



Xilinx Expands Alveo Portfolio with Industry's First Adaptable Compute, Network and Storage Accelerator Card Built for Any Server, Any Cloud

August 6, 2019

First low-profile PCIe Gen 4 card delivers dramatic improvements in throughput, latency and power efficiency for critical data center workloads

SAN JOSE, Calif., Aug. 6, 2019 /PRNewswire/ -- Xilinx, Inc. (NASDAQ: XLNX), the leader in adaptive and intelligent computing, today expanded its Alveo data center accelerator card portfolio with the launch of the Alveo™ U50. The U50 card is the industry's first low profile adaptable accelerator with PCIe Gen 4 support, uniquely designed to supercharge a broad range of critical compute, network and storage workloads, all on one reconfigurable platform.



The Alveo U50 provides customers with a programmable low profile and low-power accelerator platform built for scale-out architectures and domain-specific acceleration of any server deployment, on-premise, in the cloud and at the edge. To meet the challenges of emerging dynamic workloads such as cloud microservices, Alveo U50 delivers between 10-20x improvements in throughput, latency and power efficiency. For accelerated networking and storage workloads, the U50 card helps developers identify and eliminate latency and data movement bottlenecks by moving compute closer to the data.

Powered by the Xilinx® UltraScale+™ architecture, the Alveo U50 card is the first in the Alveo portfolio to be packaged in a half-height, half-length form factor and low 75-Watt power envelope. The card features high-bandwidth memory (HBM2), 100 gigabit per second (100 Gbps) networking connectivity, and support for the PCIe Gen 4 and CCIX interconnects. By fitting into standard PCIe server slots and using one-third the power, the Alveo U50 significantly expands the scope in which adaptable acceleration can be deployed to unlock dramatic throughput and latency improvements for demanding compute, network and storage workloads. The 8GB of HBM2 delivers over 400 Gbps data transfer speeds and the QSFP ports provide up to 100 Gbps network connectivity. The high-speed networking I/O also supports advanced applications like NVMe-oF™ solutions (NVM Express over Fabrics™), disaggregated computational storage and specialized financial services applications.

From machine learning inference, video transcoding and data analytics to computational storage, electronic trading and financial risk modeling, the Alveo U50 brings programmability, flexibility, and high throughput and low latency performance advantages to any server deployment. Unlike fixed architecture alternatives, the software and hardware programmability of the Alveo U50 allows customers to meet ever-changing demands and optimize application performance as workloads and algorithms continue to evolve.

Alveo U50 accelerated solutions deliver significant customer value across a range of applications, including:

- Deep learning inference acceleration (speech translation): delivers up to 25x lower latency, 10x higher throughput and significantly improved power efficiency per node compared to GPU-only for speech translation performance¹;
- Data analytics acceleration (database query): running the TPC-H Query benchmark, Alveo U50 delivers 4x higher throughput per hour and reduced operational costs by 3x compared to in-memory CPU²;
- Computational storage acceleration (compression): delivers 20x more compression/decompression throughput, faster Hadoop and big data analytics, and over 30 percent lower cost per node compared to CPU-only nodes³;
- Network acceleration (electronic trading): delivers 20x lower latency and sub-500ns trading time compared to CPU-only latency of 10us⁴;
- Financial modeling (grid computing): running the Monte Carlo simulation, Alveo U50 delivers 7x greater power efficiency compared to GPU-only performance⁵ for a faster time to insight, deterministic latency and reduced operational costs.

"Ever-growing demands on the data center are pushing existing infrastructure to its limit, driving the need for adaptable solutions that can optimize performance across a broad range of workloads and extend the lifecycle of existing infrastructure, ultimately reducing TCO," said Salil Rajee, executive vice president and general manager, Data Center Group, at Xilinx. "The new Alveo U50 brings an optimized form factor and unprecedented

performance and adaptability to data center workloads, and we continue to build out solution stacks with a growing ecosystem of application partners to deliver previously unthinkable capabilities to a range of industries."

Industry Support

"The forthcoming 2nd Gen AMD EPYC processor is ideally suited for data center-first accelerators like the Alveo U50 that combine compute, network and storage acceleration all on the same platform," said Raghu Nambiar, vice president & CTO of application engineering at AMD. "Taking advantage of AMD's leadership, first x86 server-class PCIe 4.0 CPU, the Alveo U50 will be the industry's first adaptable accelerator card with PCIe 4.0 support. We look forward to working with Xilinx to combine the benefits of AMD EPYC based solutions with Alveo acceleration to hyperscale and enterprise customers."

"IBM is excited about the expansion of the Xilinx Alveo portfolio with the addition of the Alveo U50 adaptable accelerator card," said Steve Fields, Chief Architect for IBM Power Systems. "We believe the combination of low-profile form-factor, HBM2 memory performance, and PCIe Gen 4 speed to interface with IBM Power processors will enable the OpenPOWER ecosystem to provide cutting edge adaptable acceleration solutions."

"With the smaller design and advanced features of the Alveo U50, Xilinx is well positioned to expand the markets for acceleration with configurable logic," said Karl Freund, senior analyst, HPC and deep learning, Moor Insights & Strategy. "The new Alveo U50 should allow them to break through the market noise with demonstrated and dramatic performance advantages in high-growth use cases."

"We are excited to be collaborating with Xilinx at FMS, showcasing the flexibility and performance of the Alveo U50 and our OpenFlex composable NVMe-oF platform," said Scott Hamilton, senior director of product management, Data Center Systems business unit at Western Digital. "Xilinx is leading the charge in fabric-based computational storage using NVMe-oF to enable full disaggregation of server resources. We believe the new Alveo U50 will be an important part of the ecosystem as organizations take a truly disaggregated approach to SDS infrastructure."

Availability:

The Alveo U50 is sampling now with OEM system qualifications in process. General availability is slated for fall 2019.

Flash Memory Summit:

Xilinx will be showcasing the Alveo U50 and other product demonstrations in booth 313 at Flash Memory Summit (FMS) 2019, taking place August 6-8 at the Santa Clara Convention Center in Santa Clara, Calif.

Additionally, Salil Raje, executive vice president and general manager, Data Center Group, at Xilinx, will be giving a keynote titled, "FPGAs: The Key to Accelerating High-Speed Storage Systems" on August 7 at 2:40 p.m. PT in the Mission City Ballroom.

About Xilinx

Xilinx develops highly flexible and adaptive processing platforms that enable rapid innovation across a variety of technologies – from the endpoint to the edge to the cloud. Xilinx is the inventor of the FPGA, hardware programmable SoCs, and the ACAP, designed to deliver the most dynamic processor technology in the industry and enable the adaptable, intelligent and connected world of the future. For more information, visit www.xilinx.com.

Footnotes:

1. Performance of Alveo U50, with both Alveo U50 and Nvidia Tesla T4 running (B=2, L=8), Tesla T4 (B=8, L=8) (estimated data)
2. Alveo U50=24ms, 150k query/hr / CPU Query time = 210ms, 34k query/hr. based on Intel Xeon Platinum 8260 Processor (35.75M Cache, 2.40 GHz) 24 core
3. Intel Skylake-SP 6152 @2.10GHz CPU (Ubuntu 16.04) CPU Query time = 210ms, 34k query/hr. Alveo U50=24ms, 150k query/hr Xilinx Alveo U50 SDAccel 2018.3 (estimate) GB/s compression per CPU core = .0229. Alveo U50 = 10GB/s (estimate)
4. Alveo U50 latency is <0.5us, CPU latency is 10us. Measured from start of packet in on Tick (Market Data) to start of packet out on the order to Start Packet Out on the Order (estimate)
5. Intel Xeon E5-2697 v4 GCC 5.4.0 Nvidia Tesla V100 16GB PCIe CUDA 10.1 / GCC 5.4.0 Intel Skylake-SP 6152 @2.10GHz CPU (Ubuntu 16.04) CPU Query time = 210ms, 34k query/hr. Alveo U50=24ms, 150k query/hr Xilinx Alveo U50 SDAccel 2018.3 (estimated data).

© Copyright 2019 Xilinx, Inc. Xilinx, the Xilinx logo, and other designated brands included herein are trademarks of Xilinx in the United States and other countries. NVMe-oF and NVM Express over Fabrics are trademarks of NVM Express, Inc. PCI, PCIe and PCI Express are trademarks of PCI-SIG and used under license. All other trademarks are the property of their respective owners.

PR Contact:

Brian Garabedian
media@xilinx.com



View original content to download multimedia:<http://www.prnewswire.com/news-releases/xilinx-expands-alveo-portfolio-with-industrys-first-adaptable-compute-network-and-storage-accelerator-card-built-for-any-server-any-cloud-300896576.html>

SOURCE Xilinx, Inc.