Xilinx Alveo Portfolio Expansion

Alveo U50 Launch

UNDER EMBARGO UNTIL TUESDAY, AUGUST 6, 2019 at 6 A.M PT/9 A.M. ET

Expanding the Alveo Platform

- Announcing Xilinx Alveo U50 - Industry's First Adaptable Compute, Networking, Storage Accelerator built for Any Server, Any Cloud
- Dramatic improvements in throughput, latency and power efficiency performance across a range of critical data center applications
- Broad and growing Alveo ecosystem of software partners and continued enhancement of developer tools to scale up Alveo solutions

XILINX

ALVEO™

FAST

Built for high throughput, ultra-low latency

Accelerate compute, networking, storage

ADAPTABLE

Deploy optimized domain-specific architectures

Adapt to changing algorithms

ACCESSIBLE

Deploy in the cloud or on-premises

Rich set of accelerated Applications

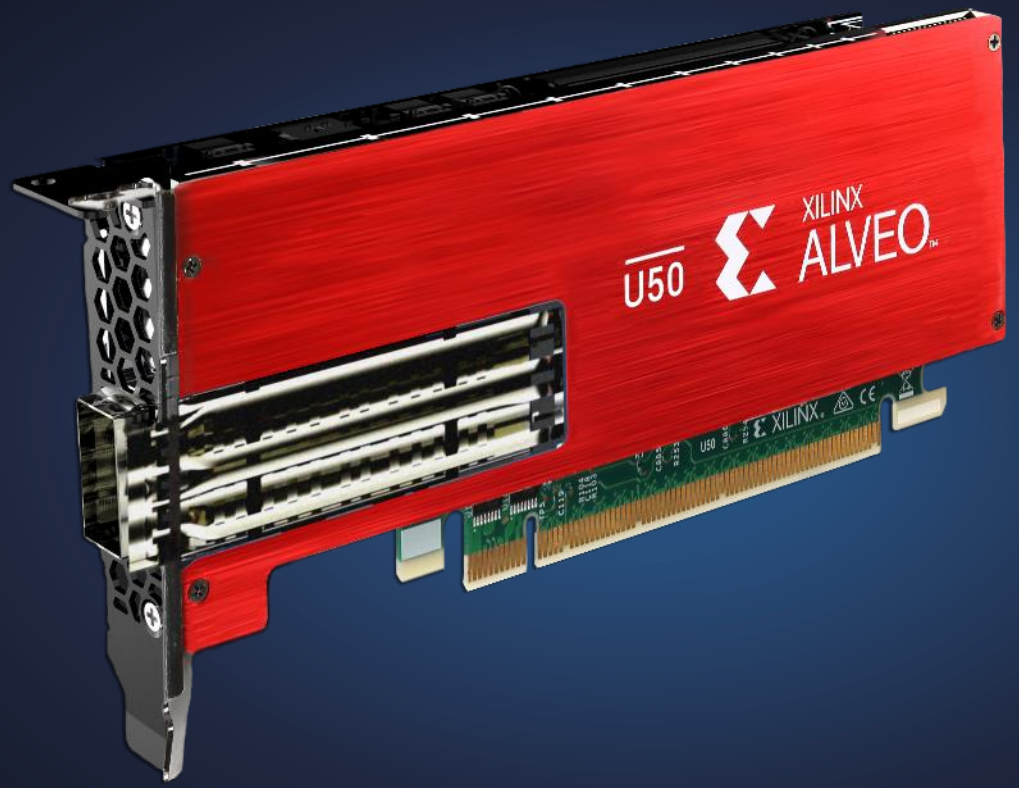
Alveo U50: Industry's First Adaptable Accelerator Built for Any Server, Any Cloud

Over 10X improvement in performance and power efficiency for critical data center applications

Most advanced, adaptable platform for accelerating compute, storage, and networking

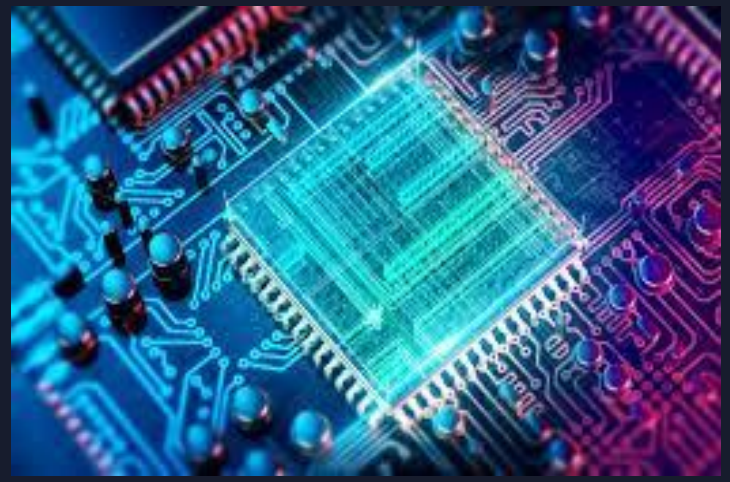
Ultra-efficient power envelope and form factor make it flexible enough for any server, any cloud

Xilinx Alveo U50 Key Specifications



- UltraScale+ Architecture
- Low-profile form factor
- 8GB HBM2 Memory, 460GB/sec
- PCIe Gen4, CCIX, PCIe Gen3
- QSFP 28 (100GbE)
- < 75W

Accelerating the Data Center



Compute

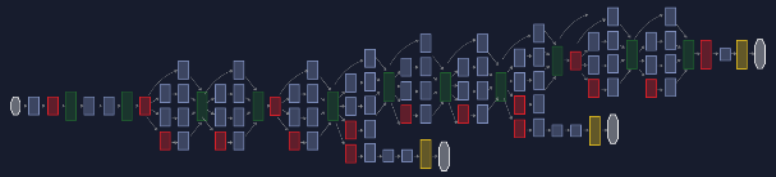
Networking

Storage

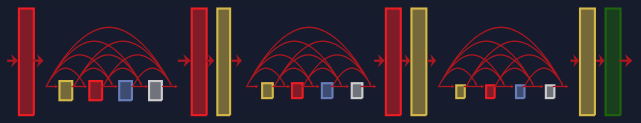
Domain Specific Architectures

AlexNet

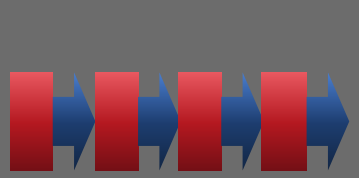
GoogLeNet

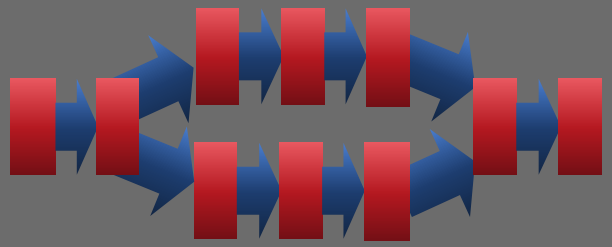


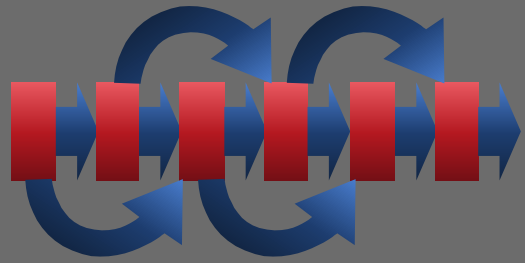
DenseNet



Highest throughput, latency, and efficiency requires different HW architecture





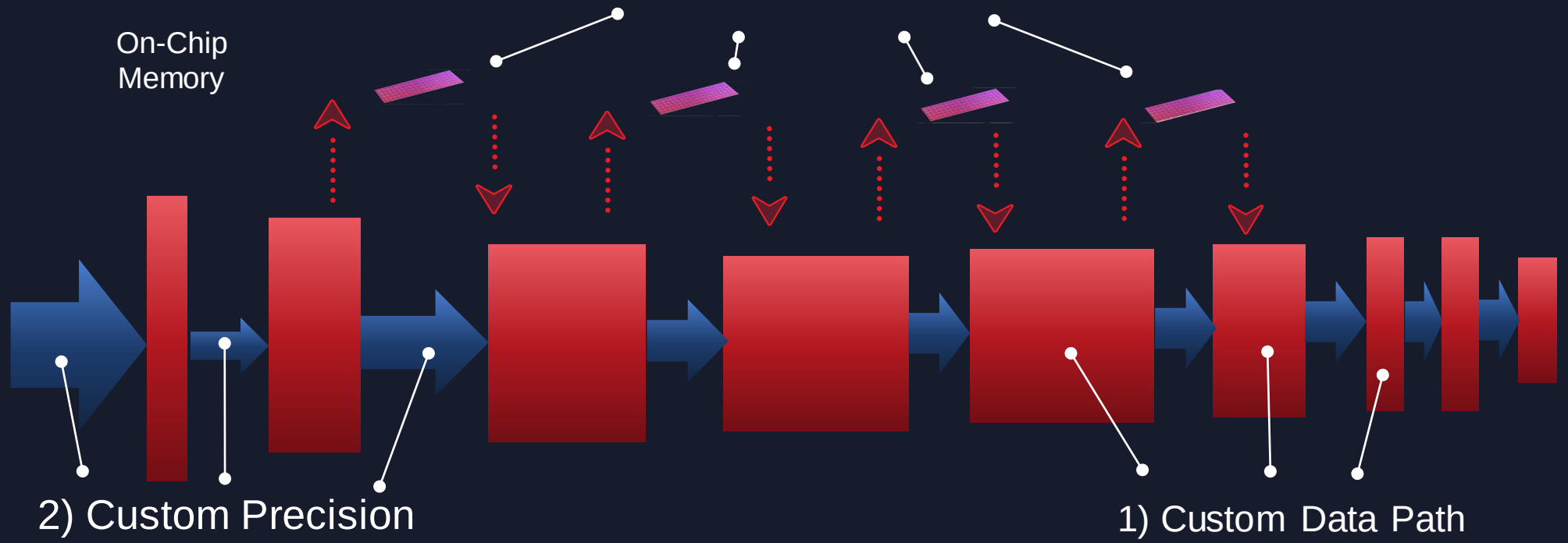


Optimized Performance

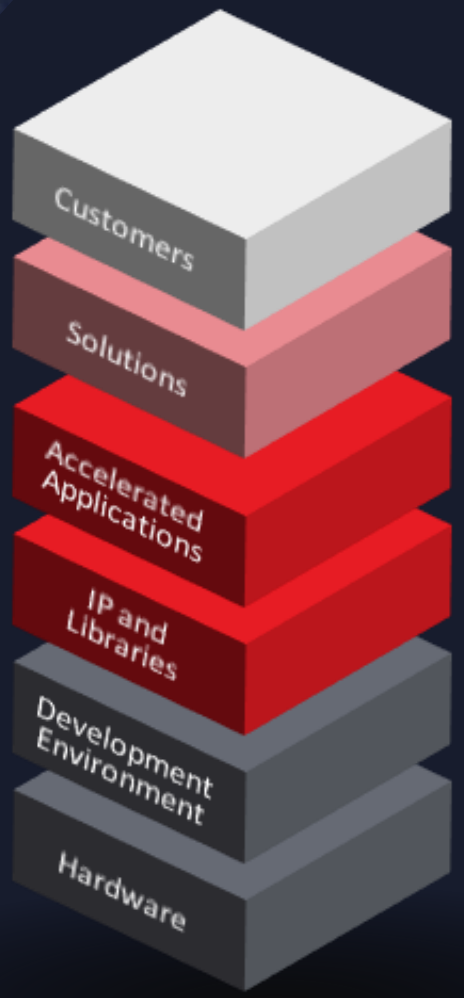
Requires Custom Memory and Datapath

Off-Chip DDR

On-Chip Memory



ALVEO Solution Stack



End Customers

Solution Providers

App & IP Developers

Channel Partners

Data Analytics	Video & Image Processing	Machine Learning	Life Science & HPC	Financial Computing

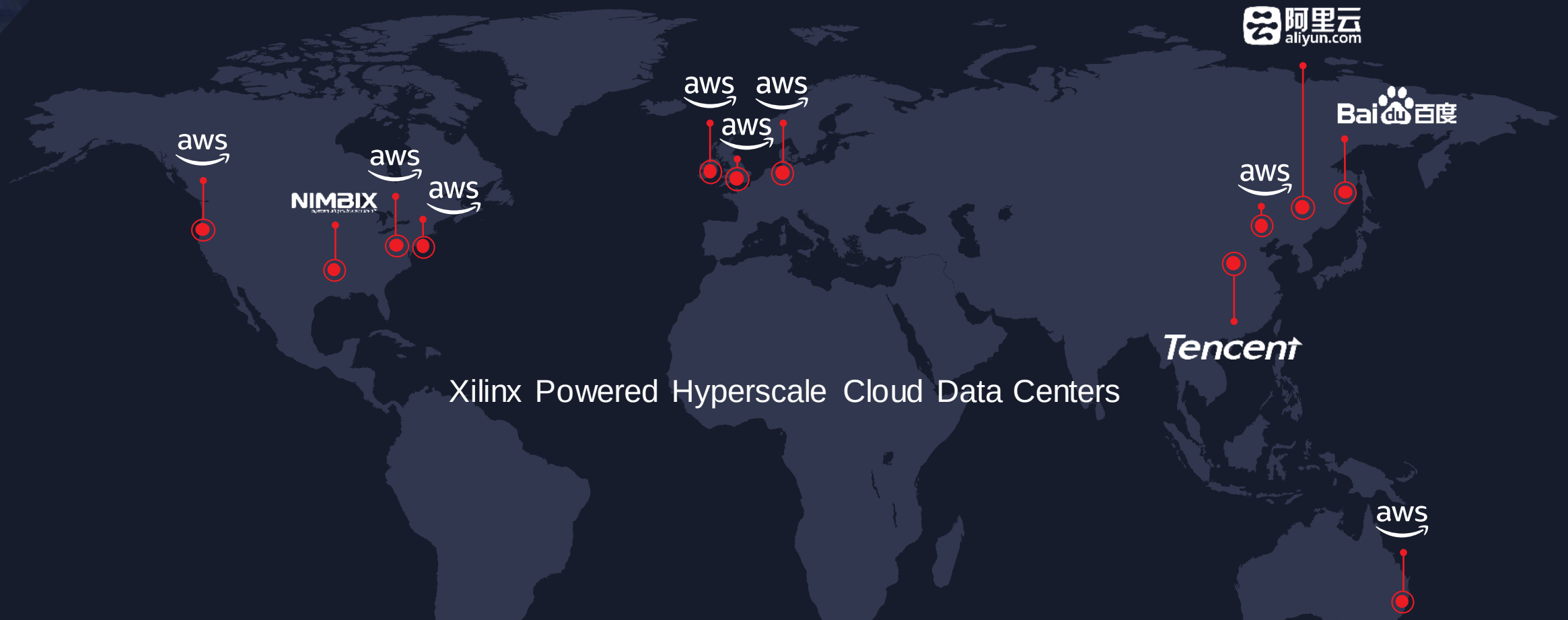
Titan IC, NGCODEC, edico genome, MAXELER Technologies, DEEPHI 深鉴科技, PLUNIFY, CTACCEL, DEEPOLY, Falcon COMPUTING, Mipsology, byte LAKE, XELERA, BLACKLYNX, VITESSE DATA, skreens, swarm64, ALGO-LOGIC, MEGH COMPUTING, boon, V-NOVA, Nextera Video, bigstream, reniac, inaccel, BigZetta Systems, LegUp, m1e

aws, Tencent Cloud, NIMBIX, SUPERMICRO, inspur, DELL, Alibaba Cloud, Baidu 百度

CLOUD

ON-PREMISE

Growing Cloud Availability



Xilinx Powered Hyperscale Cloud Data Centers

Growing Ecosystem

Data Analytics

bigstream
BLACKLYNX
swarm64
XELERA
inaccel
VITESSE DATA
BigZetta Systems
LegUp
RENIAC

Life Sciences & HPC

illumina
byte LAKE
inaccel
Falcon COMPUTING
MAXELER Technologies
deCODE Genetics

Video Processing

NGCODEC
V-NOVA
MEGH COMPUTING
Nextera Video
skreens
VYUSYNC

Machine Learning

Mipsology
DEEPHI 深鉴科技
XELERA
SumUp

Financial Computing

ALGO-LOGIC
MAXELER Technologies
SciComp

Image Processing

DEEPOLY
CTACCEL

Growth Since October 2018 Launch

2x
Published Applications

4x
Developers Trained

Deployment Stack for Scale

ALVEO CONTAINERIZED APPLICATIONS

KUBERNETES ORCHESTRATION

XILINX ALVEO

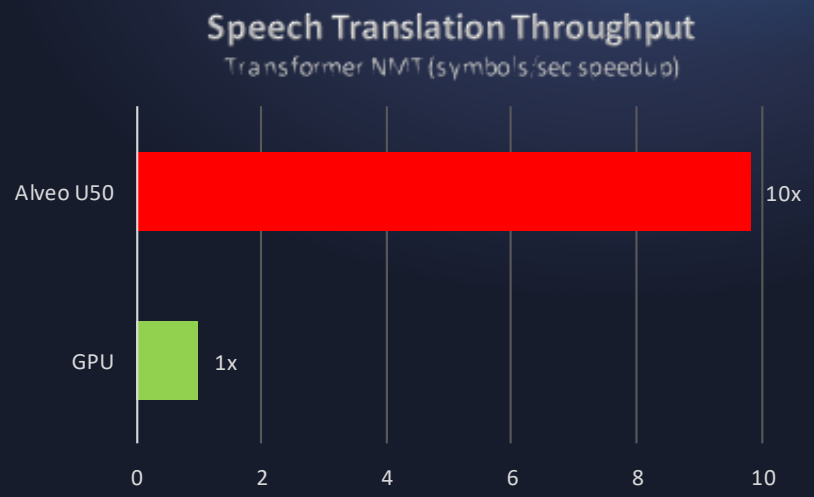
CLOUD ALVEO INSTANCES

ON-PREM ALVEO SERVERS

Speech Translation

High Throughput and Low Latency Inference Acceleration

High Throughput & Low-Latency Inference Performance Unachievable by CPUs & GPUs

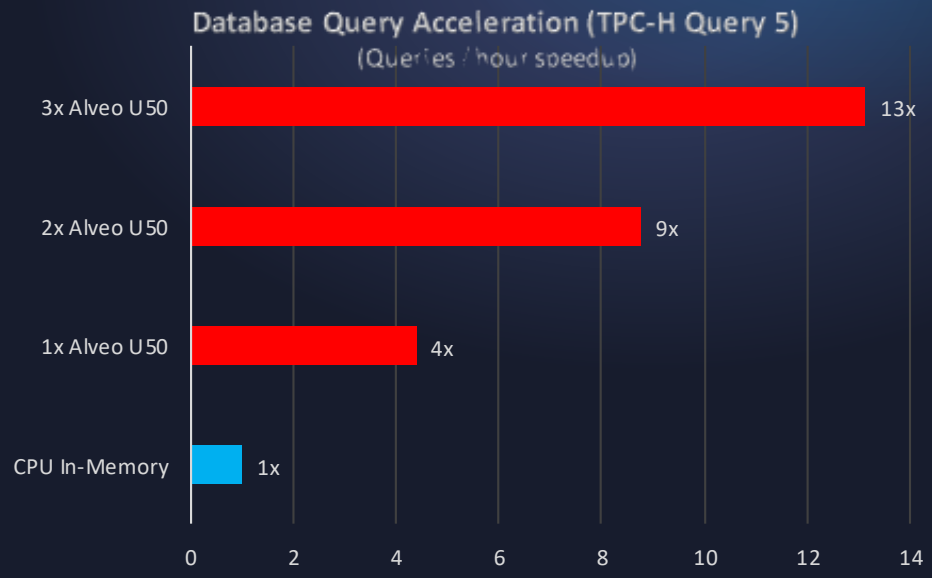


Estimated data: Alveo U50 (B=2, L=8), Tesla T4 (B=8, L=8)
https://medium.com/syncedreview/tsinghua-university-publishes-comprehensive-machine-translation-reading-list-c3f2df594218

Database Analytics

High Throughput Query Acceleration

Higher Query throughput and simplified infrastructure

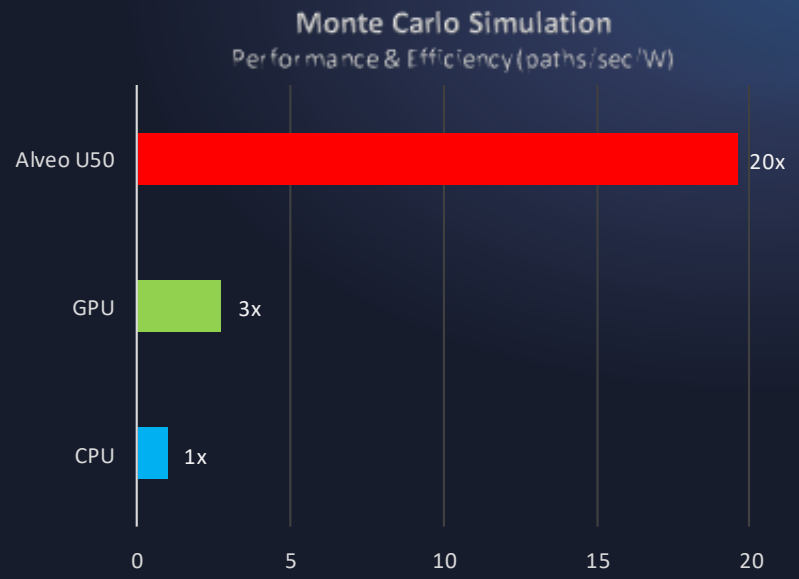


INTEL® XEON® PLATINUM 8260 PROCESSOR (35.75M Cache, 2.40 GHz) 24 core
CPU Query time = 210ms, 34k query/hr. Alveo U50=24ms, 150k query/hr

Financial Market Modeling

Hyper Efficient Derivative Pricing and Risk Modeling

Faster, more efficient time to insight at a fraction of the cost

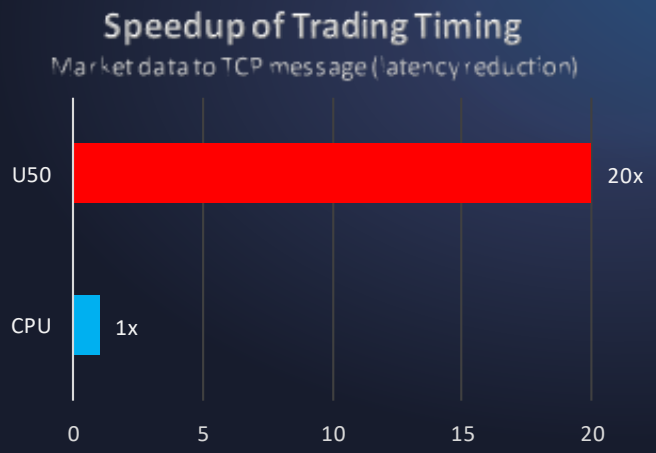


Intel Xeon E5-2697 v4 GCC 5.4.0, Nvidia Tesla V100 16GB PCIe CUDA 10.1 / GCC 5.4.0, Xilinx Alveo U50 SDAccel 2018.3

Electronic Trade Operations

Ultra low-latency Networking and Compute Acceleration

Under 500ns latency with deterministic performance



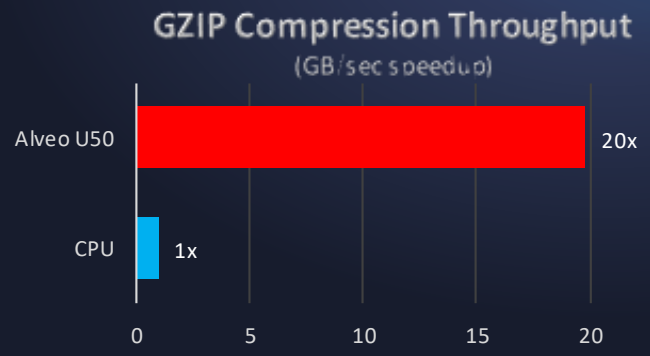
T2T latency is <0.5usec. Measured from Start of Packet in on Tick (Market Data) to Start of Packet Out on the Order

Source: Xilinx Analysis

Computational Storage

Line-rate Data Compression Acceleration

Compression, decompression, erasure coding, encryption all accelerated on one platform



Intel Skylake-SP 6152 @2.10GHz 22-core CPU (Ubuntu 16.04), GB/s compression per CPU core = .0229. Alveo U50 = 10GB/s

Source: Xilinx Analysis

Hadoop Acceleration

Line-rate Data Compression Acceleration

20x Throughput Per Node

2x Less Nodes

40% Lower Total Cost

Alveo U50 Acceleration

2x Dual CPU Servers

192TB, 1GB/sec Per Node Compression Throughput

Alveo Server with 2x Alveo U50

96TB (192TB effective), 20GB/sec Per Node Compression Throughput

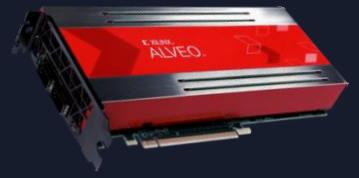
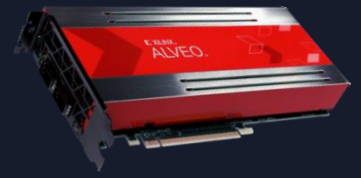
Intel Skylake-SP 6152 @2.10GHz CPU (Ubuntu 16.04), GB/s compression per CPU core = .0229. Alveo U50 = 10GB/s, Assume 2:1 compression

Key Takeaways

Expanded Market Opportunity	• First adaptable accelerator for compute, networking & storage - built for any server, any cloud
Performance & TCO Advantages	• 10-20x improvements in throughput, latency and power efficiency • First PCIe Gen 4 support with HBM2 & 100Gbps network ports
Simplified Programming & Deployment	• Growing ecosystem of software partners & accelerated solutions • Enhanced developer tools & deployment stacks for scale

APPENDIX

Xilinx Alveo Product Lineup

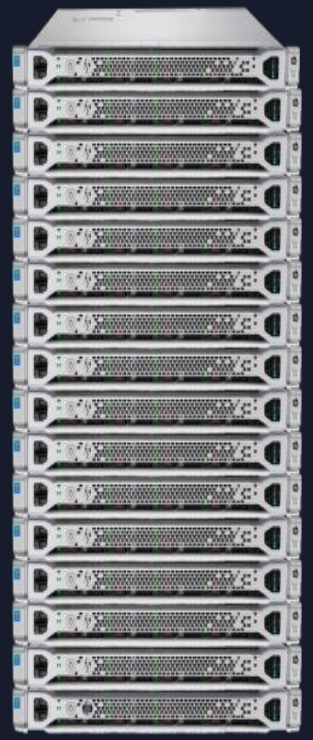
ALVEO™ U50	ALVEO™ U200	ALVEO™ U250	ALVEO™ U280
			
UltraScale+ Architecture	UltraScale+ Architecture	UltraScale+ Architecture	UltraScale+ Architecture
872k LUTs	1,182k LUTs	1,728k LUTs	1,304k LUTs
Single slot, half height	Dual slot, full height	Dual slot, full height	Dual slot, full height
8GB HBM2, 460GB/sec	64GB DDR, 77GB/sec	64GB DDR, 77GB/sec	8GB HBM2, 460GB/sec
PCIe Gen3, Gen4, CCIX	PCIe Gen3	PCIe Gen3	PCIe Gen3, Gen4, CCIX
1x QSFP 28 (100GbE)	2x QSFP 28 (100GbE)	2x QSFP 28 (100GbE)	2x QSFP 28 (100GbE)
< 75W	< 225W	< 225W	< 225W

One Platform. Broadest Acceleration

	CPU (Sequential)	GPU (Parallel)	Alveo (Sequential + Parallel)
			
3rd Party Applications	●	●	●
High Level Coding	●	●	●
Complex Memory & Datapath			●
Adaptable Hardware			●
AI Inference + Pre/Post Process			●
On-board Networking			●

Live Video Transcoding

Simplified and Lower Cost Infrastructure



Alveo U50 HEVC Video Compression

20x Throughput Per Node

8x Lower HW Cost

23x Lower Power Cost

20x Dual CPU Servers

40x Xeon Gold

H.265 very-high quality

20x 1080p30

One Alveo U50 Server

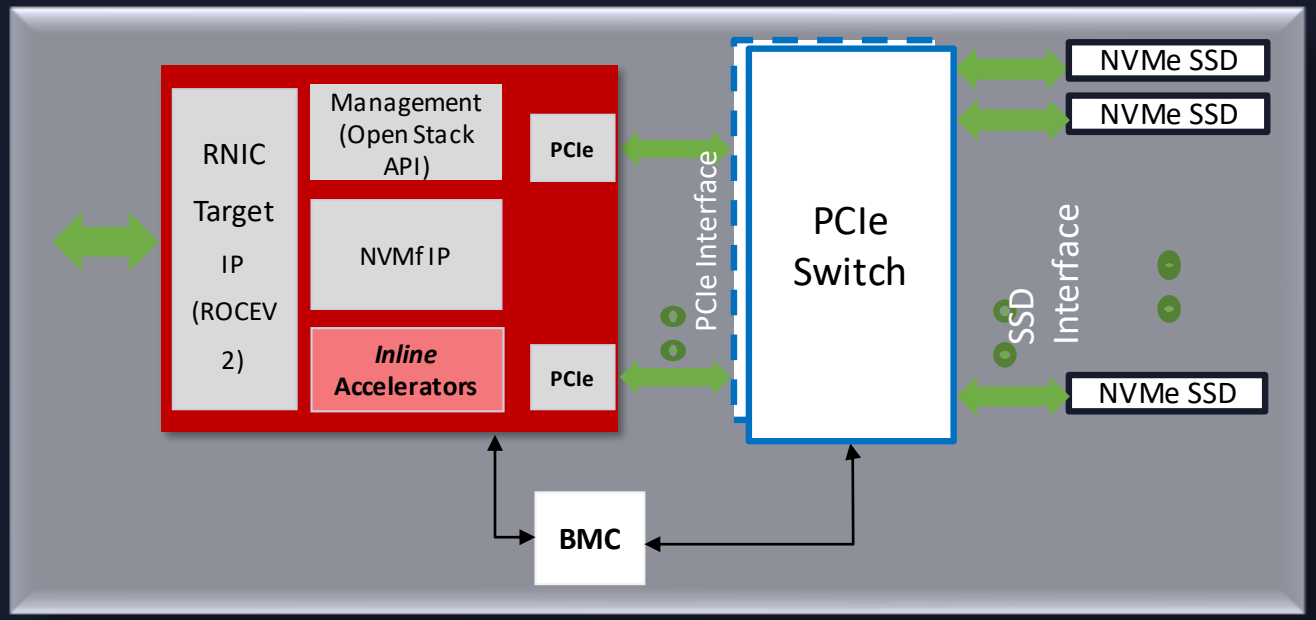
5x Alveo U50

NGCodec HEVC Very-High Quality

20x 1080p30

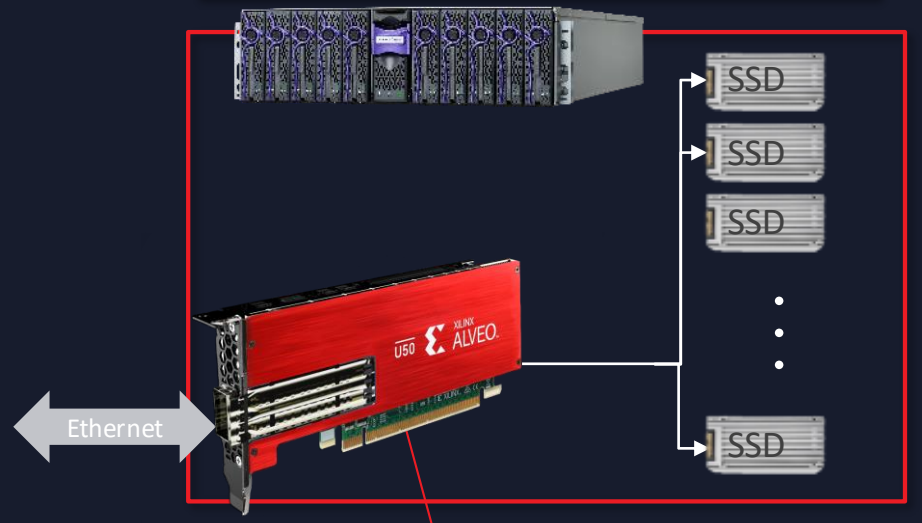
Fabric Attached Computational Storage

Bringing acceleration to NVMeoF



- Fabric connected accelerator fronts SSDs – brings the compute to the data.
- NVMeoF target offloaded to U50 supporting 2.5 Million IOPS.
- Only 1 *micro*second added latency to programable inline storage accelerators.

Technology Preview: Fabric Attached Computational Storage Array



Inline Accelerator Examples:

Storage services:

- (De)Compression
- (De)Encryption
- Data protection

Database Acceleration:

- Scan
- Filter
- Aggregate

Alveo Lineup – Detailed Specs

NEW

	Product Name	Alveo U200	Alveo U250	Alveo U280	Alveo U50
Dimensions	Width	Dual Slot	Dual Slot	Dual Slot	**Single Slot**
	Form Factor, Passive Form Factor, Active	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length	**Half Height, ½ Length**
Logic [1]	Look-Up Tables	1,182K	1,728K	1,304K	**872K**
	Registers	2,364K	3,456K	2,607K	**1,743K**
	DSP Slices	6,840	12,288	9,024	**5,952**
DRAM Memory	DDR Format	4x 16GB 72b DIMM DDR4	4x 16GB 72b DIMM DDR4	2x 16GB 72b DIMM DDR4	**–**
	DDR Total Capacity	64GB	64GB	32GB	**–**
	DDR Max Data Rate	2400MT/s	2400MT/s	2400MT/s	**–**
	DDR Total Bandwidth	77GB/s	77GB/s	38GB/s	**–**
	HBM2 Total Capacity	–	–	8GB	**8GB**
	HBM2 Total Bandwidth	–	–	460GB/s	**460GB/s**
Internal SRAM	Total Capacity	43MB	57MB	43MB	**28MB**
	Total Bandwidth	37TB/s	47TB/s	35TB/s	**24TB/s**
Interfaces	PCI Express®	Gen3 x16	Gen3 x16	Gen3 x16, 2x Gen4 x8, CCIX	**Gen3 x16, 2x Gen4 x8, CCIX**
	Network Interface	2x QSFP28	2x QSFP28	2x QSFP28	**U50[2] - 1x QSFP28 U50DD[3] - 2x SFP-DD**
Power and Thermal	Thermal Cooling	Passive, Active	Passive, Active	Passive, Active	**Passive**
	Typical Power	100W	110W	100W	**50W**
	Maximum Power	225W	225W	225W	**75W**
Time Stamp	Clock Precision	–	–	–	**IEEE Std 1588**