SOLUTION BRIEF

SUPERMICRO ENABLES ADVANCED ARTIFICIAL INTELLIGENCE TECHNOLOGIES ON VMWARE VSAN FOR INFERENCING

Innovative System Uses Supermicro X13 BigTwin ® Systems and Intel® Xeon® 4th Gen Scalable Processors with Intel® AMX

SupermicroX13 BigTwin Multi-Node Infrastructure Solutions

TABLE OF CONTENTS

Executive Summary

With the newfound popularity of AI solutions such as ChatGPT, enterprises are more motivated than ever to deploy AI use cases. By tapping into their rich data sources, enterprises seek to deliver deep real-time insights that improve operational efficiencies, lead to better product designs, improve customer satisfaction and employee productivity, and ultimately increase revenue streams.

Challenges:

Successful AI solutions begin with data and end with insights. Data is growing unchecked in both consumer segments and business segments. Consumers have increasing tools to create and share information, leading to data growth numbers that can hardly be comprehended. Businesses are doing their fair share to add to this mountain of data, hoping it can be leveraged to provide massive value as business intelligence. Even though data is in high supply, while insights can be elusive, that is not deterring enterprises from investing in AI. Worldwide spending on AI is expected to exceed $300 Billion by 2026. (According to IDC's Worldwide Artificial Intelligence Spending Guide, August 2022). As businesses look to cash in on the promise of AI, they must improve data management or how data is collected, classified, and secured so that decisions and actions can be taken in the organization's best interest.

The integration of AI and VMware vSAN is in its infancy. Developers and data practitioners are currently exploring how enterprises can use software-defined storage to utilize the valuable data stored in vSAN clusters for AI use cases. With a myriad of applications running in vSAN and the pervasive nature of AI, it's only a matter of time before advancements in vSAN, including greater scalability, improved performance, and better application management, combine with new AI tools and capabilities embedded in applications to offer the insights and innovation enterprises seek.

Value of vSAN & AI:

$300+ Billion

Global Spending on AI by 2026

VMware vSAN provides a simple path to hyper-converged infrastructure (HCI)

Compute Platform

- CPU is useful for inference, fine-tuning models, and training smaller models
- Application architectures deploy multiple instances of the application to scale as needed.
- Multi-threaded processes offer high levels of resource utilization (CPU, network, and memory)

Data Platform

- Storage clusters for large amounts of structured and unstructured data
- Backing hardware resources can be scaled as needed

Intel Advanced Matrix Extensions (Intel AMX) accelerates AI capabilities on 4th Gen Intel Xeon Scalable processors, speeding up deep learning training and inferencing without additional hardware. It is ideal for NLP, recommendation systems, and image classification and is supported out of the box in the most popular AI frameworks such as TensorFlow, PyTorch, and OpenVINO™

Supermicro X13 BigTwin for Efficiency and Performance

Investing in the Supermicro X13 BigTwin for Artificial Intelligence (AI) workloads is a strategic decision that can transform a business's operations. With its density, customers get the latest and greatest processing power of multiple nodes working simultaneously or individually to accelerate complex AI tasks like Natural Language Processing (NLP) and Image Classification. With the Supermicro X13 BigTwin, customers are unlocking possibilities to drive businesses to new heights.

The flexible options of 2U 4-Node and 2U 2-Node that the Supermicro X13 BigTwin offers enable enterprises to choose the most efficient configuration for their business needs. With its density, AI workloads demand immense computational resources, and the Supermicro X13 BigTwin delivers the computational power needed. By distributing the workloads across nodes, organizations can achieve remarkable speedups, reducing processing times from days to hours. This configuration means quicker insights, faster iterations, and a competitive edge in many industries. With the Supermicro X13 BigTwin power efficiency, by sharing resources in a single chassis, ensures that AI models reach their full potential while maximizing the return on the system investment.

As organizations' AI initiatives expand, teams won't face the headache of outgrowing their infrastructure, as this has the flexibility and scalability that AI needs. Adding nodes is a seamless process that adapts to an organizations evolving needs. Whether interested in deep learning, neural networks, or big data analysis, Supermicro X13 BigTwin future-proofs operations, allowing teams to tackle even the most demanding AI workloads.

Supermicro X13 BigTwin Systems

Highlights:

- Dual Socket 4th Gen Intel Xeon Scalable Processors Per Node
- Up to 4 Nodes per 2U
- 16 DIMM Slots per Node, Up to 4TB DDR5-4800 Memory
- Flexible Storage Options Including ALL Nvme and Hybrid Nvme/SAS3/SATA3 in 2.5” or 3.5”
- Flexible Networking Options from 1GBps to 400 Gbps

Innovations

- Supermicro X13 BigTwin flexible design with 2U 2-Node or 2U 4-Node
- Supermicro AIOM – Most Flexible, Cost-Optimized Server I/O
- Resource Sharing for Best Efficiency and TCO

Intel AMX supports two data types, INT8 and BF16, for the matrix multiplication required for AI workloads:

- INT8 is a data type used for inferencing when the precision of FP32, a single-precision floating-point format often used in AI, isn’t needed. Because the INT8 data type is lower precision, more INT8 operations can be processed per compute cycle.
- BF16 is a data type that delivers sufficient accuracy for most training. It can also deliver higher accuracy for inferencing if needed.

AI Benchmarks

Image Classification – ResNet50

This benchmark has become a standard benchmark in the deep learning community for evaluating the performance of image classification models. The Supermicro X13 BigTwin is powered by Intel 4th Gen Xeon Scalable Processors with AMX accelerators brings TCO benefits for certain AI workloads. Enterprises can make informed decisions, increase efficiency, and provide innovative solutions for their customers.

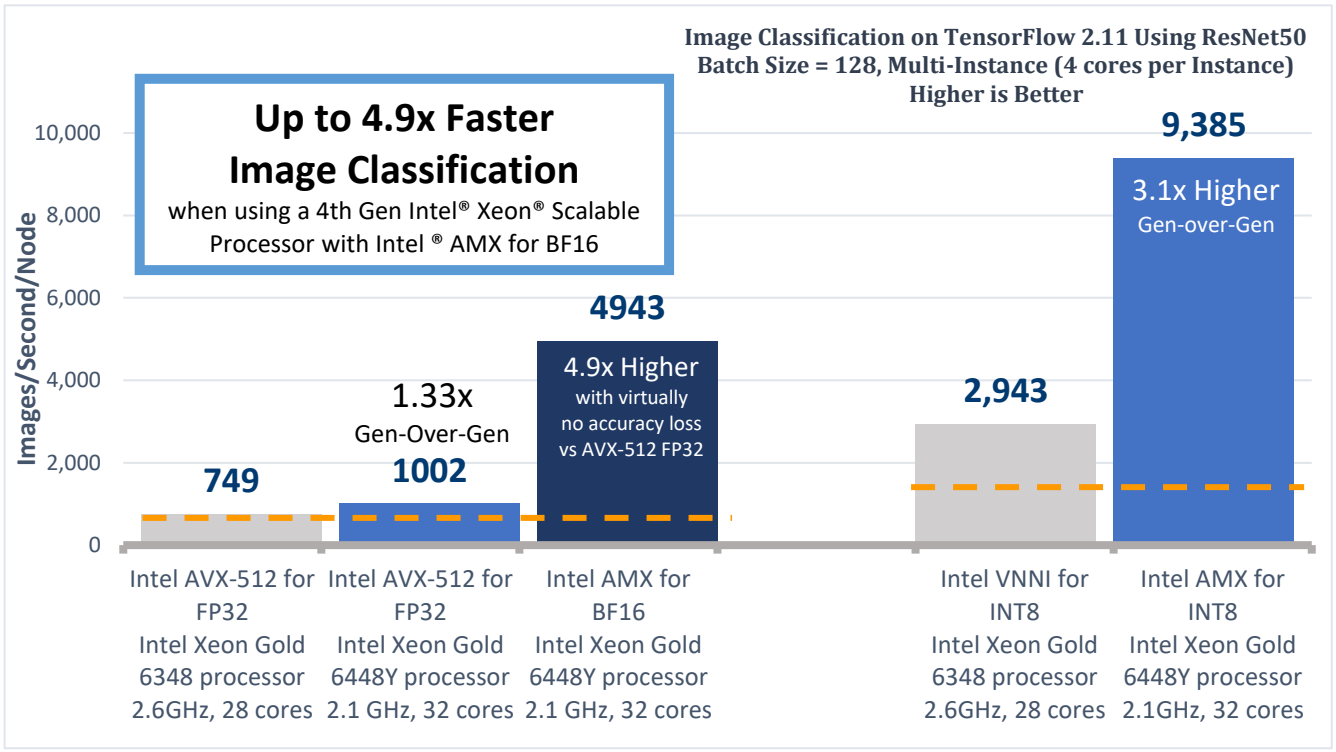

Figure 1 – Image Classification Performance Compared to the Previous Generation

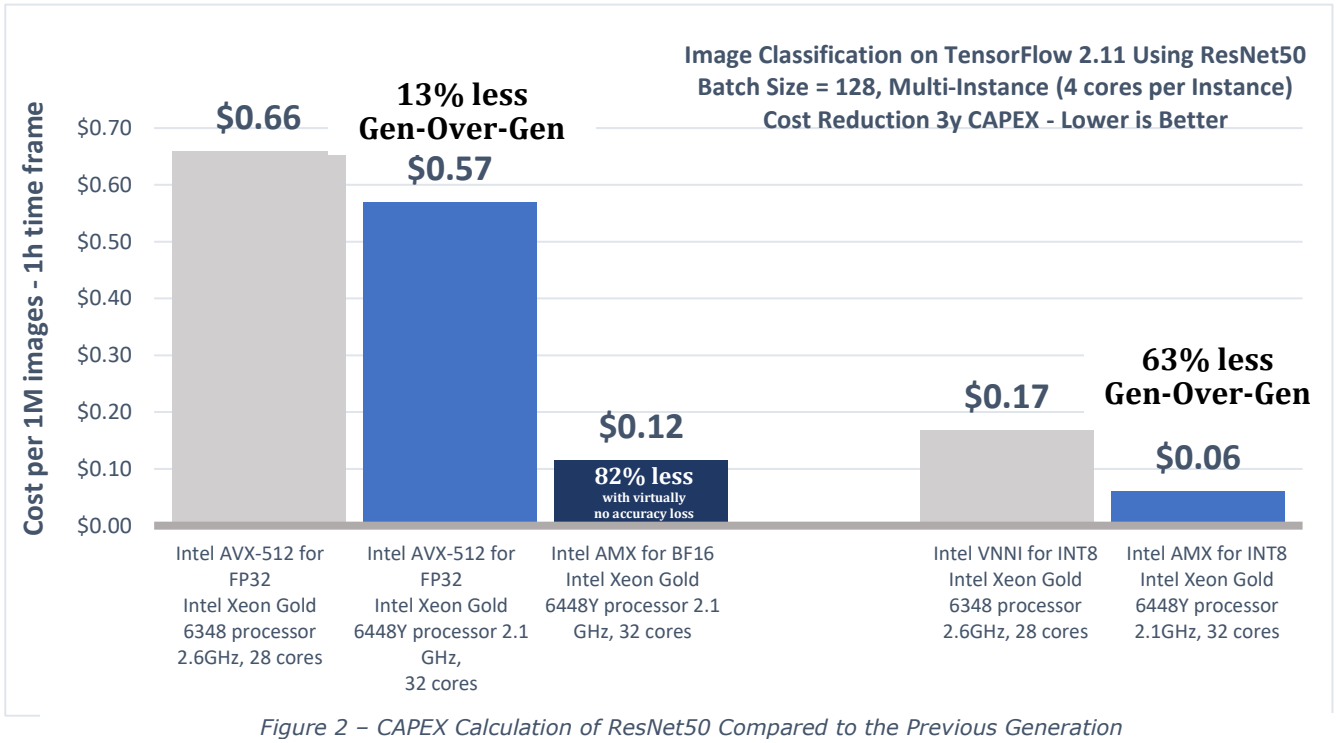

Figure 2 – CAPEX Calculation of ResNet50 Compared to the Previous Generation

Natural Language Processing – BERT-Large

This benchmark holds significant importance due to its pivotal role in evaluating and advancing language understanding models. BERT Large is derived from the BERT model capable of handling more complex language tasks and larger datasets. It can help enterprises draw insights for large and various data sources to provide personalized experiences that help increase

user engagement. The Supermicro X13 BigTwin with up to 4-Node density in a 2U form factor is well-suited for running large and complex models with the help of the Intel AMX accelerator.

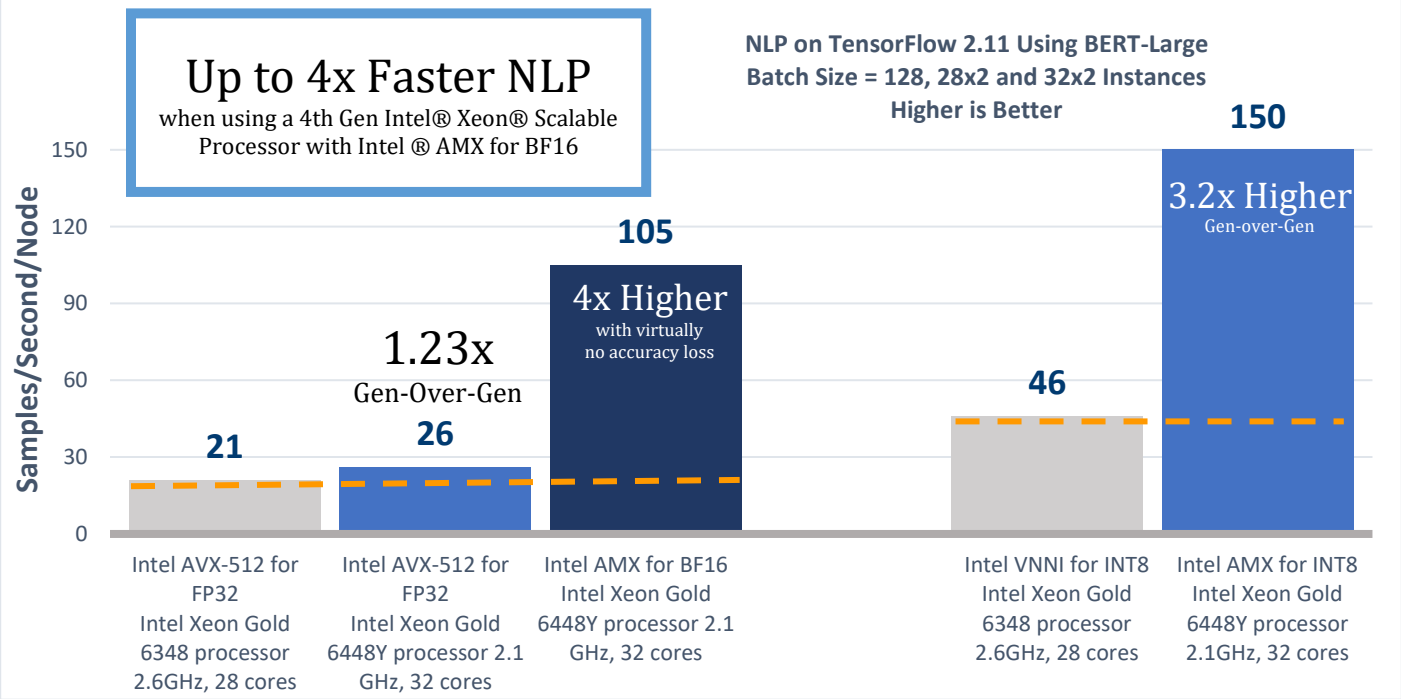

Figure 3 – Natural Language Processing Performance Compared to Previous Generation

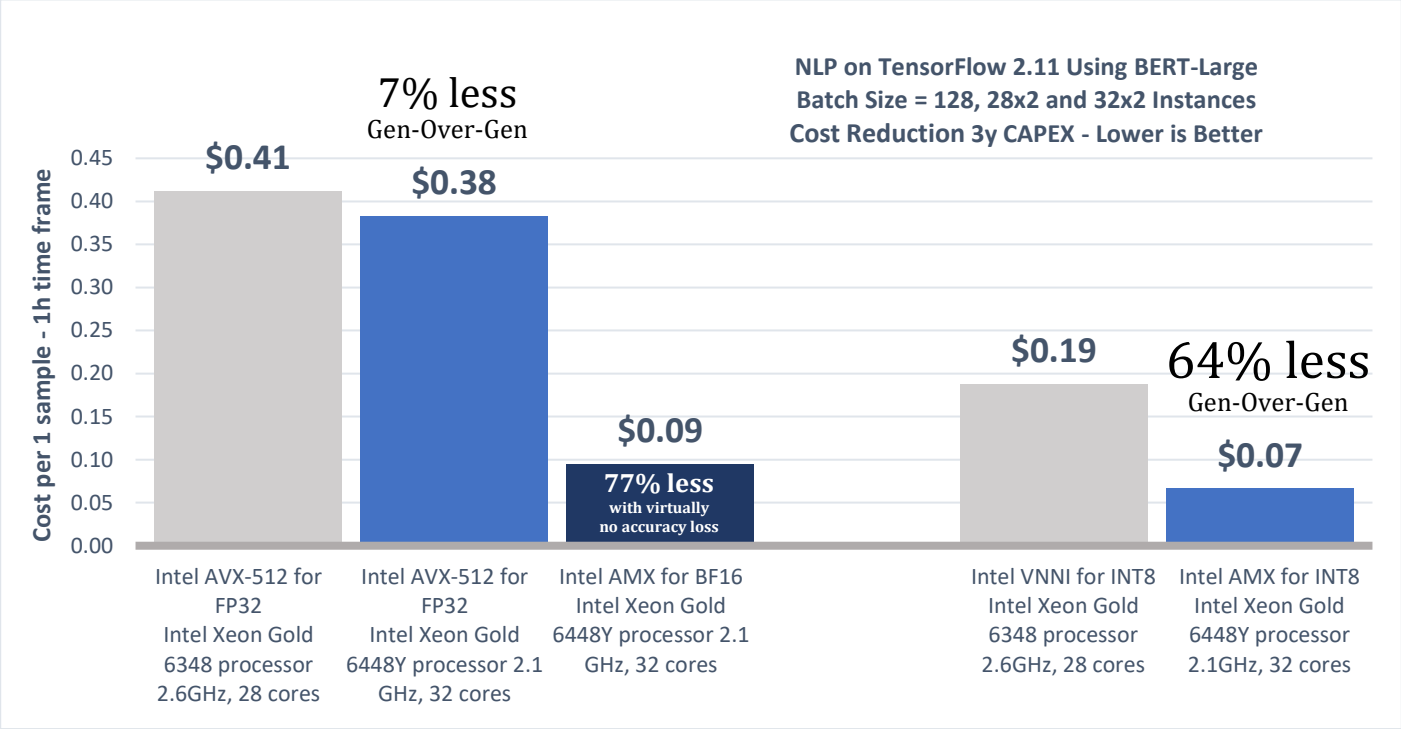

Figure 4 – CAPEX Calculation of BERT-Large Compared to Previous Generations

Configurations:
BASELINE: Intel Xeon Gold 6348 (ICX Config): 4-node cluster, Each node: 2x Intel® Xeon® Gold 6348 Processor, 1x Server Board M50CYP2UR, Total Memory 512 GB (16x 32GB DDR4 3200MHz), HyperThreading: Enable, Turbo: Enabled, NUMA noSNC, Intel VMD: Enabled, BIOS: SE5C620.86B.01.01.0008.2305172341(ucode:0xd000390), Storage (boot): 2x 80 GB Intel SSD P1600X, Storage (flat): 9x 3.84 TB Intel SSD DC P5510 Series PCIe NVMe , Network devices: 1x Intel Ethernet E810CQDA2 E810-CQDA2, at 100 GbE RoCE, Network speed: 100 GbE, OS/Software: VMware/vSAN 8.0U1, 21495797, Test by Intel as of 07/04/2023 using Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN ESA – Optimal default policy (RAID-5), Kernel 5.15,

intel-optimized-tensorflow:2.11.0, ResNet50v1.5, Batch size=128, VM=56vCPU+64GBRAM, Multi instance scenario (4 cores per instance), BERT-Large, SQuAD 1.1, Batch size=128, VM=56vCPU+64GBRAM
Intel Xeon Gold 6448Y (4th Gen Config): 4-node, 2x Intel® Xeon® Gold 6448Y, 4x SYS-221BT-DNTR X13DET-B, Total Memory 1024 GB (16x DDR5 64GB 4800MHz), HyperThreading: Enable, Turbo: Enabled, NUMA noSNC, Intel VMD: Enabled, BIOS: 1.3(ucode:0x2b000461), Storage (flat): 7x 6.4 TB Solidigim P5620 Series PCIe NVMe , Network devices: 1x AOC-S100GC-I2C (Intel E810-CAM2), at 100 GbE RoCE, Network speed: 100 GbE, OS/Software: VMware 8.0U1, 21495797 (vSAN in ESA mode), Test by Intel as of 6/23/2023 using Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN ESA – Optimal default policy (RAID-5), Kernel 5.15, intel-optimized-tensorflow:2.11.0, ResNet50v1.5, Batch size=128, VM=64vCPU+64GBRAM, Multi instance scenario (4 cores per instance), BERT-Large, SQuAD 1.1, Batch size=128, VM=64vCPU+64GBRAM

Conclusion

The convergence of AI solutions and enterprise aspirations has propelled the drive toward AI development. The exponential growth of data both in customer and business realms, presents challenges, but enterprises remain with the need for more AI solutions, evidenced by the substantial investments in AI. As the world anticipates AI spending to surpass $300 Billion by 2026, the critical role of effective data management becomes paramount and Supermicro X13 BigTwin powered by 4th Gen Xeon Scalable Processors can bring added value for this AI investments.

The integration of AI with Vmware vSAN is still in its early stages, with developers and data practitioners exploring ways to harness the potential of software-defined storage within the vSAN cluster for AI applications. The synergy between vSAN advancements and emerging AI capabilities promises to deliver the required insights and innovation to enlightened enterprises. Vmware vSAN's role as a pathway to hyper-converged infrastructure cannot be understated. Its compute and data platforms provide the necessary foundation for AI operations, offering scalable and efficient resources for various AI workloads. The Supermicro X13 BigTwin stands out as a strategic investment for AI workloads, providing the processing power and flexibility needed for complex tasks like Natural Language Processing and Image Classification which are used within inferencing, fine-tuning models, and training smaller models. The configuration options of the Supermicro X13 BigTwin, coupled with its power efficiency and scalability, ensure that organizations can meet the demands of expanding AI initiatives without outgrowing their infrastructure.

References

To learn more about VMware-vSAN solutions of Supermicro, visit

- Supermicro Vmware-vSAN solutions

To learn more about Supermicro Server Infrastructure options, please visit:

- Supermicro Data Center Server, Blade, Data Storage, AI System
- Twin Servers: High-Density Multi-Node Server Solutions | Supermicro
- Networking Devices & Hardware Products For HPC | Supermicro
- Supermicro Solution for Red Hat® OpenStack | Supermicro
- AI Infrastructure Solutions | Supermicro

Sources

1) IDC Worldwide Artificial Intelligence Spending Guide, August 2022
2) Grand View Research, "Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution, By Technology (Deep Learning, Machine Learning, Natural)"

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.

Visit www.supermicro.com

INTEL

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

Visit www.supermicro.com