Backgrounder

Xilinx Reconfigurable Acceleration Stack
Accelerates Mainstream Adoption of Xilinx FPGAs in Hyperscale Data Centers

November 14, 2016

Overview

Cloud data centers are changing. Today's CPUs have not been able keep up with today's compute-intensive applications like machine learning, data analytics, and video processing. Coupled with increasing bottlenecks in networking and storage, cloud service providers have turned to accelerators to increase the overall throughput and efficiency of their cloud data centers.

Major cloud service providers like Microsoft and Baidu have announced deployment of FPGA technology in their Hyperscale data centers to drive their services business in an extremely competitive market. FPGAs are the perfect complement to highly agile cloud computing environments because they are programmable and can be hardware-optimized for any new application or algorithm. Xilinx is accelerating the mainstream adoption of FPGAs in hyperscale datacenters.

The inherent ability of an FPGA to reconfigure and be reprogrammed over time is perhaps its greatest advantage in a fast-moving field. Using dynamic reconfiguration, FPGAs can quickly change – in less than a second – to a different design that is hardware-optimized for its next workload. As a result, Xilinx FPGAs can deliver the flexibility, application breadth, and feature velocity that complex and constantly changing hyperscale applications need – something that CPUs and custom ASICs cannot achieve. At cloud scale, the ability to rapidly create and deploy pools of reconfigurable Xilinx FPGAs maximizes accelerator utilization, lowers total cost of ownership, and delivers 2-6x the compute efficiency relative to FPGA competition.

Designed for cloud-scale applications, the new Xilinx Reconfigurable Acceleration Stack provides the fastest path for application developers and platform designers to get the full benefit from Xilinx FPGAs when deploying at cloud scale. The stack includes libraries, framework integration, a developer board with deployment reference design, OpenStack support, and a user experience consistent with industry standards.

Momentum for Xilinx FPGAs in the Data Center

The last several years have seen historic momentum for Xilinx in the data center.

Customers – Three of the top seven hyperscale cloud companies have deployed Xilinx FPGAs, including Baidu, which in October announced it had designed Xilinx UltraScale FPGA in pools to accelerate machine learning inference.

Partnerships – Both Qualcomm and IBM announced strategic collaborations with Xilinx for data center acceleration. The IBM engagement already has already resulted in a storage and networking acceleration framework, CAPI SNAP, making it easier for developers to accelerate applications such as NoSQL using Xilinx FGPAs.

Standards Leadership – Xilinx has been leading an industry initiative toward the development of an intelligent, cache coherent interconnect called CCIX. Formed in May 2016 by Xilinx along with AMD, ARM, Huawei, IBM, Mellanox, and Qualcomm, the initiative's membership has since tripled in five months.

Software Defined Tools and Products for the Data Center – The SDAccel software defined development environment for FPGA acceleration was released in 2014. In November 2016 Xilinx unveiled details for new 16nm Virtex® UltraScale™+ FPGAs with High Bandwidth Memory and CCIX Technology.

Highest Application Breadth and Accelerator Utilization

An accelerator can be fast at a certain workload, but it also must be judged on its ability to lower the overall cost of operation of a data center. Figure 1 lists the options for accelerating data center workloads: CPU, Custom ASIC, GPU, and FPGA.

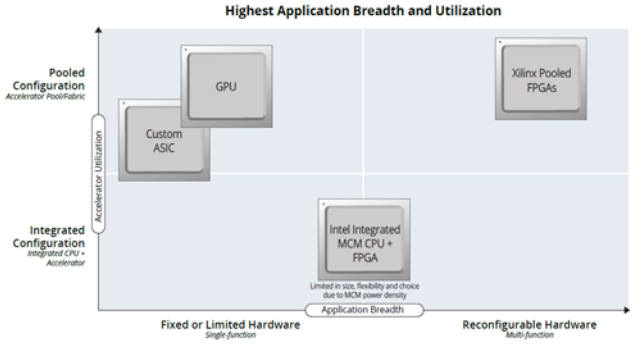


Figure 1

There are two significant factors in determining the TCO for acceleration technology: the breadth of applications (or functions) for which the accelerator is used and how easily and efficiently the accelerator can be provisioned and pooled for these applications, determining accelerator utilization.

Given their programmable and reconfigurable nature, FPGAs provide the most application breadth, spanning from compute to storage to network acceleration workloads. In addition, FPGAs are typically deployed in a pooled configuration, enabling high utilization across a full breadth of applications.

GPUs and custom ASICs are also deployed in a pooled configuration for high utilization. However, neither can support the breadth of applications. The lack of reconfigurability limits them to workloads that are well aligned to their fixed hardware architectures. Additionally, the significant design investment, risk and cost outlay to create a custom ASIC makes this approach less economical than an FPGA.

Finally, Intel's stated strategy to provide integrated CPU-FPGA designs limits application breath and accelerator utilization, placing them in 'no man's land'. These devices are constrained by power density, typically restricting the FPGA to low/mid-range devices and workloads. Being integrated on the CPU package also limits the ability to pool accelerators, dramatically lowering their utilization.

Xilinx FPGAs Deliver 2-6x the Compute Efficiency

Taking a closer look at Xilinx vs. Intel/Altera, figure 2 shows a compelling 2-6x compute efficiency advantage relative to Altera standalone FPGAs with considerably higher utilization relative to Intel integrated MCMs (as described above).

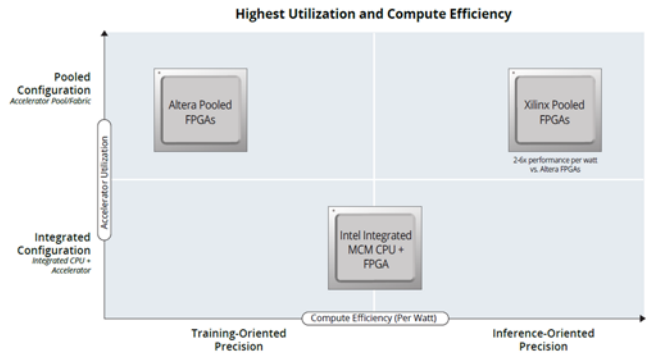


Figure 2

With respect to compute efficiency, Xilinx's advantages are derived from superior DSP architecture, memory hierarchy and silicon technology leadership. The Altera emphasis on floating point precision DSPs is a poor match for many applications including machine learning inference and, falls well short of the compute efficiency of GPUs optimized for training.

Figure 3 lists the most compute-efficient FPGAs available today, as well as sampling products from Intel/Altera. It compares the effective performance and power for machine learning deployment (using inference-oriented data types).

Shipping		TOPs	Power	GOPs/W	Efficiency
Xilinx	VU13P	19.3	66	292	6x
Xilinx	VU9P	11.1	48	231	5x
Xilinx	KU115	9.2	51	180	4x
Altera	Arria10	2.4	51	47	1x

Sampling		TOPs	Power	GOPs/W	Efficiency
Altera	Stratix10	8.3-16.6	76-122	109-136	2-3x

Figure 3

Comparison of lowest power product variants selected for maximum inference efficiency. Effective TeraOps / second for machine learning inference, using 8bit integer data types. Assumes 90% DSP utilization, 80% active cycle for all FPGAs. Stratix10 is an estimated range based on today's DSP clocks and Intel projected peak. Source: Public Intel/Altera and Xilinx product specifications and presentations.

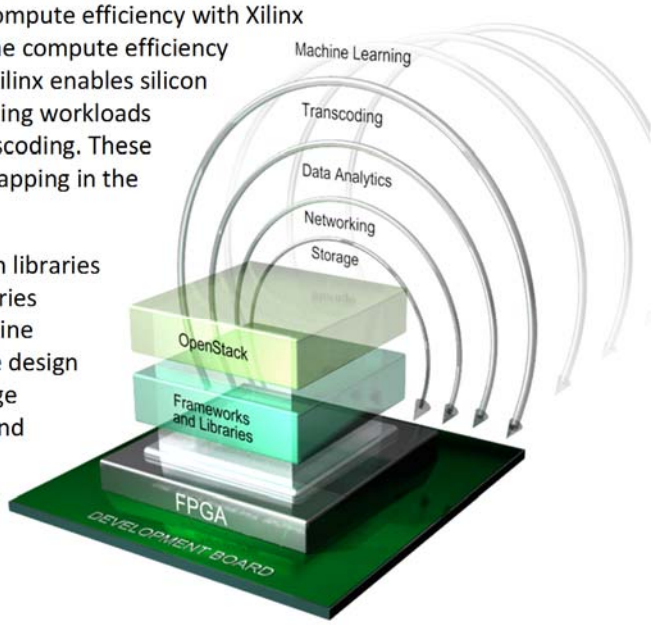
Today, Xilinx products are up to 6x more efficient and 8x higher performance in TeraOps (TOPs). Sampling products from Altera are specified at 2-3x lower compute efficiency relative to Xilinx FPGAs for machine learning inference. This significant gap in compute efficiency is derived from three advantages:

1. Xilinx's DSP architecture optimized for reduced precision integer computations, critical for applications such as machine learning inference.
2. Xilinx's superior on-chip memory hierarchy comprised of distributed RAM, Block RAM, and UltraRAM, a flexible larger capacity block that can be cascaded to create large on chip memories.
3. Xilinx's greater than one year lead in product availability of high end FPGAs at advanced process nodes.

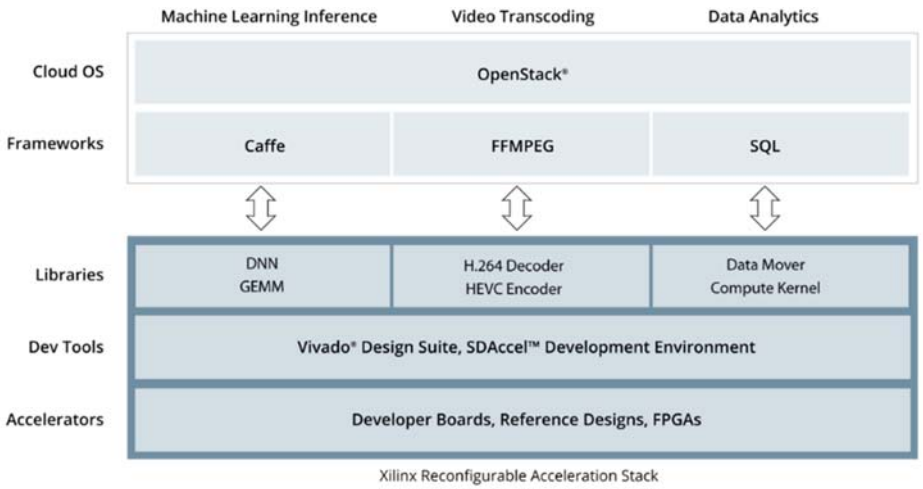
Reconfigurable Acceleration Stack: Fastest Path to Development and Deployment

Programmability and reconfigurability make Xilinx FPGAs the most cost effective and most agile of accelerators for cloud service providers. Designed for cloud scale applications, the FPGA-powered Xilinx® Reconfigurable Acceleration Stack includes libraries, framework integrations, developer boards, and OpenStack support.

The stack provides the fastest path to realize 40x better compute efficiency with Xilinx FPGAs compared to x86 server CPUs and up to six times the compute efficiency over competitive FPGAs. Using dynamic reconfiguration, Xilinx enables silicon optimization for the broadest set of performance-demanding workloads including machine learning, data analytics, and video transcoding. These workload optimizations can be done in milliseconds by swapping in the most optimal design bitstream.



The Xilinx Reconfigurable Acceleration Stack includes math libraries designed for cloud computing workloads, application libraries integrated with major frameworks, such as Caffe for machine learning, a PCIe®-based development board and reference design for high density servers, and an OpenStack support package making Xilinx FPGA-based accelerators easy to provision and manage.



Xilinx Reconfigurable Acceleration Stack

Frameworks and OpenStack Support

- **OpenStack –** Xilinx offers OpenStack support package making Xilinx FPGA-based accelerators easy to provision and manage. The package is available from Xilinx today and will be distributed in the upcoming OpenStack Ocata release in Q1 2017. OpenStack is a free and open-source software platform for cloud computing, mostly deployed as an infrastructure-as-a-service (IaaS). The software platform consists of interrelated components that control diverse, multi-vendor hardware pools of processing, storage, and networking resources throughout a data center.
- **Caffe** – Xilinx supports the Caffe deep learning framework, including compilation with the DNN application library. Caffe is a deep learning framework made with expression, speed, and modularity in mind. It is developed by the Berkeley Vision and Learning Center (BVLC) and by community contributors. Xilinx expects to support other major frameworks, including Tenserflow, starting in the first half of 2017.
- **FFmpeg** – Xilinx supports the FFmpeg multimedia framework, including integrated support for HEVC encode and decode. FFmpeg is the leading multimedia framework, able to decode, encode, transcode, mux, demux, stream, filter and play pretty much anything that humans and machines have created.
- **SQL –** Xilinx enables SQL or NoSQL database acceleration with SQL libraries that accelerate data movement and kernel computation. SQL is used to communicate with a database and is the standard language for relational database management systems.

Libraries

- **DNN** – Deep Neural Network (DNN) library from Xilinx is a highly optimized library for building deep learning inference applications. It is designed for maximum compute efficiency at 16-bit and 8-bit integer data types.
- **GEMM** – General Matrix Multiply (GEMM) library, based on the level-3 Basic Linear Algebra Subprograms (BLAS), from Xilinx delivers optimized performance at 16-bit and 8-bit integer data types and supports any matrices of any size.
- **HEVC Decoder & Encoder** – HEVC/H.265 is the latest video compression standard coming out of the MPEG and ITU standards bodies. It is the successor to H.264 and offers up to 50% bandwidth reduction. Xilinx provides two encoders – a high quality, real-time and flexible encoder to address the majority of video data center workloads and an alternate for non-camera generated content. The decoder supports all the applications for both encoders.
- **Data Mover (SQL)** – The SQL data mover library makes it easy to accelerate data analytics workloads with a Xilinx FPGA. The data mover library orchestrates standard connections to SQL

databases by sending blocks of data from the database tables to the on-chip memory of the FPGA accelerator card over PCIe. The library has been optimized to maximally utilize PCIe bandwidth between the host CPU and the accelerator functions on the FPGA device.

- **Compute Kernel (SQL)** – A library that accelerates numerous core SQL functions on the FPGA hardware such as decimal type, date type, scan, compare, filter and many others. The compute functions are optimized for exploiting the massive hardware parallelization of FPGAs.

Development Tools

- **SDAccel Development Environment** – The SDAccel™ development environment for data center application acceleration leveraging FPGAs. SDAccel, member of the SDx™ family, combines the industry's first architecturally optimizing compiler supporting any combination of OpenCL, C, and C++ kernels, along with libraries, development boards and industry standard development and run-time experience for FPGAs.

Development Board and Reference Design

- **Development Board –** Specialized reprogrammable hardware for computationally intensive applications, specifically targeting the fast-growing markets for live video transcoding, data analytics, and artificial intelligence (AI) applications using machine learning. The board is in a single slot PCIe® half-length full height form-factor and delivers 10-30x performance acceleration over traditional CPUs with a card designed to operate at 75W or less.
- **Reference Design –** The reference design is an excellent starting point for hyperscale customers who want to start early development with a solution they can pass to their ODM or other ecosystem partners to quickly move to production.

Summary

- Xilinx is Accelerating Mainstream Adoption in Hyperscale
- Xilinx Positioned with Highest Application Breadth, Utilization, and Compute Efficiency
- Xilinx Reconfigurable Acceleration Stack Provides Fastest Path to Development and Deployment

To learn more about the reconfigurable acceleration stack, visit the Xilinx Acceleration Zone at www.xilinx.com/accelerationstack.