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
ACCESSIBLE
Deploy in the cloud or on-premises
Rich set of accelerated Applications

ADAPTABLE
Deploy optimized domain-specific architectures
Adapt to changing algorithms

FAST
Built for high throughput, ultra-low latency
Accelerate compute, networking, storage
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

1) Custom Data Path
2) Custom Precision
3) Custom Memory Hierarchy

Off-Chip DDR

On-Chip Memory

Memory
# ALVEO Solution Stack

The ALVEO Solution Stack consists of various layers, each representing different types of stakeholders and solutions.

## End Customers
- Data Analytics
- Video & Image Processing
- Machine Learning
- Life Science & HPC
- Financial Computing

## Solution Providers
- Titan
- Videologic
- DEEPLY
- HPC Software
- Mipsology
- Blackfin
- PLUNIFY
- VITESSE DATA

## App & IP Developers
- Skreens
- Swarm64
- NEXTRA
- Bigstream
- FENIAC
- Inaccel
- BigZette Systems
- Mle

## Channel Partners
- AWS
- Alibaba Cloud
- Tencent Cloud
- Nimbix
- Inspur
- Dell

## Hardware

- CLOUD
- ON-PREMISE

@ Copyright 2019 Xilinx
Growing Cloud Availability

Xilinx Powered Hyperscale Cloud Data Centers
Growing Ecosystem

Data Analytics
- bigstream
- VITESSE DATA
- TENIAC

Life Sciences & HPC
- illumina
- byte LAKE
- inacel
- Falcon Computing
- MAXEKER
- delab Genetics

Video Processing
- NGCODEC
- V-NOVA
- Xilix

Machine Learning
- Mipsology
- DEEPi

Financial Computing
- HEDD ONE
- MAXEKER
- SciComp

Image Processing
- DEEPLY
- CTACCel
Growth Since October 2018 Launch

- Published Applications: 2x
- Developers Trained: 4x

Applications:
- Data Analytics
  - bigstream
  - BlackTrax
  - swarm64
  - XELERA
  - BigZette Systems
- Life Sciences & HPC
  - illumina
  - byte LAKE
  - inaccel
  - VITESSE D
  - TENSOR
- Video Processing
  - Nextera Video
  - skrerns
  - VPSync
- Machine Learning
  - Mipsology
  - DeePi
  - XELERA
  - HELD MODE
- Financial Computing
  - MAXELER
  - SciComp
- Image Processing
  - DEEPOLY
  - CTRACCLE
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-list-of-machine-translation-reading-list-c3f2df594218

@ Copyright 2019 Xilinx
Database Analytics
High Throughput Query Acceleration

Higher Query throughput and simplified infrastructure

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Query Acceleration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x CPU In-Memory</td>
<td>1x</td>
</tr>
<tr>
<td>1x Alveo U50</td>
<td>4x</td>
</tr>
<tr>
<td>2x Alveo U50</td>
<td>9x</td>
</tr>
<tr>
<td>3x Alveo U50</td>
<td>13x</td>
</tr>
</tbody>
</table>

INTEL® XEON® PLATINUM 8160 PROCESSOR (35.75M Cache, 2.40 GHz) 24 core
CPU Query time = 210ms, 34k query/hr, Alveo U50=24ms, 150k query/hr

@ Copyright 2019 Xilinx
Financial Market Modeling
Hyper Efficient Derivative Pricing and Risk Modeling

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

Monte Carlo Simulation
Performance & Efficiency (paths/sec x100)

- CPU: 1x
- GPU: 3x
- Alveo U50: 20x

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

Source: Xilinx Analysis

T2T latency is <0.5usec. Measured from Start of Packet in on Tick (Market Data) to Start of Packet Out on the Order
Computational Storage
Line-rate Data Compression Acceleration

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

Source: Xilinx Analysis

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

@ Copyright 2019 Xilinx
## Xilinx Alveo Product Lineup

<table>
<thead>
<tr>
<th>Model</th>
<th>LUTs</th>
<th>Slot Type</th>
<th>Memory</th>
<th>PCIe</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>U50</td>
<td>872k LUTs</td>
<td>Single slot, half height</td>
<td>8GB HBM2, 460GB/sec</td>
<td>PCIe Gen3, Gen4, CCIX</td>
<td>&lt; 75W</td>
</tr>
<tr>
<td>U200</td>
<td>1,182k LUTs</td>
<td>Dual slot, full height</td>
<td>64GB DDR, 77GB/sec</td>
<td>PCIe Gen3</td>
<td>&lt; 225W</td>
</tr>
<tr>
<td>U250</td>
<td>1,728k LUTs</td>
<td>Dual slot, full height</td>
<td>64GB DDR, 77GB/sec</td>
<td>PCIe Gen3</td>
<td>&lt; 225W</td>
</tr>
<tr>
<td>U280</td>
<td>1,304k LUTs</td>
<td>Dual slot, full height</td>
<td>8GB HBM2, 460GB/sec</td>
<td>PCIe Gen3, Gen4, CCIX</td>
<td>&lt; 225W</td>
</tr>
</tbody>
</table>
## One Platform. Broadest Acceleration

<table>
<thead>
<tr>
<th></th>
<th>CPU (Sequential)</th>
<th>GPU (Parallel)</th>
<th>Alveo (Sequential + Parallel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd Party Applications</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>High Level Coding</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Complex Memory &amp; Datapath</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Adaptable Hardware</td>
<td></td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>AI Inference + Pre/Post Process</td>
<td>●</td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>On-board Networking</td>
<td></td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>
Live Video Transcoding

Simplified and Lower Cost Infrastructure

- 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

Alveo U50 HEVC Video Compression
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 microsecond added latency to programmable inline storage accelerators.

Inline Accelerator Examples:
Storage services:
• (De)Compression
• (De)Encryption
• Data protection

Database Acceleration:
• Scan
• Filter
• Aggregate
# Alveo Lineup – Detailed Specs

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Alveo U200</th>
<th>Alveo U250</th>
<th>Alveo U280</th>
<th>Alveo U50</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimensions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width</td>
<td>Dual Slot</td>
<td>Dual Slot</td>
<td>Dual Slot</td>
<td>Single Slot</td>
</tr>
<tr>
<td>Form Factor, Passive Form Factor, Active</td>
<td>Full Height, ¾ Length Full Height, Full Length</td>
<td>Full Height, ¾ Length Full Height, Full Length</td>
<td>Full Height, ¾ Length Full Height, Full Length</td>
<td>Half Height, ½ Length</td>
</tr>
<tr>
<td><strong>Logic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Look-Up Tables</td>
<td>1,182K</td>
<td>1,728K</td>
<td>1,304K</td>
<td>872K</td>
</tr>
<tr>
<td>Registers</td>
<td>2,364K</td>
<td>3,456K</td>
<td>2,607K</td>
<td>1,743K</td>
</tr>
<tr>
<td>DSP Slices</td>
<td>6,840</td>
<td>12,288</td>
<td>9,024</td>
<td>5,952</td>
</tr>
<tr>
<td><strong>DRAM Memory</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDR Format</td>
<td>4x 16GB 72b DIMM DDR4</td>
<td>4x 16GB 72b DIMM DDR4</td>
<td>2x 16GB 72b DIMM DDR4</td>
<td>–</td>
</tr>
<tr>
<td>DDR Total Capacity</td>
<td>64GB</td>
<td>64GB</td>
<td>32GB</td>
<td>–</td>
</tr>
<tr>
<td>DDR Max Data Rate</td>
<td>2400MT/s</td>
<td>2400MT/s</td>
<td>2400MT/s</td>
<td>–</td>
</tr>
<tr>
<td>DDR Total Bandwidth</td>
<td>77GB/s</td>
<td>77GB/s</td>
<td>38GB/s</td>
<td>–</td>
</tr>
<tr>
<td>HBM2 Total Capacity</td>
<td>–</td>
<td>–</td>
<td>8GB</td>
<td>8GB</td>
</tr>
<tr>
<td>HBM2 Total Bandwidth</td>
<td>–</td>
<td>–</td>
<td>460GB/s</td>
<td>460GB/s</td>
</tr>
<tr>
<td><strong>Internal SRAM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Capacity</td>
<td>43MB</td>
<td>57MB</td>
<td>43MB</td>
<td>28MB</td>
</tr>
<tr>
<td>Total Bandwidth</td>
<td>37TB/s</td>
<td>47TB/s</td>
<td>35TB/s</td>
<td>24TB/s</td>
</tr>
<tr>
<td><strong>Interfaces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCI Express®</td>
<td>Gen3 x16</td>
<td>Gen3 x16</td>
<td>Gen3 x16, 2x Gen4 x8, CCIX</td>
<td>Gen3 x16, 2x Gen4 x8, CCIX</td>
</tr>
<tr>
<td>Network Interface</td>
<td>2x QSFP28</td>
<td>2x QSFP28</td>
<td>2x QSFP28</td>
<td>U502 - 1x QSFP28 U50DD3 - 2x SFP-DD</td>
</tr>
<tr>
<td><strong>Power and Thermal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermal Cooling</td>
<td>Passive, Active</td>
<td>Passive, Active</td>
<td>Passive, Active</td>
<td>Passive</td>
</tr>
<tr>
<td>Typical Power</td>
<td>100W</td>
<td>110W</td>
<td>100W</td>
<td>50W</td>
</tr>
<tr>
<td>Maximum Power</td>
<td>225W</td>
<td>225W</td>
<td>225W</td>
<td>75W</td>
</tr>
<tr>
<td><strong>Time Stamp</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clock Precision</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>IEEE Std 1588</td>
</tr>
</tbody>
</table>