XILINX RECONFIGURABLE ACCELERATION STACK TARGETS MACHINE LEARNING, DATA ANALYTICS, & VIDEO STREAMING

FPGA REPROGRAMMABILITY DELIVERS PERFORMANCE & FLEXIBILITY WHILE SPEEDING TIME-TO-INNOVATION FOR HYPERSCALE DATACENTERS

SUMMARY

Hyperscale cloud, e-commerce, and social networking datacenters are increasingly facing the challenge of accelerating workloads with complex data types such as 4K video and natural languages, the processing of which often outstrips the capacity of traditional CPUs. This issue is especially acute in the so-called “Super Seven” datacenter companies (Alibaba, Amazon, Baidu, Facebook, Google, Microsoft, and Tencent), where these new applications can demand many thousands of accelerated application servers.

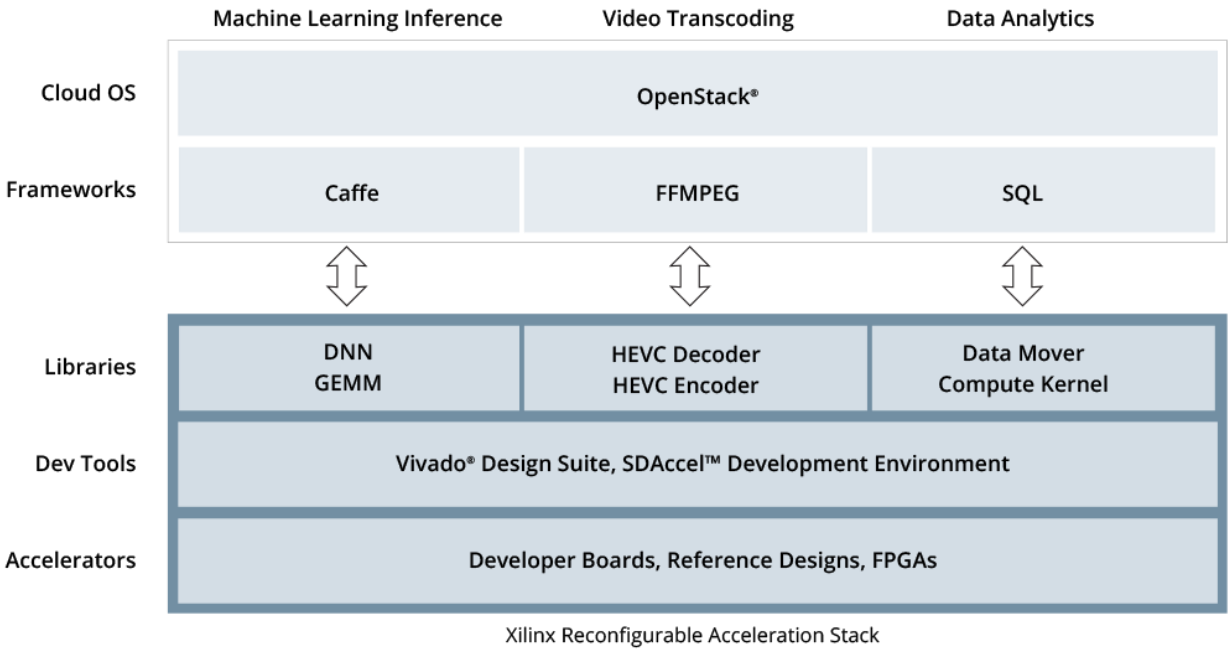
While special purpose hardware such as GPUs and application-specific integrated circuits (ASICs) can effectively accelerate these codes, the rapidly changing state-of-the-art of these algorithms can quickly obsolete a dedicated accelerator as soon it is developed, tested, and put into production. This obsolescence is especially true in the case of ASICs. As a result, many hyperscale companies are turning to field programmable gate arrays (FPGAs)—specialized reprogrammable hardware that can deliver the performance advantages of a fixed-purpose accelerator in a low power, flexible hardware platform that can enable faster time to innovation.

To better address this burgeoning market, Xilinx, a market share and technology leader in the FPGA market, has recently announced its Reconfigurable Acceleration Stack, initially for three computationally-intensive applications: machine learning, data analytics, and live video streaming. With these new offerings, Xilinx intends to reduce the programming hurdles common to FPGA adoption and speed time-to-market for Xilinx customers by providing libraries, development tools, and reference designs for some of the fastest growing workloads in hyperscale datacenters.

XILINX STRATEGY FOR DATACENTER ACCELERATORS

The Xilinx acceleration stack (Figure 1) uses UltraScale FPGAs to deliver up to 20x acceleration over traditional CPUs, according to Xilinx, with a flexible, reprogrammable platform for rapidly evolving workloads and algorithms.

FIGURE 1: XILINX ACCELERATION STACK



(Source: Xilinx)

The growth of live video has been fueled by the increasing popularity of live sporting and computer gaming events and has placed enormous demands on datacenters that provide these services. Similarly, the use of AI is ushering in a myriad of applications where massive amounts of data can be used to train a neural network to learn and recognize patterns in the data. The trained network can then be used in applications as diverse as image recognition, automated driving, search optimization, and natural language translation. In both of these areas, datacenters are increasingly turning to specialized accelerators to deliver low-latency response times to users' queries.

FPGAs are not new in the datacenter today. They are already used in networking and storage applications such as acceleration, smart NICs, NFV, compression, security, and flash arrays. Recently, Microsoft announced extensive use of FPGAs across its Bing search and Azure Cloud properties to accelerate search workloads. Microsoft points to

the flexibility and reprogrammability of FPGAs as a key reason to deploy this type of accelerator instead of fixed function ASICs or GPUs for certain workloads.

Xilinx intends to grow the datacenter footprint for FPGAs with its Reconfigurable Acceleration Stack for datacenter workload acceleration. The product is designed specifically for the world's largest datacenters such as the "Super Seven" group of Alibaba, Amazon, Baidu, Facebook, Google, Microsoft, and Tencent. Xilinx's strategy is to capitalize on the flexibility, performance, and compute efficiency of FPGAs while making the technology easier to develop, deploy, and evolve as the business needs change.

The perennial challenge with FPGAs is that they are notoriously difficult to program, requiring both software and hardware skills that are in short supply. By "acceleration stack", Xilinx means it is providing more than just the FPGA; it is also providing the optimized math and application libraries, software framework implementations, tools that support higher-level languages such as OpenCL and C/C++, tools, OpenStack support for provisioning and management, as well as the expected accelerator board reference designs. This stack is intended to enable rapid development and deployment of acceleration platforms and applications while enabling ongoing enhancements through reprogrammability and reconfigurability.

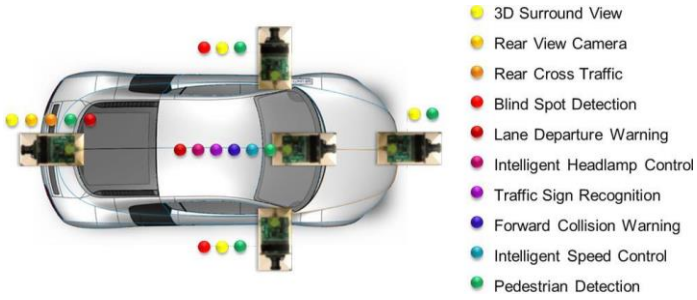
Xilinx is targeting three of the fastest growing workloads in hyperscale datacenters to launch its acceleration stack strategy. First is Machine Learning, which is growing as the underlying algorithms and new applications continue help improve advertising, customer interaction, and new consumer services. Second is high-definition video streaming, especially in gaming-as-a-service markets, where acceleration of video encoding and decoding requires more compute performance than available with standard x86 CPUs. Third, the acceleration of widely-used SQL queries in hyperscale applications.

Acceleration for Machine Learning

The use of neural networks to classify images, translate text and speech, and identify underlying patterns in unstructured data requires a two-phased approach. The first phase is to train the neural network using massive amount of tagged sample data and computation. Once a network is trained, the second phase, called inference, then processes a new data sample or query through the trained neural network to determine its likely classification. Inference is a huge workload in today's largest datacenters, as it enables a broad and growing range of important applications such as language translation, natural language interfaces, photo and video content identification, and

product selection and promotion in online market places. Two of the largest markets for these artificial intelligences will emerge as autonomous vehicles and robots begin to make their appearances on roads and in factories around the world. Here, Xilinx is already well positioned; the company's products are already deployed in Advanced Driver Assistance Systems (ADAS) in automobiles from 23 automakers and 85 models around the world (Figure 2).

FIGURE 2: AUTONOMOUS DRIVING SENSORS & PROCESSORS

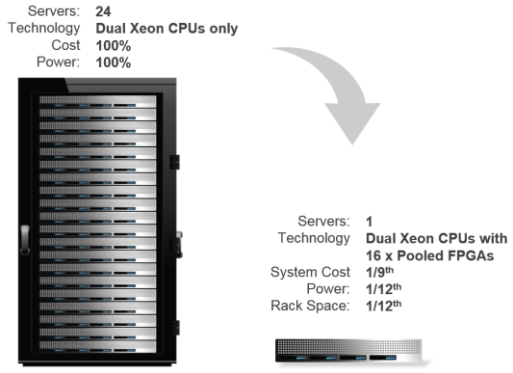


(Source: Xilinx)

The Xilinx Reconfigurable Acceleration Stack provides a solution for inference calculations in the datacenter, the second phase of neural network processing. The inference side of the operation can require billions of calculations to be performed in near real-time for each query. Behind the scenes, social networking and e-commerce sites are conducting millions of simultaneous queries on multiple neural networks. As a result, the combined computational throughput can be staggering, and the need for thousands or millions of concurrent processes requires a solution that scales with demand without using a lot of power. Xilinx believes the new platform should do quite well here, since as many as 24 accelerators can fit into a single server to handle the load, each of which can consume less than 50 watts according to the company.

Today, the majority of inference computation in the cloud is performed using Intel Xeon CPUs. Xilinx calculates that its FPGAs for deep learning can deliver over 10 times the compute efficiency of today's Intel Xeon V4 (Broadwell) CPU. According to Xilinx, a single server with 16 Xilinx accelerators will out-perform an entire rack of standard Intel Xeon servers (Figure 3).

FIGURE 3: SINGLE SERVER WITH XILINX VS RACK OF CPUS



(Source: Xilinx)

A key Xilinx differentiator that enables these performance results is the support for reduced precision arithmetic operations, namely 16-bit floating point and 8-bit integer math. Many neural networks, once trained, can produce accurate results while using lower precision operations, effectively doubling or quadrupling performance at the same power envelope compared to a CPU, which natively performs these operations on 32-bit numbers. These lower-precision ALUs (arithmetic logical units) should theoretically give Xilinx FPGAs an effective 2-6x advantage in computing efficiency compared to FPGAs without this capability, according to Xilinx.

The Xilinx stack for machine learning includes an optimized release of the popular Caffe deep learning framework for convolutional networks, common in image classification. XDNN (Xilinx's DNN primitive library) and GEMM (General Matrix to Matrix Multiplication) are also available for those needing to build their own machine learning applications. According to Xilinx, work is underway to port and optimize additional deep learning frameworks such as Google TensorFlow. Xilinx also provides the SDAccel Design Environment, Vivado Design Suite, IP blocks, as well as reference PCI accelerator board designs to streamline the initial evaluation and development process. By providing these pre-optimized tools and libraries, Xilinx hopes to increase adoption of more workloads that can benefit from acceleration.

ACCELERATION FOR VIDEO TRANSCODING

Xilinx is already well established in the markets for acceleration of many networking and video broadcast functions, and the new acceleration stack for video transcoding is a logical outgrowth from that experience. As many consumers are switching to over-the-top services, especially for viewing live sporting events and game streaming, the

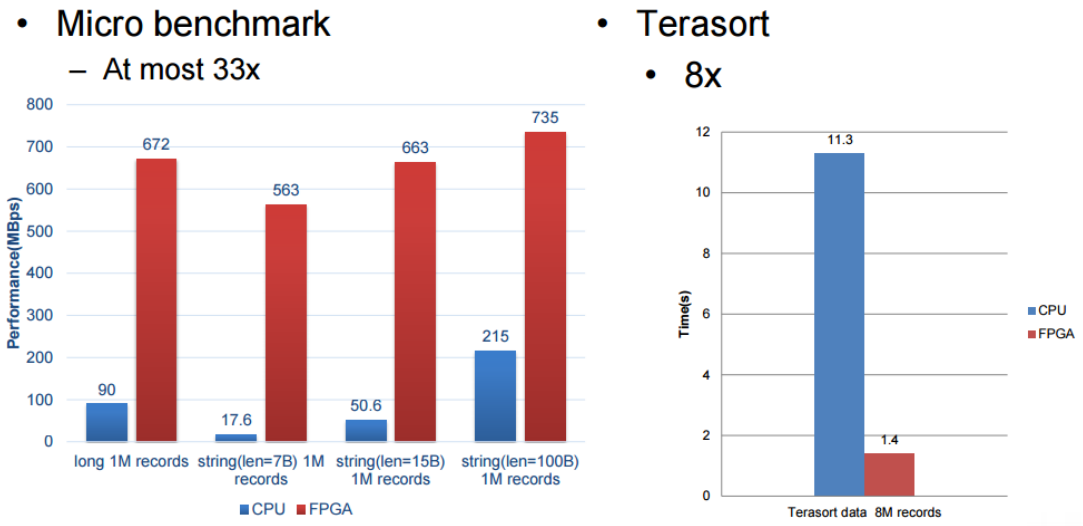
demand for a more efficient platform has grown significantly. Today, most video streaming in the datacenter runs natively on Intel Xeon-based servers, using software-based transcoding solutions. As these events move to higher definition 4K video platforms and H.265 encoding algorithms, the need for accelerated solutions is becoming acute, since the CPU solution is challenged to keep up with the frame rates required for live event broadcasting at 4K resolutions.

Today's FPGA-based transcoders for HD quality video can deliver real-time performance at 60 frames-per-second at 40x better computing efficiency and more than 90% of the encoding quality compared to Xeon E5 CPUs at 1920x1080 resolution, according to data provided by Xilinx. With the new generation of 16-nm UltraScale+ FPGAs, coupled with optimizations to the broadcast-quality encoder, Xilinx says users should be able to realize similar benefits at 4K Ultra HD resolution, enabling a quality of experience that cloud service providers and viewers demand. Factoring in the relative costs of the two alternatives, the Xilinx FPGAs could potentially deliver 20 times the throughput per server rack and roughly 12 times the transcoding throughput per dollar, according to data provided by Xilinx. The video libraries enabled by this stack are available from a Xilinx partner, who also provides acceleration appliances to broadcasters.

Acceleration for Data Analytics

Recently, FPGAs have been gaining traction in hyperscale datacenters to accelerate big data analytics, an important area since SQL database queries are used across the hyperscale infrastructure. Xilinx has prepared a stack for this market as well, and is experiencing excellent early results. Xilinx provides an SQL data mover library that offloads data to an FPGA accelerator. The application leverages the FPGA by constructing data blocks from the appropriate row and columns of an SQL or NoSQL database, then accelerating customer specific analytics and queries. Xilinx has also created TPC Query 6 kernel acceleration examples for the Postgres open source database which can deliver up to a 15x acceleration over Xeon EP CPUs for these kernels, according to the company. At the recent Hot Chips conference, Baidu presented a project outlining how they achieved a 25x acceleration result on the Xilinx FPGA (Figure 4). Baidu noted that the solution performance is limited by memory bandwidth to the FPGA, which Xilinx intends to further address when they add their recently announced solution for integrating High Bandwidth Memory (HBM2) and FPGAs in a 3D device.

FIGURE 4: FPGA ACCELERATION FOR SQL QUERIES



(Source: Baidu)

THE VALUE OF HARDWARE RECONFIGURABILITY

While the focus of discussions surrounding FPGAs is frequently centered on their performance and low power, and the challenging programming model, the inherent ability to reconfigure and reprogram the hardware over time is perhaps its greatest advantage in a fast-moving field. An accelerator such as a GPU, or especially an ASIC, has a fixed set of functional units for processing data in parallel. This initial design represents the development team's understanding of the underlying computational and data requirements at the time they began the work on the new chip. If researchers subsequently want to try a new technique, or discover a new algorithm that demands a different set of functions or utilize a different data type, a team using an FPGA can adapt both the hardware features and the software. The alternative is to build an ASIC, an effort than can take years and can cost tens of millions of dollars. And the result of such a large effort may be rendered obsolete even before it becomes available if the underlying technology and algorithms are rapidly evolving, as is the case in artificial intelligence and machine learning. FPGAs avoid this dilemma through their inherent re-programmability and reconfigurability.

There are two primary benefits of hardware reconfigurability in datacenter application acceleration: **flexibility** and **time to market**. Through reprogrammable, reconfigurable hardware, FPGAs enable cloud leaders to solve today's challenges quickly and adapt as technology and business needs shift. And since the FPGA is connected through

standard interfaces to the CPU, the design team is free to select the right CPU architecture to pair with the FPGA, including x86 from Intel or AMD, ARM SOCs, and IBM OpenPOWER. And by enabling reprogramming of the hardware as well as the higher-level software, the implementation can continue to yield significant performance and TCO advantages over time, protecting the initial investment.

Compared to building an ASIC, the FPGA also yields a significant time to market advantage. Developers can innovate quickly through software libraries that are tuned and optimized to the platform and application needs. This effort typically results in a platform that can then adapt to future changes and enhancements quickly, without going through the multi-month development process of producing an ASIC. This capability is especially well suited for environments where large numbers of algorithms are being updated weekly to drive more innovation in cloud applications and services.

An example of the benefit of reprogrammability in the machine learning discipline is the recent emergence of lower-precision integer (8-bit for inference) and floating point (16-bit for training) arithmetic. This approach can double or even quadruple the performance of the deep learning algorithm if and only if the underlying hardware is able to perform these calculations natively in an optimized arithmetic unit or logic. By using an FPGA with the new Xilinx Reconfigurable Acceleration Stack, a project team may be able to implement their latest innovations in hardware more quickly, potentially improving time to market, while reusing and extending the life span of their existing hardware.

CONCLUSIONS

Based on recent announcements from companies such as Microsoft and Baidu, it appears that FPGA-based acceleration is finally getting significant adoption thanks to the combination of performance, power consumption, and reprogrammability. With the new Reconfigurable Acceleration Stack, Xilinx has made progress in reducing some of the programming hurdles to further FPGA adoption for three of the hottest markets in hyperscale datacenters: machine learning, live high-definition video transcoding, and accelerated SQL queries. This stack offers a set of logic, libraries, software frameworks, reference designs, and tools that are designed to enable developers to more quickly capture acceleration opportunities and shrink time to market.

Of the three approaches to acceleration technologies available in the market today, GPUs, ASICs, and FPGAs, only FPGAs offer the benefits of hardware reprogrammability. And the new platforms being launched by Xilinx are focused on areas experiencing rapid innovation and change, where the ability to reconfigure the

hardware to evolve along with the software can provide additional flexibility and speed time to market. That being said, for certain applications such as floating point intensive codes with significant opportunity for parallelism, the GPU may be the preferred solution. And for very large-scale environments where millions of acceleration units for a specific code are foreseen, the investment in developing an ASIC may be worthwhile to realize lower per-unit costs and potentially higher performance. But for rapidly changing environments and broad ranges of target applications, the FPGA provides an attractive solution to datacenter acceleration.

By providing robust reconfigurable acceleration technology for early adopters and innovators, and by focusing on the “Super Seven” datacenters, Xilinx can make a strong case to establish the FPGA as a leading contender for hyperscale datacenters.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR
Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER
Patrick Moorhead, President & Principal Analyst at Moor Insights & Strategy

EDITOR
Scott McCutcheon, Director of Research at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
This paper was commissioned by Xilinx, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.