


Accelerate AI Workloads on VMware vSphere® and VMware vSAN™ Using 4th Gen Intel® Xeon® Scalable Processors with Intel® AMX

Business Challenge: Are you getting the best possible business value from your data? vSAN clusters are home to increasingly large and valuable data stores. Users can take advantage of AI capabilities that are built into many enterprise applications and analytics tools to help accelerate innovation, improve customer experiences, and optimize operations.



Up to
5x Faster¹
and Still Accurate
Image Classification
using BF16 compared to FP32

Solution Overview and Summary

Solution: Intel and VMware collaborated to provide significant performance improvements, underscoring the need for ongoing data center modernization. VMware vSAN™ includes the new Express Storage Architecture™, an optional, alternative architecture that can process and store data with elevated levels of efficiency, scalability, and performance. When combined with 4th Generation Intel® Xeon® Scalable processors, vSphere and vSAN software can power the most demanding workloads.

While the latest generation of Intel® hardware and VMware software together support a variety of applications, the benefits for AI workloads like image classification and natural language processing (NLP) are particularly compelling. The built-in AI acceleration from Intel® Advanced Matrix Extensions (Intel® AMX) is specifically designed to provide massive speedup to the low-precision math operations that underpin AI inference. Mainstream applications already running on vSAN and Intel Xeon processors—such as databases, analytics, business-critical and collaboration applications, and IT automation tools—are being enhanced with AI algorithms and can benefit from Intel AMX. The result is a completely optimized pipeline on a single hardware and software platform that can scale from data center to cloud to edge. Scale AI everywhere by using the broad, open software ecosystem and unique Intel tools. Customers can utilize their large and valuable vSAN data store on standard Intel Xeon processor-based servers, while gaining the efficiency and performance of a built-in AI accelerator.

Intel AMX supports INT8 and BF16 data types, augmenting the optimizations from Intel® Advanced Vector Extensions 512 and Intel® Deep Learning Boost, to enable fast and efficient AI and deep learning across a broad range of industries and use cases.

Results: The testing highlighted in this document demonstrates the performance benefits of vSphere and vSAN on 4th Gen Intel Xeon Scalable processors with Intel AMX, compared to previous generations of hardware, as well as the speedup provided by using BF16 versus FP32. For example, image classification and NLP are both about 3x faster gen-over-gen at INT8 precision. Additionally, using BF16 instead of FP32 delivers 4.1x to 5x faster inference, with nearly no loss in precision.¹ See the next page for a full discussion of testing results.

“Customers can utilize their large and valuable vSAN data store on standard Intel Xeon processor-based servers, while gaining the efficiency and performance of a built-in AI accelerator.”

Key Components

Intel® Optimization for TensorFlow 2.11 with the Intel® oneAPI Deep Neural Network Library (oneDNN). This library can natively take advantage of Intel® AMX instructions to accelerate AI workloads.

Pre-built containers and validated VMs from [Model Zoo](#), a publicly available repository created by Intel and now available on GitHub. In particular, we used:

- [ResNet50 containers](#) on Docker Hub
- [BERT-Large containers](#) on Docker Hub

Test Methodology

The testing methodology included tests for NLP use cases (using the BERT-Large model) and object detection and image classification use cases (using the ResNet-50v1.5 model). We benchmarked these two models for AI inference performance at various precisions (FP32, BF16 and INT8). See the “ ” section for more details on how the tests were configured.

FP32 is a standard 32-bit floating-point data type used to train deep learning models and for inferencing. This data type is more computationally demanding than other data types, but typically achieves higher accuracies. BF16 is a truncated version of FP32 and is used for both training and inference. It offers similar accuracy but faster computation. INT8 offers higher performance and is the least computationally demanding data type, which is ideal for constrained environments. It has minimal impact on accuracy.

We experimented with several batch sizes before determining 128 was ideal. See the sidebar for a list of key components of the solution.

Results

Many deep learning workloads are mixed-precision and 4th Gen Intel Xeon Scalable processors can seamlessly transition between Intel AMX and Intel AVX-512 to use the most efficient instruction set. Intel publicly distributes pre-optimized versions of the most popular deep learning models, which are compatible with Intel® processors at various precisions. Intel also publishes best known methods for achieving the highest performance on Intel® architecture, including this solution design brief.

The table below summarizes the notable results. Figures 1 and 2 illustrate the gen-over-gen and precision comparisons for image classification and NLP. BF16 precision (not available on older processors) can provide up to 5x faster performance with virtually no loss of accuracy compared to FP32¹. Running on newer processors, AI is up to 3.2x faster. If time to insight matters to your business, it’s time to upgrade both your hardware and software.

Speedup by Model Type	Image Classification	Natural Language Processing (NLP)
Gen-Over-Gen FP32	1.23x	1.24x
Gen-Over-Gen INT8	3x	3.2x
Same-Gen BF16 vs. FP32 (almost no loss in accuracy)	5x	4.1x

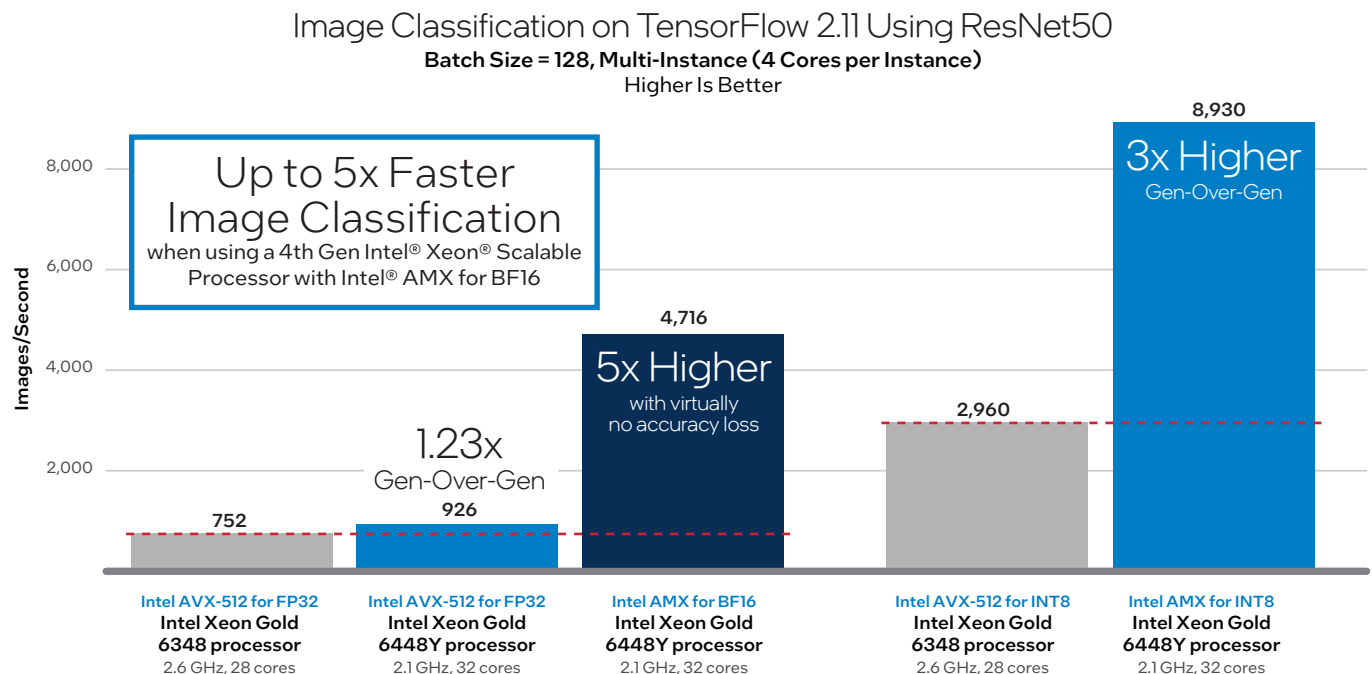


Figure 1. The latest generation of Intel® Xeon® Scalable processor with Intel® AMX delivers up to 3x higher INT8 throughput for image classification inference (compared to older hardware). On the same processors, BF16 provides up to 5x faster inference than FP32 with nearly the same accuracy.¹

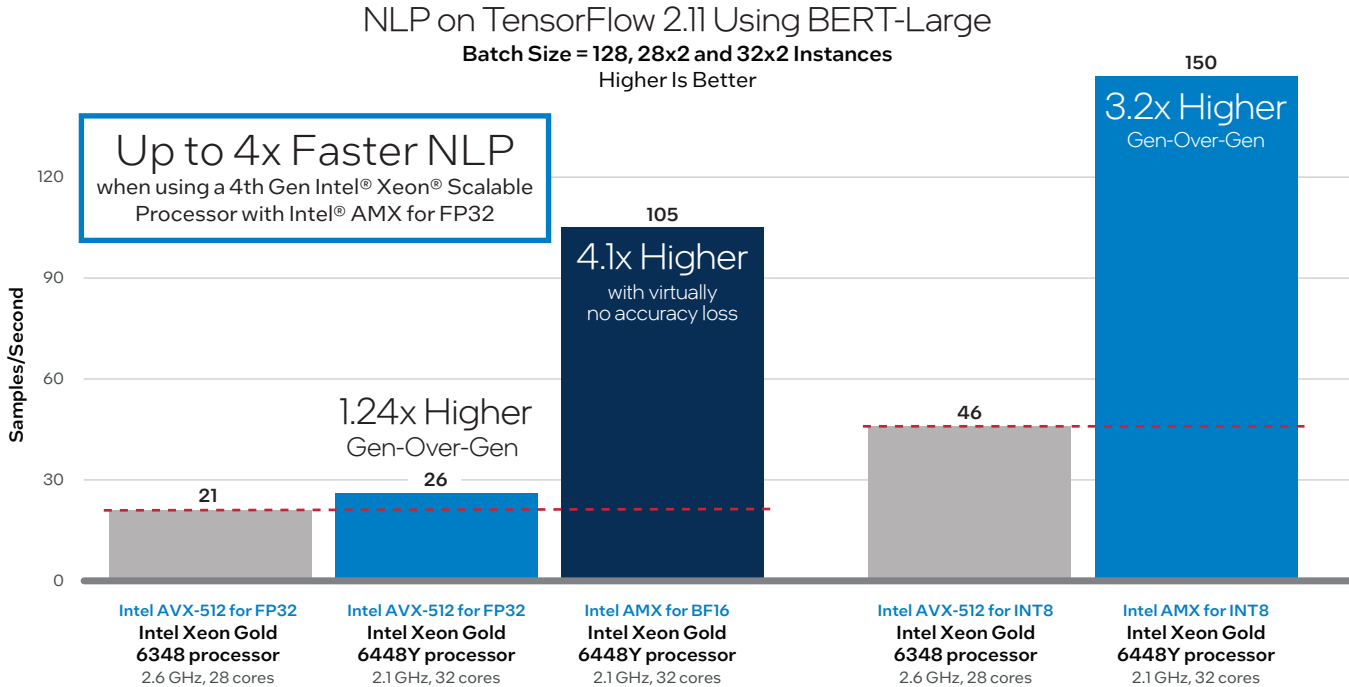


Figure 2. The latest generation of Intel® Xeon® Scalable processor with Intel® AMX delivers up to 3.2x higher INT8 throughput for NLP inference (compared to older hardware) and up to 4.1x higher throughput than FP32.¹

Configuration Details

The following tables provide information about components and settings of the infrastructure used during testing. These tables reflect typical types of configurations used for each generation of processor.

	3rd Gen Intel® Xeon® Scalable Processor Configuration	4th Gen Intel Xeon Scalable Processor Configuration
Server	4x Intel® Server Board M50CYP2UR	4x Intel Server Board M50FCP2SBSTD
Processor	2x Intel Xeon Gold 6348 processor per node (28 cores, 2.0 GHz base/2.6 GHz All-Core Turbo ^a)	2x Intel Xeon Gold 6448Y processor per node (32 cores, 2.1 GHz base/3.0 GHz All-Core Turbo ^a)
Memory	512 GB (16x 32 GB DDR4 3,200 MHz)	512 GB (16x 32 GB DDR5 4,800 MHz)
Storage (ESA Flat Tier)	9x Solidigm D7-P5510 SSD 3.84 TB	
Network Card	1x Intel® Ethernet Network Adapter E810-CQDA2, Dual-Port at 100 GbE with RDMA using RoCEv2	
Network Switch	Arista DCS-7170-64C	

^a All-Core Turbo Boost frequency range: Intel® Turbo Boost Technology automatically provides opportunistic performance improvement by allowing individual cores to operate at a higher frequency. This results in higher frequency in both single- and multi-threaded applications when headroom is available. For more information on the All-Core Turbo Boost frequency range, see section 2.1 and 2.3 at <https://builders.intel.com/docs/networkbuilders/power-management-technology-overview-technology-guide.pdf>.

Software Versions (same for both configurations)	
Hypervisor	vSphere ESXi 8.0GA, 20513097
vSAN Mode	ESA, optimal default policy RAID-5
Benchmark Tools	Intel® Optimization for TensorFlow 2.11

Important System Settings	
Number of Nodes	4
Power and Performance Policy	Performance
Frequency Governor	Native (OS control)
Max C-State	c0/c1
Prefetchers	L2 HW, L2 Adj., DCU HW, DCU IP
NUMA	Enabled (no sub-NUMA clustering)
IRQ Balance	Enabled
RDMA	Enabled

Accelerator Technologies Enabled	
Intel® Speed Select Technology Performance Profile 2.0 ^a	
Intel® Advanced Matrix Extensions	
Intel® Deep Learning Boost	
Intel® Advanced Vector Extensions 512	
Intel® Hyper-Threading Technology	
Intel® Turbo Boost Technology	

^a Refer to <https://builders.intel.com/docs/networkbuilders/power-management-technology-overview-technology-guide.pdf>.

Profiles and Workloads

Residual Network (ResNet) is a popular deep learning model for image recognition. We selected ResNet50v1.5 for our benchmarking and ran the tests using synthetic data. Bidirectional Encoder Representations from Transformers (BERT-Large) is a transformer model that is pretrained on BookCorpus and English Wikipedia data in a self-supervised fashion. BERT has become the ubiquitous baseline for handling various NLP tasks. We used pretrained BERT-Large for inferencing with the Stanford Question Answering Dataset (SQuAD) to measure performance.

To prepare the benchmark, we used a single VM with 56 vCPUs on the 3rd Gen Intel Xeon Scalable processor system and 64 vCPUs on the 4th Gen Intel Xeon Scalable processor system. Intel® Hyper-Threading Technology was enabled. The OS was Ubuntu 22.04, with Docker installed. We used a native (OS control) CPUFreq governor and set the BIOS CPU settings to Performance. Tests were conducted in containers downloaded from the relevant repository. Refer to the [ResNet50 documentation](#) and the [BERT-Large documentation](#) for all necessary steps. The following tables provide details about the various test configurations.

ResNet50 Configurations

	3rd Gen Intel® Xeon® Scalable Processor	4th Gen Intel Xeon Scalable Processor
Tested Precisions	INT8, FP32	INT8, BF16, FP32
Instances	14	16
OMP_NUM_THREADS ^a	4	4

^a This is the number of threads per instance. Multiplying the number of instances by the number of threads results in the number of physical cores on a server.

BERT-Large Configurations

	3rd Gen Intel Xeon Scalable Processor	4th Gen Intel Xeon Scalable Processor
Tested Precisions	INT8, FP32	INT8, BF16, FP32
Instances	2	2
OMP_NUM_THREADS ^a	28	32

^a This is the number of threads per instance. Multiplying the number of instances by the number of threads results in the number of physical cores on a server.

Conclusion

Running your AI workloads on vSphere and vSAN with 4th Gen Intel Xeon Scalable processors can improve both time to insight and time to market for AI solutions (compared to previous-generation Intel processors). Intel and VMware worked together to ensure that the built-in acceleration of AI computation provided by Intel AMX works out-of-the-box, without any vSAN configuration changes. Plus, many pretrained AI models and container images are available from Intel to the general public. These resources are compatible with 4th Gen Intel Xeon Scalable processors and Intel AMX and include support for a variety of precision levels. The solution highlighted in this brief consists of a pre-validated hardware and software configuration, benchmarked and tested so you can run your enterprise AI workloads on VMware with confidence.

Further Information

- [4th Gen Intel® Xeon® Scalable processors](#)
- [Intel® Ethernet 800 Series](#)
- [vSphere](#)
- [vSAN with ESA](#)
- [ResNet50 model inference execution guidance on Model Zoo](#)
- [BERT-Large model inference execution guidance on Model Zoo](#)

Authors

Ewelina Kamyszczek, Cloud Solutions Engineer
DCAI/CESG Intel Group

Patryk Wolsza, Cloud Solutions Architect, vExpert
DCAI/CESG Intel Group



Learn more about the
[Intel and VMware Partnership](#)
and [Data Center solutions](#).



Contact your Intel
representative to learn more
about this solution.

Solution Provided By:



3rd Gen Intel® Xeon® Scalable platform configuration: Test by Intel as of March 2023. 4-node cluster. Each node: 2x Intel® Xeon® Gold 6348 processor, 1x Intel® Server Board M50CYP2UR, total memory 512 GB (16x 32 GB DDR4 3200 MHz), Intel® Hyper-Threading Technology = enabled, Intel® Turbo Boost Technology = enabled, NUMA enabled noSNC, Intel® Volume Management Device (Intel® VMD) = enabled, BIOS: SE5C620.86B.01.01.0006.2207150335 (microcode:0xd000375), storage (boot) = 2x 80 GB Intel® Optane™ SSD P1600X, storage (ESA flat tier): 9x 3.84 TB Solidigm D7-P5510 SSD, network devices: 1x Intel® Ethernet E810-CQDA2, FW 4.0, at 100 GbE with RDMA using RoCEv2, network speed: 100 GbE, OS/Software: vSphere/vSAN 8.0, 20513097, Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN ESA – optimal default policy RAID-5, Kernel 5.15, Intel® Optimization for TensorFlow 2.11.0, ResNet50v1.5, batch size=128, VM=56 vCPU+64 GB RAM, multi-instance scenario (4 threads per instance), BERT-Large, SQuAD 1.1, batch size=128, VM=56 vCPU+64 GB RAM, multi-instance scenario (28 threads per instance).

4th Gen Intel Xeon Scalable platform configuration: Test by Intel as of March 2023. 4-node cluster. Each node: 2x Intel Xeon Gold 6448Y processor QS pre-production, 1x Intel Server Board M50FCP2SBSTD, total memory 512 GB (16x 32 GB DDR5 4800 MHz), Intel Hyper-Threading Technology = enabled, Intel Turbo Boost Technology = enabled, NUMA enabled noSNC, Intel VMD = enabled, BIOS: SE5C741.86B.01.01.0002.2212220608 (microcode:0x2b000161), storage (boot) = 2x 240 GB Solidigm D3-S4520 SSD, storage (ESA flat tier): 9x 3.84 TB Solidigm D7-P5510 SSD, network devices: 1x Intel Ethernet E810-CQDA2, FW 4.0, at 100 GbE with RDMA using RoCEv2, network speed: 100 GbE, OS/Software: vSphere/vSAN 8.0, 20513097, Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN ESA – optimal default policy RAID-5, Kernel 5.15, Intel Optimization for TensorFlow 2.11.0, ResNet50v1.5, batch size=128, VM=64 vCPU+64 GB RAM, multi-instance scenario (4 threads per instance), BERT-Large, SQuAD 1.1, batch size=128, VM=64 vCPU+64 GB RAM, multi-instance scenario (32 threads per instance).

Performance varies by use, configuration, and other factors. Learn more on the [Performance Index](#) site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.

Intel and the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries.

Other names and brands may be claimed as the property of others. © Intel Corporation 1123/JCAP/KC/PDF