



AMD ALVEO™ U200 | U250

Adaptable Accelerator Cards for Data Center Workloads

OVERVIEW

AMD Alveo™ Data Center accelerator cards are designed to meet the constantly changing needs of the modern Data Center, providing up to 90X performance increase over CPUs for common workloads, including machine learning inference, video transcoding, and database search and analytics.

With complex algorithms evolving faster than silicon design cycles, it's impossible for fixed function GPU and ASIC devices to keep pace. Built on AMD 16nm UltraScale™ architecture, Alveo U200 accelerator cards provide reconfigurable acceleration that can adapt to continual algorithm optimizations, supporting any workload type while reducing overall cost of ownership.

Enabling Alveo accelerator cards is a growing ecosystem of AMD and partner applications for common Data Center workloads. For custom solutions, AMD's Application Developer tool (**Vitis™**) and **Machine Learning Suite** provide the tools for developers to bring differentiated **applications** to market.

HIGHLIGHTS

Fast – Highest Performance

- > Up to 90X higher performance than CPUs¹ on key workloads at 1/3 the cost²
- > Over 4X higher inference throughput³ and 3X latency advantage over PU-based solutions⁴

Adaptable – Accelerate Any Workload

- > Machine learning inference to video processing to any workload using the same accelerator card
- > As workloads algorithms evolve, use reconfigurable hardware to adapt faster than fixed-function accelerator card product cycles

Accessible – Cloud On-Premises Mobility

- > Deploy solutions in the cloud or on-premises interchangeably, scalable to application requirements
- > **Applications available** for common workloads, or build your own with the **application developer tool**

1: BlackLynx Elasticsearch on Alveo versus EC2 c4.8xlarge
 2: Based on CapEx & OpEx savings for DNN inference on Alveo vs dual-socket Intel Xeon Platinum servers
 3: **Accelerating DNNs with Alveo Accelerator Cards White Paper**
 4: Measured on CNN+BLSTM Speech-to-Text ML inference against Nvidia P4



ADAPTABLE TO ANY WORKLOAD

- > Database Search and Analytics
- > Financial Computing
- > Machine Learning
- > Storage Compression
- > Video Processing/Transcoding
- > Genomics

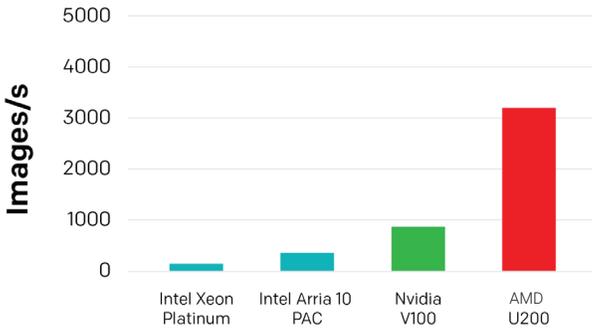
BENCHMARKS

Adapt and Accelerate Any Workload

Area	Partner Workload	AMD Alveo Acceleration vs CPU
Database Search and Analytics	BlackLynx Unstructured Data Elasticsearch	90X
Financial Computing	Maxeler Value-at-Risk (VAR) Calculation	89X
Machine Learning	Xilinx Real-Time Machine Learning Inference	20X
Video Processing / Transcoding	NGCodec HEVC Video Encoding	12X
Genomics	Falcon Computing Genome Sequencing	10X

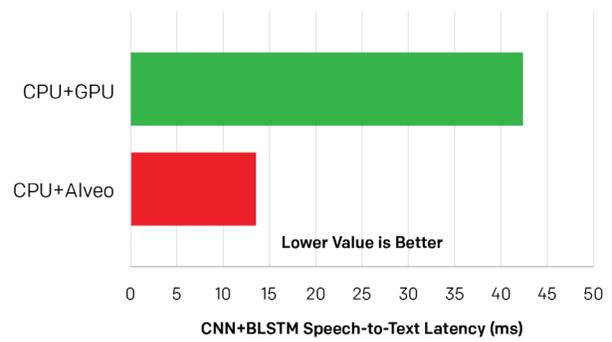
CPU Comparisons: Xeon c4.8xlarge AWS | Xeon E5-2643 v4 3.4GHz | Xeon Platinum c5.18xlarge AWS | Dual Socket E5-2680 v3 2.5GHz | Xeon f1.16xlarge

Increase Real-Time Machine Learning* Throughput by 20X



*GoogLeNet V1 : Accelerating DNNs with Xilinx Alveo Accelerator Cards White Paper

Reduce ML Inference Latency by 3X



CPU+GPU: Nvidia P4 + Xeon CPU E5-2690 v4 @2.60GHz (56 Cores)
 CPU+Alveo: Alveo U200 or U250 + Xeon CPU E5-2686 v4 @2.3GHz (8 Cores)

Features	AMD Alveo™ u200 Accelerator Cards
Peak INT8 TOPs	18.6
DDR Memory Bandwidth	77GB/s
Internal SRAM Bandwidth	31TB/s
Look-up Tables (LUTs)	892,000
Thermal Options	Passive or Active

TAKE THE NEXT STEP

Visit www.xilinx.com/alveo for more information. Buy AMD Alveo™ through **AMD Accelerated Computing Partners**.

DISCLAIMERS

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale.

© Copyright 2023 Xilinx, Inc. Xilinx, the Xilinx logo, Artix, ISE, Kintex, Spartan, Virtex, Vivado, Zynq, and other designated brands included herein are trademarks of Xilinx in the United States and other countries. AMBA, AMBA Designer, ARM, ARM1176JZ-S, CoreSight, Cortex, and PrimeCell are trademarks of ARM in the EU and other countries. PCIe, and PCI Express are trademarks of PCI-SIG and used under license. All other trademarks are the property of their respective owners. Printed in the U.S.A. AC0219A