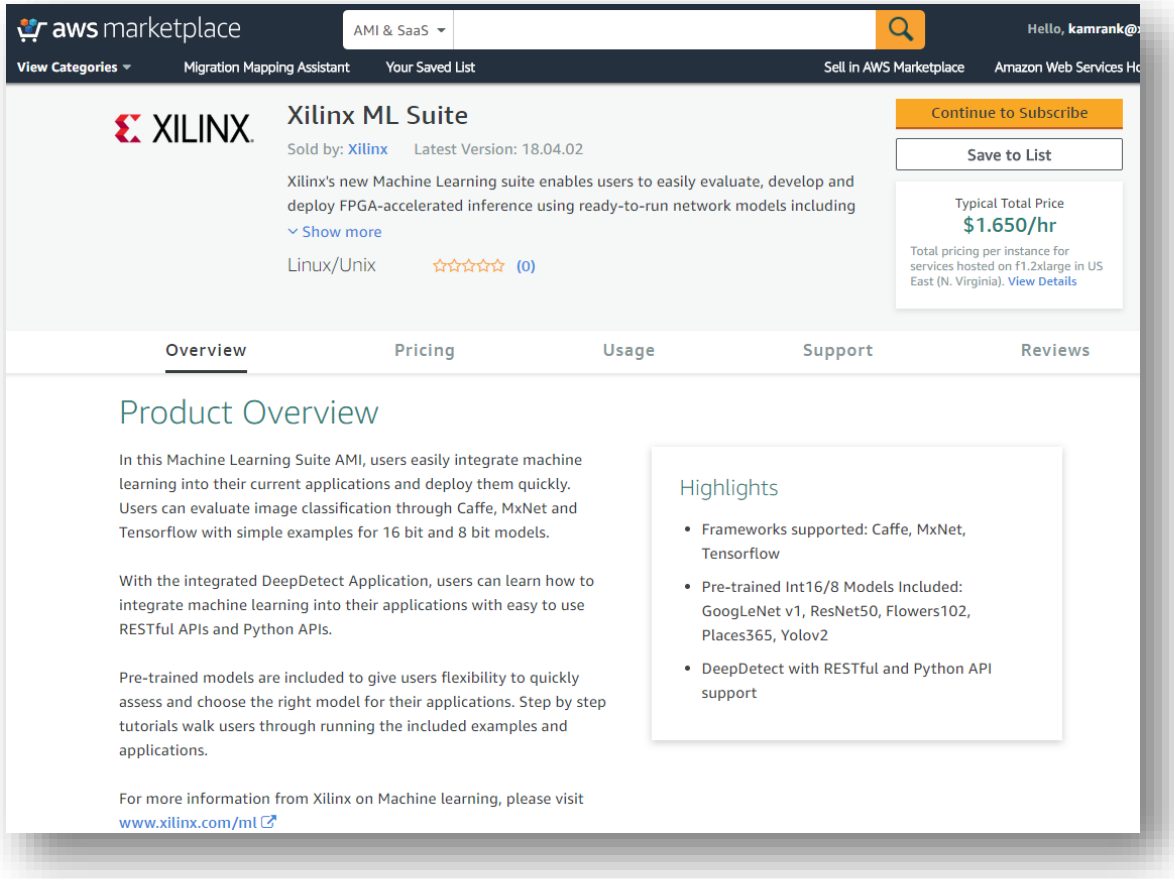
Alibaba Cloud
aliyun.com

HUAWEI

XILINX

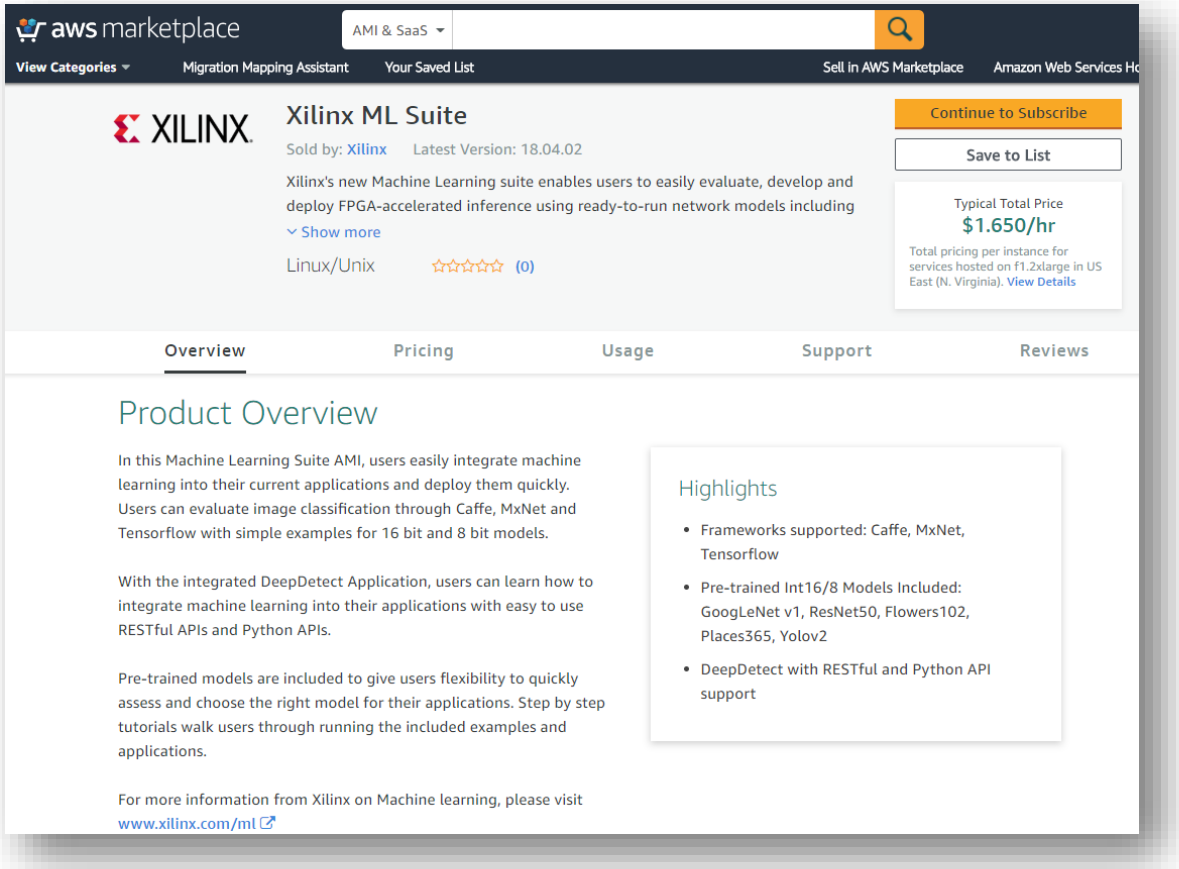
Want to try the out Xilinx ML Suite ?

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



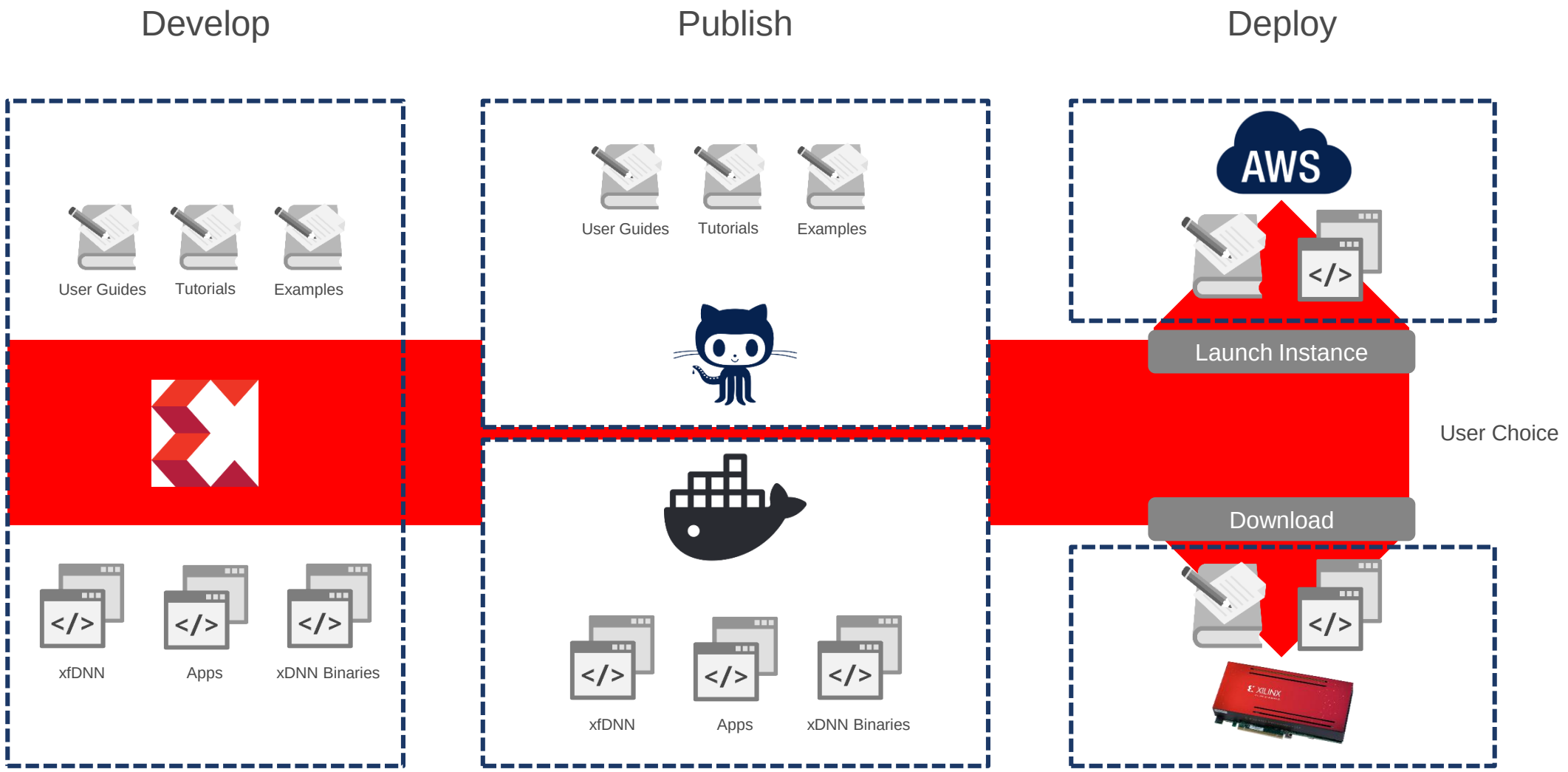
Xilinx ML Suite - AWS Marketplace

- **ML Suite**
  - Supported Frameworks:
    - Caffe
    - MxNet
    - Tensorflow
    - Keras
    - 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



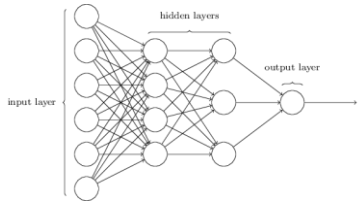
https://aws.amazon.com/marketplace/pp/B077FM2JNS?qid=1544477354556&sr=0-2&ref_=srh_res_product_title

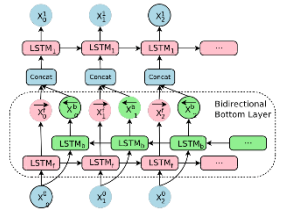
Unified Simple User Experience from Cloud to Alveo



XILINX

Deep Learning Models





Multi-Layer Perceptron

- Classification
- Universal Function Approximator
- Autoencoder

Convolutional Neural Network

- Feature Extraction
- Object Detection
- Image Segmentation

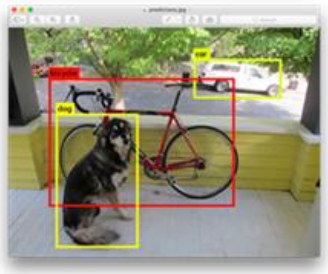
Recurrent Neural Network

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

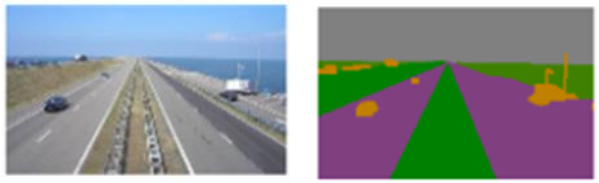
Classification

“Dog”

Object Detection



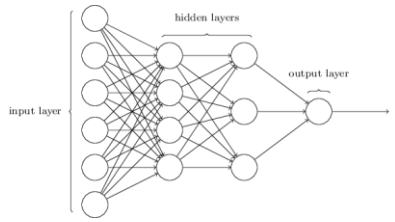
Segmentation



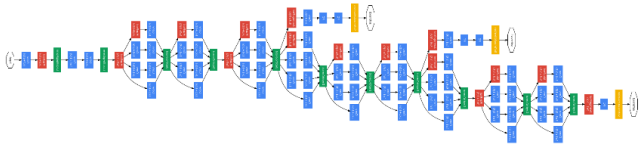
XILINX.

Overlay Architecture
Custom Processors Exploiting Xilinx FPGA Flexibility

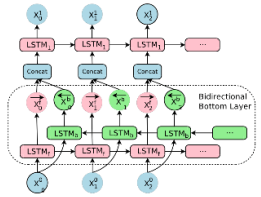
- Customized overlays with ISA architecture for optimized implementation
- Easy plug and play with Software Stack



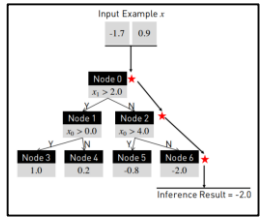
MLP Engine
Scalable sparse and dense implementation



xDNN – CNN Engine for Large 16 nm Xilinx Devices
Deephi DPU – Flexible CNN Engine with Embedded Focus



Deephi ESE
LSTM Speech to Text engine
Available on AWS



Random Forest
Configurable RF classification

XILINX

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

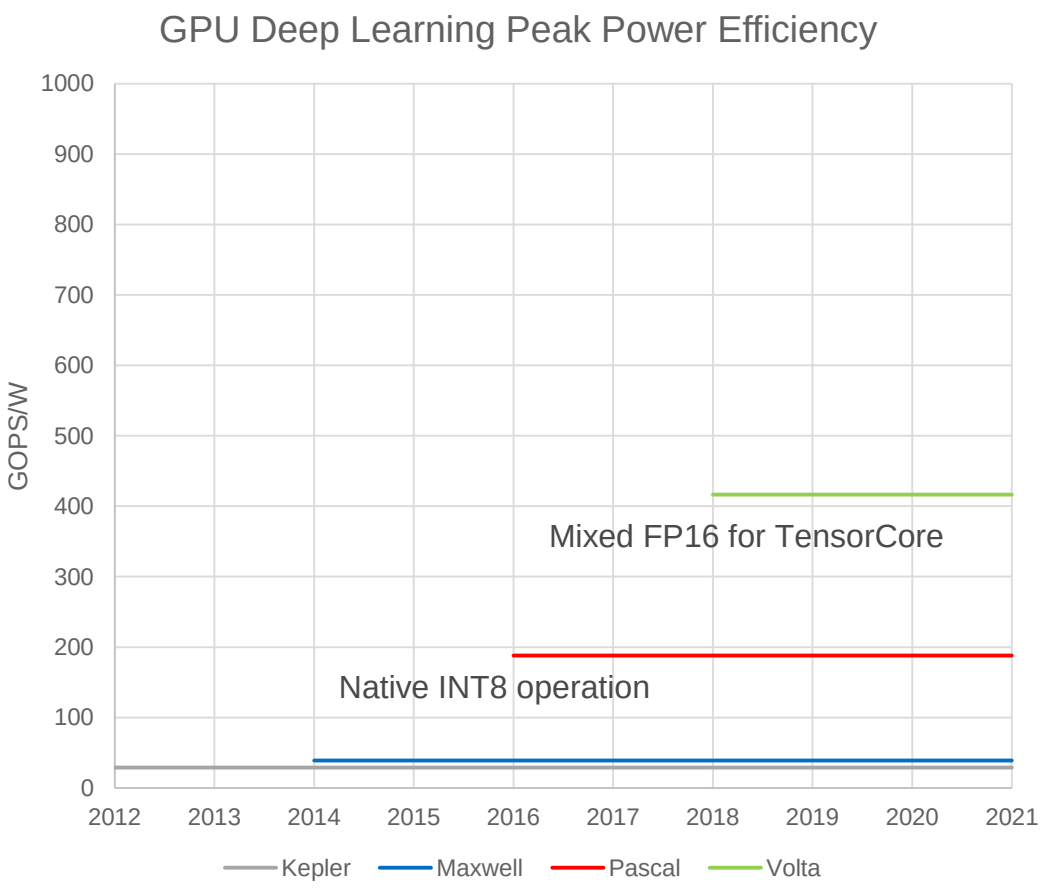
- 500 MHz
- All xDNN-v1 features
- DDR Caching: Larger Image, CNN Networks
- Instructions: Depth wise Convolution, Deconvolution, Convolution, Transpose Upsampling
- Rectangular Kernels
- Q2CY18

xDNN-v3

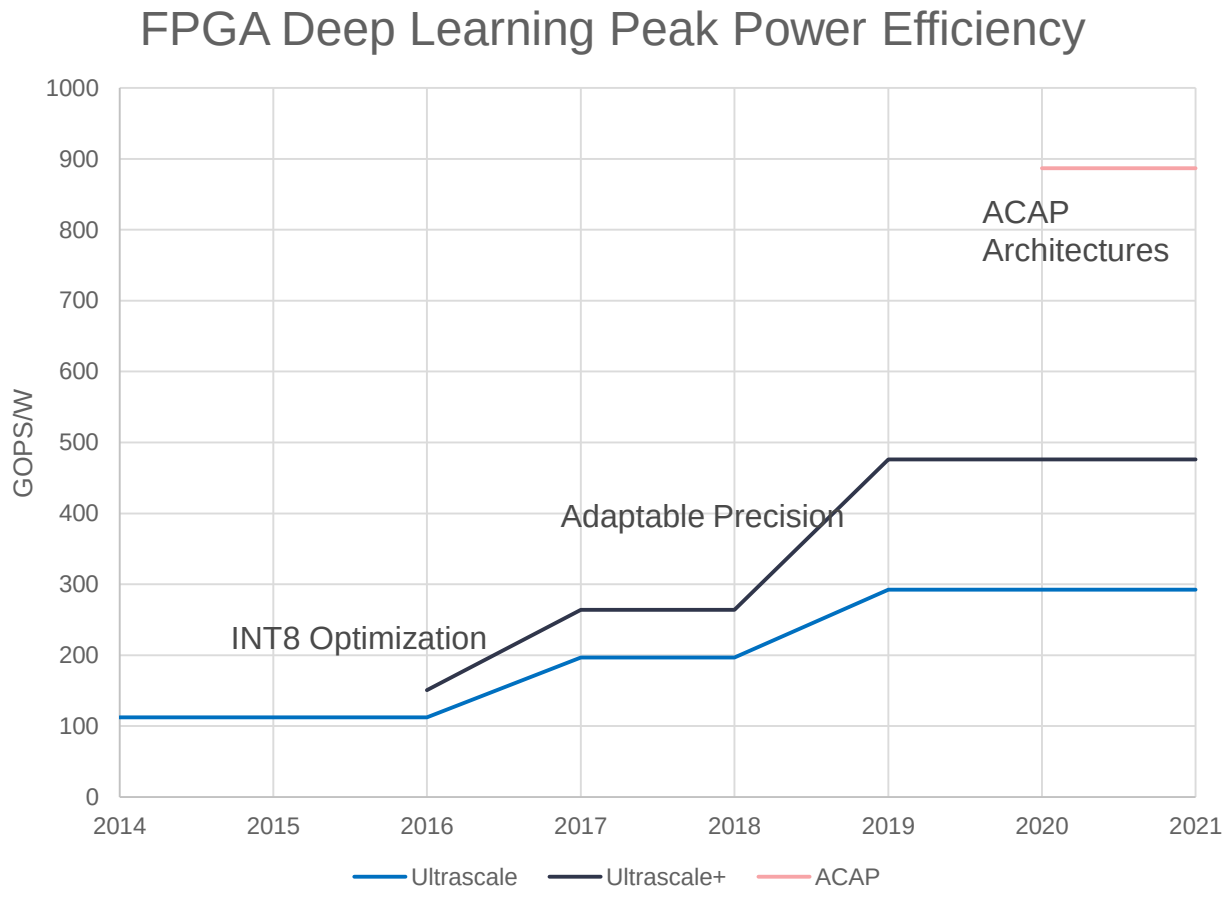
- 700 MHz
- Feature compatible with xDNN-v2
- New Systolic Array Implementation: 50% Higher FMAX and 2.2x time lower latency
- Batch of 1 for 8 bit implementation
- Non-blocking Caching and Pooling
- Q4CY18

Break Through on Peak Performance

- GPU: Introduce new architectures and silicon

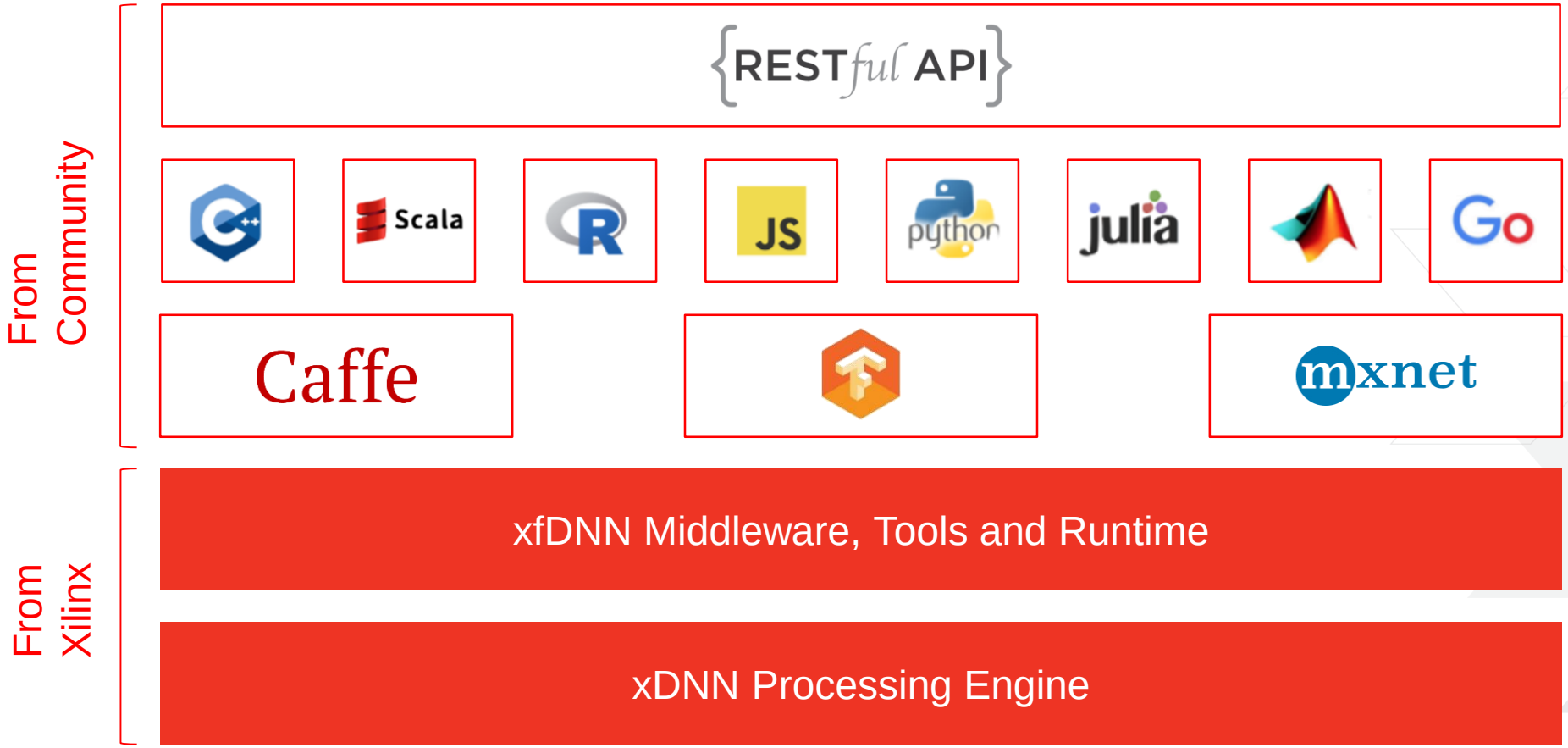


- Xilinx: Adapt the break through of emerging domain knowledge

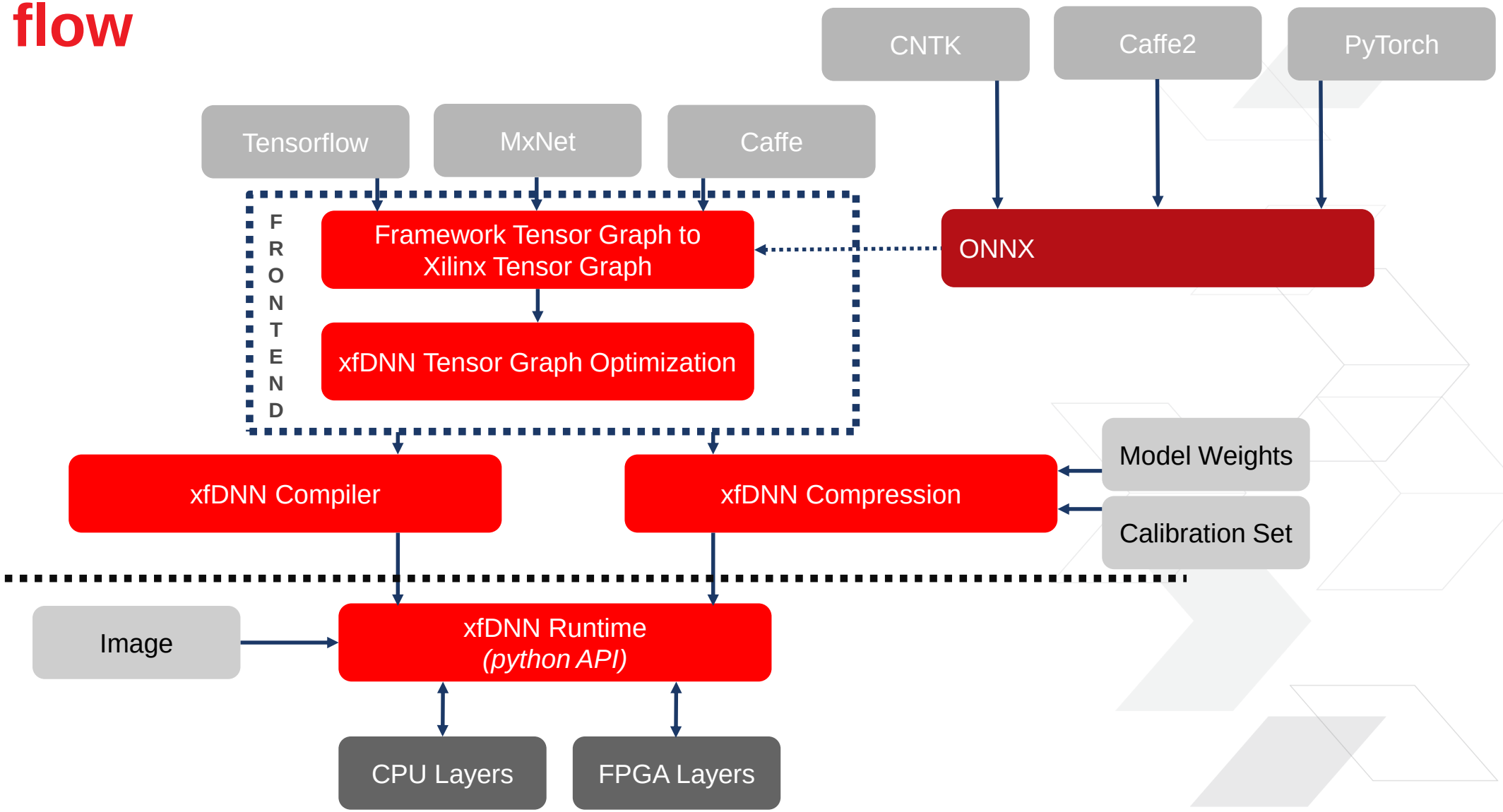


XILINX.

Seamless Deployment with Open Source Software



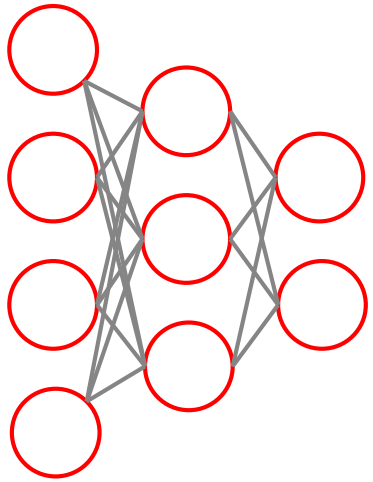
xfDNN flow



XILINX

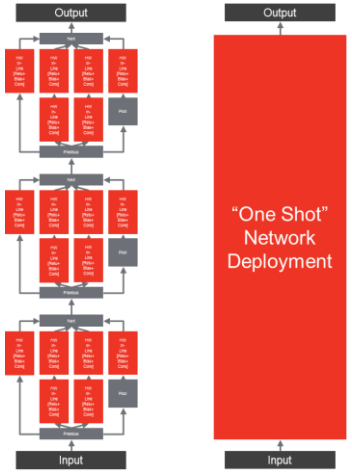
xfDNN Inference Toolbox

Graph Compiler



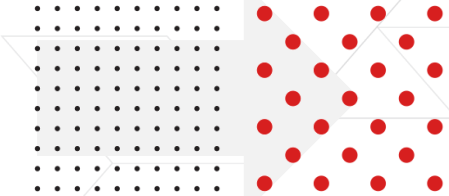
- Python tools to quickly compile networks from common Frameworks – Caffe, MxNet and Tensorflow

Network Optimization



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

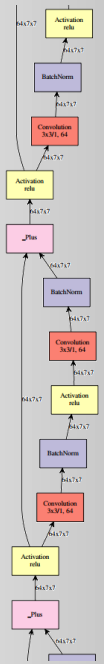
xfDNN Quantizer



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

xfDNN Graph Compiler

Pass in a Network

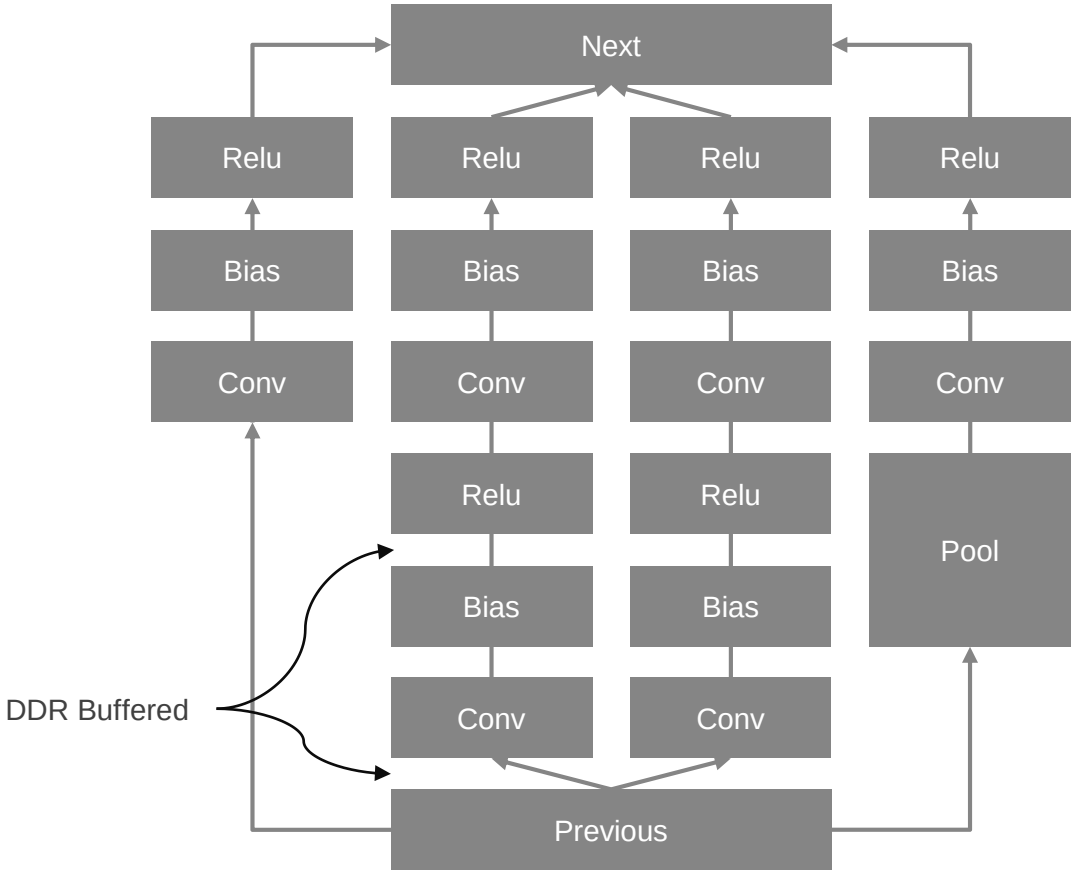


xfDNN
Graph Compiler

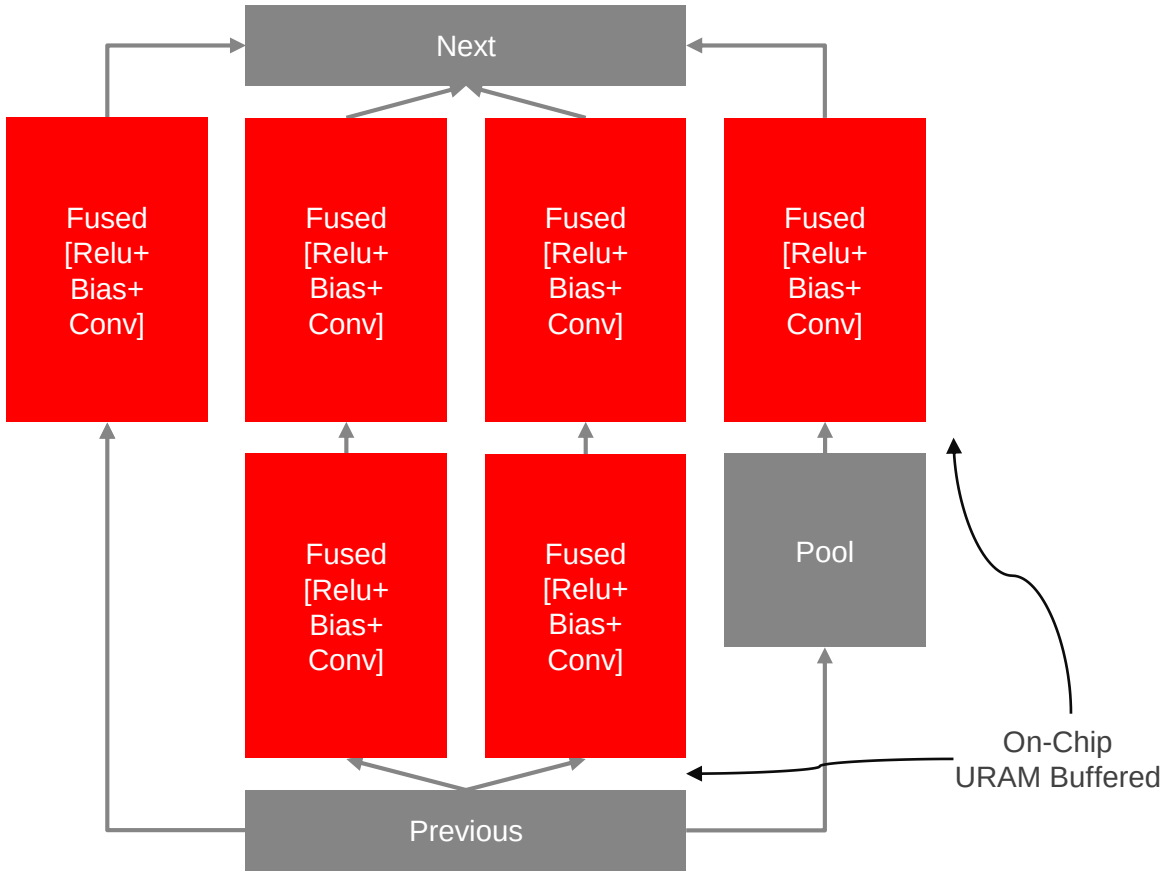
Microcode for xDNN is Produced

XILINX

xfDNN Network Optimization
Layer to Layer



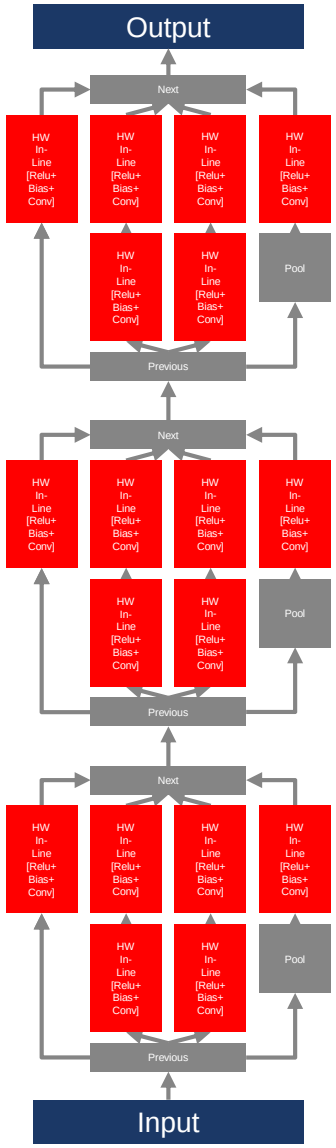
Unoptimized Model



xfDNN Intelligently Fused layers
Streaming optimized for URAM

XILINX.

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

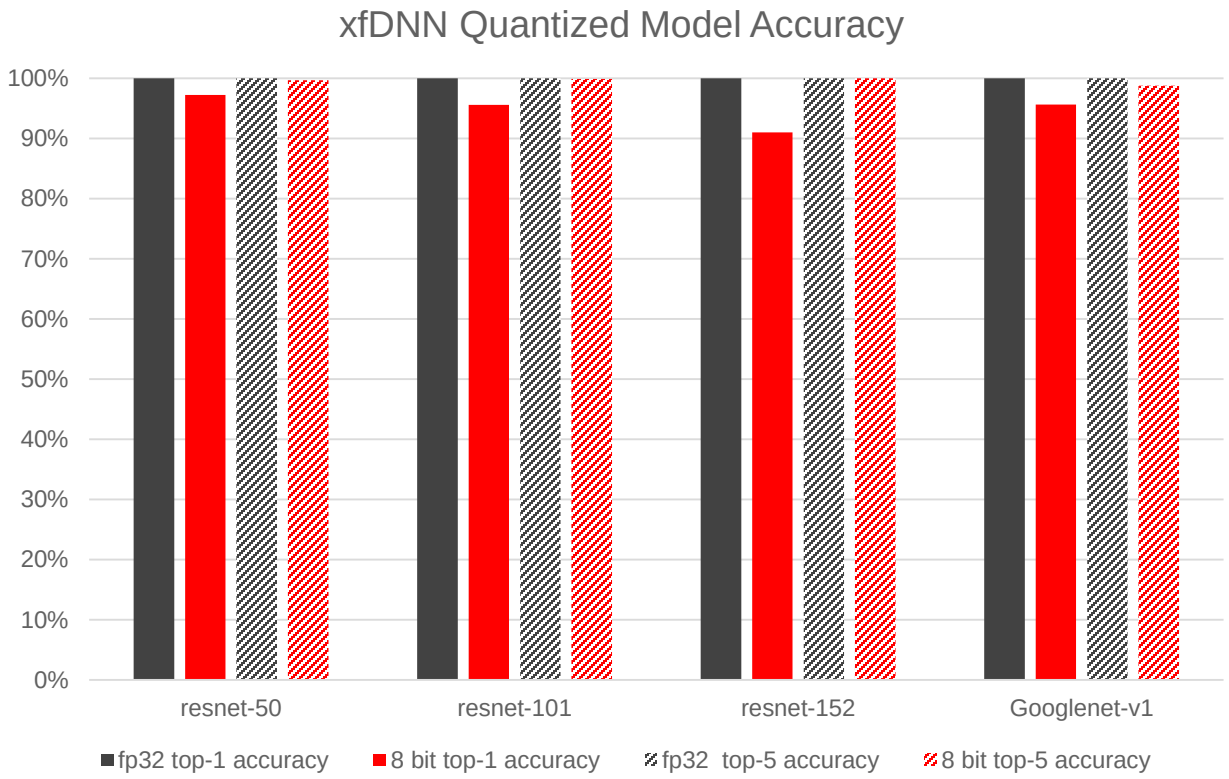
- xfDNN analyzes network memory needs and optimizes scheduler
  - For Fused and "One Shot" Deployment

"One Shot" deploys entire network to FPGA

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

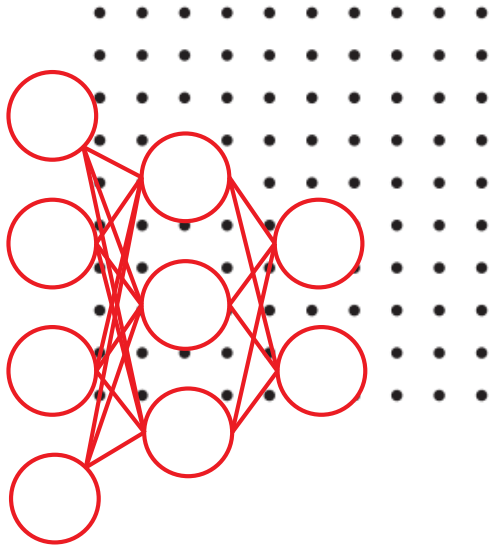
xfDNN Quantizer: FP to Fixed-Point Quantization

- 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 looses 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

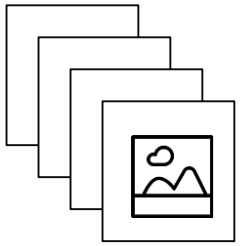


XILINX

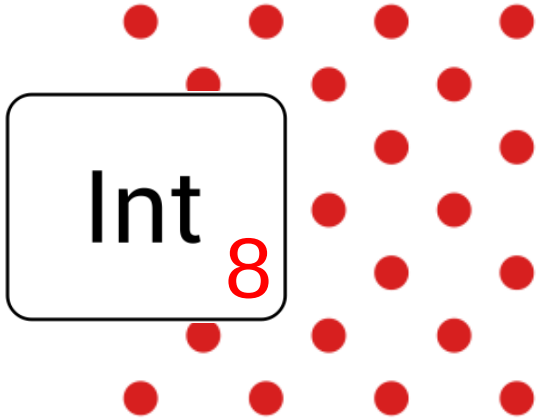
xfDNN Quantizer: Fast and Easy



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



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



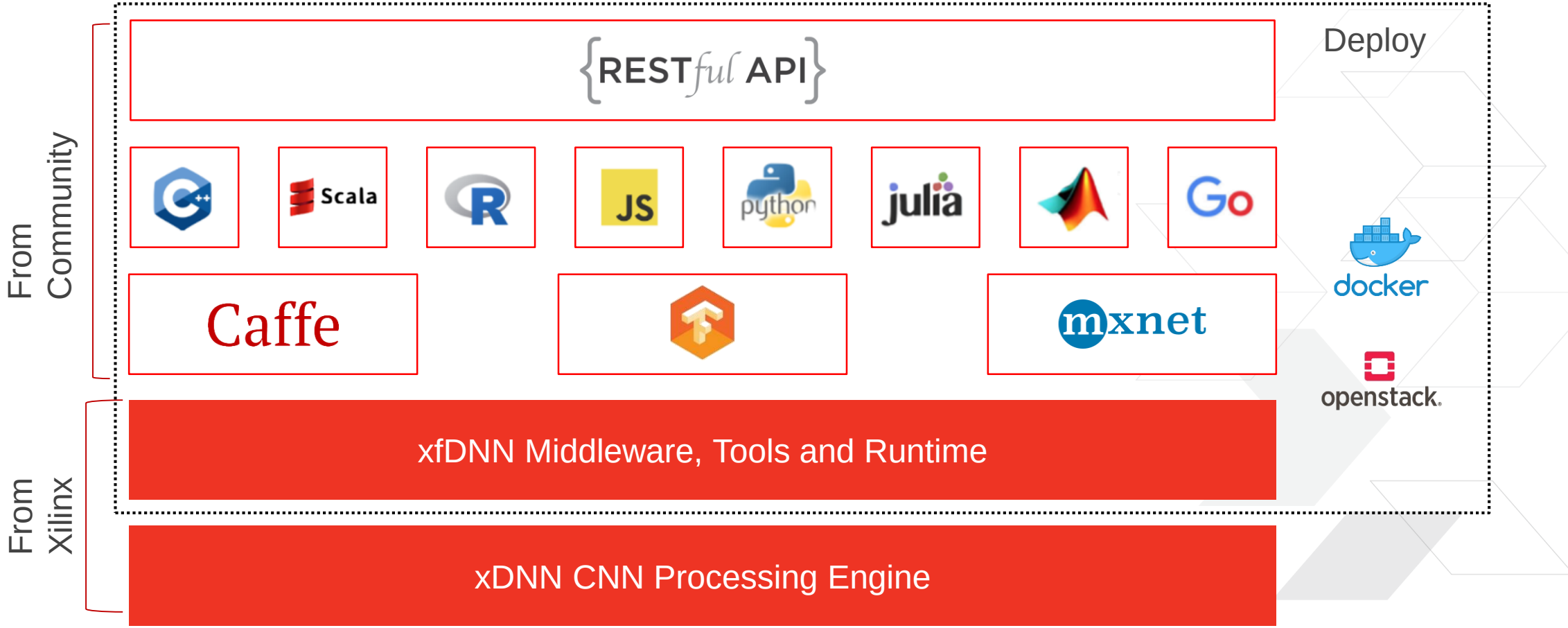
3) Specify desired precision
   - Quantizes to <8 bits to match Xilinx's DSP

Xilinx 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
- 700+MHz DSP Freq (VU9P)
- Custom Network Acceleration

Seamless Deployment with Open Source Software



*TensorFlow Q4 2017

XILINX

ML Suite Overlays with xDNN Processing Engines

Adaptable

> AI algorithms are changing rapidly

> Adjacent acceleration opportunities

Realtime

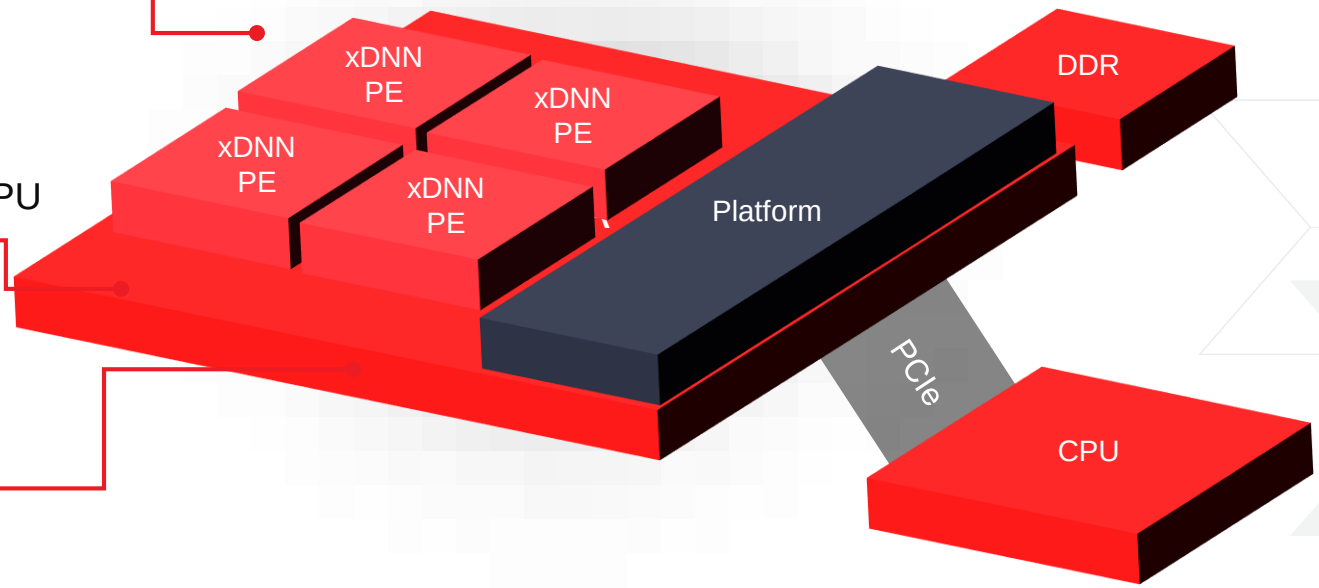
> 10x Low latency than CPU and GPU

> Data flow processing

Efficient

> Performance/watt

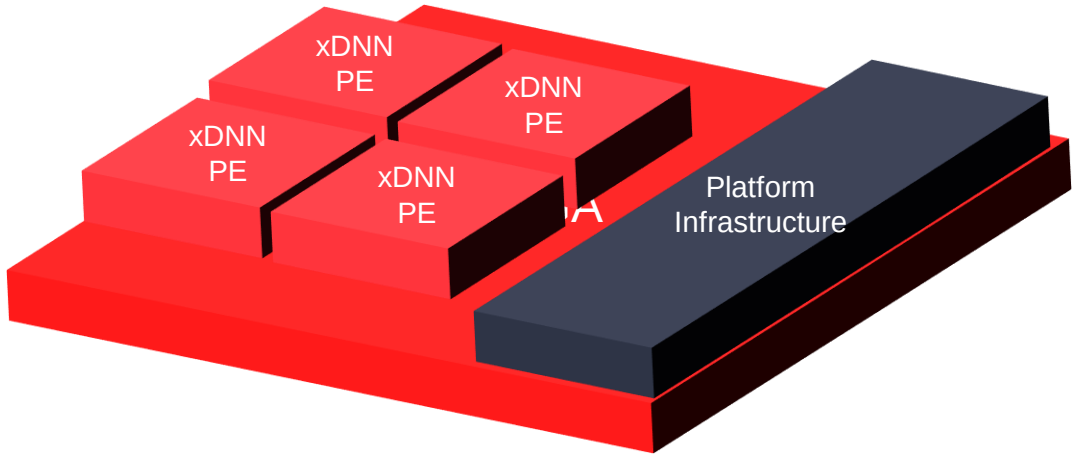
> Low Power



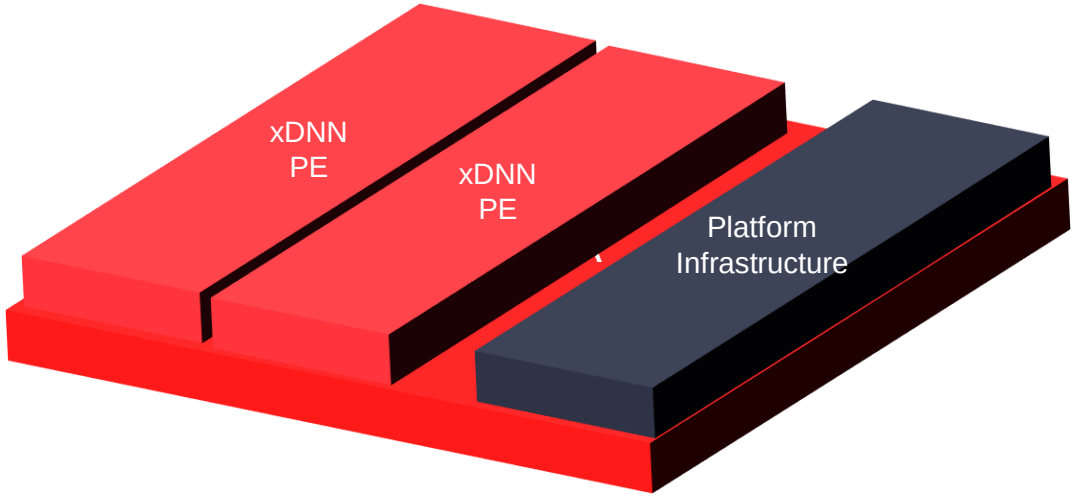
XILINX.

xDNN PEs Optimized for Your Cloud Applications

Throughput, Multi-Network Optimized

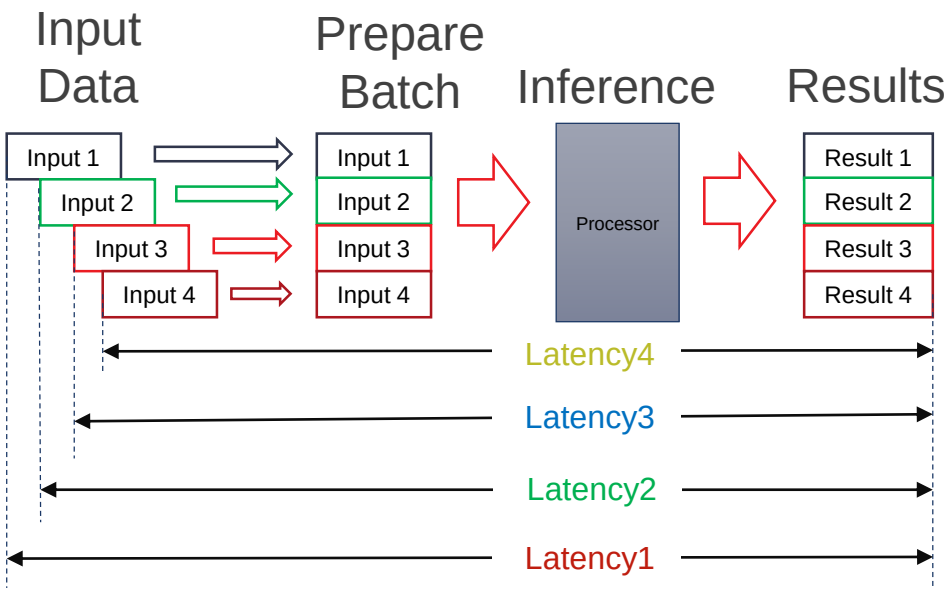


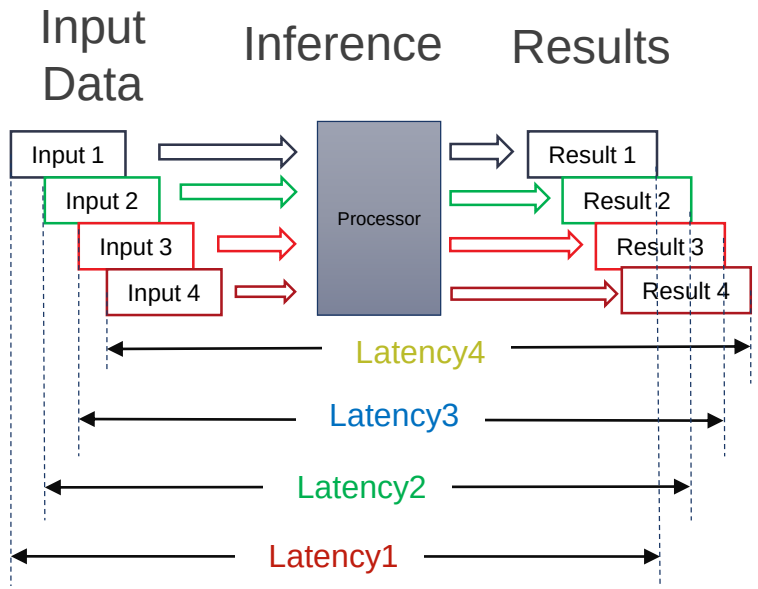
Latency, High Res Optimized



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





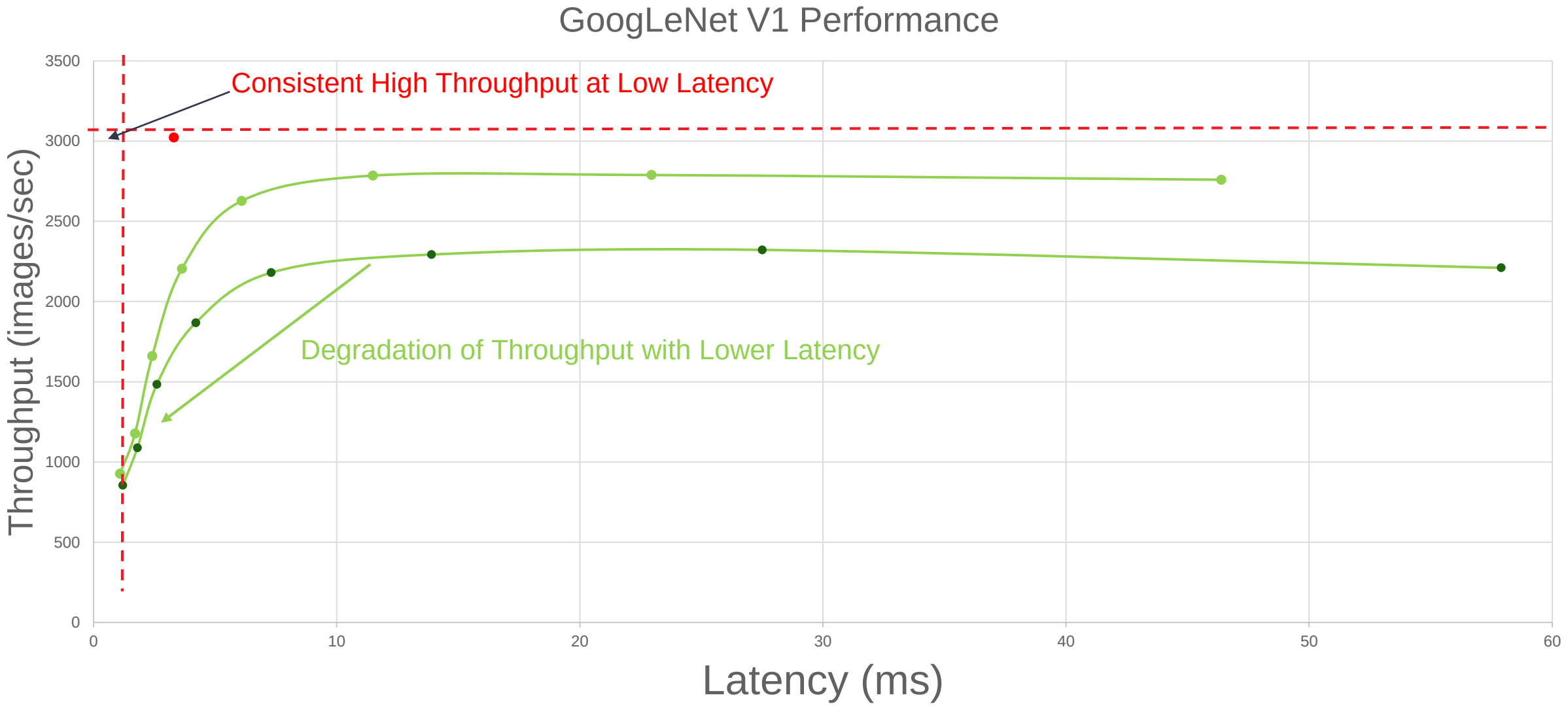
Inference with batches

- 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

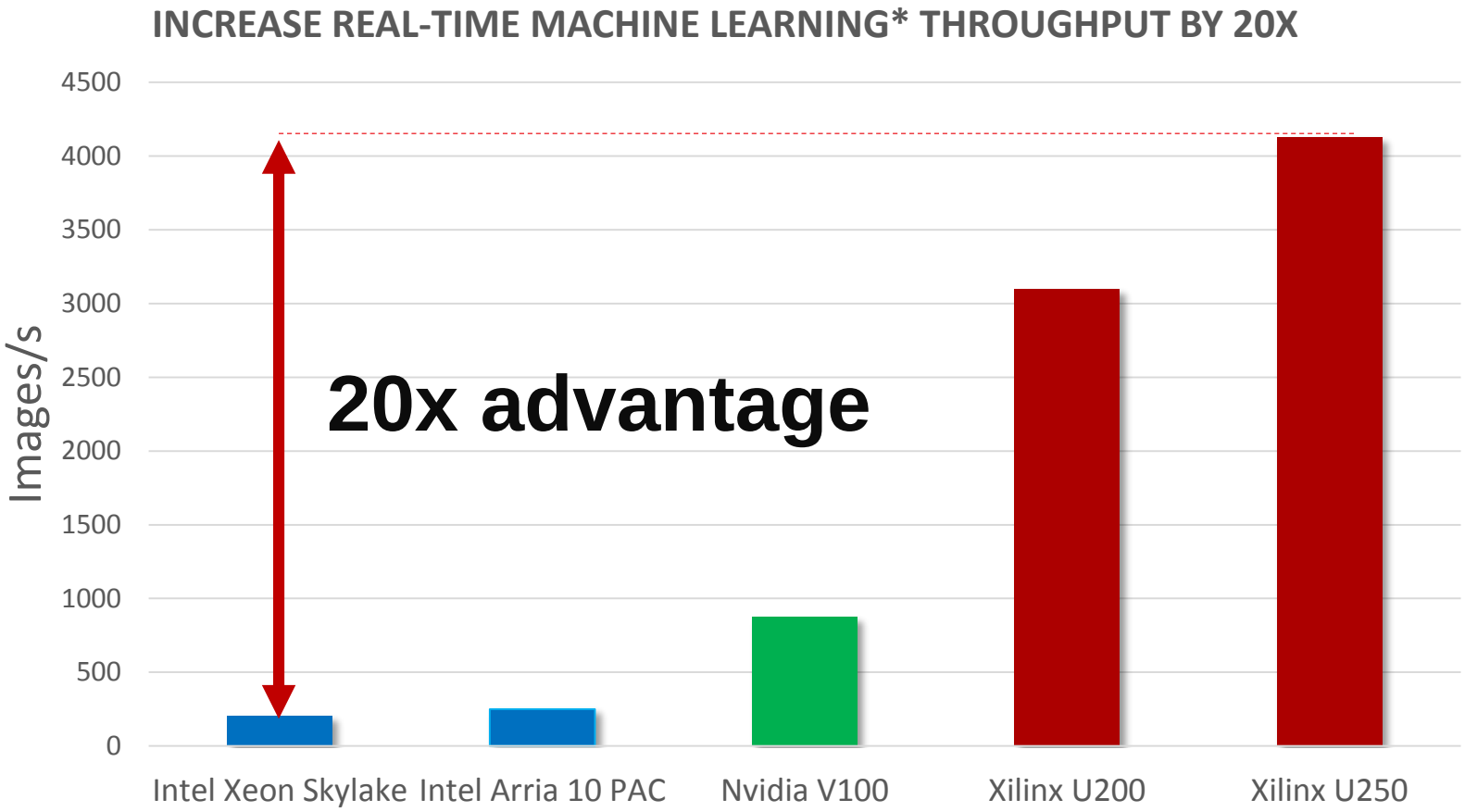
- No requirement for batch input data
- Throughput less related to batch size
- Low and deterministic latency
- Consistent compute efficiency

Xilinx - High Throughput at Real-Time



XILINX.

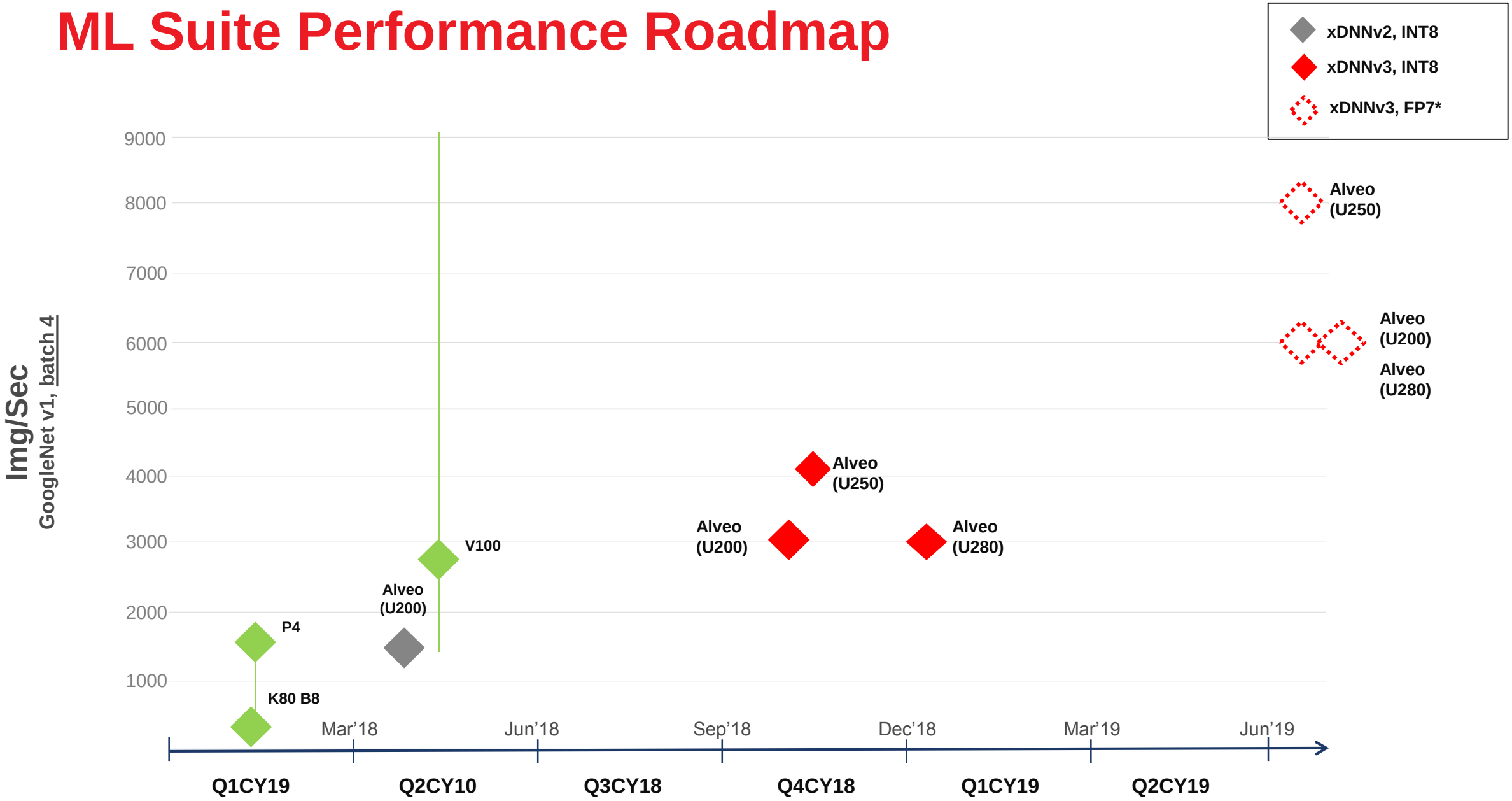
Fast
Advantages in Machine Learning Inference



* Source: Accelerating DNNs with Xilinx Alveo Accelerator Cards White Paper

XILINX

ML Suite Performance Roadmap



XILINX

Alveo Overview

Jim Heaton
Sr. FAE

XILINX.

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

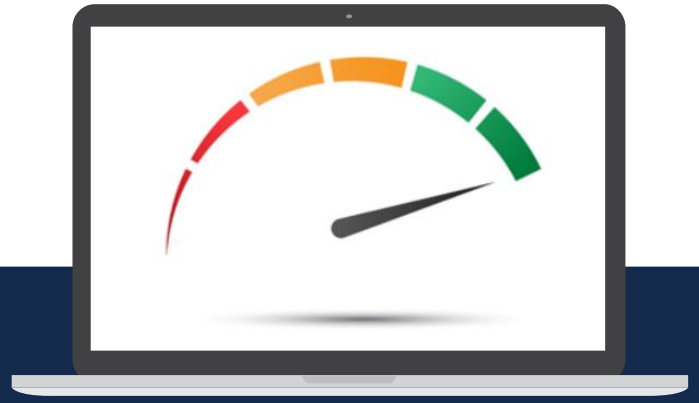
PCIe Gen3x16

Server OEM Support
- Major OEMs in Qualification

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

XILINX

Alveo Accelerator Card Value Proposition



Fast

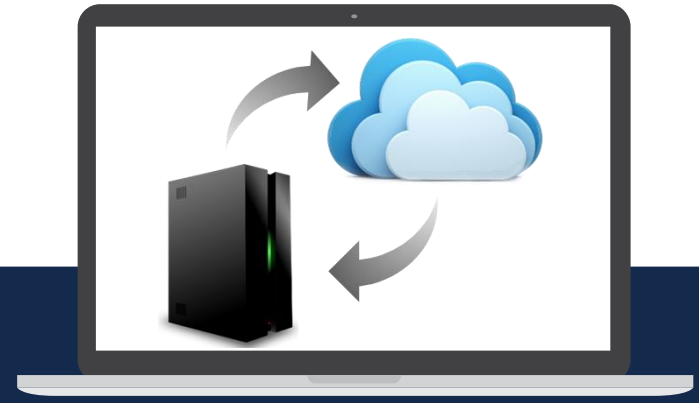
Highest Performance

- Faster than CPUs & GPUs
- Latency advantage over GPUs

Adaptable

Accelerate Any Workload

- Optimize for any workload
- Adapt to changing algorithms



Accessible

Cloud ↔ On-Premises Mobility

- Deploy in the cloud or on-premises
- Applications available now

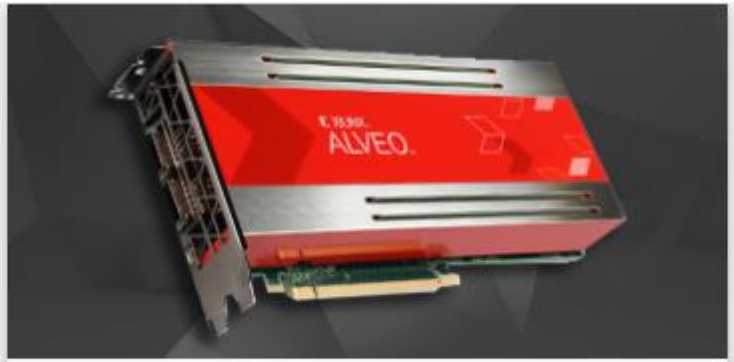
Accelerator Cards That Fit Your Performance Needs

Alveo U200

- **18.6** Peak INT8 TOPs
- **77GB/s** DDR Memory Bandwidth
- **31TB/s** Internal SRAM Bandwidth
- **892,000** LUTs

Buy Now

Product Brief >



Alveo U250

- **33.3** Peak INT8 TOPs
- **77GB/s** DDR Memory Bandwidth
- **38TB/s** Internal SRAM Bandwidth
- **1,341,000** LUTs

Buy Now

Product Brief >

Alveo U280

- **24.5** Peak INT8 TOPs
- **460GB/s** HBM2 Memory Bandwidth
- **30TB/s** Internal SRAM Bandwidth
- **1,079,000** LUTs

Learn More

Coming Soon

XILINX.

Expanding Accelerator Card Portfolio

16nm UltraScale+™

7nm Versal

Performance & Capability

U250
Available Now

U200
Available Now

U280

SmartNIC

Broad Portfolio of Acceleration Cards

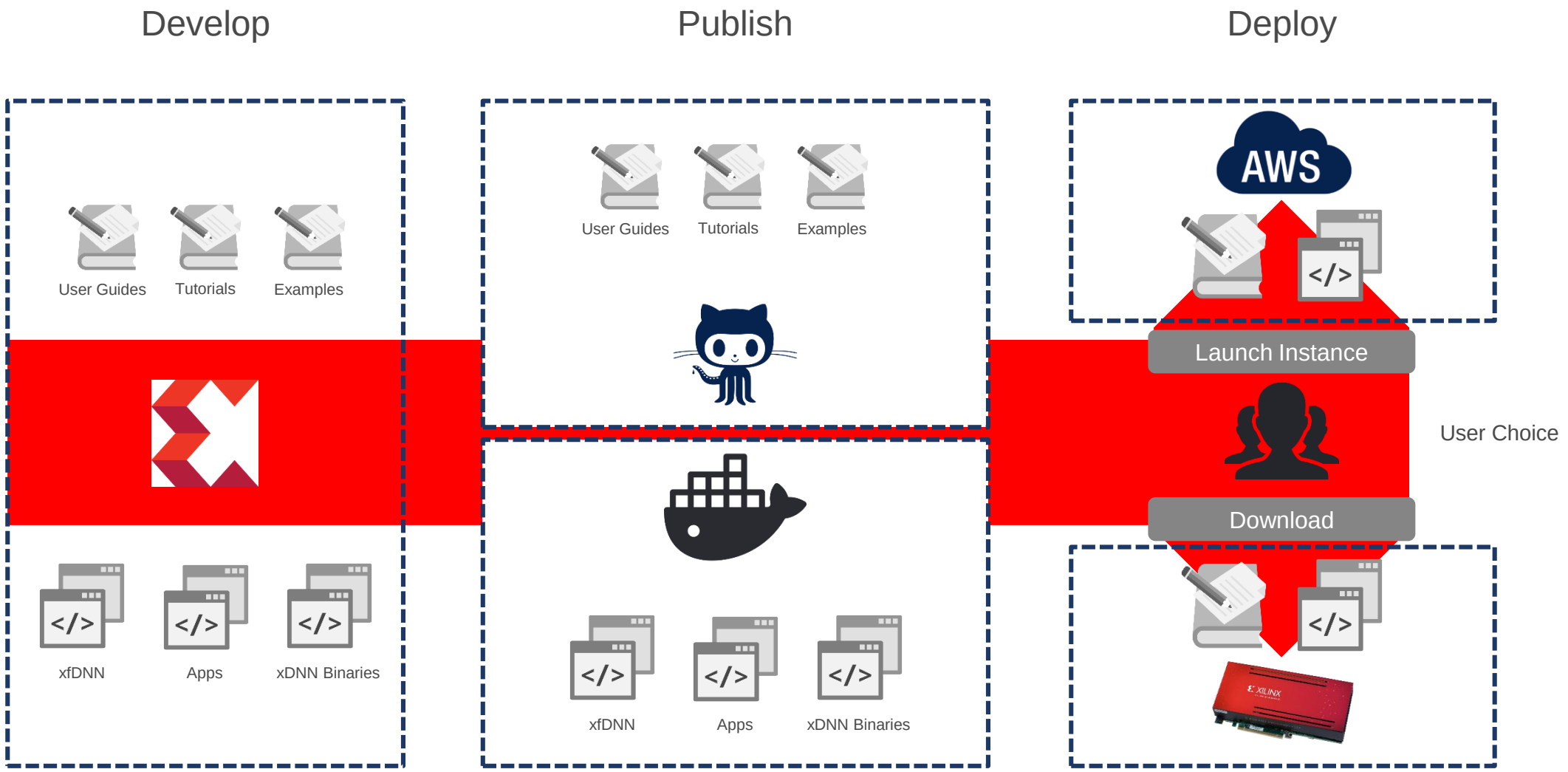
2018

2019

Planned

XILINX

Accessible
Unified Simple User Experience from Cloud to Alveo

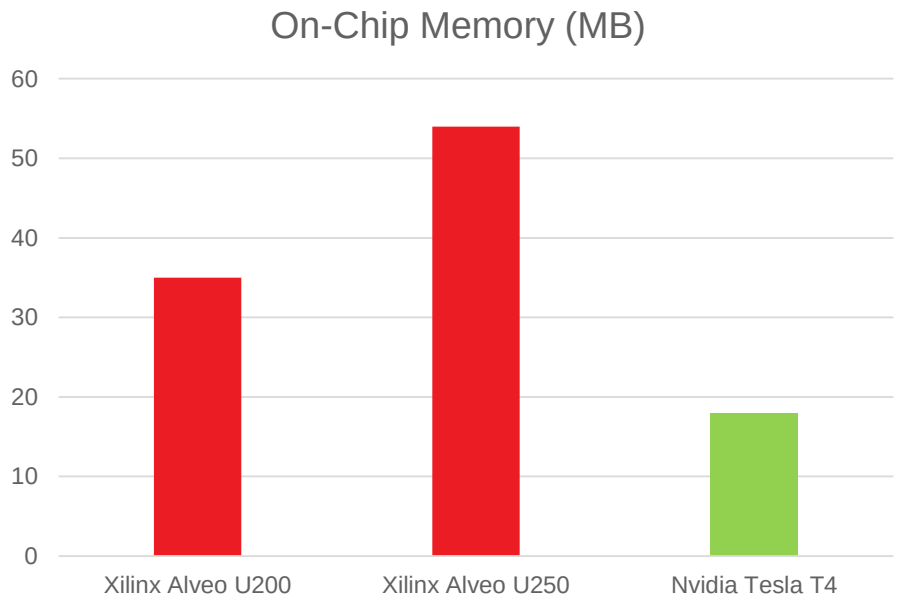


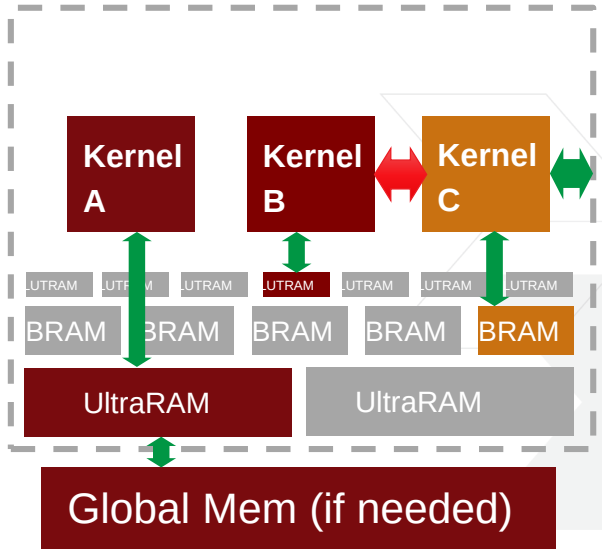
XILINX

Adaptable - On-chip memory

The critical asset on-chip to assist and feed the compute
Store parameters, buffer intermediate activations, and move data around

- **Alveo: Highest on-chip memory capacity**
- **Alveo: Most adaptable memory architecture**





Memory Access of Alveo

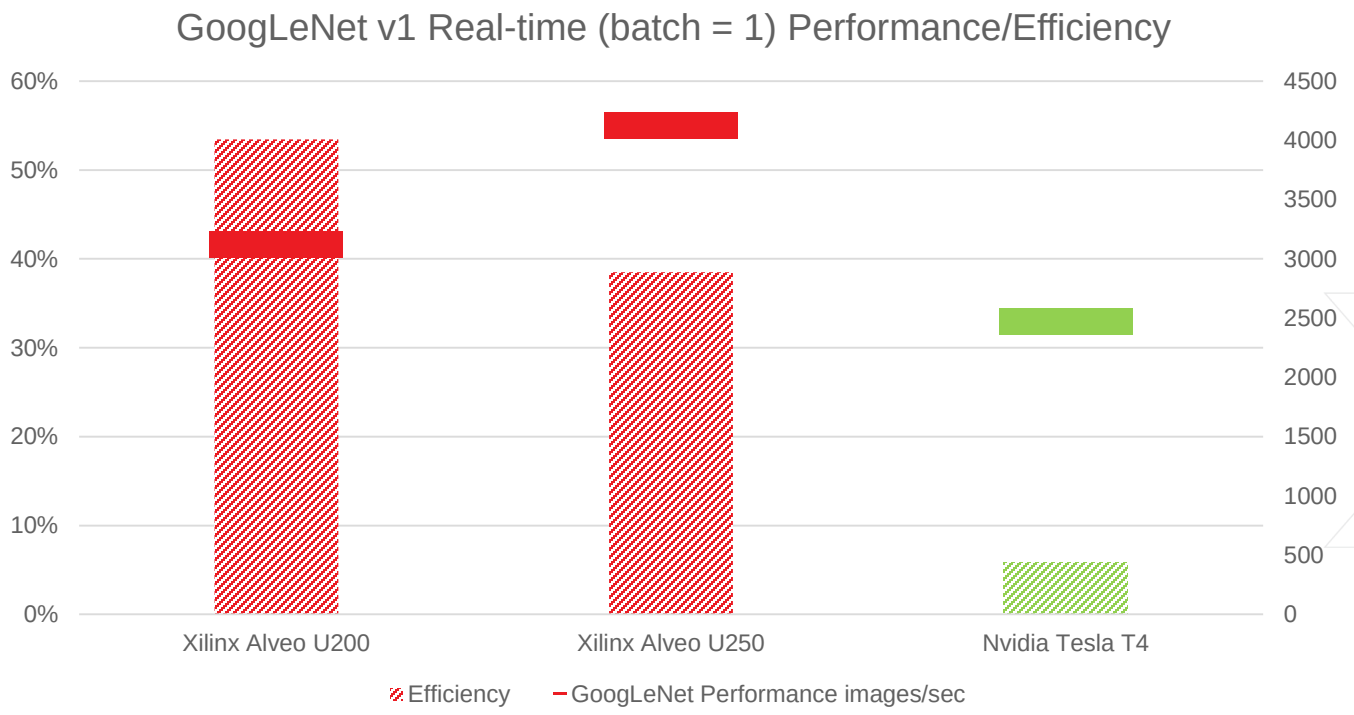
XILINX

Adaptable - Neural Network Inference Efficiency

*T4 efficiency assumes 2x power efficiency improvement vs. P4 as claimed in Nvidia whitepaper: https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/technologies/turing-architecture/NVIDIA-Turing-Architecture-Whitepaper.pdf

XILINX.

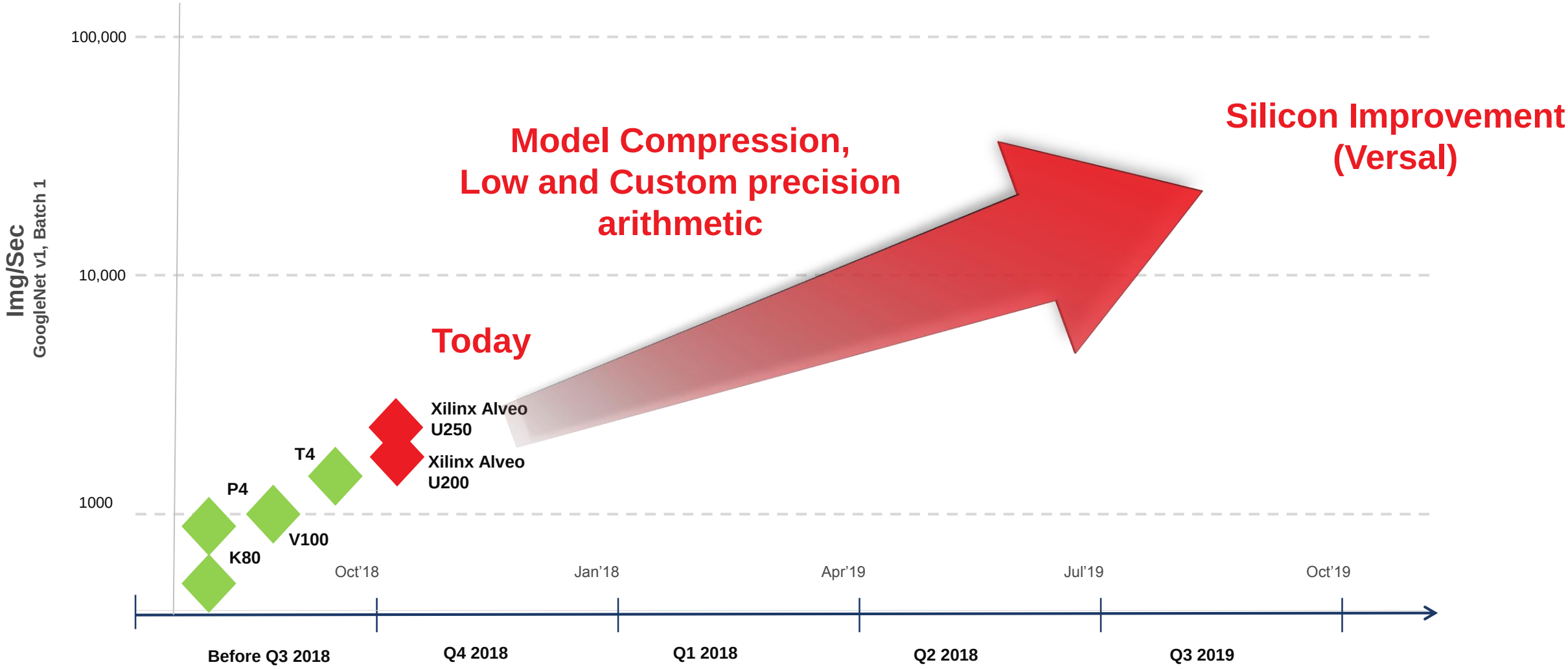
Fast
Real-time Performance



*T4 Performance and efficiency assumes 2x power efficiency improvement vs. P4 as claimed in Nvidia whitepaper:
https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/technologies/turing-architecture/NVIDIA-Turing-Architecture-Whitepaper.pdf

XILINX.

Xilinx Performance Roadmap



XILINX.

Beyond Machine Learning

XILINX

Fast

Advantages Across Many Workload Types

Database	Financial	Machine Learning	Video	HPC & Life Sciences
90x	89x	20x	12x	10x

XILINX.

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 Accelerated Computing — Image Processing

Financial

- MAXELER Technologies MAXIMUM PERFORMANCE COMPUTING — Real-Time Risk Dashboard
- MAXELER Technologies MAXIMUM PERFORMANCE COMPUTING — SIMM Calculation

HPE & Life Sciences

- Falcon COMPUTING — Accelerated Genomics Pipelines

Application Ecosystem Continues to Grow

Infuse Machine Learning with other accelerations

Database

Machine Learning

HPC & Life Sciences

Video

Financial

XILINX

Solution Stack

Accelerated Solutions

Developer Package

Platforms

DEVELOPERS

100%

Growth of Published Applications

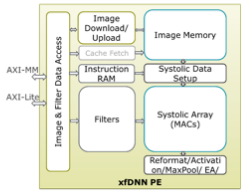
Hundreds

of Developers Trained

RTL, C, C++, OpenCL

SDAccel™ Environment

Xilinx ML Suite

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

End user

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

HPC &LIFE SCIENCES	**FINANCIAL**	**VIDEO**
MACHINE LEARNING		**DATABASE**

Solutions Xilinx ISVs

Platform

Cloud

FPGA as a Service (FaaS)

On-premise

XILINX

Xilinx is Qualifying with Major Server OEMs

Server Support & Qualification Strategy

Interop Level Qual	Stringent Qual	Very Stringent Qual
Xilinx Validated	**3rd Party Option (3PO) Qualified**	**OEM Factory Qualified**
• Dell R730 • Dell R740 • SuperMicro SYS-4028GR-TR • SuperMicro SYS-4029GP-TRT • SuperMicro SYS-7049GP-TRT • HPE ProLiant DL380 G10 COMPLETE	IN PROGRESS	IN PROGRESS

Ordering Information

A	U###	P	64G	PQ	G
Alveo	**Kit Name** U200 U250	**Cooling** P: Passive A: Active	**Memory** 64G: 64GB	**Solution Qualification** ESx: Engineering Sample PQ: Production Qualified	**RoHS Indicator** G: RoHS 6/6

Two Cooling Options

Passive
No Fan

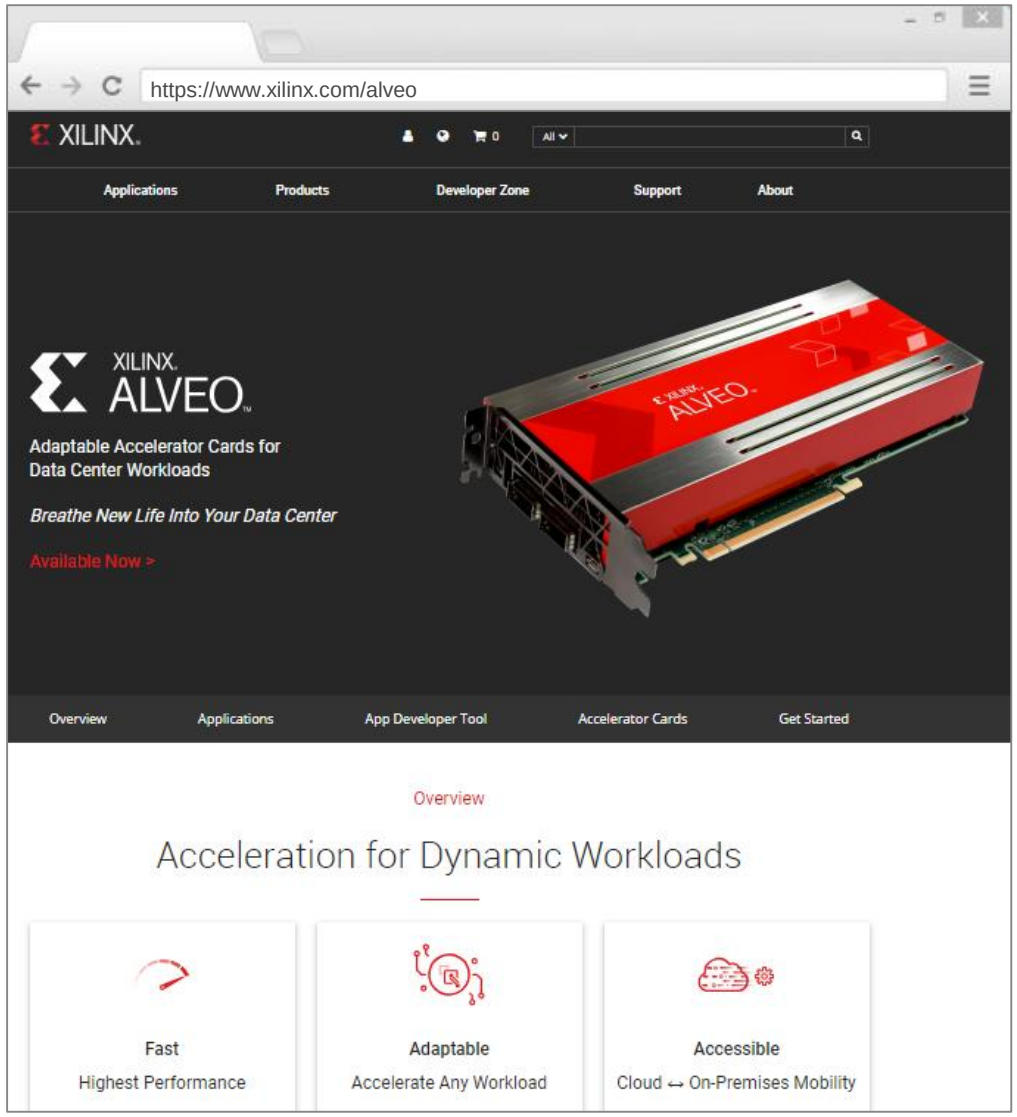
Active
Built in Fan

Alveo Accelerator Card	Part Number	SRP 1pc
U200	A-U200-A64G-PQ-G A-U200-P64G-PQ-G	$8,995
U250	A-U250-A64G-PQ-G A-U250-P64G-PQ-G	$12,995

- **Buy via standard quote/PO process from Xilinx or Avnet**
- **Buy via eCOM (at SRP) from Xilinx, Avnet, DigiKey, EBV, or Premier Farnell**
- **Developer discount available**
  - Requires qualification through Accelerator Program

XILINX

More Information Available on Xilinx.com



Xilinx.com
Product Brief
Product Selection Guide
Getting Started Guide
Data Sheet
ML Solution Brief
SDAccel Solution Brief
ABR Transcoding Solution Brief
Accelerating DNNs with Alveo White Paper
Applications Directory

Adaptable.
Intelligent.

XILINX