Solution Design Brief

Data Centers | AI Workloads with Intel® Advanced Matrix Extensions
Intel Accelerated Solution

Accelerate AI Inferencing Workloads and Boost Security on Azure Stack HCI with Intel® AMX and Intel® TME

Business Challenge: How can organizations run flexible and efficient AI on their hyperconverged infrastructure (HCI) while maintaining security?

Microsoft Azure

Up to
9.3x
Boost in AI Inferencing
With Intel® AMX

Solution Overview and Summary

Solution: Enterprises that are modernizing their data centers and edge environments with hyperconverged infrastructure (HCI) are now turning to AI to shorten the time to value from their data. Faster AI inferencing fuels real-time insights that can transform their businesses. A modern, hybrid cloud model with built-in security helps cut costs and simplify management.

Microsoft Azure Stack HCI supports hosting virtualized workloads on-premises, enables seamless integration with Azure services and delivers unified management across the hybrid cloud environment with Azure Arc. Azure Stack HCI enables customers to improve cloud efficiencies while modernizing on-premises and edge infrastructure. Additionally, enterprises can design and deploy AI projects quickly and efficiently with optimized training and inferencing.

This solution brief shows how companies can use Azure Stack HCI and technologies that are built into 4th Generation Intel® Xeon® Scalable processors to accelerate AI workloads and boost data security. Specifically, Intel® Advanced Matrix Extensions (Intel® AMX) increases AI inference throughput, while Intel® Total Memory Encryption (Intel® TME) encrypts a computer's entire memory system.

Inferencing at the edge or in the data center is an intensive operation. Extensive hardware and software optimizations that enable AI everywhere without adding more compute resources are delivered by Intel AMX, including the following:

- Accelerate INT8 and BF16 data types.
- Augment optimizations from Intel® Advanced Vector Extensions 512 (Intel® AVX-512) and Intel® Deep Learning Boost (Intel® DL Boost) from previous generations of Intel Xeon Scalable processors.
- Enable fast and efficient AI for a range of use cases, including video analytics, industrial machine vision and natural language processing (NLP).

Although running AI at the edge can help speed time to insight and decrease network bandwidth costs, the edge environment may not have the same security options that are available in the centralized data center. This solution uses Intel TME and the optional multi-key version—Intel® Total Memory Encryption–Multi-Key (Intel® TME-MK)—to encrypt VM memory and processes to protect data.

Results: Testing reveals that using this solution can boost AI inferencing by up to 9.3x, depending on the AI model, precision and cluster size.[1] See the Results section on page 2 for a full discussion of the test results.

Test Methodology

The testing was conducted using two AI models: ResNet50 for image classification and BERT-Large for NLP. We tested three precisions: FP32 with Intel AVX-512, Bfloat 16 (BF16) and INT8 with Intel AMX. Intel TME was enabled for all tests.[2] We tested a two-node cluster, which is ideal for edge deployments, and a four-node cluster that is a typical starting size for data center deployments.

Note that although the tests ran on the two-node and four-node clusters, the AI workload itself ran on one node for simplicity's sake. (We expect linear scaling across multiple VMs; the relative performance would remain the same.) A single VM was created with the same number of virtual cores as physical cores in the node. Two instances of the workload were started, with each instance locked to one of the sockets in the node.

Precisions Used for Testing

- **FP32** is a standard 32-bit floating point data type used to train AI models and for inferencing—more computationally demanding but typically achieves higher accuracy than other precisions.
- **BF16** is a truncated version of FP32, used for both training and inference, that offers similar accuracy but faster computation.
- **INT8** offers higher performance and is least computationally demanding for constrained environments, with minimal impact on accuracy.

Many AI 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.

Results

Overall, testing results prove that AI throughput is significantly increased by using a lower precision (such as BF16 or INT8) and a more powerful processor. For example, ResNet50 inferencing throughput is measured by the number of images processed per second. Figure 1 illustrates the increase in ResNet50 throughput when using Intel Xeon Scalable processors with Intel AMX:

- For a 4th Gen Intel Xeon Gold processor, dropping precision to INT8 enables up to an 8.9x increase in ResNet50 throughput compared to FP32 on the same Gold processor with minimal impact to accuracy.[3]
- A Platinum-level 4th Gen Intel Xeon processor delivers up to a 9.3x increase at INT8 for the same workload, compared to FP32 on that same Platinum processor. That is, the higher-level SKU provides more speedup.[3]

Testing with BERT-Large yielded similar results. BERT-Large inferencing throughput is measured by the number of samples processed per second. Figure 2 illustrates the increase in BERT-Large throughput when using Intel Xeon Scalable processors with Intel AMX:

- For a 4th Gen Intel Xeon Gold processor, dropping precision to INT8 enables up to a 6.9x increase in BERT-Large throughput, compared to FP32 on the same Gold processor with minimal impact to accuracy.[3]
- A Platinum-level 4th Gen Intel Xeon processor delivers up to a 5.7x increase at INT8 for the same workload, compared to FP32 on that same Platinum processor. That is, the higher-level SKU provides more speedup.[3]

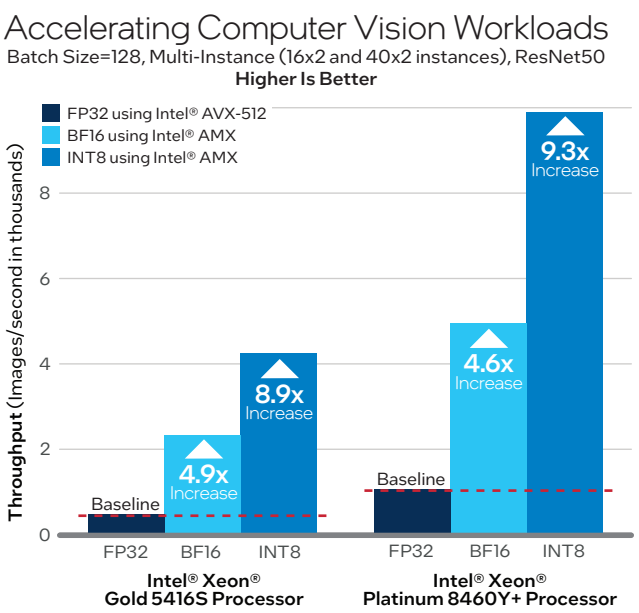

Figure 1. Acceleration of computer vision workloads by using a lower precision or a more powerful 4th Gen Intel® Xeon® Scalable processor.[3]

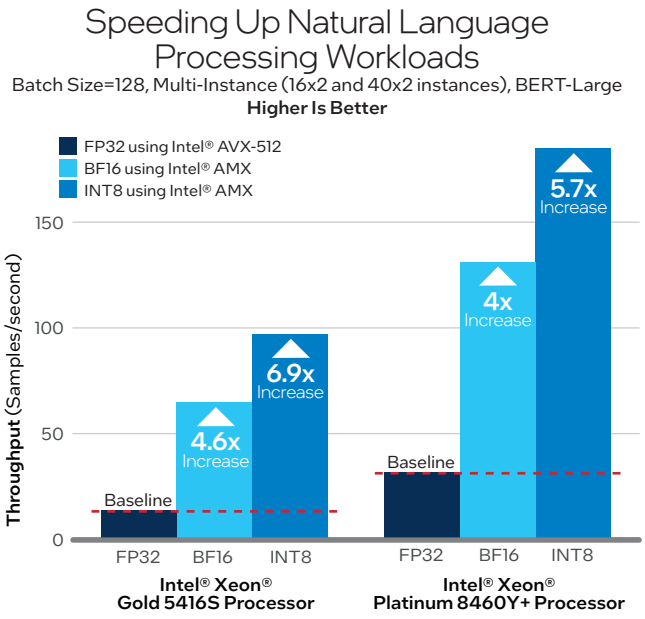

Figure 2. Speedup of natural language processing (NLP) workloads by using a lower precision or a more powerful 4th Gen Intel® Xeon® Scalable processor.[3]

Configuration Details

The following tables provide information about components and settings of the infrastructure used for performance analysis and characterization testing.

Hardware Configurations		
	2-Node Cluster (Edge)	4-Node Cluster (Data Center)
Server	QuantaGrid D54Q-2U	QuantaGrid D54Q-2U
Processor	2x Intel® Xeon® Gold 5416S processor (16 cores, 2.0 GHz)	2x Intel Xeon Platinum 8460Y+ processor (40 cores, 2.0 GHz)
Memory	512 GB (16x 32 GB DDR5 4800 MT/s)	512 GB (16x 32 GB DDR5 4800 MT/s)
Storage	4x Solidigm D7-P5510 Series (3.84 TB, 2.5in PCIe 4.0 x4, TLC, NVMe)	6x Solidigm D7-P5510 Series (3.84 TB, 2.5in PCIe 4.0 x4, TLC, NVMe)
Network Card	Intel® Ethernet Network Adapter E810-C-Q2 @ 25 Gbps	Intel® Ethernet Network Adapter E810-XXV-4 @ 25 Gbps

Important System Settings	
Number of Nodes	2x edge/4x data center (one node on each cluster was used for testing)
Power Setting	Performance
BIOS	SE5C620.86B.01.01.0006.2207150335
Microcode	0x2b0001b0
Max C-State	C0/C1
Frequency Governor	Performance
Turbo	Enabled
Intel® Total Memory Encryption (Intel® TME)	Enabled
IRQ Balance	Enabled
Prefetchers	L2 HW, L2 Adj., DCU HW, DCU IP

Software Versions	
HCI Software	Microsoft Azure Stack HCI 22H2, 10.0.20349
Hypervisor	Hyper-V
AI Framework	Intel® Optimization for TensorFlow 2.11
Image Processing	ResNet50 v1.5 with synthetic dataset
NLP	BERT-Large with SQuAD 1.1 dataset

Accelerator Technologies Enabled
Intel® DL Boost for FP32
Intel® AMX for BF16 and INT8
Intel® AVX-512
Intel® Hyper-Threading Technology
Intel® Turbo Boost Technology

Profiles and Workloads

The following table describes the workloads used in testing.

VM Profiles			
Workload	**vCPUs**	**RAM**	**RAM Reserved**
ResNet50	32/80	96 GB	96 GB
BERT-Large	32/80	96 GB	96 GB

Conclusion

AI capabilities are being integrated into a wide range of workloads, both at the edge and in the data center. Examples of workloads include cybersecurity, manufacturing quality control, equipment maintenance, fraud detection and enhanced customer experiences. Combining the flexibility of Azure Stack HCI with the AI acceleration and security benefits of 4th Generation Intel Xeon Scalable processors enables organizations to more quickly gain the insights they need to meet their business goals—all while securing their valuable and sensitive data.

Further Information

- 4th Gen Intel® Xeon® Scalable processors
- Intel® Advanced Matrix Extensions
- Microsoft Azure Stack HCI
- Unify Operations Across Hybrid and Multi-Cloud Environments white paper

Authors

Nagesh Dn, Data Center Platform Application Engineer
Darren Freimuth, Senior Cloud & Enterprise Solutions Architect

Learn more about Intel and Microsoft's Shared Cloud Vision.

Contact your Intel representative to learn more about this solution.

Solution Provided By:

Microsoft intel

[1] **Edge – 2-Node Azure Stack HCI Configuration:** Tested by Intel as of April 28, 2023. Two nodes with AI workload running on one node, 2x Intel® Xeon® Gold 5416S processor QS pre-production (16 cores, 2.0 GHz), 1x Intel® Server Board (QuantaGrid D54Q-2U), total memory: 512 GB (16x 32 GB 4800 MHz DDR5 DIMM), Intel® Hyper-Threading Technology = ON, Intel® Turbo Boost Technology = ON, BIOS = SE5C620.86B.01.01.0006.2207150335, microcode = 0x2b0001b0, storage (boot): 1x Solidigm DC S4610 (960 GB), storage: 4x Solidigm D7-P5510 Series (3.84 TB), network devices: 1x 25 GbE Intel® Ethernet Network Adapter E810-C-Q2 @ 25 GbE, 1x 10 GbE Intel® Ethernet Converged Network Adapter X550-T2 @ 1 GbE, OS/Software: Microsoft Azure Stack HCI build 20385 with Ubuntu Server 2022.

Data Center – 4-Node Azure Stack HCI Configuration: Tested by Intel as of April 28, 2023. Four nodes with AI workload running on one node, 2x Intel® Xeon® Platinum 8460Y+ processor QS pre-production (40 cores, 2.0 GHz), 1x Intel® Server Board (QuantaGrid D54Q-2U), total memory: 512 GB (16x 32 GB 4800 MHz DDR5 DIMM), Intel® Hyper-Threading Technology = ON, Intel® Turbo Boost Technology = ON, BIOS = SE5C741.86B.01.01.0002.2212220608, microcode = 0x2b0001b0 storage (boot): 1x Solidigm DC S4610 (960 GB), storage: 6x Solidigm D7-P5510 Series (3.84 TB), network devices: 1x 25 GbE Intel® Ethernet Network Adapter E810-XXV-4 @ 25 GbE, 1x 10 GbE Intel® Ethernet Converged Network Adapter X550-T2 @ 1 GbE, OS/Software: Microsoft Azure Stack HCI build 20385 with Ubuntu Server 2022.

[2] To enable Intel® TME in the BIOS, go to System settings --> Processors --> Total Memory Encryption (TME) and enable the option, then save your changes.

[3] See endnote 1.

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. No product or component can be absolutely secure. Results using 4th Gen Intel Xeon Processor are based on pre-production processors and systems. Your costs and results may vary. Intel technologies may require enabled hardware, software, or service activation. Intel, 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 0923/JCAP/KC/PDF