WP499 (v1.0) November 19, 2018

Breathe New Life into Your Data Center with Alveo Adaptable Accelerator Cards

Alveo Data Center accelerator cards with their ready to go applications deliver a much-needed increase in compute capability, at lowest TCO, for the broadest range of workloads.

ABSTRACT

It is well documented that the staple of the data center—the CPU—has run out of steam.

The increase in performance offered by next-generation CPUs is minimal, but demand for compute resources, across a large array of workloads and applications, is expanding exponentially. Adaptable compute accelerators are clearly needed to meet broad compute demand and to manage scale out and operating costs.

Xilinx® Alveo™ Data Center accelerator cards, with a broad array of ready to go applications, offer rapid production deployment of adaptable accelerators, resulting in a dramatic increase in compute capability and TCO savings of up to 65%.

© Copyright 2018 Xilinx, Inc. Xilinx, the Xilinx logo, Alveo, Artix, ISE, Kintex, Spartan, Versal, Virtex, Vivado, Zynq, and other designated brands included herein are trademarks of Xilinx in the United States and other countries. PCI, PCIe, and PCI Express are trademarks of PCI-SIG and used under license. All other trademarks are the property of their respective owners.

Introduction

Intelligent leverage of massive compute capability and vast datasets—e.g., big data analytics and machine learning—has become a key differentiator in almost every industry, driving broad-based demand for higher performance, more cost-effective compute solutions in the Data Center.

Meeting demand using only CPUs (or multicore CPUs) is not viable, due to cost, power consumption, and scale of the CPU-only solution. Furthermore, for many workloads, throwing more CPU-based servers at the problem, simply will not deliver the required performance. As Moore's law grinds to a halt, next-generation CPUs offer little in the way of hope. Adaptable compute accelerators are clearly needed to meet broad compute demand and to manage scale out and operating costs.

Fast, Accessible, and Adaptable

To address the growing gap between compute demand and the capability of CPU-only Data Centers, Xilinx has developed the Alveo portfolio of PCIe® form factor, production-ready, adaptable Data Center accelerator cards. Figure 1 shows an illustration of the passively cooled variant of Alveo U200 and U250 accelerator cards. More details on the Alveo portfolio are available on Xilinx's Alveo website.

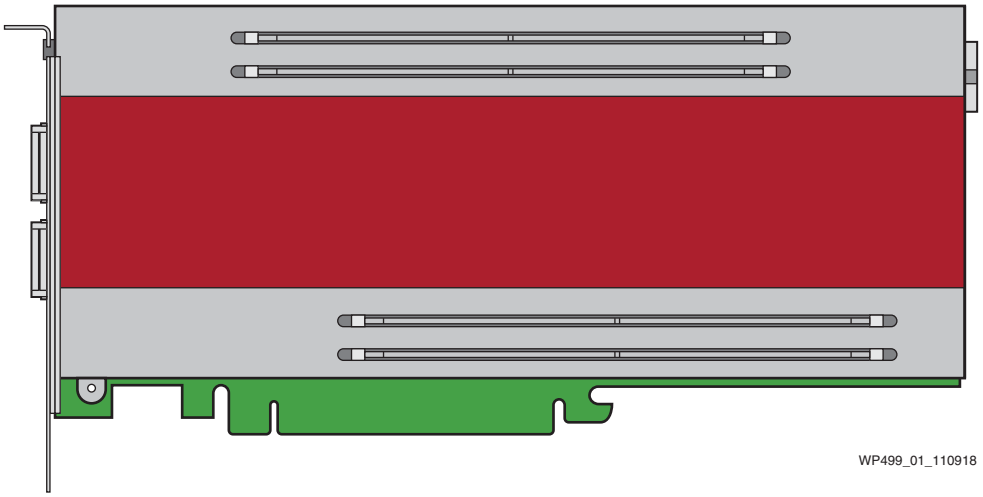


Figure 1: **Alveo U200 and U250 Data Center Accelerator Cards with Passive Cooling**

The Alveo Data Center accelerator card leverages the benefits of Xilinx devices (More details on the benefits of Xilinx devices are detailed in Xilinx white paper WP492, *Xilinx All Programmable Devices: a Superior Platform for Compute-Intensive Systems*) to deliver a solution for a broad array of key workloads that is:

- Fast - Delivers the highest throughput and lowest latency
- Adaptable - Meet the existing and evolving needs of many workloads
- Accessible - Offers ready-to-go applications for on-premise or cloud-based deployment

These benefits enable the Alveo cards to deliver unparalleled value, including:

- Application breadth - e.g., Machine learning inference, database search and analytics, video, financial computing, and genomics
- Best performance and efficiency vs. CPU/GPU - of up to 90X vs. CPU and 5X vs. GPU
- Lowest TCO - Up to 65% TCO savings
- Future-proof solution - Alveo cards adapt as workloads evolve

The Alveo U200 and U250 accelerator cards are available for order today with the U280 accelerator card coming soon. Future Alveo cards will incorporate the recently announced Versal™ adaptive compute acceleration platform (ACAP) devices for even greater levels of compute performance and efficiency. For instance, the AI Engines in Versal devices will increase ML inference performance by 20X. More details on Versal is available on Xilinx's Versal ACAP website and WP505, Xilinx *Versal: The First Adaptive Compute Acceleration Platform (ACAP)*.

Fast

Alveo accelerator cards are unique in delivering both the highest throughput and lowest latency, for a broad array of Data Center workloads. CPUs and GPUs typically must trade-off latency versus throughput due to the limitations of their rigid architectures and software driven dataflow. Figure 2 details some of the acceleration benefits of Alveo cards relative to CPUs for a range of key data center workloads. More details on the fast benefits of Alveo relative to CPUs and GPUs, will be described later in this paper.

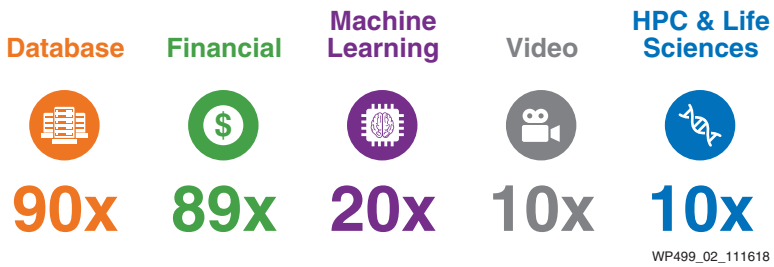


Figure 2: **Fast - Acceleration Offered by Alveo Data Center Acceleration Cards**

Adaptable

Alveo accelerator cards extend Xilinx's legacy of delivering products with unrivaled adaptability. Unlike CPUs and GPUs, Alveo cards offer hardware adaptability, meaning that the underlining hardware can be configured (and reconfigured) to map to any workload. CPUs and GPUs only offer flexibility in the software domain, with their underlining hardware remaining fixed. Therefore, if workloads do not map well to the underlining hardware architecture, only poor performance is achieved. Some examples of the impact lack of hardware adaptability has will be described later in this white paper. Alveo accelerator cards' hardware adaptability delivers:

- Highest Data Center hardware utilization - Accelerate the broadest array of application with less hardware
- Future-proof hardware - Adapts to meet evolving workloads without need for new HW (as illustrated in Figure 3).

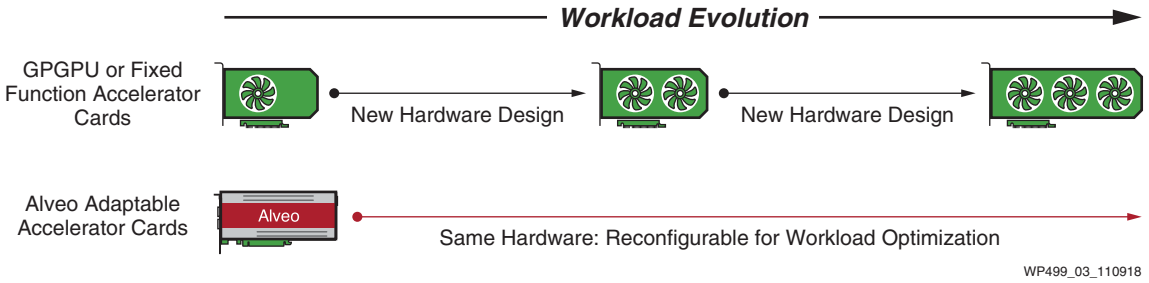


Figure 3: **Alveo Accelerator Cards' Adaptability Means Fewer Hardware Refreshes**

For developers who want to build and adapt their own applications targeted at Alveo accelerator cards, Xilinx has a comprehensive and easy-to-use development environment called the SDAccel™ development environment.

Accessible

To enable users to immediately leverage the acceleration capabilities of the Alveo Data Center accelerator card, Xilinx and its partners, have developed an array of ready-to-go applications. These applications cover a broad range of key workloads, including machine learning inference, video transcoding, database search and analytics, genomics, and financial computing. Table 1 details a small sampling of the applications available, along with their acceleration speed up. The complete range of applications is available on Alveo accelerator card's application website.

Table 1: **Example Applications for Alveo Accelerator Cards with Achieved Speed Up**

Domain	Offering	Vendor	Card	Speed Up
Data Analytics	Search and Analytics Toolkit	Blacklynx	U200	90X
Machine Learning	Machine Learning Suite for Inference	Xilinx	U200 U250	20X
Video	Adaptive Bit Rate Video Transcoding Application	Xilinx and NGCodec	U200	10X
Genomics	Accelerated Genomics Pipelines	Falcon Computing	U200	10X
Financial Computing	Real-Time Risk Dashboard	Maxeler	U200	100X

Recognizing acceleration benefits of Xilinx devices, cloud service providers, such as AWS, Nimbix, Alibaba, Baidu, Tencent, and Huawei, have deployed Xilinx-based accelerators in their cloud as FPGA as a Service (FaaS) offerings. These cloud solutions offer customers an easy and accessible way to evaluate accelerated applications like those detailed above and to explore the benefits of Xilinx-based acceleration before deploying Alveo accelerator cards on-premises. On-premise deployment is commonly required for a variety of reasons, including for applications dealing with sensitive data or requiring the highest level of performance, as in the case of latency sensitive workloads.

More details on FaaS offerings based on Xilinx devices are available on Xilinx's accelerated cloud services website.

Data Center Workloads Examples

To explore the fast, accessible, and adaptable benefits of Alveo Data Center accelerator cards in more detail, three key workloads are analyzed:

- Machine Learning inference
- Database search and analytics
- Video

Accelerated Machine Learning Inference with Alveo

Numerous industries have realized dramatic benefits from adopting machine learning (ML) technology, e.g., medical diagnostics, video surveillance, satellite imaging, network security, and autonomous vehicles. ML itself is an extremely broad and dynamic area embodying many types of ML networks, each of which is continually evolving. In addition, ML has two unique stages:

- Training - Train a machine learning network offline to perform a specific task
- Inference - Deploy the trained network to perform the task in real-time, e.g., translation

Both training and inference are extremely compute intensive. As a result, solutions based on CPUs only are not viable as they cannot achieve the required throughput and latency in any reasonable cost or power envelope.

The GPU architecture has excelled at ML training workloads, mainly because of massive floating-point compute capability and high compute to data transfer ratios. However, GPUs have struggled to demonstrate their suitability for ML inference. There are some very critical characteristics of ML inference that differentiates it from training workloads:

- Latency is critical for inference
- Compute to data transfer ratio is much lower for inference
- Lower precision data types can be used in inference

The combination of the need for low latency with higher ratios of data transfers to compute results in ML inference being a real challenge for GPUs. Nvidia's published data shows that their flagship Tesla V100 has only 870 images/s for GoogleNetv1 at Batch=1, representing only ~2% of the V100 theoretical GoogleNet Inference performance of ~40k images/sec, even with 900GB/s of HBM memory bandwidth[1]. In contrast, Alveo accelerator cards achieve about 50% of its theoretical performance for GoogleNet v1 inference at Batch=1.

As ML networks evolve, the GPU architecture might become even more challenged. For example, GPUs only support certain data types. If a lower or custom precision data type is desired to maximize throughput and efficiency, the user must wait to for a new card to become available before they can take advantage of a reduced or custom precision implementation. As a result, and

1. Nvidia white paper, NVIDIA AI INFERENCE PLATFORM https://www.nvidia.com/en-us/data-center/resources/inference-technical-overview/ last accessed on 16-Oct-2018.

as described in the Adaptable section of this white paper and illustrated in Figure 3, GPUs accelerated system require frequent hardware upgrades, resulting in higher equipment costs.

Alveo Data Center accelerator cards are the optimum choice for ML inference. They deliver:

- Ultra-low deterministic latency
- Ability to adapt hardware to the needs of any ML network
- Ability to reconfigure hardware as ML workloads evolve without the need to buy new hardware
- Support any precision data types without the need to buy new HW
- Ability to connect directly to the network without a host CPU in the loop

To enable users to take advantage of the many benefits of Alveo Data Center accelerator cards in ML inference, a broad range of applications have been developed by Xilinx and its partners. Table 2 captures some of these applications as well as their acceleration benefits relative to CPUs.

Table 2: **Machine Learning Inference Application Examples for Alveo Cards**

Machine Learning Application Description	Vendor	Alveo Card	Speed Up
Machine Learning Suite for Inference (Tensorflow, Caffe and MXNet)	Xilinx	U200 U250	20X
Neural Network Inference for Image Classification	Mipsology	U200 U250	100X–500X
Unstructured Text Analysis Acceleration	SumUp Analytics	U200	100X

To further understand the benefits of Alveo Data Center accelerator cards for ML inference, the following section provides details on the Xilinx's ML suite.

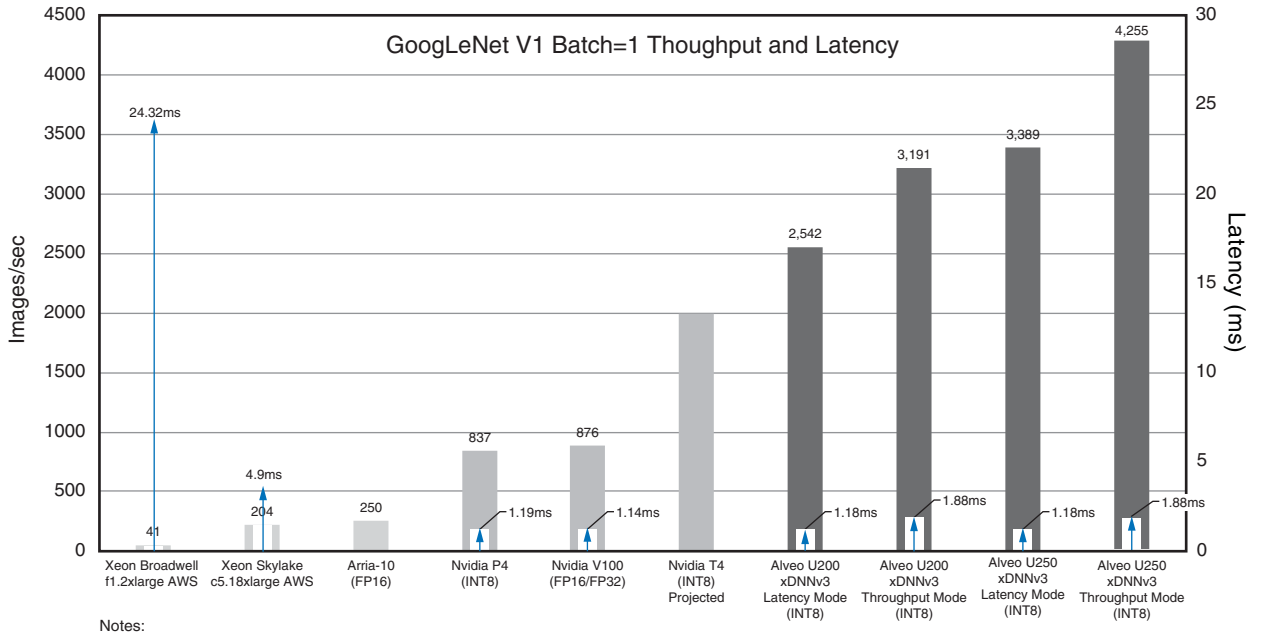
Xilinx ML Suite

Xilinx has developed its ML Suite, to enable rapid deployment of accelerated convolution neural network (CNN) inference (a type of ML network used in image classification and object detection), on Xilinx devices and accelerator cards. The ML Suite includes xDNN processing engines implemented on the Xilinx device, and xfDNN software tools, giving users the ability to connect to higher level ML frameworks and use C++/Python APIs. More details on the Xilinx's ML suite features etc. can be found on Xilinx's Machine Learning website.

Fast - Highest Throughput - Lowest Latency Inference

ML Suite can be used to generate optimized CNNs inference applications for Alveo Data Center accelerator cards. As detailed in WP504, *Accelerating DNNs with Xilinx UltraScale+ FPGAs*, Xilinx's ML Suite can achieve the lowest latency, highest throughput, and greatest energy efficiency across all available accelerators. Figure 4 illustrates the capability of Xilinx's ML suite on the Alveo U200 and U250 cards relative to CPU only, GPU, and competing FPGA offerings. The U250 card achieves 4,100 images/sec with latency of only 1.8ms and only 110W of power. This represents a >20X increase in throughput relative to a CPU, over 15X in throughput relative to the Arria 10 FPGA and nearly 5X increase in throughput relative to Nvidia's flagship Tesla V100 solution. It should also be

noted that the results achieved on the Alveo cards with the ML suite represents a 50% improvement relative to what is achievable on the AWS F1.



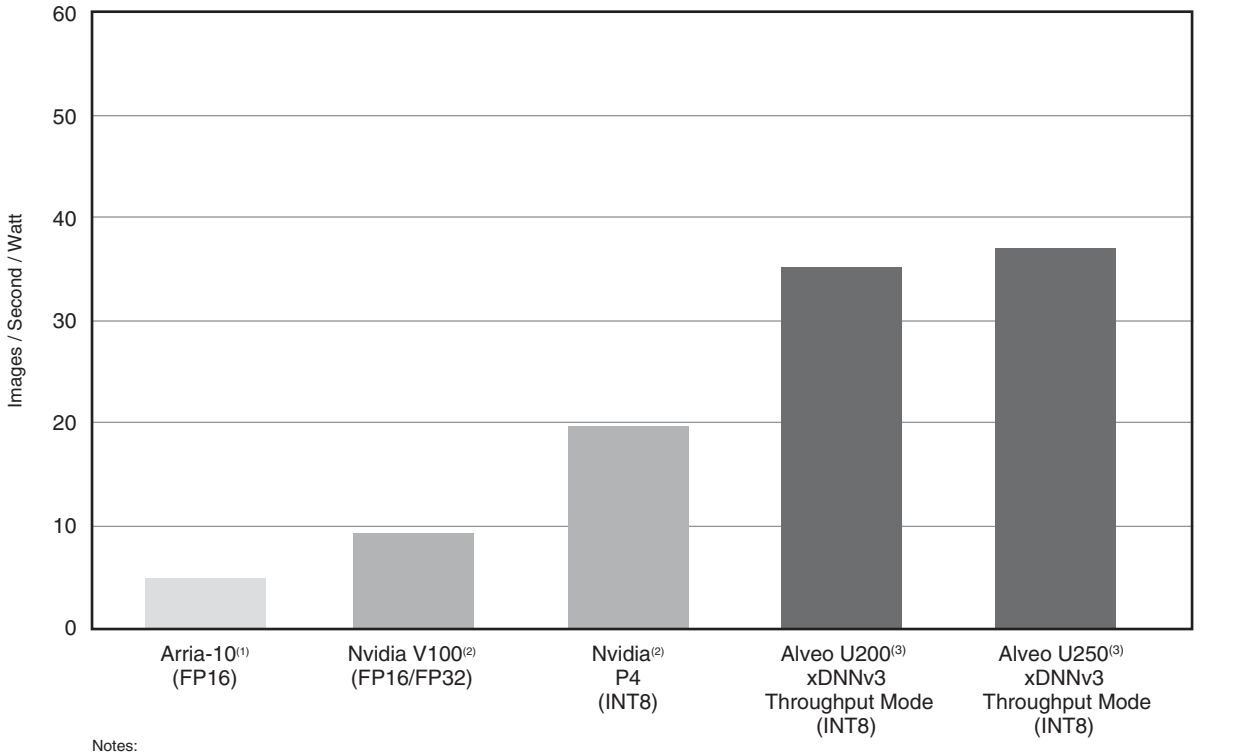
Notes:

1. Xeon E5-2696 v4 f1.2xlarge AWS instance, Ubuntu 16.04LTS, amd64 xenial image built on 2018-08-14, Intel Caffe (https://github.com/intel/caffe), Git Version: a3d5b02, run_benchmark.py w/ Batch=1 modification.
2. Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Ubuntu 16.04LTS, amd64 xenial image built on 2018-08-14, Intel Caffe, Git Version: a3d5b02, run_benchmark.py w/ Batch=1 modification.
3. Arria-10 numbers taken Intel White Paper, "Accelerating Deep Learning with the OpenCL™ Platform and Intel Stratix 10 FPGAs." https://builders.intel.com/docs/aibuilders/accelerating-deep-learning-with-the-opencl-platform-and-intel-stratix-10-fpgas.pdf. Arria latency figures have not been published.
4. Nvidia P4 and V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services, from the Data Center to the Network's Edge." https://images.nvidia.com/content/pdf/inference-technical-overview.pdf. Data retrieved on September 3, 2018.
5. Nvidia T4 projection based on current available published benchmark. GoogLeNet Batch=1 performance range between 1700-2000 images/sec based on early power efficiency benchmarks.
6. Alveo U200 numbers measured Intel Xeon CPU E5-2650v4 2.2GHz, 2400MHz DDR4, Ubuntu 16.04.2 LTS Instance running on OpenStack Pike, Centos 7.4, Pre-release Version of MLSuite, streaming_classify.py, synthetic data, MLSuite DSA Thin Shell, FC and SoftMax layers running on Xeon Host and operations not included in compute totals (0.06% of overall compute).
7. Alveo U250 numbers measured Intel Xeon Silver 4110 CPU @ 2.10GHz, CentOS Linux release 7.4.1708, Pre-release version of MLSuite, streaming_classify.py, synthetic data, DSA: ML Thin Shell, FC and SoftMax layers running on Xeon Host and Operations not included in compute totals (0.06% of overall compute).
8. Alveo accelerator cards throughput latency represents worst-case latency (2X latency of large systolic kernel delay), though measured latency might be less.

WP499_04_110718

Figure 4: **GoogleNet v1 Throughput and Latency, Batch=1**

Energy efficiency is another critical parameter, as it drives the overall power consumption (large component of OpEx) and the amount of compute that can be added to a given rack. Figure 5 illustrates the 80% plus energy efficiency savings offered by the U200 and U250 cards for low latency/ real-time inference, relative to competing FPGAs and GPUs.



Notes:

1. Arria-10 numbers taken Intel White Paper, "Accelerating Deep Learning with the OpenCL™ Platform and Intel® Stratix® 10 FPGAs," https://builders.intel.com/docs/aibuilders/accelerating-deep-learning-with-the-opencl-platform-and-intel-stratix-10-fpgas.pdf.
2. Nvidia P4 and Nvidia V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services, from the Data Center to the Network's Edge." https://images.nvidia.com/content/pdf/inference-technical-overview.pdf. Data retrieved on September 3, 2018.
3. Board power figures reported by Xilinx run-time software during benchmark execution.

WP499_05_110718

Figure 5: **GoogleNet v1 Energy Efficiency, Batch=1**

To further understand the benefit of the Alveo accelerator cards's adaptability, the Alveo U200 and U250 cards measured results shown in Figure 4 and Figure 5, represent 50% of the theoretical capability of the Alveo cards. Unlike GPUs, the underling hardware is configured to suit the workload resulting in higher utilization and higher performance.

Xilinx's Versal ACAPs will further extend its leadership in Data Center real-time inference, by delivering up to a 20X increase in the machine learning inference performance. As Versal devices become available, the Alveo portfolio will be expanded.

Accelerated Database Search and Analytics with Alveo Cards

Data analytics is now a critical tool for enterprise. At its core, data analytics is about retrieving relevant data (search), mainly stored in vast databases, and processing/analyzing that data to gain insight and intelligence. Being able to process and analyze large and unstructured datasets, faster, leads to differentiation for an enterprise. The cost and scale of CPU-only solutions is becoming a key challenge to advancements in data analytics.

GPUs have been used by some to accelerate a limited set of database search and analytics sub-functions. However, the acceleration benefits of GPUs are severely limited by their architecture. Many database search and analytics workload characteristics are at odds with today's GPU architecture:

- Low compute to data transfer ratio
- Wide variation in dataflow patterns
- Direct network access a key acceleration advantage

In contrast, the ability for Alveo Data Center accelerator cards to configure (and re-configure) the underlining hardware (e.g., configure hardware for individual queries) along with their ability connect to the network, have shown them to be an ideal choice for accelerating a broad range of database search and analytics workloads. These acceleration benefits result in faster solutions that require less hardware and power (cost).

A range of ready-to-go applications are available for Alveo Data Center accelerator cards. Table 3 provides details on some of the available applications along with the massive acceleration benefits of the application relative to a CPU. Similar energy efficiency improvements are also achieved. The breath of applications and their associated acceleration benefits illustrate the value of the adaptability of Alveo cards.

Table 3: **Database Search and Analytics Accelerated Applications**

Database Search and Analytics Application Description	Vendor	Alveo Card	Speed Up
High-Speed Search	BlackLynx	U200	90X
Low Latency Key-Value Store (KVS)	Algologic	U200 U250	20X
Hyper-Acceleration Layer for ETL, Streaming Analytics, and SQL Analytics	Bigstream	U200	30X
Hyperion 10G RegEx File Scan	Titan Ic	U200	N/A
Deepgreen DB - Scalable MPP Data Warehouse Solution Derived from Greenplum Database	Vitesse Data	U200	86X
Spark-MLlib Random Forest Acceleration	Xelera	U200	25X
Accelerated Apache Spark MLlib	InAccel	U200	12X

Accelerated Video with Alveo

By 2021, video streaming is projected to grow at 31% CAGR to approximately 160PB/month, or around 81% of consumer Internet traffic(1). The sheer volume of video data is stressing existing network and storage infrastructure. The highest quality video codecs are needed to help to reduce the size of a given video stream without sacrificing video quality, thus reducing the stress and costs associated with transporting and storing video streams.

Hard video codecs, such as the video codecs that can be found in some ASICs or in some GPUs simply do not offer the quality that is achievable by using the latest soft codec running on a CPU. This increase in quality has made a soft implementation on a CPU the de facto standard for Data Center video transcoding.

1. Cisco white paper, Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html, last accessed 25-Oct-2018.

However, given the massive increase in video demand in the Data Center and video codecs becoming more complex, scaling to meet demand using CPUs only is not viable due to the cost, power, and scale of the required solution. Also, CPUs have not being able to deliver real-time video transcoding. For live event streaming and video surveillance, for example, real-time video transcoding is a key requirement. Real-time video streaming is said to grow by 15X by 2021 to represent 13% of Internet video traffic.(1)

Alveo Data Center accelerator cards offer a solution to the video challenge:

- High compute density - Replace many CPUs with a single Alveo card
- Unrivaled adaptability - Evolve the underling hardware to support the most up-to-date codec
- Real-time performance - Support real-time video transcoding

To enable users to immediately leverage the above benefits, a range of video applications have been developed by Xilinx and its partners.

Table 4 details some of the video applications available, along with the expected acceleration benefits of the application relative to a CPU.

Table 4: **Accelerated Video Applications**

Video Application Description	Vendor	Alveo Card	Speed Up
Adaptive Bit Rate Video Transcoding Application	Xilinx, NGCodec, and VYUSync	U200	10X
Personalized Streaming Video Engine	Skreens	U200	5X
Dynamic Image Processsing and Transformation	CTACCEL	U200	10X

Example TCO Saving

The acceleration and energy efficiency benefits of Alveo Data Center accelerator cards enable a significant reduction in capital and operating expenditure, relative to CPU-only solutions. These savings are a result of lower hardware costs (CapEx) and lower power consumption and cooling costs (OpEx), because a smaller number of more energy-efficient Alveo accelerator cards are required to achieve a given capability.

In addition, TCO savings can be achieved for the three example workloads described earlier: ML inference, video and database search, and analytics. Solutions for the three example workloads that previously required 10,000 CPU-based servers can now be replaced by only 2600, 2800, and 2700 Alveo accelerated servers respectively, resulting in a 58–65% equipment cost (CapEx) reduction. Similarly, the Alveo accelerator card's superior energy efficiency ensures a 59–65% reduction in power and cooling costs (OpEx), resulting in an overall TCO savings of 58–65%. These savings are illustrated in Figure 6.

1. Cisco white paper, Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html, last accessed 25-Oct-2018.

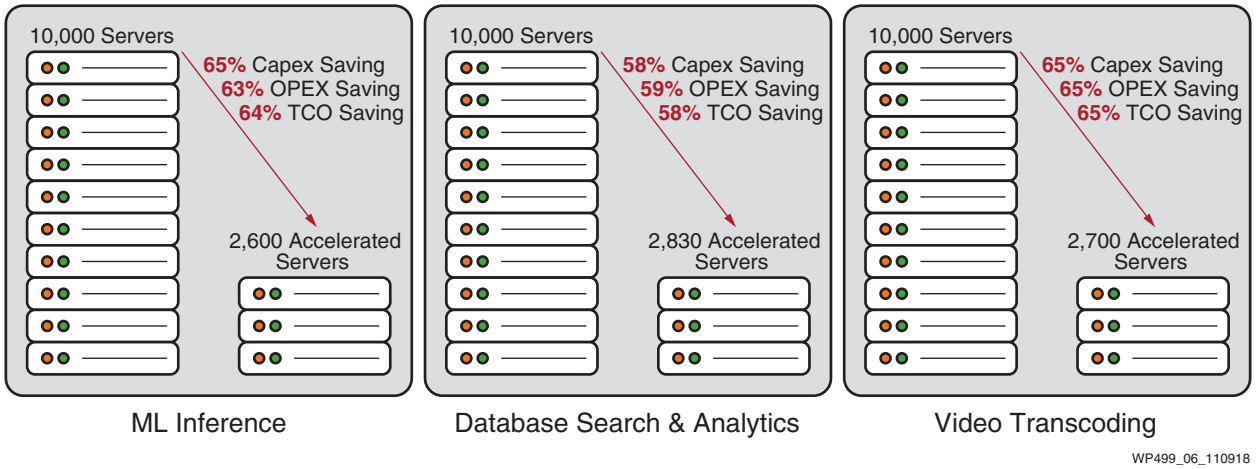


Figure 6: **Video Transcoding TCO Savings**

One additional advantage not considered in the above analysis is that the Alveo cards are futureproofed. As described in the Adaptable section of this white paper and illustrated in Figure 3, as workloads evolve, the Alveo card's underlying hardware can be reconfigured to adapt to the changes in the workload. This adaptability means Alveo Data Center accelerator cards can be replaced less often, resulting in a further reduction in equipment/hardware costs.

Conclusion

Alveo Data Center accelerator cards, with their range of ready to go accelerated applications, are:

- Fast
- Accessible
- Adaptable

These qualities enable solution architects and products owners to breathe new life into data centers and the workloads and applications. In addition, Alveo Data Center accelerator cards enable very significant TCO savings of 65% of more by reducing hardware & power/cooling costs.

More information of acceleration benefits can be found on Xilinx's Alveo Data Center accelerator card landing page.

Explore the benefits of Xilinx FPGA-based acceleration by evaluating solutions on the various FaaS offerings, or visit the website to order an Alveo accelerator card today.

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
11/19/2018	1.0	Initial Xilinx release.

Disclaimer

The information disclosed to you hereunder (the "Materials") is provided solely for the selection and use of Xilinx products. To the maximum extent permitted by applicable law: (1) Materials are made available "AS IS" and with all faults, Xilinx hereby DISCLAIMS ALL WARRANTIES AND CONDITIONS, EXPRESS, IMPLIED, OR STATUTORY, INCLUDING BUT NOT LIMITED TO WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, OR FITNESS FOR ANY PARTICULAR PURPOSE; and (2) Xilinx shall not be liable (whether in contract or tort, including negligence, or under any other theory of liability) for any loss or damage of any kind or nature related to, arising under, or in connection with, the Materials (including your use of the Materials), including for any direct, indirect, special, incidental, or consequential loss or damage (including loss of data, profits, goodwill, or any type of loss or damage suffered as a result of any action brought by a third party) even if such damage or loss was reasonably foreseeable or Xilinx had been advised of the possibility of the same. Xilinx assumes no obligation to correct any errors contained in the Materials or to notify you of updates to the Materials or to product specifications. You may not reproduce, modify, distribute, or publicly display the Materials without prior written consent. Certain products are subject to the terms and conditions of Xilinx's limited warranty, please refer to Xilinx's Terms of Sale which can be viewed at http://www.xilinx.com/legal.htm#tos; IP cores may be subject to warranty and support terms contained in a license issued to you by Xilinx. Xilinx products are not designed or intended to be fail-safe or for use in any application requiring fail-safe performance; you assume sole risk and liability for use of Xilinx products in such critical applications, please refer to Xilinx's Terms of Sale which can be viewed at http://www.xilinx.com/legal.htm#tos.

Automotive Applications Disclaimer

XILINX PRODUCTS ARE NOT DESIGNED OR INTENDED TO BE FAIL-SAFE, OR FOR USE IN ANY APPLICATION REQUIRING FAIL-SAFE PERFORMANCE, SUCH AS APPLICATIONS RELATED TO: (I) THE DEPLOYMENT OF AIRBAGS, (II) CONTROL OF A VEHICLE, UNLESS THERE IS A FAIL-SAFE OR REDUNDANCY FEATURE (WHICH DOES NOT INCLUDE USE OF SOFTWARE IN THE XILINX DEVICE TO IMPLEMENT THE REDUNDANCY) AND A WARNING SIGNAL UPON FAILURE TO THE OPERATOR, OR (III) USES THAT COULD LEAD TO DEATH OR PERSONAL INJURY. CUSTOMER ASSUMES THE SOLE RISK AND LIABILITY OF ANY USE OF XILINX PRODUCTS IN SUCH APPLICATIONS.