Solution Brief

Data Center | Hybrid Cloud

Accelerate Hybrid Cloud AI Workloads

Ease your journey to hybrid/multicloud with a reference architecture for Intel® technology and VMware Cloud Foundation

Solution Benefits

Intel's VMware Cloud Foundation reference architecture takes advantage of Intel® compute, memory, storage, and networking innovations to help enable software-defined data centers and hybrid/multicloud adoption.

- **Fast AI inference.** AI workloads can benefit from innovations from Intel such as Intel® DL Boost.
- **Flexibility and portability.** VMware Cloud Foundation helps enable enterprises to run their workloads where it makes most sense, whether that's on-premises, in a public cloud, or in several clouds at once.

Executive Summary

To remain competitive in today's world, organizations need a modern data center. Companies using these data centers must accelerate their product development, compete more successfully at a lower cost, and reduce their downtime and maintenance overhead. Technology must move and change with the times—solutions for hosting applications and services must innovate and change as well.

Companies with older and outdated data centers will want to meet these challenges by upgrading to hybrid cloud solutions, where the data center can easily and seamlessly interface between on-premises and cloud systems. Based on VMware Cloud Foundation with VMware Tanzu, Intel addressed these requirements by offering a hybrid/multicloud reference architecture—available in a Base and Plus configuration—that is easily deployable and manageable for virtual machines (VMs) and containers.

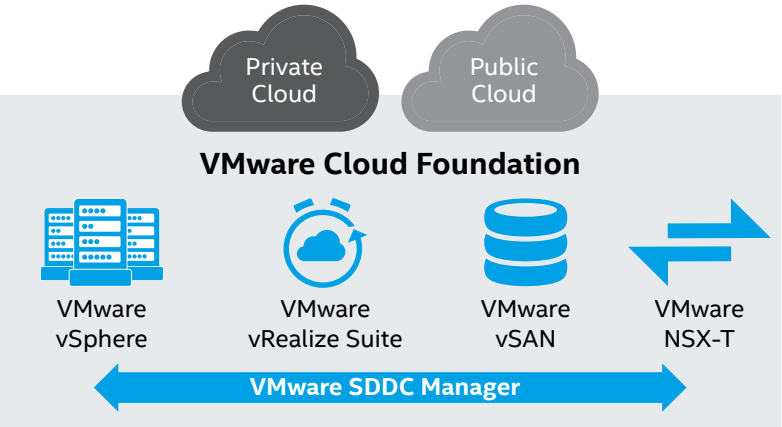

Figure 1. VMware Cloud Foundation supports software-defined data centers that can benefit greatly from Intel® compute, memory, storage, and networking technologies.

Business Challenge: Building a Hybrid Cloud Machine-Learning Architecture

A data center and the machine learning that a data center enables needs to run optimally for a business to stay competitive. Balancing high performance and cost is a continual challenge. Enterprises seek infrastructure that is characterized by less 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 hybrid/multicloud environment. Such data centers cannot 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 solutions that can accelerate the entire software and hardware provisioning, deployment, and maintenance lifecycle along with application development, testing, and delivery. But, especially for machine learning, companies may be daunted by assembling and maintaining hybrid cloud infrastructure. Machine learning requires large datasets that are difficult to get into the cloud, and machine-learning models must be continuously retrained and updated. In addition, data for machine learning can be sensitive, highly regulated, or may contain intellectual property, which raises the data's security concerns. Intel and VMware have teamed up to help take the guesswork out of building a machine-learning solution. VMware Cloud Foundation is a hybrid cloud platform that runs on Intel® hardware, offering an easily deployable and manageable hybrid/multicloud platform for managing VMs and orchestrating containers.

Typical Workloads in the Hybrid/Multicloud Data Center

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

- **Machine-learning training.** Image classification is one of the most popular use cases for deep learning. Training such models can be time-consuming and, without the right tools, requires specialized skills. The VMware Cloud Foundation platform works with various machine-learning frameworks, including DataRobot, which is a popular automated machine-learning platform that takes advantage of optimizations for Intel® architecture. With a library of hundreds of powerful open-source machine-learning algorithms, the DataRobot platform applies many best practices to machine learning and helps to accelerate and scale data science capabilities while increasing transparency, accuracy, and collaboration.
- **Machine-learning inference.** Once a model is trained, it can be run on new data sets to uncover hidden insights. Inference is compute-intensive, and can benefit from innovations from Intel such as Intel® Deep Learning Boost (Intel® DL Boost) with Vector Neural Network Instructions (VNNI)—available starting with vSphere 7 and ESXi 7.0, which are foundational components of the VMware Cloud Foundation 4 platform.
- **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.

A Closer Look at Intel® DL Boost and VNNI

2nd Gen Intel® Xeon® Scalable processors offer something unique that is not available with any other processor on the market: Intel® DL Boost with VNNI. This technology takes advantage of, and improves upon, Intel® AVX-512. VNNI improves AI performance by combining three instructions into one—thereby optimizing the use of compute resources and utilizing the cache more effectively and avoiding potential bandwidth bottlenecks. In Intel benchmarks, VNNI speeds the delivery of inference results by up to 30x, compared to the previous-generation Intel Xeon Scalable processor.[1]

Solution Value: High Performance in the Hybrid/Multicloud Environment

VMware Cloud Foundation is a full-stack HCI solution that helps accelerate adoption of hybrid/multicloud environments. When combined with Intel technology, VMware Cloud Foundation provides consistently high performance, reduced data center footprint, and efficient operations management.

With the end-to-end solution that Intel and VMware offer, enterprises can 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.

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 migrate workloads from on-premises to the cloud and back.

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 new 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 services and optimize their infrastructure investments.

Intel Optane technology can be deployed in two different ways:

- Intel® Optane™ persistent memory (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® 3D NAND SSDs in the capacity tier.

This solution provides infrastructure and operations across private and public clouds with excellent performance and reliability from Intel® hardware components.

Users can also take advantage of Intel DL Boost with VNNI. Intel conducted experiments to show the improvement of inference performance with an Intel architecture-optimized container stack that uses the new VNNI instruction set.

Our tests benchmarked the ResNet50 v1.5 topology with int8 and fp32 precision, using the Intel® Optimization for TensorFlow container stack with Intel's Model Zoo pretrained models. We ran three tests:[3]

- **Compare the performance improvement of Intel DL Boost with VNNI using int8 precision against fp32 precision.** As shown in Figure 2, int8 precision enabled a 4.1x improvement for the Base configuration and a 4.38x improvement for the Plus configuration. For a small decrease in precision, performance quadrupled.
- **Compare throughput from the default TensorFlow container against a container using the Intel Optimization for TensorFlow.** Framework optimizations from the Intel Optimization for TensorFlow can provide 2.33x improvement for the Base configuration and 2.61x performance improvement for the Plus configuration.
- **Compare the results of running VMware Cloud Foundation 4.0.1 (which takes advantage of Intel DL Boost and VNNI) against the reference architecture for VMware Cloud Foundation 3.9 (which does not use Intel DL Boost or VNNI).** The newer system provided a 1.53x improvement over the older system for the Base configuration and a 1.64x improvement for the Plus configuration (see Figure 3).

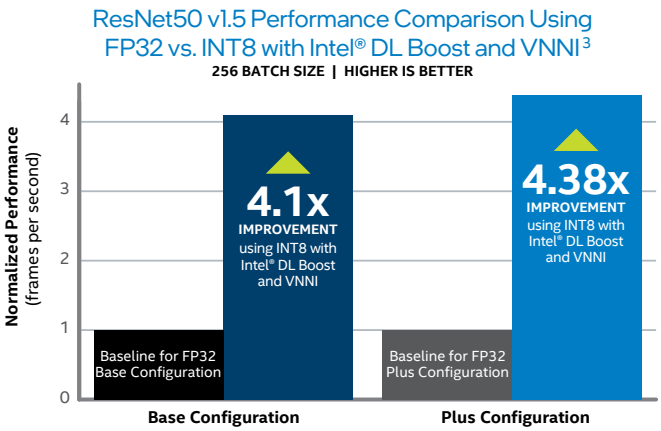

Figure 2. Using int8 precision instead of fp32 precision more than quadrupled the performance of the ResNet 50 v1.5 topology.

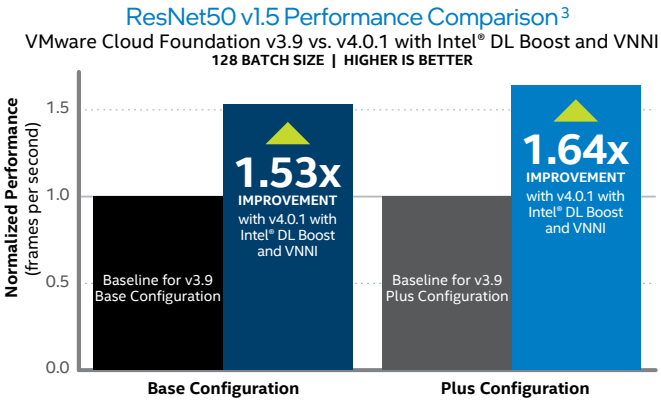

Figure 3. Using VMware Cloud Foundation 4.0.1 with Intel® DL Boost and VNNI improves the performance of the ResNet50 v1.5 topology for both the Base configuration and the Plus configuration.

As the results show, hardware and software optimizations have a huge impact on improving the performance of inference. VMware Cloud Foundation 4 is an excellent example of how software can take advantage of hardware innovations like Intel DL Boost and VNNI to deliver significantly better performance results.

High-Performance Hybrid/Multicloud Data Center

VMware Cloud Foundation provides a simplified path to the hybrid cloud through an integrated software platform for both private and public cloud environments. It offers a complete set of software-defined services for compute, memory, storage, network, and security, along with application-focused cloud management capabilities.[4] The result is a simple, security-enabled, and agile cloud infrastructure for on-premises and as-a-service public cloud environments.

The reference architecture from Intel includes several main VMware components: VMware vSphere with Kubernetes, VMware TKG Service for vSphere, VMware vSAN, VMware vRealize Suite, VMware NSX-T Data Center, and VMware Software-Defined Data Center Manager to provide infrastructure-as-a-service capabilities. VMware Hybrid Cloud Extension (HCX) enables VM migration, workload rebalancing, and protection between on-premises and cloud environments. Underlying the software components of VMware Cloud Foundation are 2nd Generation Intel® Xeon® Scalable processors, Intel Optane PMem, Intel Optane SSDs and Intel 3D NAND SSDs, and Intel® Ethernet products (see Figure 4).

Intel® oneAPI Initiative

Modern workload diversity necessitates architectural diversity; no single architecture is best for every workload. Intel® oneAPI products will deliver the tools needed to deploy applications and solutions across various architectures, including CPUs, GPUs, FPGAs and other accelerators. Its set of complementary toolkits—a base kit and specialty add-ons—simplify programming and help developers improve efficiency and innovation. The core Intel oneAPI DPC++ Compiler and libraries implement the oneAPI industry specifications available at oneapi.com.

Intel oneAPI Base Toolkit (Beta) is a foundational kit that enables developers of all types to build, test, and deploy performance-driven, data-centric applications across CPUs, GPUs, and FPGAs.

In addition, there are domain-specific toolkits that can be used for specialized workloads that are powered or based on the oneAPI Base Toolkit. Examples include:

- Intel® AI Analytics Toolkit (Beta) for accelerating end-to-end machine-learning and data science pipelines.
- Intel® Distribution of OpenVINO™ Toolkit for deploying high-performance inference applications from device to cloud.
- Intel oneAPI DL Framework Developer Toolkit (Beta) for building deep learning frameworks or customizing existing ones.

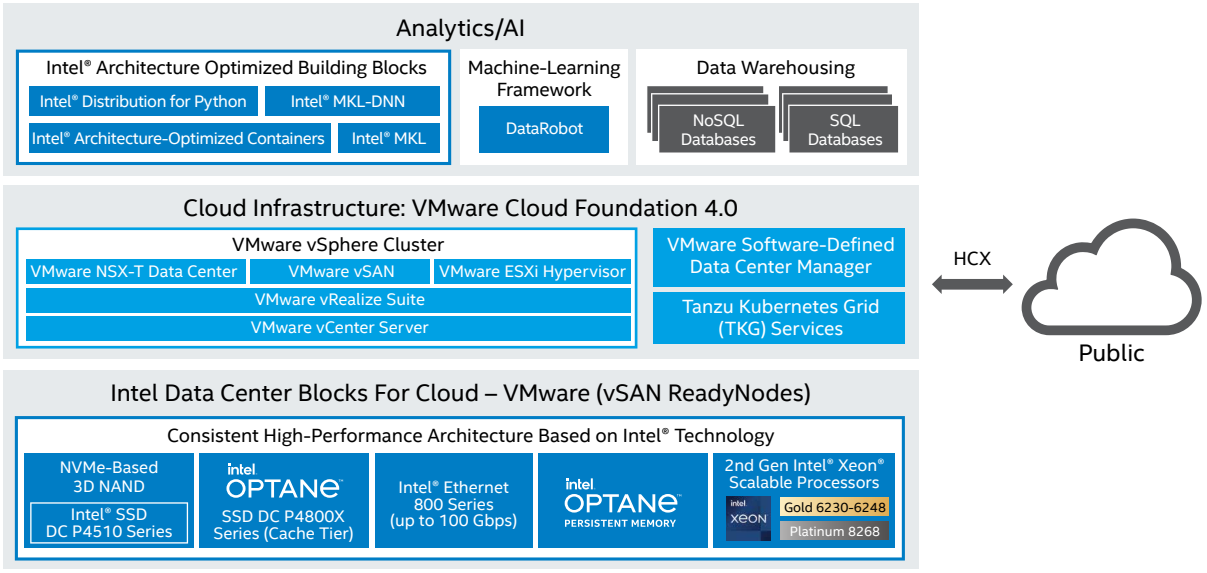

Figure 4. VMware and Intel provide the building blocks for the VMware hybrid/multicloud platform.

Conclusion

Enterprises are seeking to improve their data analytics and AI performance and modernize their data center architecture. Moving to a hybrid/multicloud environment helps them migrate workloads to and from private and public clouds, and also utilize their local infrastructure seamlessly. Using Intel's reference architecture for VMware Cloud Foundation 4, enterprises can have a single, easy-to-manage architecture, on their own premises or in the cloud.

The VMware hybrid/multicloud reference architecture combines the best of Intel hardware and VMware virtualization software. With this end-to-end solution that is ready to deploy, enterprises are poised to run both their traditional data analytics workloads and the AI and machine-learning workloads of the future.

Find the solution that is right for your organization. Contact your Intel representative or visit the Intel and VMware Partnership website.

Learn More

You may also find the following resources useful:

- Intel® Optane™ SSDs
- Intel® Optane™ persistent memory
- 2nd Generation Intel® Xeon® Scalable processors
- Intel® Deep Learning Boost
- VMware Cloud Foundation

Solution Provided By:

vmware® | intel®

[1] Intel® Deep Learning Boost configuration: Tested by Intel as of 2/26/2019. Platform: 2-socket Intel® Xeon® Platinum 9282 processor (56 cores per socket), Intel® Hyper-Threading Technology ON, Intel® Turbo Boost Technology ON, Total Memory 768 GB (24 slots/32 GB/2933 MHz), BIOS: SE5C620.86B.0D.01.0241.112020180249, Centos 7 Kernel 3.10.0-957.5.1.el7. x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd-2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer DummyData: 3x224x224, 56 instance/2 socket, Datatype: INT8.

Baseline configuration: Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 processor (28 cores per socket), Intel Hyper-Threading Technology OFF, Intel Turbo Boost Technology OFF, scaling governor set to "performance" via Intel® P-state driver, 384 GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 (800 GB, 2.5in SATA 6Gb/s, 25nm, MLC. Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (https://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50), Intel C++ compiler ver. 17.0.2 20170213, Intel® Math Kernel Library (Intel® MKL) small libraries version 2018.0.20170425. Caffe run with "numactl -l".

[2] Source: Intel-tested as of November 15, 2018. 4K 70/30 read/write performance at low queue depth. Measured using FIO 3.1. Common configuration: Intel® 2U Server System, OS: CentOS 7.5, kernel 4.17.6-1.el7.x86_64, 2x Intel® Xeon® Gold 6154 processor at 3.0 GHz (18 cores), 256 GB DDR4 RAM at 2666 MHz. Configuration: 375 GB Intel® Optane™ SSD DC P4800X compared to 1.6 TB Intel® SSD DC P4600. Intel microcode: 0x2000043; system BIOS: 00.01.0013; Intel® ME firmware: 04.00.04.294; BMC firmware: 1.43.91f76955; FRUSDR: 1.43.

[3] Testing by Intel, August-September 2020. Each cluster (Base, Plus, and the Management cluster that was used for workload generation) consists of four machines. All machines are identical within the given cluster.

Base 4-node cluster configuration: Intel® Xeon® Gold 6248 processor (2.5 GHz, 20 cores); 384 GB DRAM (12x 32 GB 2933 MHz); Intel® Hyper-Threading Technology ON; Intel® Turbo Boost Technology ON; Storage: VMware vSAN, disk group: 1x Intel® Optane™ SSD DC P4800X 375 GB and 2x Intel® SSD DC P4510 2 TB, two disk groups per node; BIOS = 02.01.0010; microcode = 0x0500002c; OS = VMware ESXi 7.0b 16386292; 1x Intel® Ethernet Adapter XXV710-DA2.

Plus 4-node cluster configuration: Intel® Xeon® Platinum 8268 processor (2.9 GHz, 24 cores); 384 GB DRAM (12x 32 GB 2933 MHz) plus 1,536 GB Intel® Optane™ persistent memory (12x 128 GB); Intel Hyper-Threading Technology ON; Intel Turbo Boost Technology ON; Storage: VMware vSAN, disk group: 1x Intel Optane SSD DC P4800X 375 GB and 3x Intel SSD DC P4510 2 TB, two disk groups per node; BIOS = 02.01.0010; microcode = 0x0500002c; OS = VMware ESXi 7.0b 16386292; 1x Intel Ethernet Adapter XXV710-DA2; PMem used a 2-2-2 configuration in App Direct Mode; PMem firmware version = 01.02.00.5417.

Workload: ResNet 50 v1.5 topology at int8, fp32 precision. Default TensorFlow tensorflow/tensorflow:1.15.0-py3; optimized TensorFlow: clearlinux/stacks-dlrs-mkl:v0.5.0.

Base Fat VM configuration: 80 vCPUs; OS=CentOS Linux release 8.2.2004; kernel=4.18.0-93.14.2.el8_2.x86_64; other SW: VMware Cloud Foundation 4.0.1, VMware ESXi hypervisor 7.0b Plus Fat VM configuration: 96 vCPUs; otherwise, identical to the Base Fat VM configuration.

[4] According to the VMware Cloud Foundation 4.0.1 release notes, HCX is not officially supported on VMware Cloud Foundation 4.0.1. VMware states that it plans to add HCX support in the future.

Performance varies by use, configuration and other factors. Learn more at 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.