PRODUCT BRIEF

SUPERMICRO'S LATEST SERVERS DEMONSTRATE SUBSTANTIAL PERFORMANCE IMPROVEMENTS OVER PREVIOUS GENERATIONS

Supermicro Servers with 5th Gen Intel® Xeon® Processors Show Remarkable Improvement Running BERT-Large, ResNet, and SPEC Benchmarks

TABLE OF CONTENTS

Executive Summary

Data center operators constantly evaluate new server hardware to give employees and customers a faster and better experience at lower costs. The recent release of the latest Supermicro X13 server product families with the 5th Gen Intel® Xeon® processors enables enterprises and cloud providers to upgrade their existing infrastructure to new servers and quickly realize performance gains. Data center refresh cycles are about five years, which means that when comparing performance, the latest systems should be compared to those in a current data center. Performance gains from new CPU capabilities can be significant in the cases discussed below. The case for upgrading to the latest generation of Supermicro's Intel-based servers has never been stronger.

Supermicro servers continue to give customers leading-edge performance for a wide range of workloads. Supermicro's application-optimized product families are designed to maximize performance and reduce energy usage through shared components and advanced design.

Artificial Intelligence Inferencing Benchmark Performance

AI is becoming an indispensable technology enterprises rely on to improve business processes and provide a competitive advantage. The introduction of BERT (Bidirectional Encoder Representations from Transformers)-Large has significantly propelled Natural Language Processing (NLP) into a new era, delivering substantial performance enhancements in large-scale language processing tasks. With the release of the Supermicro X13 product families with 5th Gen Intel® Xeon® processors, Supermicro is offering a groundbreaking solution that attains a remarkable 23x performance improvement in a generation-over-generation (2nd Gen Intel® Xeon® Scalable processors to 5th Gen Intel® Xeon® processors) comparison, as shown in the BERT Large benchmark below. In many cases, dedicated GPUs within a server are not needed to run AI inferencing tasks, and can now be accomplished with CPUs and in some cases using dedicated accelerators on the CPU.

By significantly reducing the time it takes to execute an application, Supermicro empowers organizations to conduct more inferencing tasks in each timeframe than ever before. With a shorter time to completion, more complex models can be trained, or more models can be trained in a given timeframe. Furthermore, there can be a reduction in the number of servers, reducing the number of racks and real estate costs. The new Supermicro servers with 5th Gen Intel® Xeon® processors offer enhanced security, more cores, shorter time to run apps, smaller footprint, and lower energy consumption per unit of work, resulting in a lower total cost of ownership.

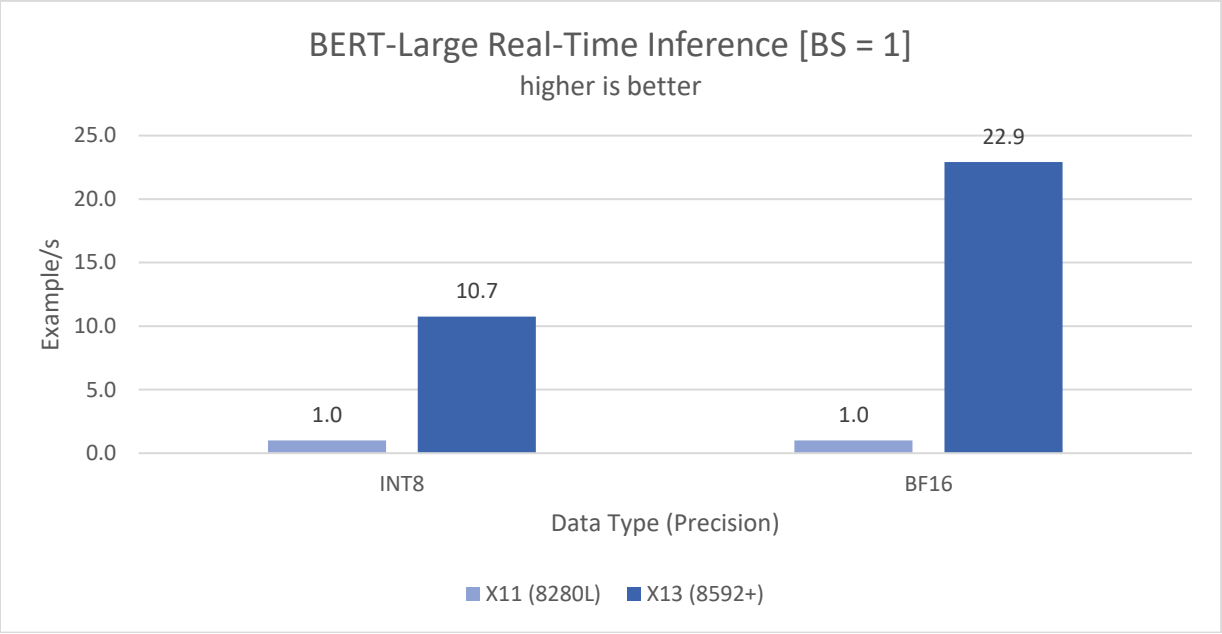
Benchmark Setup

The goal of the BERT-Large and ResNet50 v1.5 benchmarks is to compare the system performance of a Supermicro server containing 2nd Gen Intel® Xeon® processors with a Supermicro server using the new 5th Gen Intel® Xeon® processors. BERT-Large is used for many scenarios where fast decisions need to be made for question answering, sentiment analysis, document classification, and sentence similarity. ResNet can be used to classify images, to identify objects within an image or within a video. The benchmarks used below are based on a standard configuration from Supermicro.

The systems that were used for these benchmarks that compare Gen-over-Gen-over-Gen-over-Gen performance:

X11 System Tested	X13 System Tested
• Supermicro Server: SYS-6019U-TN4RT (Ultra) • CPU:2x Intel® Xeon® Platinum 8280L @ 2.6 GHz (base) • Memory: 12x32GB DDR4-2933 MHz, 384GB • System Drive: 1x Intel 4510 1TB NVMe • Operating System: Ubuntu 22.04.03 LTS • BIOS Version: 4.0 (BIOS Build Time: 06/20/2023) • Kernel Version: Linux 5.15.0-89-generic • Cooling: Air Cooled	• Supermicro Server: SYS-221H-TNR (Hyper) • CPU: 2x Intel® Xeon® Platinum 8592+ @1.9GHz (base) • Memory:16x64GB Samsung DDR5-5600 MHz, 1024GB • System Drive: 1x 7.68TB Samsung MZQL27T6HBLA-00A07 • Operating System: Ubuntu 22.04.03 LTS • BIOS Version: 2.0 (BIOS Build Time:11/04/2023) • Kernel Version: Linux 5.15.0-89-generic • Cooling: Air Cooled

Artificial Intelligence Benchmark Results:

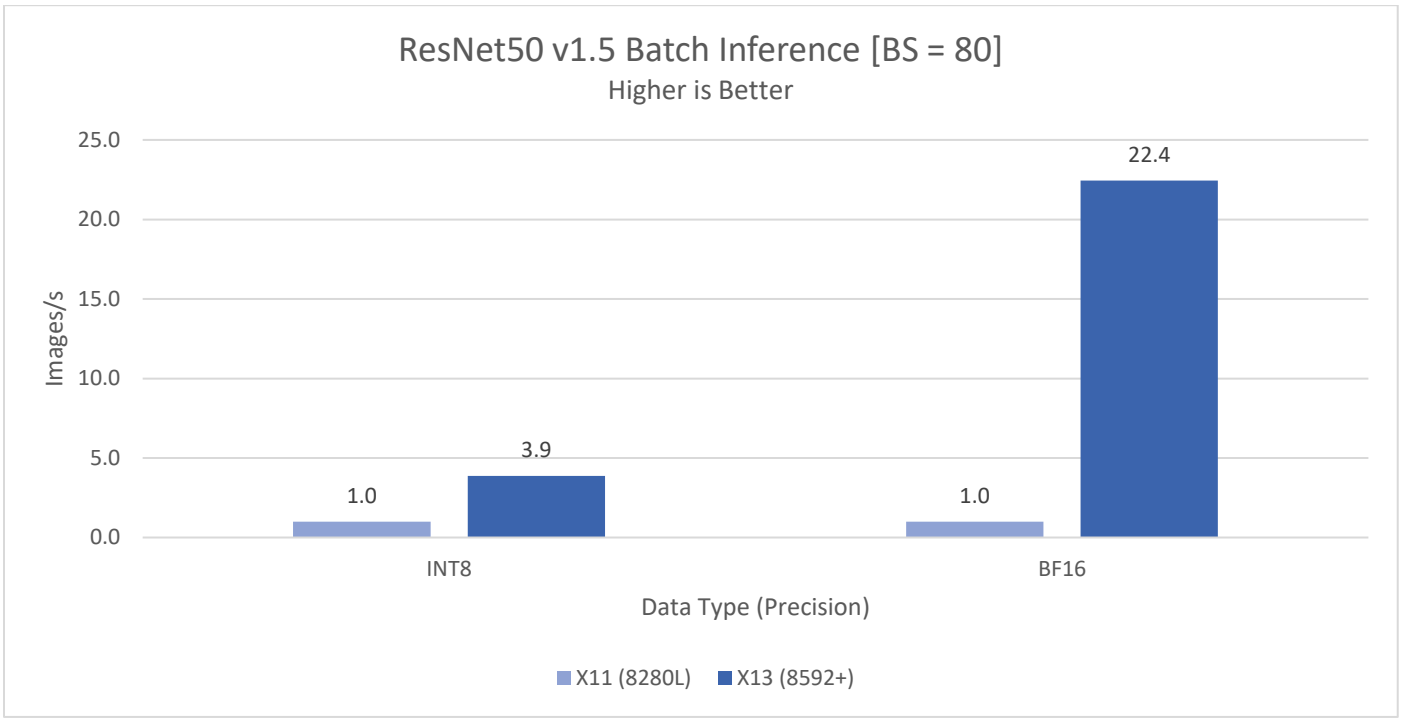


X11 System uses FP32 to emulate the BF16 data operations

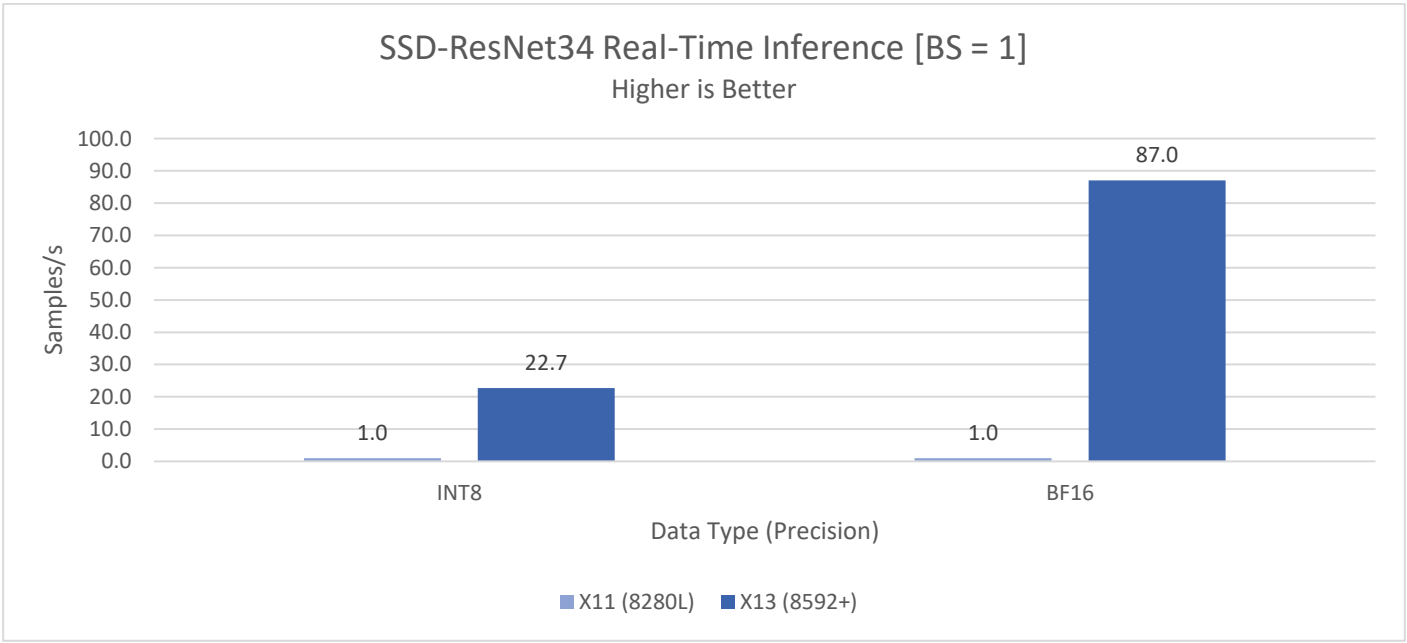
The Supermicro X13 system with 5th Gen Intel® Xeon® processors shows a 10.7x improvement (with INT8 representation) and over a 22x improvement with BFloat16 (BF16) representation. For the X13 generation system, Intel ® Advanced Matrix Extensions (Intel® AMX) were activated, which helped achieve such high performance. Intel Advanced Matrix Extensions is a built-in accelerator that significantly improves AI inference performance.

The ResNet benchmark comes in various versions, referred to as ResNet Architecture, such as ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. They are all designed to focus on similar computer vision concentration.

- ResNet50 v1.5 focuses on AI image classification workloads.
- SSD-ResNet34 focuses on AI object detection workloads.



X11 System uses FP32 to emulate the BF16 data operations



X11 System uses FP32 to emulate the BF16 data operations

The performance gains for the ResNet50 v1.5 Batch Inference benchmark were almost 4x to 22.4x faster when using the X13 system compared to an X11 system. The SSD-ResNet34 benchmark showed a performance gain from about 23x using INT8 to 87x when using the Bfloat16 data representation.

SPEC Benchmarks

SPEC Performance – The Standard Performance Evaluation Corporation (SPEC) was founded in 1988 by a few workstation vendors who realized that the marketplace desperately needed realistic, standardized performance tests. The key realization was that an ounce of honest data was worth more than a pound of marketing hype. SPEC has grown to become one of the more successful performance standardization bodies with more than 60 member companies. SPEC publishes many performance results each quarter, spanning various system performance disciplines. See www.spec.org for more information.

The following SPEC benchmarks have been run on the latest Supermicro systems, using the 5th Gen Intel® Xeon® processors. The SPEC CPU values indicate the performance of a given system compared to a base system for running applications that are floating-point based, which relate to the system performance of HPC type applications. In addition, the performance reported is of a single copy running (speed) or many copies running (rate) on the given system.

Benchmark	2nd Gen Intel® Xeon®	5th Gen Intel Xeon	Increase from 2nd Gen to 5th Gen
	Intel® Xeon® 8280	Intel® Xeon® 8592+	
CPU 2017 FP Speed (2 socket)	158	387	2.45x
	Intel® Xeon® 6258R	Intel® Xeon® 8592+	
CPU 2017 FP Rate (2 sockets)	320	1220	3.81x

For more information, visit www.spec.org

Summary

These benchmarks demonstrate that applications run significantly faster with the latest CPU and system technology using the latest Supermicro servers with 5th Gen Intel® Xeon® processors. The new system technology available with Supermicro's X13 product families shows a significant improvement over previous generations of servers. The performance gains over multiple generations of servers are substantial, as shown when looking at the results of popular benchmarks. These examples from the AI and HPC domains indicate that significantly more work can be done in a specific time frame, leading to better training or more complex simulations.

Supermicro's X13 platforms with 5th Gen Intel Xeon offer up to 3.8x higher performance on general compute workloads, and the data collected demonstrated up to an 87x performance gain in AI inference workloads. This puts Supermicro's X13 platforms as a very compelling solution for enterprises running AI inference or mix load workloads and strongly suggests this is the right time to upgrade older platforms and start realizing OpEx cost reductions.

When upgrading to the latest generation of Supermicro servers, customer will experience enhanced security, more cores, shorter time to run apps, smaller footprint, and lower energy consumption per unit of work, resulting in a lower total cost of ownership.

For More Information

Supermicro X13 Systems – www.supermicro.com/X13

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. See www.supermicro.com