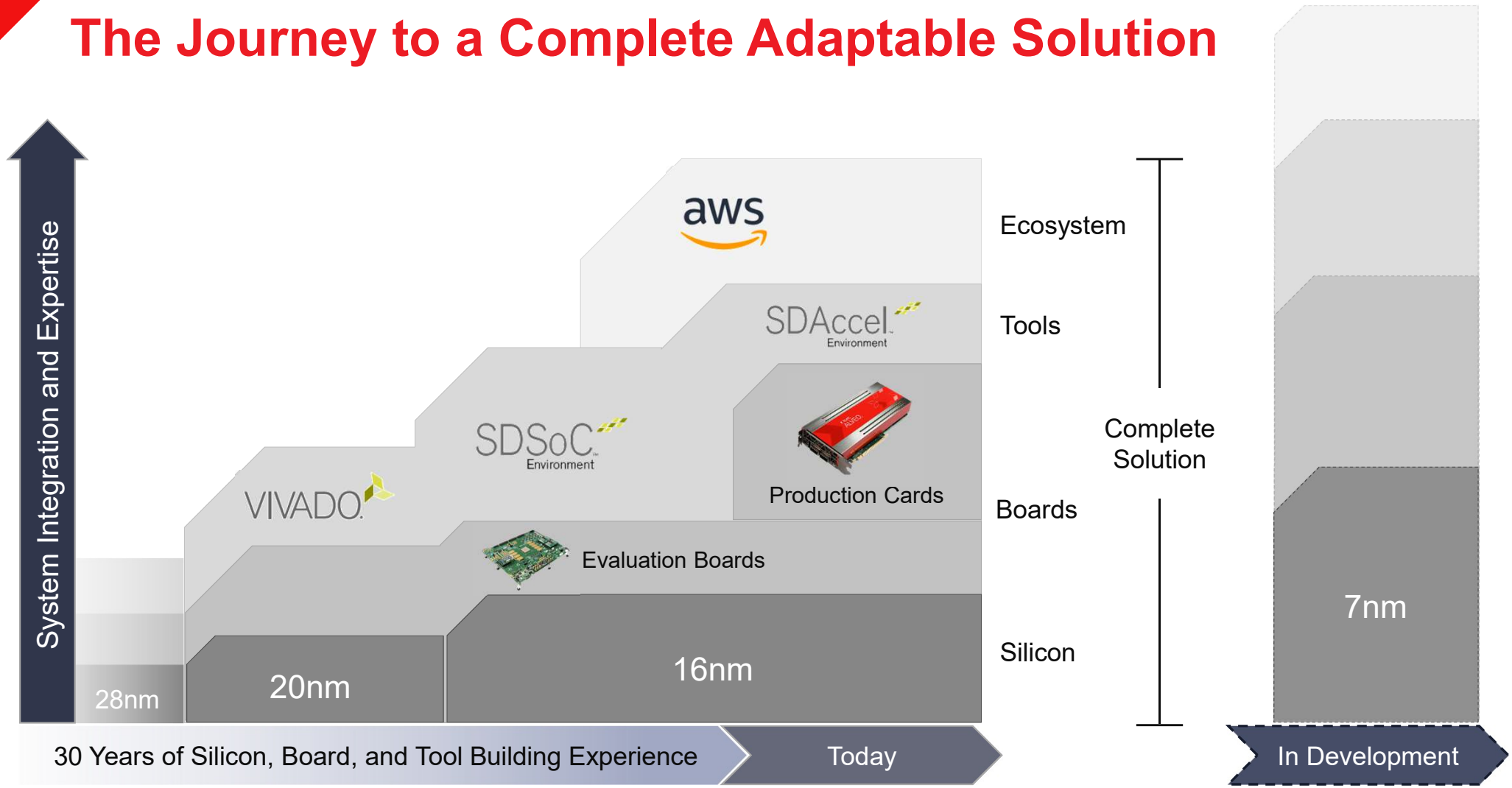

XILINX
ALVEO
XILINX

Adaptable Machine Learning with Alveo Data Center Acceleration Cards

Yao Fu
System Architect – Data Center Acceleration

XILINX

The Journey to a Complete Adaptable Solution



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

Alveo Accelerator Card Value Proposition

Fast

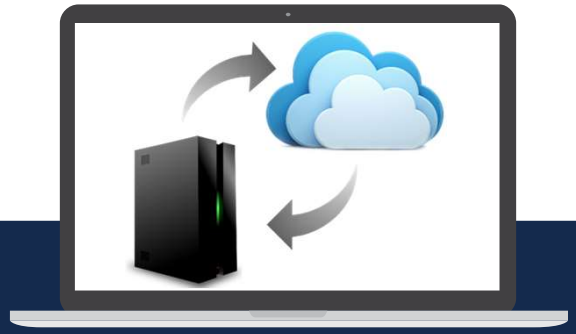
Highest Performance

- 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

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

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

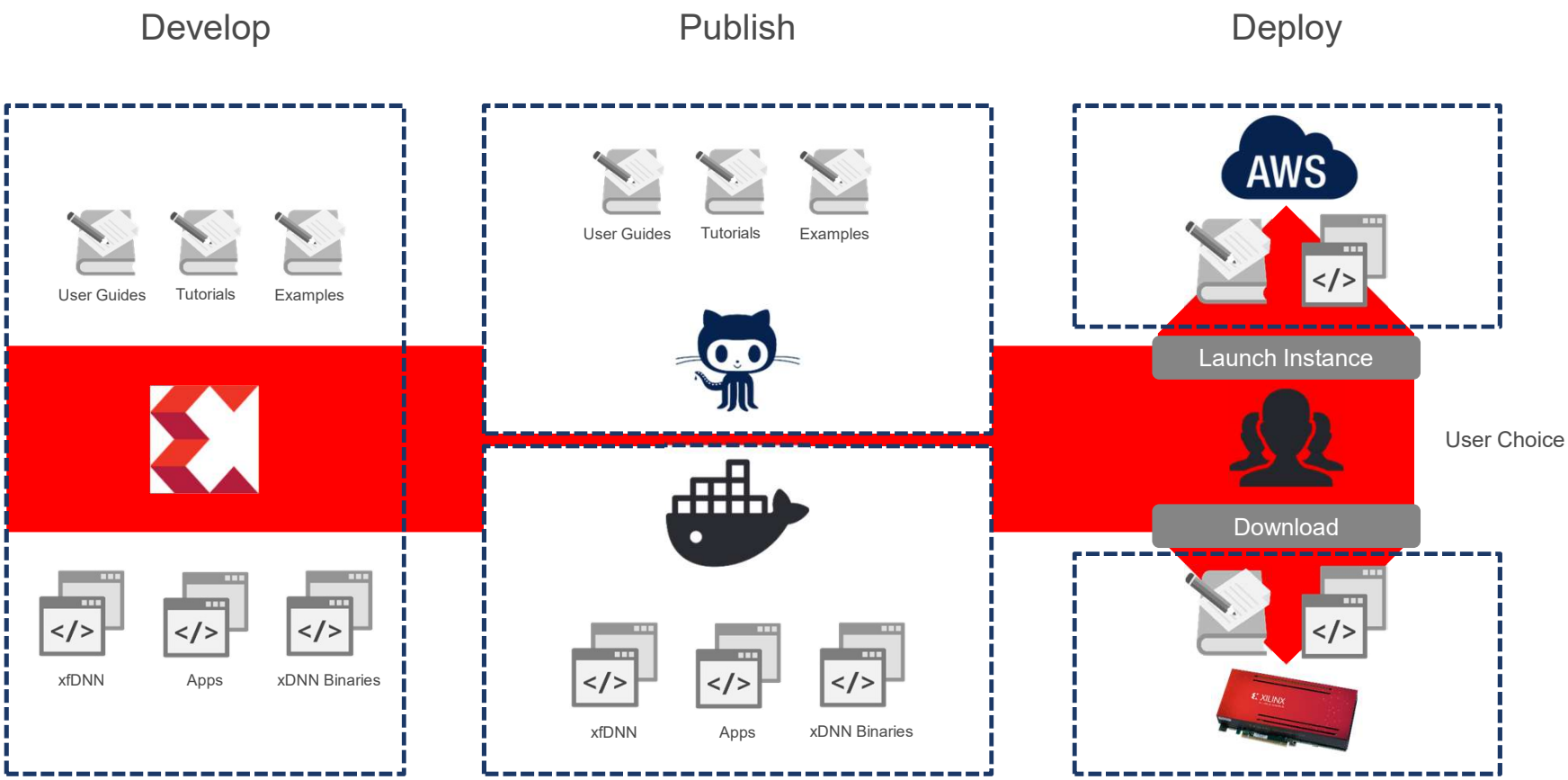
Coming Soon

XILINX.

Adaptable Machine Learning with Alveo

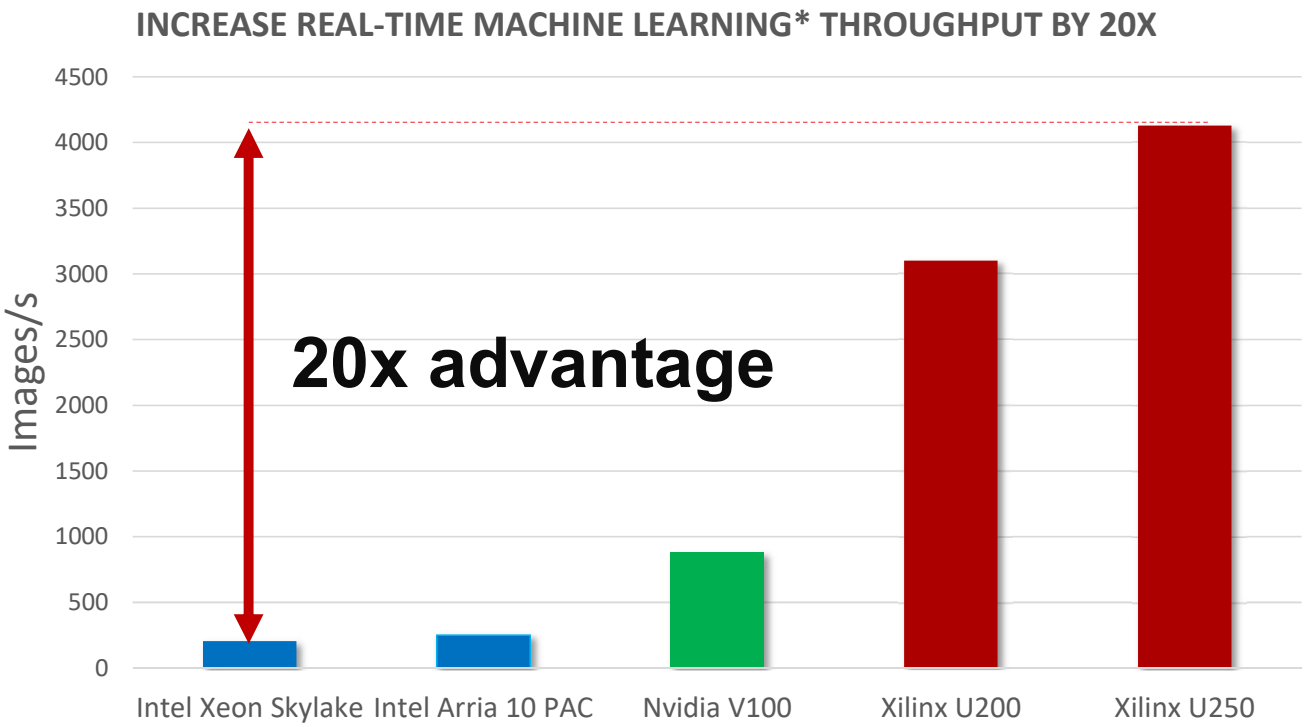
XILINX

Accessible
Unified Simple User Experience from Cloud to Alveo



XILINX

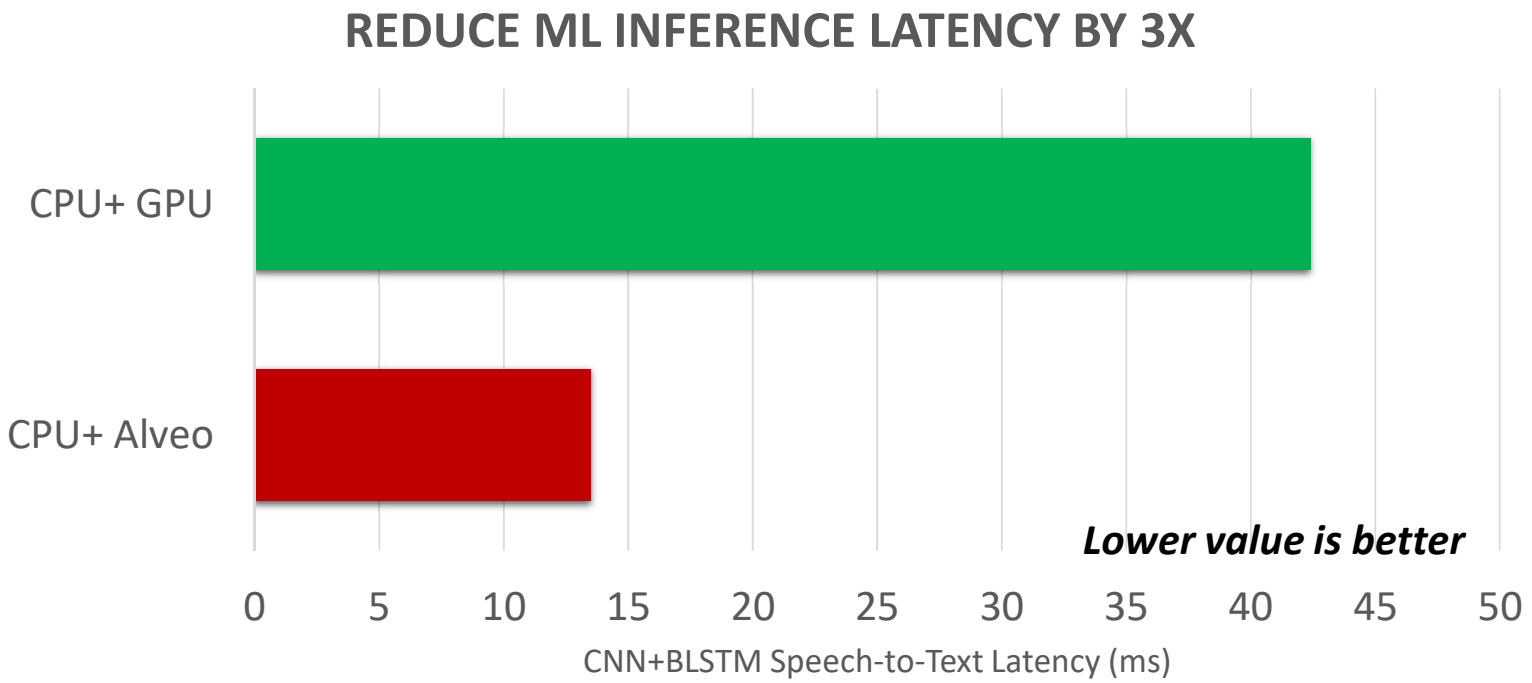
Fast
Advantages in Machine Learning Inference



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

XILINX.

Fast
Advantages in Latency



Alveo Provides Massive Parallel Compute with Lowest Latency vs GPUs

XILINX.

Accessible
Xilinx Machine Learning Suite

From Community

{RESTful API}

C++ | Scala | R | JS | python | julia | MATLAB | Go

Caffe | TensorFlow | mxnet

Xilinx ML Suite

xfDNN Middleware, Tools and Runtime

xDNN Processing Engine

HUAWEI | NIMBIX | Alibaba Cloud aliyun.com | awsmarketplace

Xilinx U200, U250

- 4x higher real time inference over GPUs
- Highest Perf/Watt
- Easy to Use ML Frameworks and APIs

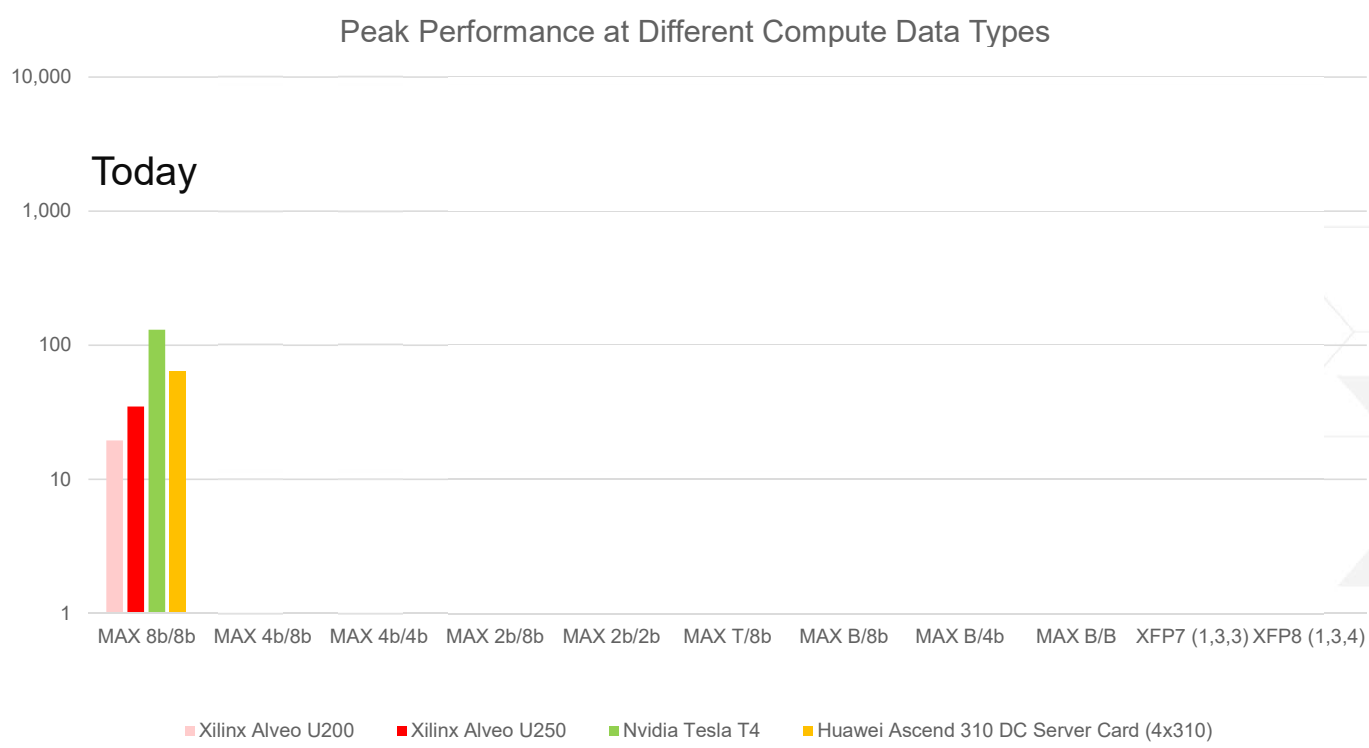
https://www.xilinx.com/ml

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

XILINX

Adaptable - Compute Capability

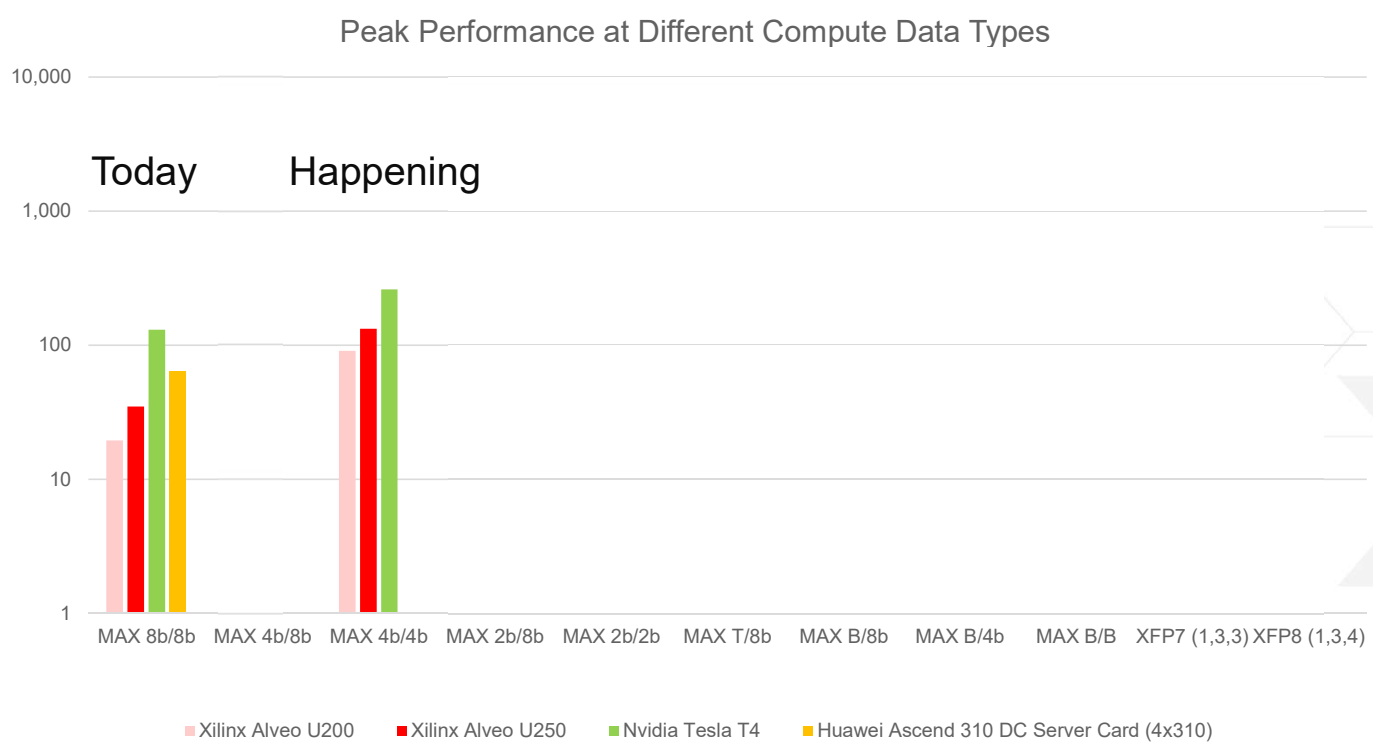
> **INT8 is becoming standard for inference devices**



XILINX

Adaptable - Compute Capability

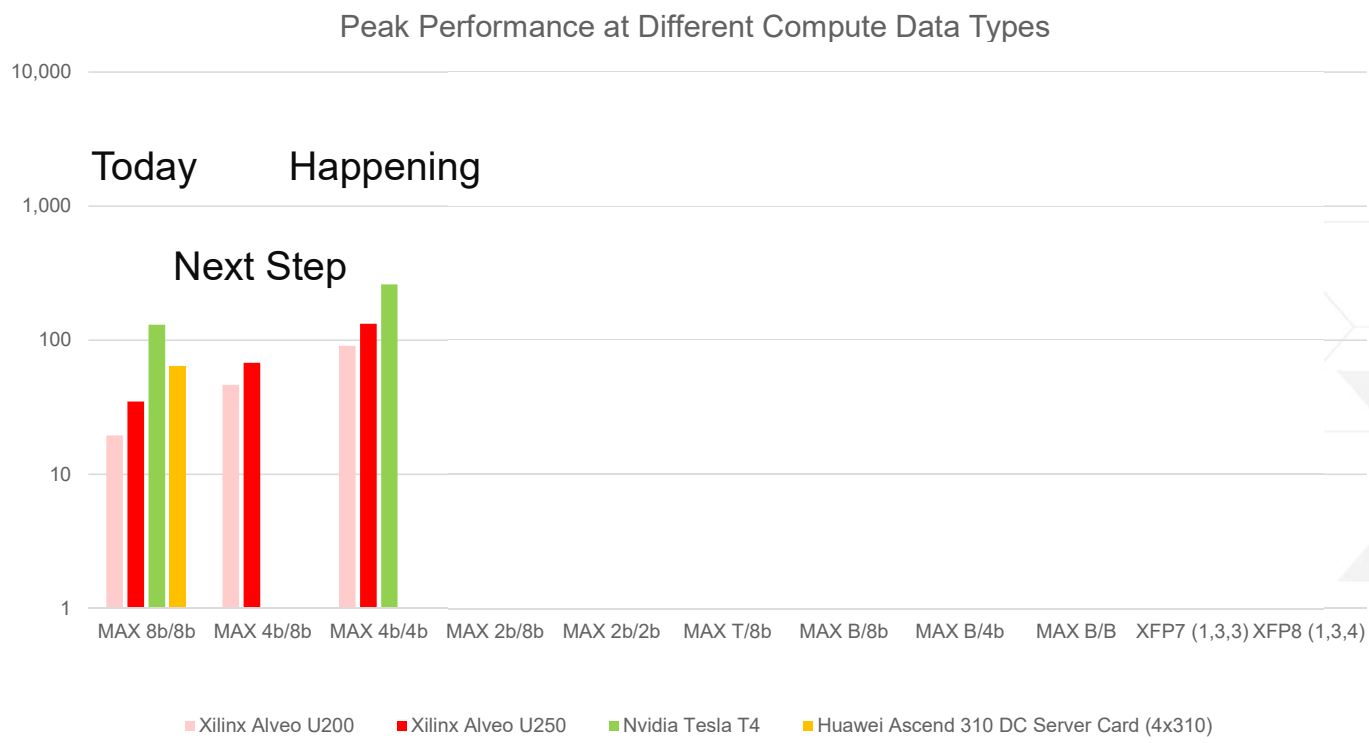
> **Trends are moving toward lower precision, e.g. INT4 is emerging**



XILINX

Adaptable - Compute Capability

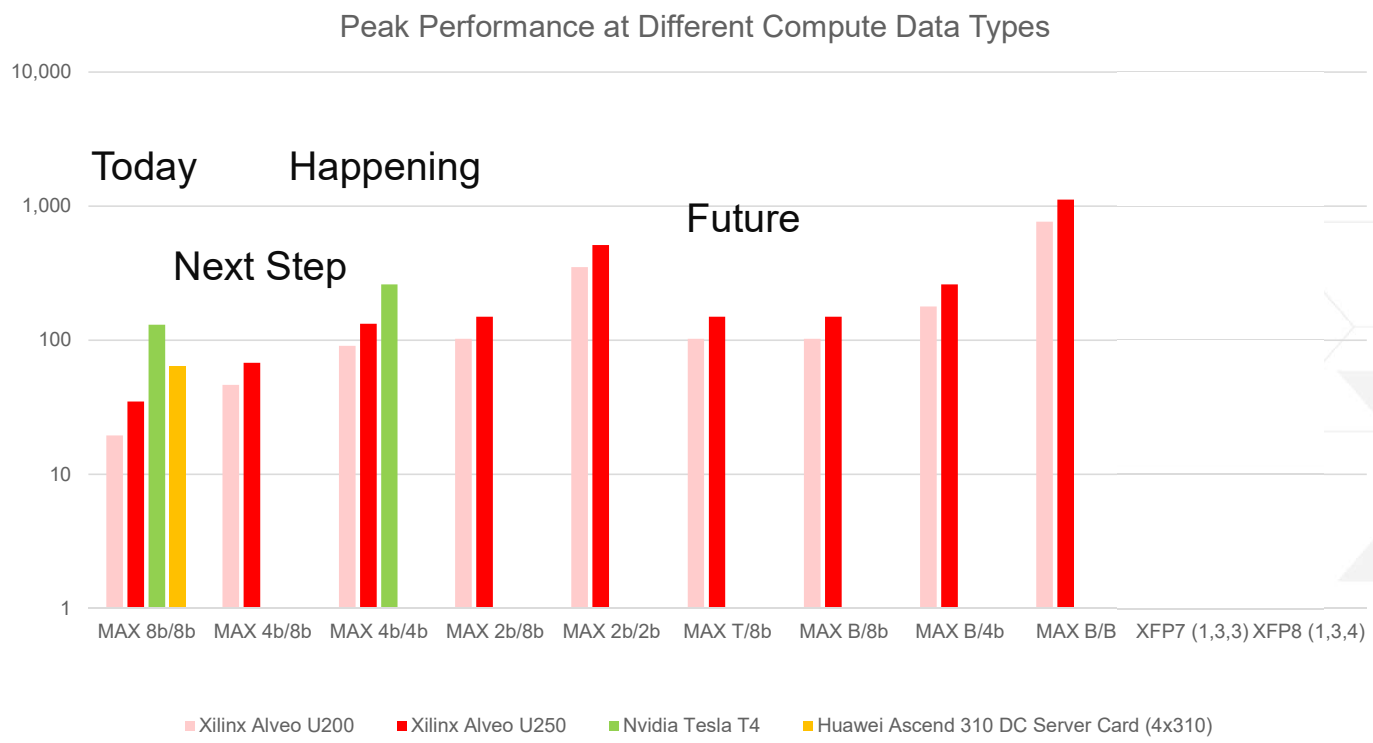
> **Mixed weight and activation precisions provide additional optimization for inference while keeping high accuracy**



XILINX

Adaptable - Compute Capability

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

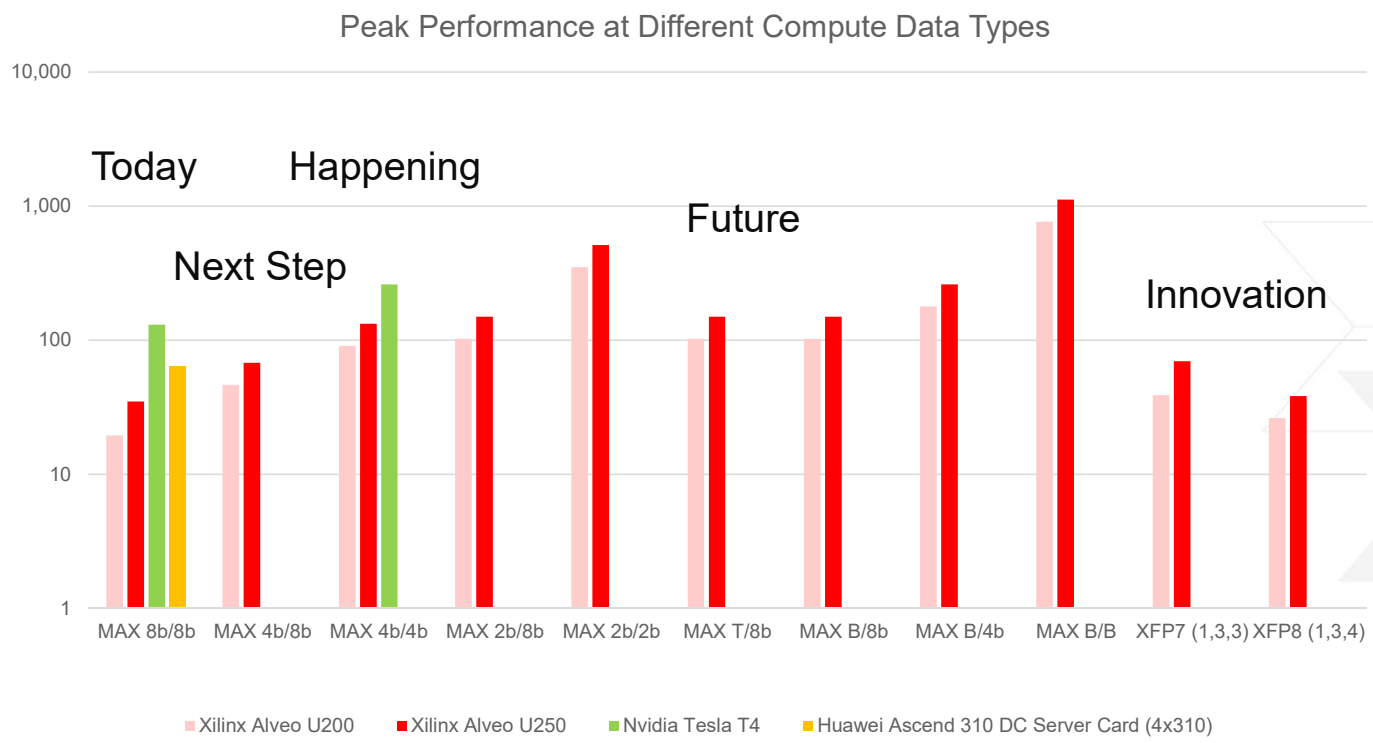


XILINX

Adaptable - Compute Capability

> **Alveo: Most future proof architecture**

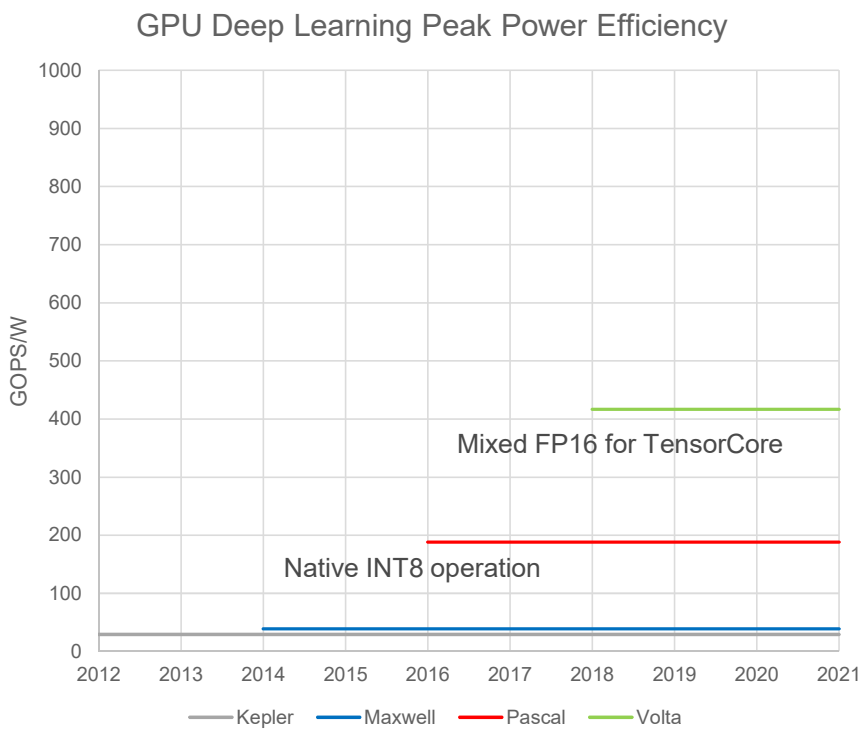
> **Scalable performance for any data type**

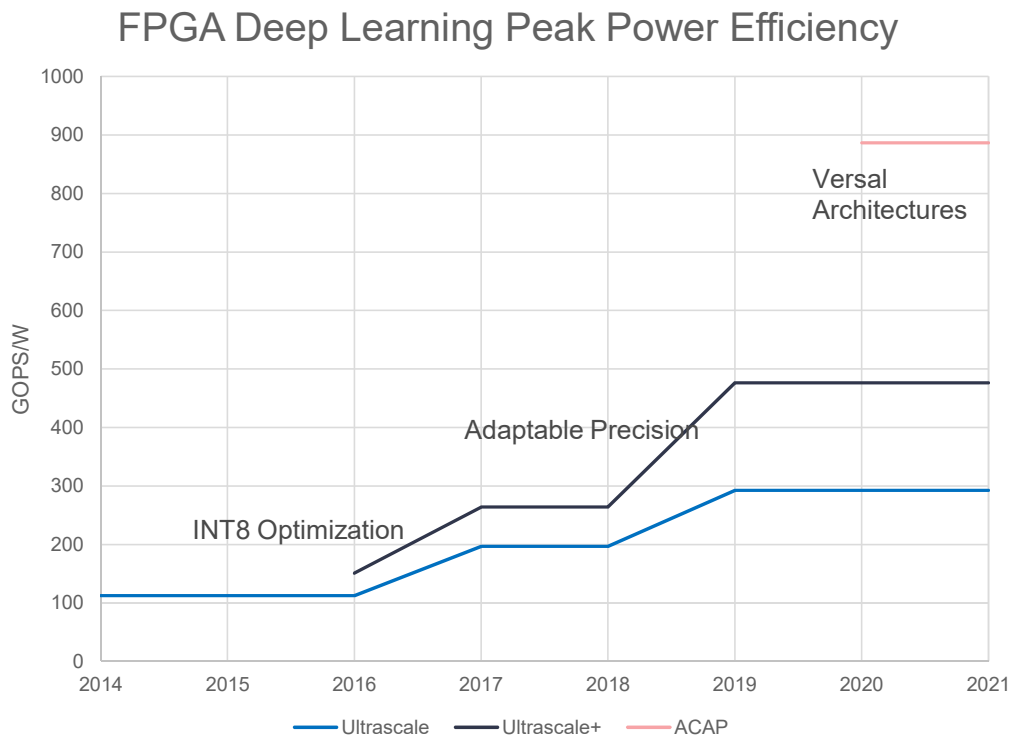


XILINX

Adaptable - Break Through on Peak Performance

- GPU: Introduce new architectures and silicon
- Xilinx: Adapt the break through of emerging domain knowledge





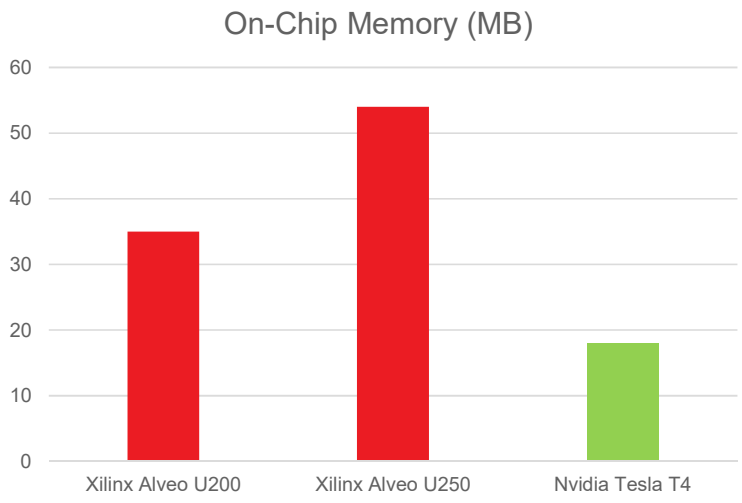
XILINX.

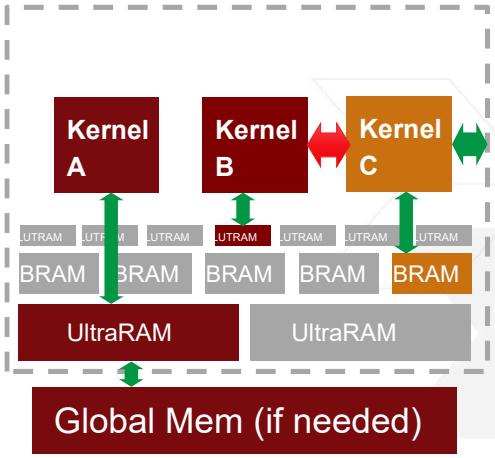
Adaptable - On-chip memory

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**

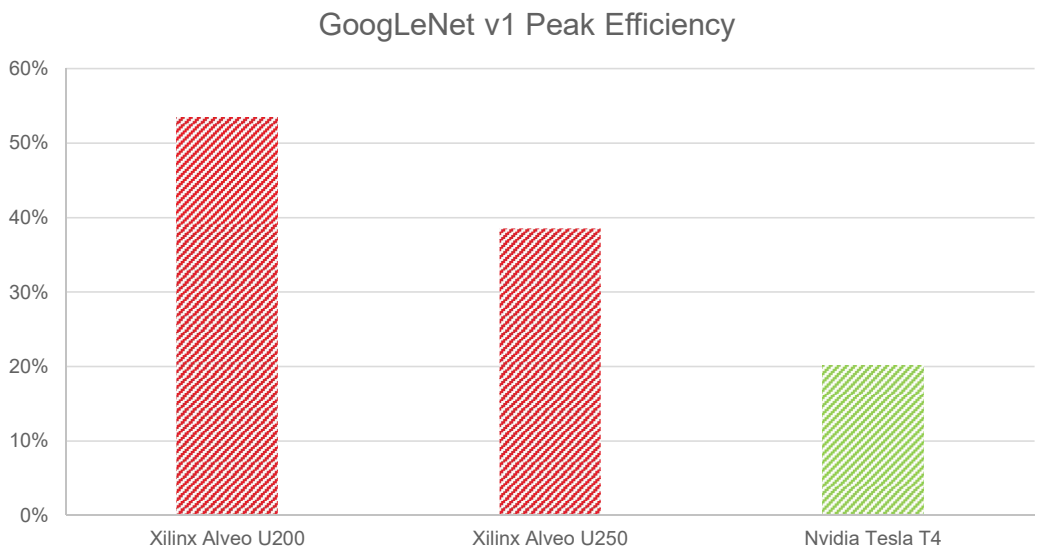




Memory Access of Alveo

XILINX

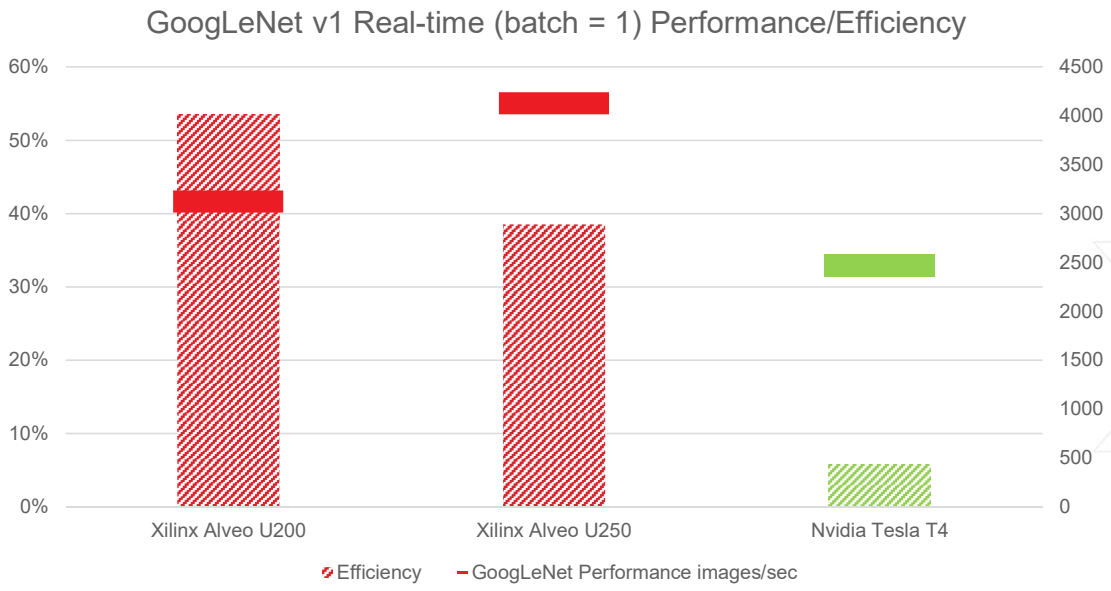
Adaptable - Neural Network Inference Efficiency



*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

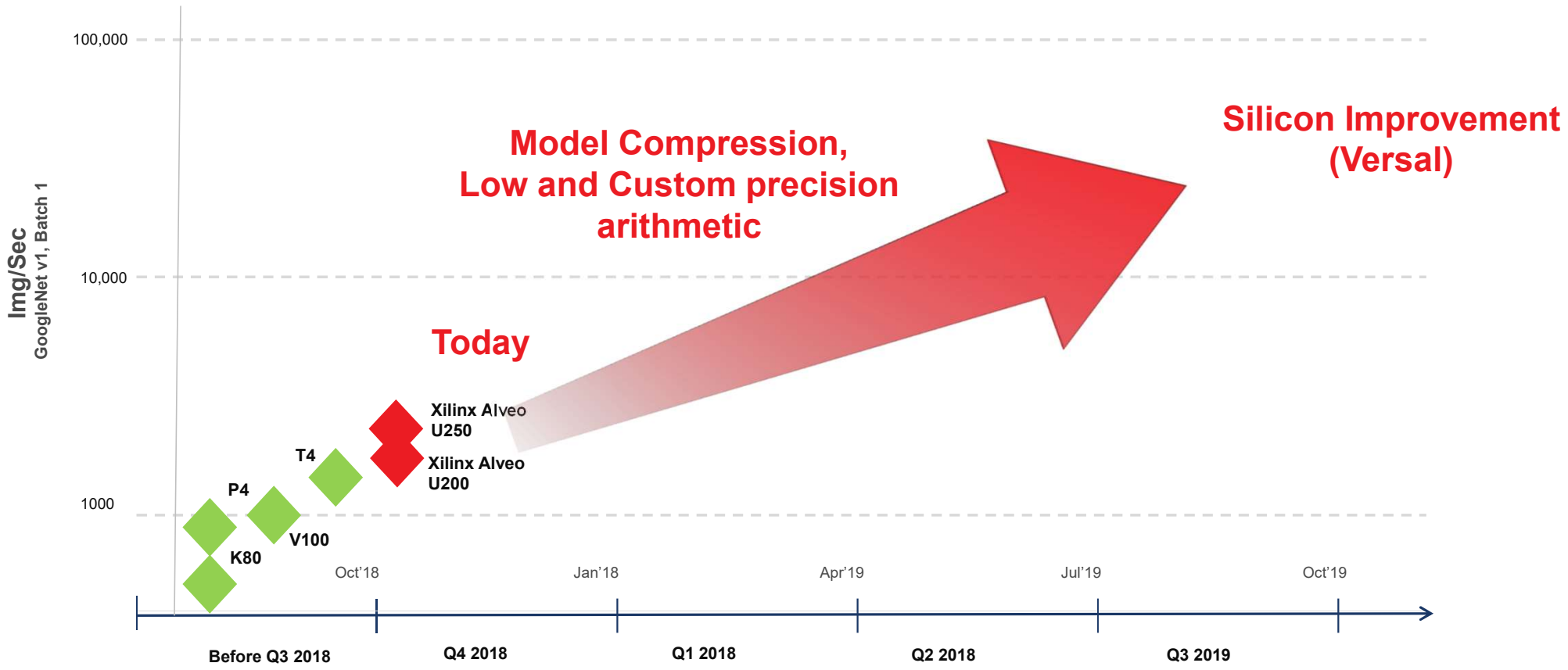
XILINX

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 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.

Application Adaptability

Strong Acceleration
Some Acceleration
No Acceleration

Machine Learning | Video | Database | HPC & Life Sciences | Financial

FPGA

GPU

ASIC

More

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

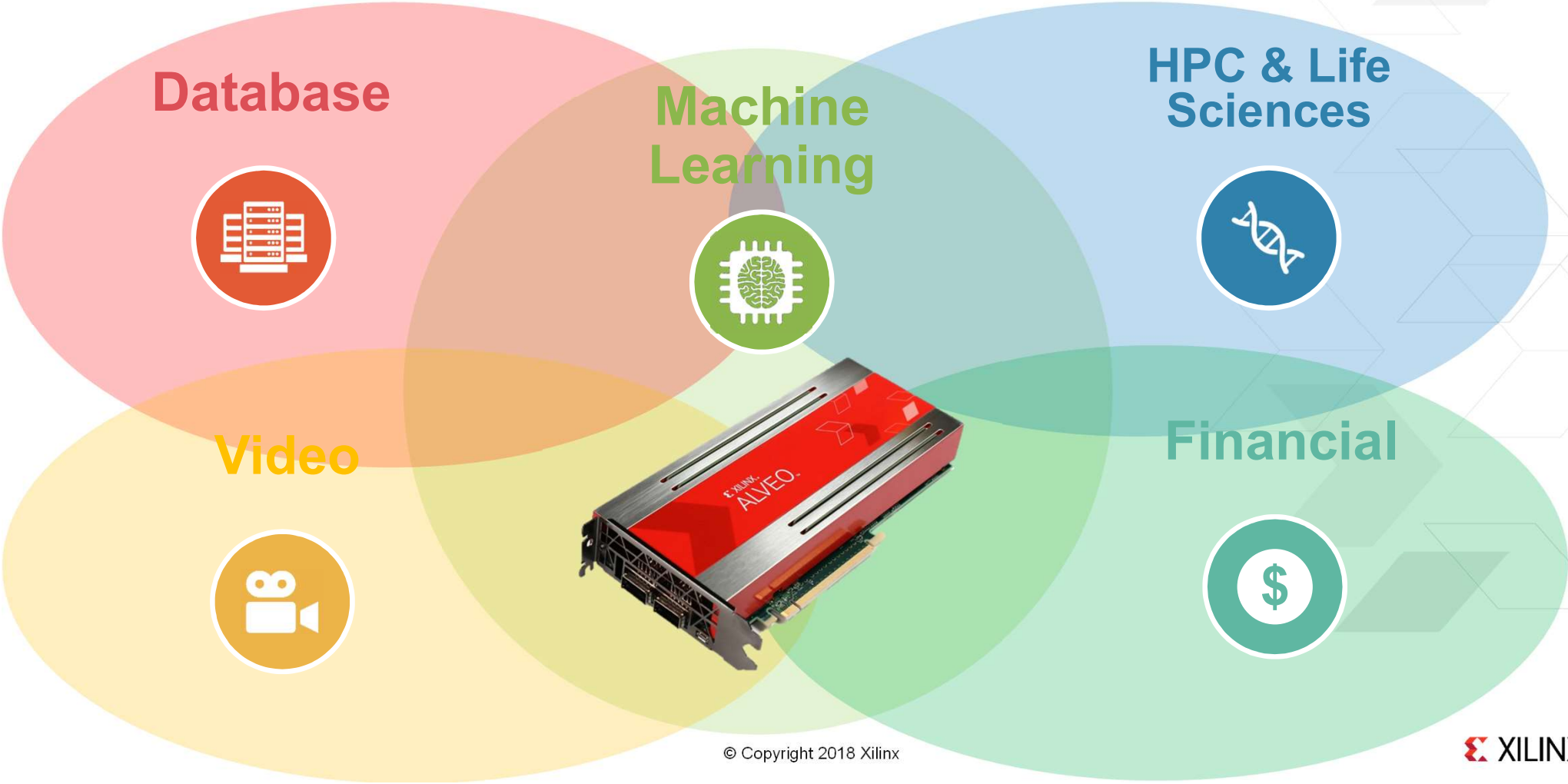
HPE & Life Sciences

- Falcon COMPUTING — Accelerated Genomics Pipelines

Application Ecosystem Continues to Grow

XILINX

Infuse Machine Learning with other accelerations



XILINX

Solution Stack

Accelerated Solutions

Developer Package

Platforms

DEVELOPERS

100%

Growth of Published Applications

Hundreds

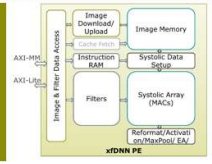
of Developers Trained

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

FPGA as a Service (FaaS)

Cloud

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

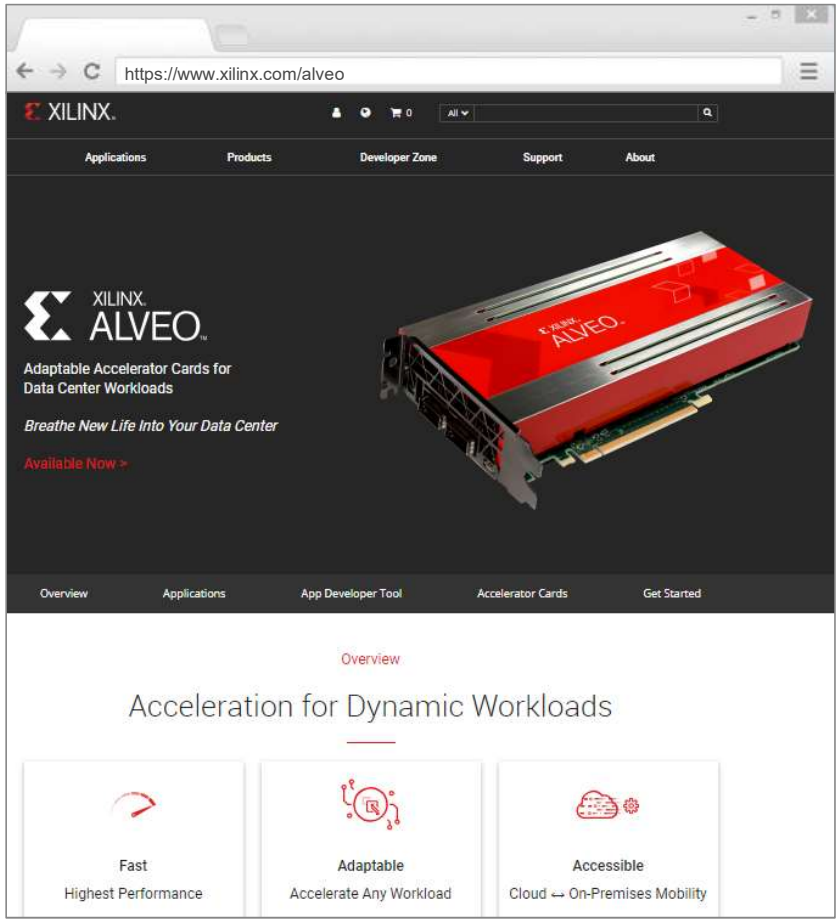
Alveo Accelerator Card	Part Number	SRP 1pc	Developer Discount 1-2pc limit[1]
U200	A-U200-A64G-PQ-G A-U200-P64G-PQ-G	$8,995	$3,995
U250	A-U250-A64G-PQ-G A-U250-P64G-PQ-G	$12,995	$6,495

1 – Requires qualification through Accelerator Program

- **Buy via standard quote/PO process from Xilinx or Avnet**
- **Buy via eCOM (at SRP) from Xilinx, Avnet, DigiKey, EBV, or Premier Farnell**

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

XILINX.

Adaptable.
Intelligent.

XILINX