Solution Brief

December 2021

Benefits of a Multicloud Analytics Solution with VMware Cloud Foundation

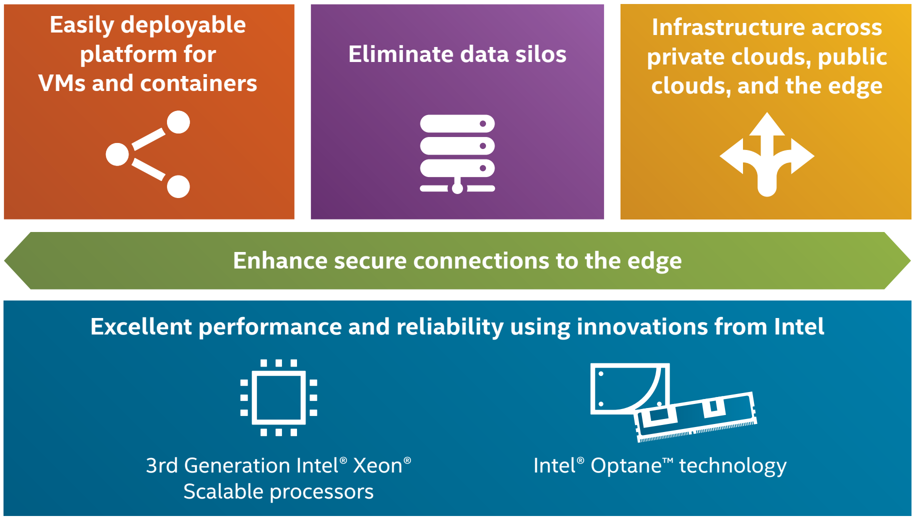
Deploy and manage data-intensive workloads from edge to cloud, taking advantage of high-performance 3rd Generation Intel® Xeon® Scalable processors and software optimized for Intel® architecture

Executive Summary

A modern compute environment is key to remaining competitive. The traditional approach for deploying applications and services cannot deliver innovation at the pace today's businesses require. In addition, as data volumes grow, enterprises struggle to get more value out of their data. Data silos and cumbersome data management and analytics processes hinder discovering business insights that can drive competitive advantage. What's more, as applications move to the edge in industries such as retail, establishing secure connectivity between the core data center, the cloud, and the edge becomes crucial to success.

Addressing these challenges involves replacing legacy hardware and software with modern, multicloud-capable solutions that can accelerate and streamline the entire software and hardware provisioning, deployment, and maintenance lifecycle. Simultaneously, companies need a platform that natively supports containerization for efficient data-intensive workloads like AI and machine learning.

Intel's flexible Multicloud Analytics Solution, based on VMware Cloud Foundation, offers an easily deployable platform for managing VMs and orchestrating containers. This solution helps eliminate data silos and provides security-enabled infrastructure, operations, and connectivity across private clouds, public clouds, and the edge. The solution delivers excellent performance and reliability using innovations from Intel like 3rd Generation Intel® Xeon® Scalable processors and Intel® Optane™ technology. For example, deep-learning inference throughput increases by up to 6x using a TensorFlow framework optimized for Intel architecture.[1]



Easily deployable platform for VMs and containers

Eliminate data silos

Infrastructure across private clouds, public clouds, and the edge

Enhance secure connections to the edge

Excellent performance and reliability using innovations from Intel

3rd Generation Intel® Xeon® Scalable processors

Intel® Optane™ technology

Business Challenge

Today's enterprises want the flexibility to run analytics workloads where they make most sense—in the core data center, in one or more public clouds (multicloud), and/or at the edge. But to make this flexibility operationally feasible, there must be a way to efficiently manage all the analytics workloads, wherever they reside. Without a single pane of glass, management costs rapidly spiral out of control, application development becomes inconsistent, and performance may suffer.

Enterprises seek analytics infrastructure that is characterized by reduced downtime, less setup time, easier maintenance, and lower overhead costs—without sacrificing performance. Legacy data centers cannot take advantage of the cost efficiencies and new technologies available in a multicloud analytics environment. Nor can such data centers adapt to changing workload requirements quickly and nimbly.

For companies with outdated data center technologies, meeting these challenges involves replacing legacy hardware and software with modern, hybrid-cloud-capable analytics solutions. These solutions can accelerate the entire software and hardware provisioning, deployment, and maintenance lifecycle along with application development, testing, and delivery. But, whether it's an on-premises machine-learning cluster or a remote branch office analytics cluster, companies may find assembling and maintaining multicloud infrastructure daunting.

Solution Value

Intel and VMware have teamed up to offer the Multicloud Analytics Solution to help take the guesswork out of building multicloud and edge analytics solutions. The Multicloud Analytics Solution combines VMware Cloud Foundation with innovative Intel technology to provide a unified Software-Defined Data Center (SDDC) platform for running and managing private cloud, multicloud, and edge containerized analytics workloads.

VMware Cloud Foundation is a full-stack hyperconverged infrastructure (HCI) solution that simplifies the path to and helps accelerate adoption of hybrid/multicloud analytics environments. It offers a complete set of software-defined services for compute, memory, storage, network, and security, along with application-focused cloud management capabilities. When combined with Intel technology, VMware Cloud Foundation provides consistently high-performance analytics, reduced data center footprint, and efficient operations management.

Enterprises can use the end-to-end Multicloud Analytics Solution to quickly launch database processing and AI, and scale workloads to accommodate future needs. The unified cloud solution presented in this solution brief can run containerized applications and traditional VMs that are located in an on-premises data center as well as in the public cloud, such as on Amazon Web Services (AWS) and Microsoft Azure.

In short, the Multicloud Analytics Solution is a simple, security-enabled, and agile cloud infrastructure for on-premises, as-a-service public cloud, and edge analytics workloads.

Solution Benefits

- **Unified platform** for running, managing, and seamlessly connecting VMs and containers across private cloud, multicloud, and edge environments
- **Accelerated analytics deployment** with a verified, end-to-end solution for a wide range of workloads
- Agile, scalable, and **security-enabled infrastructure** with excellent analytics performance
- **Increased throughput** with Intel architecture-optimized deep-learning frameworks[1]

Solution Architecture Highlights

The Multicloud Analytics Solution reference architecture from Intel includes several main VMware components: VMware vSphere with Kubernetes, VMware Secure Access Service Edge (SASE) with VMware Software-Defined WAN (SD-WAN), VMware Tanzu Mission Control, VMware vSAN, VMware NSX-T, VMware SDDC Manager, and VMware vRealize Suite to provide infrastructure-as-a-service capabilities. It also includes VMware services on public clouds—VMware Cloud on AWS (VMC) and Azure VMware Solution (AVS). Container provisioning and lifecycle management are provided by VMware Tanzu Kubernetes Grid (TKG).

The hybrid/multicloud structure of the solution allows enterprises to extend available resources and easily distribute analytics workloads between on-premises, public cloud, and the edge. VMware SD-WAN is used to provide reliable and secure network connectivity over public internet from any to any location (on-premises to the edge and to public cloud and vice versa).

VMware Cloud Foundation includes access to the Tanzu Application Catalog, which contains more than 70 Kubernetes applications and components from the Bitnami collection that are maintained and verified for use in production environments. Among these applications are popular analytics tools like TensorFlow, MxNet, PyTorch, and many others.

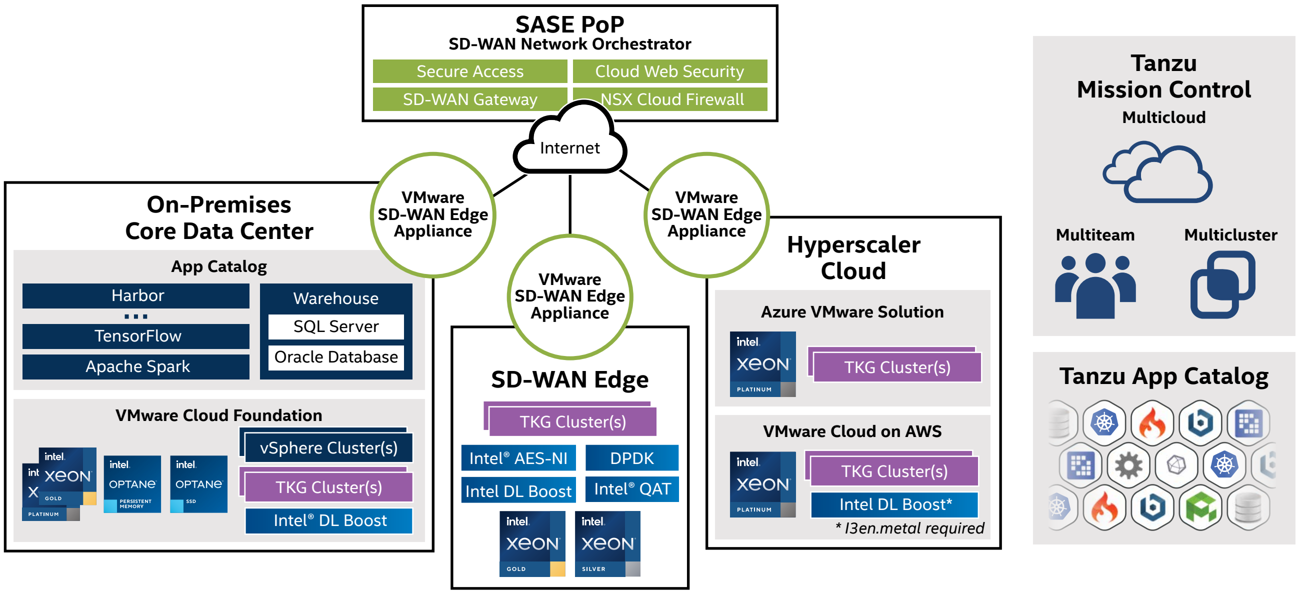


Figure 1. VMware and Intel provide the building blocks for the Multicloud Analytics Solution.

Underlying the software components of VMware Cloud Foundation in the on-premises core data center are 3rd Generation Intel® Xeon® Scalable processors, Intel® Optane™ persistent memory (PMem), Intel Optane SSDs, Intel® SSD D7 and D5 Series, and Intel® Ethernet products (see Figure 1).

Enterprises can use Intel Optane technology to boost their VMware Cloud Foundation workload performance by placing data closer to the CPU. This technology is a class of non-volatile memory and storage media that fills the gap between high-performing volatile memory and lower-performing NAND storage and HDDs. By placing data closer to the CPU, Intel Optane technology helps architects to confidently deploy an agile, high-performing infrastructure that helps organizations create innovative analytics services and optimize their infrastructure investments.

Intel Optane technology can be deployed in two different ways (see Figure 2):

- **Intel Optane PMem** gives enterprises the ability to extract more from larger datasets by combining more capacity and native persistence in a DIMM form factor. Data can be accessed, processed, and analyzed in near real time to deliver deep insights, improve operations, and create new revenue streams.
- **Intel Optane SSDs** help remove data bottlenecks to accelerate transactions and time to insights, so users get what they need, when they need it. With high quality of service and at least 6x faster performance than NAND SSDs at low queue depths, Intel Optane SSDs deliver fast, predictable performance—even in the most demanding environments.[2] For tiered storage like vSAN, it is recommended to use Intel Optane SSDs in the cache tier and Intel SSD D7 or D5 Series in the capacity tier.

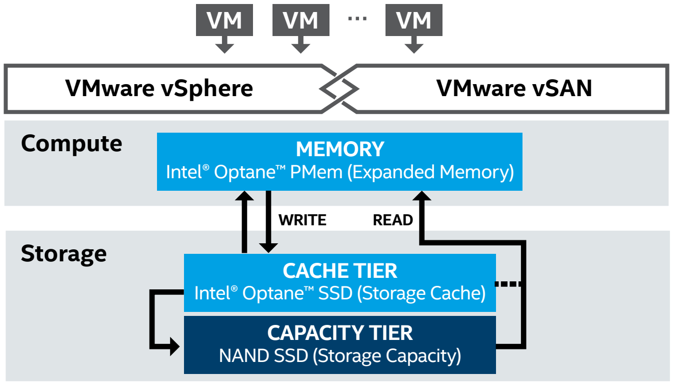

VM
VM
...
VM
VMware vSphere
VMware vSAN
Compute
MEMORY
Intel® Optane™ PMem (Expanded Memory)
WRITE
READ
Storage
CACHE TIER
Intel® Optane™ SSD (Storage Cache)
CAPACITY TIER
NAND SSD (Storage Capacity)

Figure 2. The placement of Intel® Optane™ SSDs and Intel® Optane™ persistent memory within the architecture.

A Closer Look at VMware Cloud Foundation 4.3

VMware Cloud Foundation 4.3 introduces several new features and enhancements that help customers deploy scalable, flexible infrastructure:

- **Enhanced workload domain deployment and lifecycle management** to support large-scale VM and container architectures.
- **Integration with VMware vSphere 7 Update 2** to deliver AI- and developer-ready infrastructure, boost data security, and help simplify operations.
- **Integration with VMware vSAN 7 Update 2**, which provides enhancements to the vSAN Data Persistence Platform for improved cloud-native storage and persistent services support.
- **Enhanced networking automation** provides faster expansion and better scaling of NSX-T Edge clusters.
- **Enhanced security operations** include stronger security mechanisms to improve the management and administration of security settings within VMware Cloud Foundation.

For more details about what's new in VMware Cloud Foundation 4.3, visit the release announcement.

Use Cases

The combination of VMware Cloud Foundation and Intel technology running on VMs or in containers can support a wide variety of use cases:

Deep-Learning Inference

Inference is compute-intensive and can benefit from innovations such as Intel® Deep Learning Boost (Intel® DL Boost) with Vector Neural Network Instructions (VNNI)—a special instruction set that speeds up inference—available starting with vSphere 7 and ESXi 7.0, which are foundational components of the VMware Cloud Foundation 4.3 platform.

Enterprises need high-performance data analytics and AI to remain competitive. They require flexible solutions that can run traditional data analytics and AI applications. The VMware multicloud platform includes components that take advantage of performance optimizations for Intel hardware. Intel supports developing machine-learning workloads at multiple layers in the solution stack. These building blocks enable enterprises to quickly operationalize analytics, AI, and machine-learning workloads because they are already optimized for Intel architecture and have been verified with multiple production deployments. Therefore, enterprises can immediately begin to use them.

The proposed use case showcases a solution that can increase performance in deep-learning inference workloads. This use case shows the improvement of inference performance with an Intel architecture-optimized container stack that uses the special VNNI instruction set and unlocks the full potential of 3rd Gen Intel Xeon Scalable processors.

See the "Results" section for illustration of how Intel technology and software optimization for that technology can significantly boost deep-learning inference throughput.

Retail at the Edge

For retail stores, healthcare, and smart industry, running workloads closer to customers and closer to the sources of the data can improve performance, which can lead to increased customer satisfaction. VMware Cloud Foundation makes it easy to deploy and manage remote workloads, using the same technology that is used for public and private cloud workloads.

The proposed use case showcases a solution that can increase retail customer engagement and improve the shopping experience. We include three scenarios:

- **Product recommendations.** When the client shows interest in a specific area or department, we can use a machine-learning algorithm to send personalized product recommendations. Based on people's similar choices and the customer's position in the store, the algorithm creates a list of the most relevant products. The customer is notified and can check the personalized recommendations using a mobile application. The process occurs every time the system discovers a new customer interest.
- **Presence detection.** We use deep-learning techniques and image recognition algorithms to detect customers in the Customer Service area. Cameras installed in the store send images to the deep-learning pipeline. When such an event occurs, the store staff is informed.
- **Hesitance detection.** When a customer is wandering around the store with no apparent purpose, without stopping, the business rules engine assumes the customer is looking for something, is lost, or may need assistance. A notification—including the customer's name, age, gender, and position in the store—is sent to the store staff so they can quickly find and identify a person in need.

Data Warehousing and Analytics

Data warehouses are considered one of the core components of business intelligence. They are a central location to store data from one or more disparate sources as well as current and historical data. The VMware hybrid/multicloud platform supports data warehousing, including industry-proven solutions based on Microsoft SQL Server 2019 or Oracle Database 19c.

Results: Deep-Learning Inference

Image classification is one of the most popular use cases for deep learning. Our tests benchmarked the ResNet50 v1.5 topology with int8 and fp32 precision, using the TensorFlow distribution from the Intel architecture-optimized container stack with Intel's Model Zoo pretrained models.

We ran two tests (see Appendix A for software used in testing):

- Performance comparison of default TensorFlow container versus the Intel architecture-optimized TensorFlow container
- Performance comparison of fp32 precision versus int8 precision (both using Intel DL Boost with VNNI and the Intel architecture-optimized TensorFlow container)

As the following graphs illustrate, the hardware and software optimizations for inference have a substantial impact on improving the performance of inference. In this use case, the optimizations enabled a significant increase in throughput (frames per second). The VMware Cloud Foundation 4.3 platform is an excellent example of how software can take advantage of hardware innovations like Intel DL Boost and VNNI to deliver faster insights.

Up to 6x Better Throughput by Optimizing TensorFlow for Intel Architecture

In this benchmark, we compared throughput performance of the default TensorFlow container against a container using the Intel® Optimization for TensorFlow, which is optimized to take advantage of Intel DL Boost and VNNI. Both containers used fp32 precision. As Figure 3 shows, framework optimizations from the Intel Optimization for TensorFlow can provide up to a 5.56x throughput improvement for the Base design and up to a 6.14x throughput improvement for the Plus design.[3]

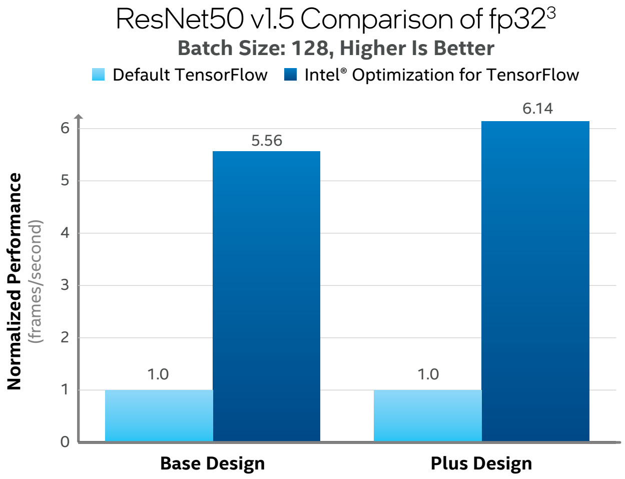


Figure 3. The Intel® Optimization for TensorFlow provides up to a 6.14x throughput improvement compared to the default TensorFlow framework.

Up to 4x Better Throughput with int8 Precision

In this benchmark, we compared throughput performance of Intel DL Boost with VNNI using int8 precision against fp32 precision. Both containers used the Intel Optimization for TensorFlow. As shown in Figure 4, a small reduction in precision enabled up to a 3.44x throughput improvement for the Base design and up to a 4x throughput improvement for the Plus design.[4]

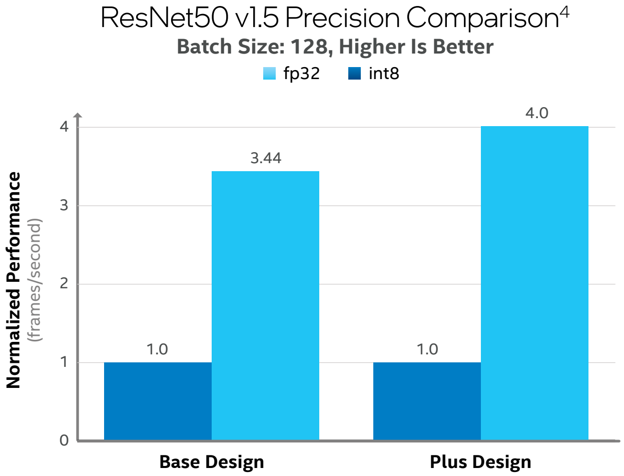


Figure 4. The Intel® Optimization for TensorFlow with int8 precision provides up to a 3.44x throughput improvement for the Base design and up to a 4x throughput improvement in the Plus design, compared to fp32 precision.

Learn More

- 3rd Gen Intel® Xeon® Scalable processors
- Intel® Ethernet products
- Intel® Optane™ persistent memory
- Intel® Optane™ SSDs
- VMware Cloud Foundation

Contact your Intel representative or visit the **Intel and VMware Partnership website**.

Appendix A: Testing Software

Table A1. Software Versions Used to Test Deep-Learning Inference

	BASE	PLUS
Guest OS	Ubuntu Server 20.04.3 LTS	
Guest OS Kernel	5.4.0-88-generic	
Containers	intel/intel-optimized-tensorflow:2.5.0-ubuntu-18.04 tensorflow/tensorflow:2.5.0	
AI Precision	int8, fp32	
Other Software	VMware Cloud Foundation 4.3; VMware vSAN 7.0 U2a; VMware vSAN 7.0 U2a; VMware vCenter Server 7.0 U2c; VMware NSX-T 3.1.3	
Other Software (hypervisor)	VMware ESXi 7.0 U2a (build 17867351)	
VM vCPU	42	56
VM vRAM	256 GB	256 GB
Framework/Toolkit included version	TensorFlow	
Framework URL	TensorFlow Docker images used: intel/intel-optimized-tensorflow:2.5.0-ubuntu-18.04 and tensorflow/tensorflow:2.5.0	
Topology or ML algorithm	ResNet50v1.5	
Compiler	Not compiled, used Docker images	
Libraries	Container with TensorFlow optimized with oneAPI Deep Neural Network Library (oneDNN)	
Dataset	Synthetic data (autogenerated, --benchmark-only parameter)	
Precision	int8, fp32	
Build Flags	Not compiled, used Docker images	
KMP AFFINITY	granularity=fine,verbose,compact,1,0 'verbose,warnings,respect,granularity=fine,compact,1,0'	
NUMACTL	Not used	
OMP_NUM_THREADS	42/56	
Command Line Used	`python3 /tf/intel-models/benchmarks/launch_benchmark.py --in-graph ${IN_GRAPH} --model-name ${MODEL_NAME} --framework tensorflow --precision ${PRECISION} --mode inference --batch-size ${BATCH_SIZE} --benchmark-only`	

[1] Testing by Intel as of October 11, 2021. Results may vary. See Appendix A for details on the software used for testing.

Base Configuration: Single-node, 2x Intel® Xeon® Gold 6342 processor, 1x Intel® Server Board M50CYP2UR, Total Memory = 512 GB (16 slots/32 GB/3200 MHz), Intel® Hyper-Threading Technology = OFF, Intel® Turbo Boost Technology = ON, BIOS: SE5C6200.86B.0022.D64.2105220049 (ucode: 0x0d0002b1), Storage (boot): 1x Intel® Optane™ P1600X 118 GB, Storage (cache): 2x Intel® Optane™ SSD P5800X 400 GB, Storage (capacity): 4x Intel® SSD D7-P5510 3.84 TB, Network devices: 1x Intel® Ethernet Adapter E810-CQDA2 (100 GbE), Hypervisor: VMware ESXi 7.0 U2a (build 17867351) and VMware Cloud Foundation 4.3, OS/Software: Ubuntu Server 20.04.3 LTS, 5.4.0-88-generic, 42 vCPU, 256 GB vRAM, Containers: intel/intel-optimized-tensorflow:2.5.0-ubuntu-18.04, tensorflow/tensorflow:2.5.0, Model Zoo for Intel: https://github.com/IntelAI/models/tree/v2.4.0, ResNet50 v1.5, fp32, batch size = 128

Plus Configuration: Single-node, 2x Intel® Xeon® Platinum 8362 processor, 1x Intel® Server Board M50CYP2UR, Total Memory = 1024 GB (2LM) - 256 GB (16 slots/16 GB/3200 MHz) + 1024 GB Intel® Optane™ persistent memory 200 series (8 slots/128 GB/3200 MHz), Intel Hyper-Threading Technology = OFF, Intel Turbo Boost Technology = ON, BIOS: SE5C6200.86B.0022.D64.2105220049 (ucode: 0x0d0002b1), Storage (boot): 1x Intel Optane P1600X 118 GB, Storage (cache): 2x Intel Optane SSD P5800X 800 GB, Storage (capacity): 6x Intel SSD D7-P5510 3.84 TB, Network devices: 2x Intel Ethernet Adapter E810-CQDA2 (100 GbE), Hypervisor: VMware ESXi 7.0 U2a (build 17867351) and VMware Cloud Foundation 4.3, OS/Software: Ubuntu Server 20.04.3 LTS, 5.4.0-88-generic, 56 vCPU, 256 GB vRAM, Containers: intel/intel-optimized-tensorflow:2.5.0-ubuntu-18.04, tensorflow/tensorflow:2.5.0, Model Zoo for Intel: https://github.com/IntelAI/models/tree/v2.4.0, ResNet50 v1.5, fp32, batch size = 128

[2] Intel, https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/data-center-ssds/optane-ssd-p5800x-p5801x-brief.html

[3] See endnote 1.

[4] Testing by Intel as of October 10, 2021. Results may vary. See Appendix A for details on the software used for testing.

Base Configuration: Single-node, 2x Intel® Xeon® Gold 6342 processor, 1x Intel® Server Board M50CYP2UR, Total Memory = 512 GB (16 slots/32 GB/3200 MHz), Intel® Hyper-Threading Technology = OFF, Intel® Turbo Boost Technology = ON, BIOS: SE5C6200.86B.0022.D64.2105220049 (ucode: 0x0d0002b1), Storage (boot): 1x Intel® Optane™ P1600X 118 GB, Storage (cache): 2x Intel® Optane™ SSD P5800X 400 GB, Storage (capacity): 4x Intel® SSD D7-P5510 3.84 TB , Network devices: 1x Intel® Ethernet Adapter E810-CQDA2 (100 GbE), Hypervisor: VMware ESXi 7.0 U2a (build 17867351) and VMware Cloud Foundation 4.3, OS/Software: Ubuntu Server 20.04.3 LTS, 5.4.0-88-generic, 42 vCPU, 256GB vRAM, Containers: intel/intel-optimized-tensorflow:2.5.0-ubuntu-18.04, Model Zoo for Intel: https://github.com/IntelAI/models/tree/v2.4.0, ResNet50 v1.5, int8 and fp32, batch size = 128

Plus Configuration: Single-node, 2x Intel® Xeon® Platinum 8362 processor, 1x Intel® Server Board M50CYP2UR, Total Memory = 1024 GB (2LM) - 256 GB (16 slots/16 GB/3200 MHz) + 1024 GB Intel® Optane™ persistent memory 200 series (8 slots/128 GB/3200 MHz), Intel Hyper-Threading Technology = OFF, Intel Turbo Boost Technology = ON, BIOS: SE5C6200.86B.0022.D64.2105220049 (ucode: 0x0d0002b1), Storage (boot): 1x Intel Optane P1600X 118 GB, Storage (cache): 2x Intel Optane SSD P5800X 800 GB, Storage (capacity): 6x Intel SSD D7-P5510 3.84 TB, Network devices: 2x Intel Ethernet Adapter E810-CQDA2 (100 GbE), Hypervisor: VMware ESXi 7.0 U2a (build 17867351) and VMware Cloud Foundation 4.3, OS/Software: Ubuntu Server 20.04.3 LTS, 5.4.0-88-generic, 56 vCPU, 256GB vRAM, Containers: intel/intel-optimized-tensorflow:2.5.0-ubuntu-18.04, Model Zoo for Intel: https://github.com/IntelAI/models/tree/v2.4.0, ResNet50 v1.5, int8 and fp32, batch size = 128

Performance varies by use, configuration and other factors. Learn more at intel.com/PerformanceIndex. Performance results are based on testing by Intel and may not reflect all publicly available security updates. See configuration disclosures for 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, 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 1221/JSTA/KC/PDF