Solution Snapshot

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

The Challenge

Data is the most valuable asset of any organization, but deriving value from that data is becoming more difficult as those organizations wrestle with ever-increasing amounts of data from diverse sources. It's a constant struggle to properly discover, classify, analyze, and protect data so that it can be best utilized to drive improved insights and outcomes for the business.

While artificial intelligence (AI) is an emerging use case on vSAN, the speed with which AI and deep learning (DL) are evolving means they will soon be built into nearly every enterprise application and analytics tool. Today's vSAN users have increasingly large and valuable stores of data in their vSAN clusters, and using these new AI tools and capabilities will be essential for leveraging the data on those clusters to accelerate innovation, improve customer experiences, and optimize operations.

Use Cases:

- Infrastructure Modernization
- **AI Workloads with Intel® Advanced Matrix Extensions**
- SQL Server
- Virtual Desktop Infrastructure (VDI)

Why Run AI Workloads on VMware vSAN?

Any AI Code, Every Workload
AI is now being weaved into all of your modern business applications, from CRM and finance applications to security and infrastructure tools. 4th Gen Intel® Xeon® processors feature Intel® Advanced Matrix Extensions (AMX), which gives your AI-enabled apps the ability to deliver flexible and efficient performance.

Build and Deploy Everywhere
Design and deploy AI projects quickly and efficiently with optimized training and inferencing. Intel® AMX brings extensive hardware and software optimizations to:

- Accelerate INT8 and BF16 data types
- Augment optimizations from Intel® AVX-512 and Intel® DL Boost in previous generations
- Enable fast and efficient AI for a range of use cases, including video analytics, industrial machine vision, and natural language processing

Data infrastructure is already optimized for effective ingest using Intel® technology. With the built-in AI accelerator, the result is a completely optimized pipeline on a single hardware and software platform that scales from data center to cloud to edge. Customers can scale AI everywhere by leveraging the broad, open-software ecosystem and unique Intel® tools.

Implement Pre-Built Solutions
Extensive Intel® AI product options and partnerships can help accelerate time to insights. As a result, Intel is uniquely positioned to provide an ecosystem that caters to the demands of the ever-expanding frontiers of artificial intelligence.

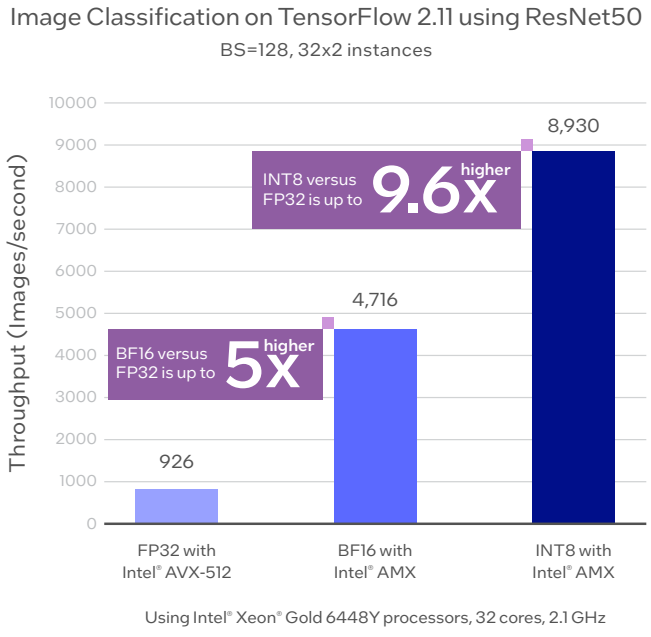
Customers tend to use the mainstream distributions of the most popular AI frameworks and tools. To simplify the use of these new accelerator engines and extract the best performance, Intel's AI experts have been working for years with the AI community to co-develop and optimize a broad range of open and free-to-use tools, optimized libraries, and industry frameworks to deliver the best out-of-the-box performance and end-to-end productivity.

Proof Points

Intel can run and accelerate a range of AI workloads, including image classification, speech recognition, language translation, and object detection. To illustrate the capabilities, we benchmarked both image classification and natural language processing (NLP) workloads running on vSAN.

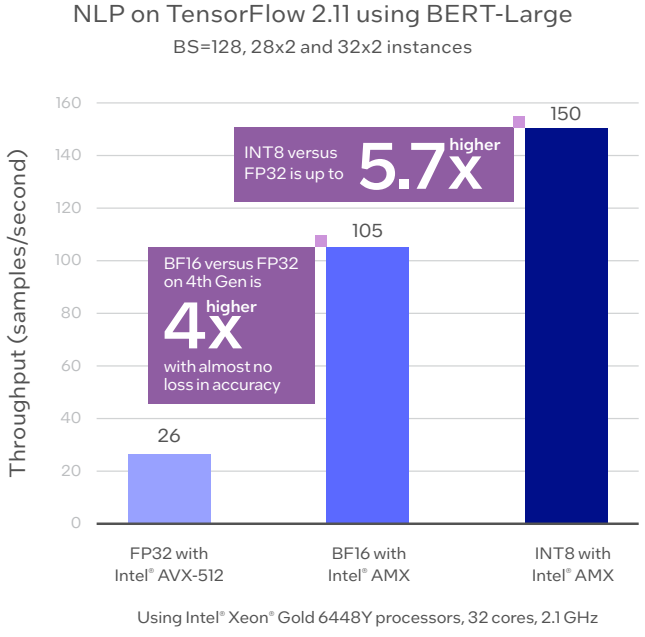
- FP32 is a standard 32-bit floating point data type used to train DL models and for inferencing — more computationally demanding but typically achieving higher accuracy
- Bfloat16 is a truncated version of 32-bit floating point, used for both training and inference, offering similar accuracy but faster computation
- INT8 offers higher performance and is least computationally demanding for constrained environments, with minimal impact on accuracy
- Many DL 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

Image Classification on vSphere/vSAN 8.0 Using a 4th Gen Intel® Xeon® Scalable Processor with Intel® AMX[1]



Using Intel® Xeon® Gold 6448Y processors, 32 cores, 2.1 GHz

Natural Language Processing (NLP) on vSphere/vSAN 8.0 Using a 4th Gen Intel® Xeon® Scalable Processor with Intel® AMX[1]



Using Intel® Xeon® Gold 6448Y processors, 32 cores, 2.1 GHz

1 See Configurations for details.

4th Gen Intel® Xeon® Scalable platform configuration: 4-node cluster, Each node: 2x Intel® Xeon® Gold 6448Y Processor QS pre-production, 1x Server Board M50FCP2SBSTD, Total Memory 512 GB (16x DDR5 32GB 4800MHz), HyperThreading: Enable, Turbo: Enabled, NUMA noSNC, Intel® VMD: Enabled, BIOS: SE5C741.86B.01.01.0002.2212220608 (ucode:0x2b000161), Storage (boot): 2x240GB Solidigm S4520, Storage (data): 6x 3.84 TB Solidigm SSD DC P5510 Series PCIe NVMe, Network devices: 1x Intel® Ethernet E810CQDA2 E810-CQDA2, FW 4.0, at 100 GbE RoCE v2, Network speed: 100 GbE, OS/Software: VMware/vSAN 8.0, 20513097, Test by Intel as of 03/13/2023 using Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN ESA – Optimal default policy (RAID-5, flat), Kernel 5.15, Intel® Optimization for 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

Benefits of AI Workloads on vSAN with Intel Technologies

- **Leverage the valuable stores of data already on your vSAN clusters**
- **Mainstream apps already running on vSAN are being augmented with AI algorithms**
- **The flexibility of a standard Intel® Xeon® Scalable processor server, with the efficiency and performance of a built-in AI accelerator**

Want More Information?

Contact your account representative for more information on Intel and VMware vSAN AI Solutions. VMware Workloads

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

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 + vmware®