SOLUTION BRIEF

SUPERMICRO AND XILINX ACCELERATE PERFORMANCE FOR DATA CENTER WORKLOADS

A Phenomenal Increase of 90x Compared to CPU's for many Datacenter Tasks

TABLE OF CONTENTS

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Executive Summary

Many of the common data center workloads involving machine learning inference, video transcoding, and database search and analytics traditionally rely on auxiliary GPU technology to supplement or increase CPU footprint as an extension. However, these solutions often contribute to increased costs, management complexity, and negative impact on the environment via increased thermals. Supermicro's solution based on Xilinx Alveo Data Center accelerator cards helps address all these problems while providing up to 90x performance increase over CPUs employed for the same tasks.

Solution Features and Benefits

Today's data center workload requirements are more dynamic and are evolving faster than the traditional refresh cycles of the infrastructure. Having a fixed function GPU and CPU addresses the complexities of algorithms is becoming an unfaltering and challenging task. Supermicro's flexible Ultra servers and Xilinx UltraScale architecture provide reconfigurable acceleration that scales to changing algorithm optimizations customers have come to expect. This, in turn, has a significant impact in keeping the costs and schedules under control while boosting productivity and performance multi-fold.

2U ULTRA SERVER SERIES

Many of the workloads commonly encountered in the data center, the database search & analytics, fintech, machine learning, video processing, or HPC, such as genome sequencing, oil & gas, and basic research, require huge compute and efficient storage capabilities. Supermicro systems are ideal for these tasks, with plenty of data storage capacity and flexible rear I/O to support various network cards for data delivery. The Supermicro Xilinx Accelerator Solution is based on the flexible, scalable, and versatile Ultra Platform. Supermicro's 2U Ultra 2029U series and Supermicro's 6029U series are flagship enterprise-grade rackmount servers that provide a balance of high-end compute, storage, and expansion all in one system. These systems support dual 2nd Gen Intel® Xeon® Scalable processors, with up to 24 DIMM slots. The Supermicro 2029U series supports up to 24 of the 2.5" hot-swap drive bays, and the Supermicro 6029U series supports up to 12 of the 3.5" hot-swap drive bays. Both systems contain flexible onboard Ethernet options and up to 8 PCI-E slots for various add-on cards. Other available Ultra 2U systems, such as the SYS-2029U-TN24R4T, supports up to 24 of the 2.5" NVMe drives that can enhance system storage with higher throughput and IOPS while lowering the overall latency.

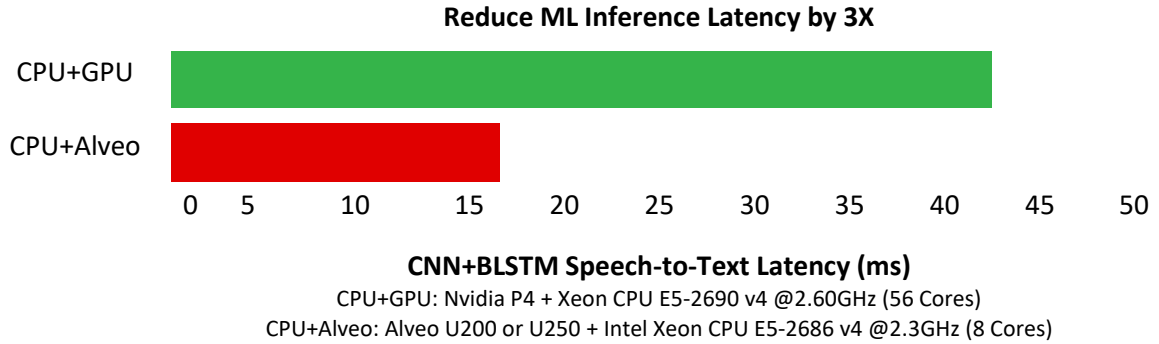
Additionally, Supermicro Ultra 2U's highly modular design offers a wide range of configurations that fit the varying requirements of the data centers.

FPGAs such as the Xilinx Alveo U200 & U250 have shown significant improvement in power consumption and performance in Deep Neural Networks (DNNs) applications, which offer high accuracy for important image classification tasks. The Xilinx Alveo U200 & U250 have the potential to be resource/power-efficient. It can accelerate network performance without making a significant investment in new hardware.

By leveraging the Supermicro Ultra 2U 2.5" system, customers can utilize the capacity and performance of 24 drive bays. The last four drive bays are hybrid for future storage upgradeability, configured for NVMe support (with optional parts) for faster data delivery of up to 6X than traditional SATA drives. The expansion support offers a default of 8 PCI-E slots configured to x16 or x8 devices. By replacing the default riser with an RSC-W2-66, Supermicro can take 4 PCI-E x8 lanes and convert them to 2 PCI-E x16 lanes. This allows Supermicro to support Xilinx Alveo U200 or U250 and a Mellanox MCX516A-CCAT (dual-port 100G) network adapter combination.

FPGAs are well-known for their power efficiency. With the constant demand of the modern Data Center, Xilinx® Alveo™ U200 & U250 data center accelerator cards help offload critical workloads from CPUs, including applications as machine learning

inference, video transcoding, and database search and analytics. Built on the Xilinx Alveo U200 & U250 16nm UltraScale ™ architecture, Xilinx Alveo U200 & U250 provides up to 90X higher performance than traditional CPUs for key workloads and 3X higher inference throughput and 3X latency advantage over GPU-based solutions. The Xilinx Alveo U200 & U250 is designed to meet the dynamic changes in acceleration requirements and algorithm standards while reducing the overall cost of ownership. The chart below compares a CPU+GPU solution with a CPU + Alveo solution when running an inferencing benchmark.



Combined with the Xilinx Alveo U200 or U250, a Supermicro Ultra system accelerates real-time AI inferencing, which provides a higher throughput performance and better power efficiency for AI-based speech systems. Also, a higher throughput system for video analytics pipelines as compared to GPU-based systems. A comparison of different workloads and the performance of the Xilinx Alveo accelerator is shown below.

Adapt and Accelerate Any Workload

AREA	PARTNER WORKLOAD	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

FEATURES	ALVEO U200 Accelerator Card	ALVEO U250 Accelerator Cards
Peak INT8 TOPs	18.6	33.3
DDR Memory Bandwidth	77GB/s	77GB/s
Internal SRAM Bandwidth	31TB/s	38TB/s
Look-up Tables (LUTs)	892,000	1,341,000
Thermal Options	Passive or Active	Passive or Active

Summary

Supermicro Xilinx accelerator solution based on Ultra 2U platform provides the flexibility, scalability, and simplicity to run multiple different data center workloads without extensive outlays into fixed function accelerator supports changing technology, allowing customers to consolidate many additional requirements into a single platform, thus reducing overall TCO.