PRODUCT BRIEF

SUPERMICRO X13 SYSTEMS OFFER POTENTIAL BEYOND GENERAL COMPUTING

Support for AI Inferencing with Supermicro using 5th Gen Intel® Xeon® Processors with Intel AMX

Supermicro Hyper System

Supermicro BigTwin System

TABLE OF CONTENTS

Executive Summary

Artificial intelligence has emerged as a pivotal focus within the computer industry in recent years. Among the notable applications, GPT (Generative Pre-trained Transformer) has catalyzed a transformative shift in the landscape of AI applications. As the field of natural language processing continues to evolve, the demand for high-performance server systems has reached unprecedented levels. The internet search activity ends here for those seeking a solution to empower the AI inference workload in their daily tasks.

Supermicro X13 systems represent the cutting edge of advancements in server solutions. The Intel AMX technology in Supermicro X13 systems enhances the general compute system with three times better performance on the machine learning workload compared with the previous generation of the system. In the tests performed, the powerful Hyper (SYS-221H-TNR) and BigTwin (SYS-221BT-HNTR) systems were used to set up an OpenShift cluster ready for Inferencing in a Kubernetes environment. These state-of-the-art systems are meticulously engineered to redefine the paradigm of how organizations leverage the capabilities of BERT (Bidirectional Encoder Representations from Transformers) for natural large language model

processing. The X13 systems not only establishes new benchmark numbers but also showcases the remarkable potential of processors in executing AI tasks, delivering unprecedented performance enhancements over its predecessors. These results solidify its position as the pinnacle of next-generation systems.

The Supermicro X13 systems leverage the 5th Gen Intel® Xeon processors that are optimized to increase performance per watt and lower total cost of ownership across critical workloads like AI, HPC, storage, networking, and beyond. It stands out in the embedded acceleration, offering technologies like Intel® AMX and Intel® QAT for tailored performance boosts in compression, encryption, data analytics, and low-latency operations in large language models. Its adaptability with various accelerator technologies adds flexibility for workload-specific requirements.

Solution Architecture

In the era of cloud computing, many applications are embracing cloud-native technologies to enhance the efficiency of development, testing, deployment, and operations through container platforms. Supermicro simulated the customer environment and ran the BERT Large benchmark test using the Red Hat OpenShift Container Platform.

By seamlessly integrating both Supermicro X13 BigTwin and Hyper in the same OpenShift cluster, Supermicro built a flexible and scalable environment specially designed to meet the demanding requirement of BERT Large benchmarking testing conducted with computational power. This thoughtful combination ensures a cohesive orchestration of resources, allowing for optimal utilization and performance enhancement.

The X13 systems work smoothly with Red Hat OpenShift, forming a solid foundation to handle the intricate tasks of processing BERT Large language models. This integration exemplifies our commitment to cutting-edge solutions and provides a seamless and powerful experience for organizations venturing into advanced language processing and AI workloads conducted with 5th Gen Intel Xeon processors.

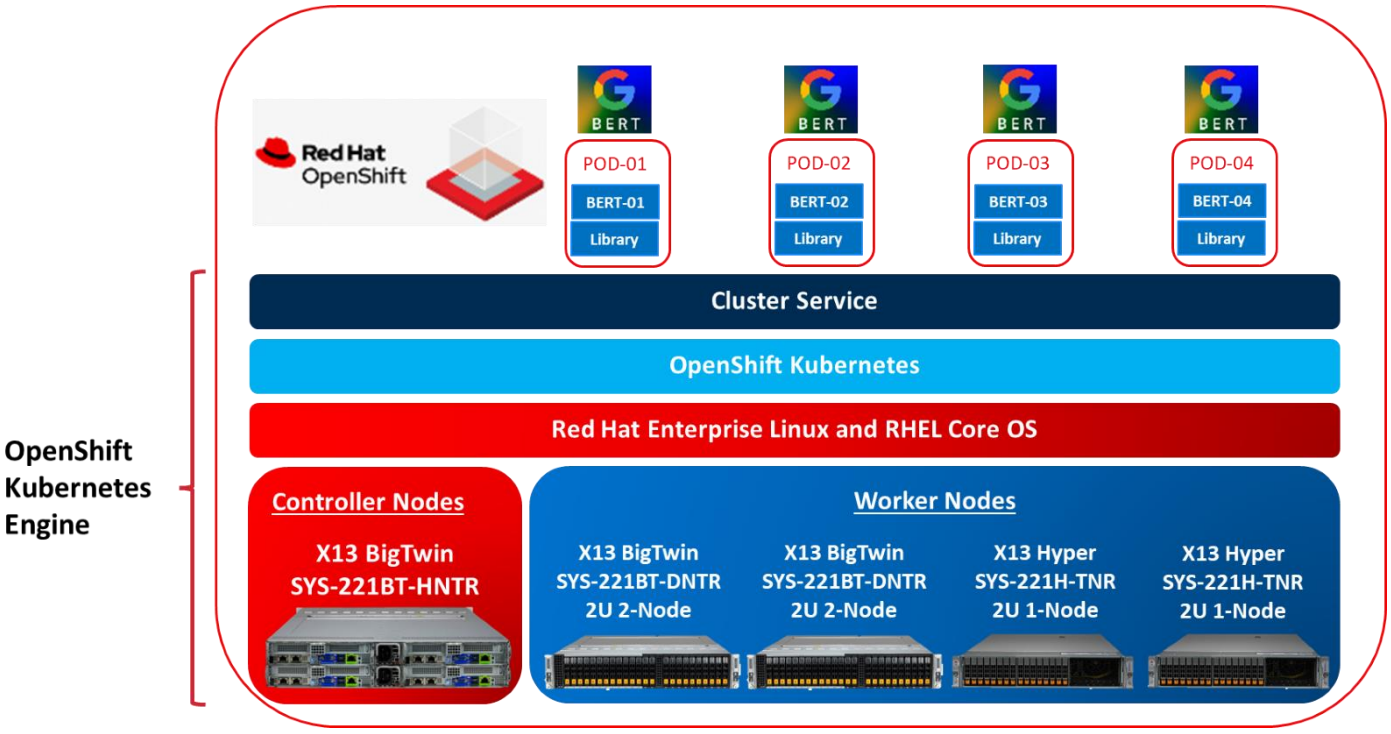

Figure 1 - OpenShift test environment with X13 BigTwin and X13 Hyper systems

For instance, the BigTwin multi-node system is a dense and efficient system capable of creating an OpenShift cluster in a single machine with a maximum of 4 nodes in a 2U chassis, making it ideal for small clusters. However, its versatility extends since it can also be the perfect system to host the controller nodes in an OpenShift environment or add 2U 2-Node BigTwin as worker nodes supporting top bin processors. On the other hand, the Hyper system can have 32 DIMM slots to support 4TB of memory and use a high TDP processor for the best performance available. Combining these two systems within a single cluster positions them as an ideal solution for handling demanding inferencing workloads effectively.

Performance on Large Language Models Benchmarks

Supermicro X13 Systems utilize 5th Gen Intel® Xeon processors. This processor has improved CPU cores and frequencies compared to previous generations. However, what sets them apart are the latest accelerators designed to assist specific workloads that offload the CPU cores and enable more efficient task execution. One notable accelerator is the Intel Advance Matrix Extensions, which are designed to optimize some AI workloads within the processor.

- Intel AMX (Intel Advanced Matrix Extensions): Improves the performance of matrix multiplication, which is pertinent to artificial intelligence, machine learning, and scientific computing, by reducing the memory access and compute cycles required.

Supermicro conducted comparative analyses using the previous generation systems as a reference point to gauge the performance enhancements in the BERT Large benchmark.

Supermicro X13 Systems with 5th Gen vs 4th Gen Intel® Xeon Processors

In contrast to its predecessor, the Supermicro X13 systems offer enhanced capabilities, showcasing notable advancements over the previous generation system.

- Using the most recent version of Intel® AMX technology from the 5th Gen Intel® Xeon Processor, the X13 BigTwin systems demonstrated an impressive 149% improvement in the inference performance of BERT-Large using INT8 precision over the previous 4th Gen processor.

- Supermicro X13 Hyper Systems feature a distinctive BIOS configuration that unlocks a significant improvement in CPU performance efficiency, getting 58% for FP32.

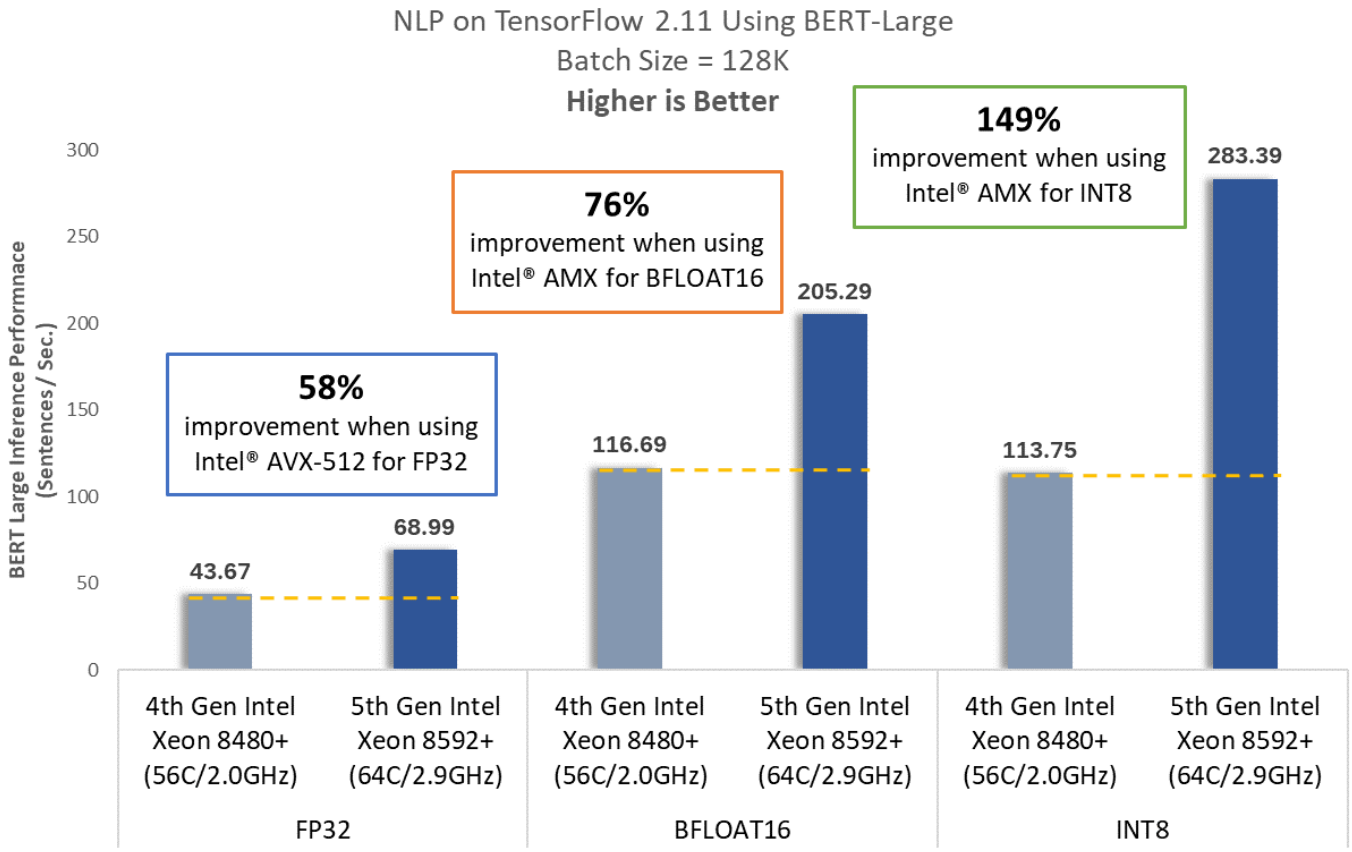

Figure 2 – Bert-Large results with SYS-221H-TNR and SYS-221BT-HNTR Systems

X13 with the 5th Gen Intel® Xeon vs X12 with the 3rd Gen Intel® Xeon

Traditionally, customers opt to upgrade their server infrastructure to improve their performance capabilities required in the AI era. With this in mind, our comparative analysis incorporates older server models for a comprehensive evaluation. The X13

system stands out in this comparison, showcasing an impressive 339% performance improvement with the BERT-Large (INT8) benchmark. This represents a substantial advancement in efficiency and technological capability to leverage computational power for AI tasks.

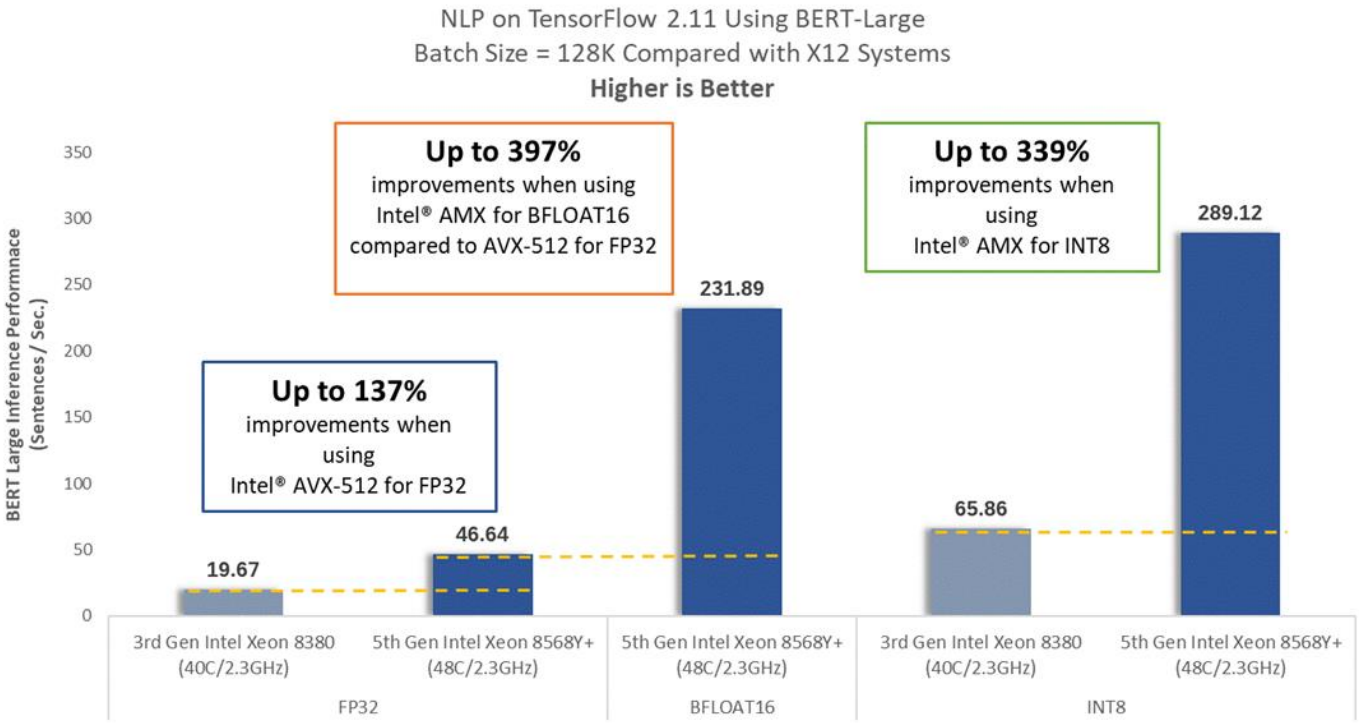

Figure 3 - BERT-Large Results with SYS-221H-TNR and SYS-221BT-HNTR compared to previous Supermicro X12 servers.

Summary

In the rapidly evolving landscape of artificial intelligence, the demand for robust systems capable of efficiently managing inference workloads has become paramount. As organizations navigate the complexities of the AI era, the Supermicro X13 systems emerge as an indispensable ally, offering unparalleled performance and reliability. It provides a foundational entry into AI inference capabilities via its CPU, with the option to integrate GPUs for enhanced performance as needed. This flexibility allows customers to begin with a cost-effective, modest setup and scale up to more powerful configurations as they explore and realize the full advantages of AI inference. Designed to meet the rigorous demands of modern AI-driven environments, the X13 systems stand as a beacon of innovation, empowering businesses to unlock new levels of efficiency and productivity. With its advanced features and cutting-edge technology, the X13 systems represent a strategic investment in the future, positioning organizations for success in the dynamic world of artificial intelligence.

Appendix - System Configurations for Benchmarks

Type	Description	System QTY
Barebone	X13 Hyper 2U, with 8x 2.5" drives	1
CPU	5th Gen Intel Xeon 8568Y+	2
CPU	4th Gen Intel Xeon 8480+	2
MEMORY	64GB DDR5-5600	16
Drive	400GB NVMe PCIe 4.0 M.2 22x80mm	2
Drive	6.4TB NVMe PCIe 4.0 U.2	10
AIOM NETWORK	AIOM 4-port 10GbE	1
AOC NETWORK	100GbE 2-port QSFP56	1

Supermicro X13 Hyper System

Type	Description	System QTY
Barebone	X13 BigTwin 2U 2-Node with 12x 2.5” drives	1
CPU	5th Gen Intel Xeon 8568Y+	4
CPU	4th Gen Intel Xeon 8480+	4
MEMORY	64GB DDR5-5600	32
Drive	400GB NVMe PCIe 4.0 M.2 22x80mm	4
Drive	6.4TB NVMe PCIe 4.0 U.2	14
AIOM NETWORK	AIOM 2-port 10GbE	2
AOC NETWORK	100GbE 2-port QSFP56	2

Note: Only 1 Node in the BigTwin was used in this benchmark

Supermicro X13 BigTwin (2U 2N) System

Type	Description	System QTY
Barebone	X12 Hyper 2U with 24x 2.5"drives	1
CPU	3rd Gen Intel Xeon 8380	2
MEMORY	32GB DDR4	32
Drive	480GB NVMe PCIe 4.0 M.2 22x80mm	2
AIOM NETWORK	AIOM 4-port GbE	1
AOC NETWORK	100GbE 2-port QSFP56	1

Type	Description	Node QTY
Barebone	X12 BigTwin with 24x 2.5" drives	1
CPU	3rd Gen Intel Xeon 8380	2
MEMORY	64GB DDR4	16
Drive	1TB NVMe	1
Drive M.2	1TB NVMe M.2 22x80mm	1
AOC NETWORK	100GbE 2-port QSFP56	1

Note: Only 1 Node in the BigTwin was used in this benchmark

For More Information:

Supermicro X13 Systems – www.supermicro.com/x13

Supermicro X13 Hyper System Information: https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr

Supermicro X13 BigTwin System Information: https://www.supermicro.com/en/products/system/bigtwin/2u/sys-221bt-hntr

Intel AMX Description: https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html

Red Hat OpenShift https://www.redhat.com/en/technologies/cloud-computing/openshift

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