White Paper

Hybrid Cloud Modernization
Intel® Xeon® Scalable Platform

Scaling Data Center Modernization for Hybrid Cloud

Accelerate workload deployment, simplify management, and boost competitive edge with the ideal combination of Intel-VMware technologies which is designed, deployed, migrated, and managed by Tech Mahindra.

Authors

Intel:

Narinder Sharma
Industry Technical Specialist

Tech Mahindra:

Abhinandan Chakraborty
Technical Alliance Manager – ICS

Table of Contents

Executive Summary

A modern data center is key to remaining competitive. All companies must overcome challenges to accelerate product development, reduce downtime and maintenance overhead, and compete more successfully at lower costs. Technology is now a centerpiece of every new change; the traditional approach for hosting applications and services cannot deliver innovation at the pace businesses require.

For companies with outdated data center technologies, meeting these challenges involves replacing legacy hardware and software with modern, hybrid-cloud capable solutions that can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. End-user self-service solutions are expected to reduce the time to market even more.

Tech Mahindra, Intel, and VMware join hands to address these new market requirements by offering an easily deployable and manageable hybrid/multi-cloud platform for managing virtual machines (VMs) and orchestrating containers. This solution provides infrastructure and operations across private and public clouds with excellent performance and reliability powered by Intel® technologies.

Because the term "cloud computing" is now often associated with both VMs and the use of containerization, this reference architecture illustrates a variety of applications. These include VM-based solutions, as well as container-based analytics and artificial intelligence (AI) workloads. This reference architecture describes the advantage of relevant Intel® technology such as Intel® Optane™ persistent memory and Intel® Deep Learning Boost (Intel® DL Boost), and VMware Cloud Foundation 4.0 environment connected to VMware Cloud on Amazon Web Services (AWS).

Why VMware Cloud Foundation for Your Hybrid/Multi-Cloud?

A hybrid/multi-cloud infrastructure combines the benefits of an on-premises infrastructure with the flexibility and instant availability of one or more public clouds. This approach is gaining popularity as businesses adapt their existing computing resources to ever-changing demands and market needs. While there are many examples of services that are ideal for the public cloud, some workloads are better suited to staying on-premises (such as sensitive data that is required for machine-learning models). Increasingly, enterprises want a hybrid/multi-cloud option for this flexibility and business agility. Hybrid and multi-cloud solutions are especially important as AI and machine learning workloads become increasingly prevalent.

Solution Overview

This reference architecture provides detailed configuration information for building a hybrid/multi-cloud environment. At a high level, the reference architecture consists of an optimized combination of Intel® hardware and VMware software.

With the end-to-end solution that Tech Mahindra, 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 reference architecture (see Figure 1) can run containerized applications and traditional VMs, located in an on-premises data center as well as in the public cloud, such as on AWS. The hybrid/multi-cloud nature of the solution allows enterprises to extend available resources and easily migrate workloads from on-premises to the cloud and back.

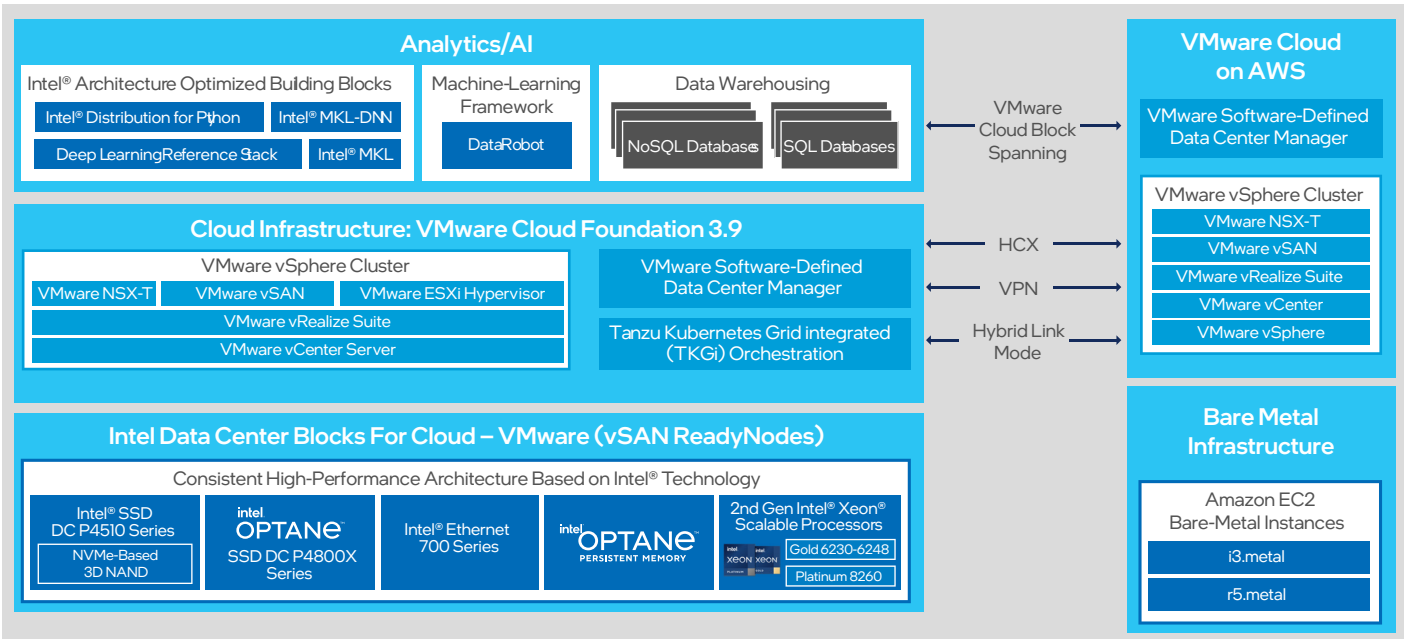

Figure 1. Reference architecture building blocks for the VMware hybrid/multi-cloud platform.

Building a Powerful Foundation with Intel® Hardware

The hardware stack for the solution is built on Intel® server platforms and hence can be built on any OEM platform while choosing same hardware configurations. The platforms include the latest generation of Intel® Xeon® Gold and Platinum processors. 2nd Gen Intel® Xeon® Scalable processors are workload-optimized and provide the compute power that can meet the data-centric demands of analytics, artificial intelligence, and in-memory database workloads. These processors support Intel® Deep Learning Boost (Intel® DL Boost), which uses Vector Neural Network Instructions (VNNI) to boost AI inferencing performance. For high-performance all-flash software-defined storage, the reference architecture includes Intel® Optane™ SSD DC P4800X and NVMe-based Intel® SSD DC P4510 combined with Intel® Optane™ Persistent memory (Intel® Optane™ PMem). Intel® Optane™ PMem introduces innovative memory technology that delivers large-capacity system memory and persistence. For accelerated software-defined network, the platform uses 25 Gb/s Intel® Ethernet Converged Network Adapters X700 series.

The reference architecture can be deployed in two configurations - "Base" and "Plus". The Base configuration uses the Intel® Xeon® Gold 6248 processor and optimally balances price and performance for mainstream workloads. The Plus configuration uses the Intel® Xeon® Platinum 8268 processor, which can efficiently handle high-density deployments and data-intensive, latency-sensitive workloads. Enterprises that need even higher performance can replace the default CPU with a higher-number SKU in either configuration.

Intel architecture provides a foundational industry-standard hardware infrastructure that supports extensible virtualized networking and security functions for VMs and containers. In the same manner, Intel and VMware have co-engineered optimizations that leverage the Intel developer tool set, including the following:

- Intel® QuickAssist Technology (Intel® QAT) provides a software-enabled foundation for security, authentication and compression to increase overall performance and efficiency.
- Intel® AES New Instructions (Intel® AES-NI) accelerates key parts of the encryption algorithm in hardware, making pervasive, end-to-end encryption possible without degrading performance. NSX-T Data Center also benefits from Intel AES-NI to accelerate processor-intensive encryption and decryption routines in hardware, helping to maintain pervasive encryption as workloads and topologies scale.
- Intel® Trusted Execution Technology (Intel® TXT) moves the root of trust from software to hardware, checking the execution environment against a known-good image at

startup to verify that no unauthorized changes have been made that could jeopardize the security of application workloads.

- Intel® Deep Learning Boost (Intel® DL Boost) with Vector Neural Network Instructions (VNNI) takes advantage of, and improves upon, Intel® Advanced Vector Extensions 512 (Intel® AVX-512). VNNI improves AI performance by combining three instructions into one thereby maximizing the use of compute resources and utilizing the cache more effectively and avoiding potential bandwidth bottlenecks.
- Intel® Optane™ PMem represents a new class of memory and storage technology. It can work in two different operating modes (Memory Mode and App Direct Mode), depending on business need and/or application support. Regardless of the mode used, Intel® Optane™ PMem provides capacity that is unavailable when using traditional DRAM modules.

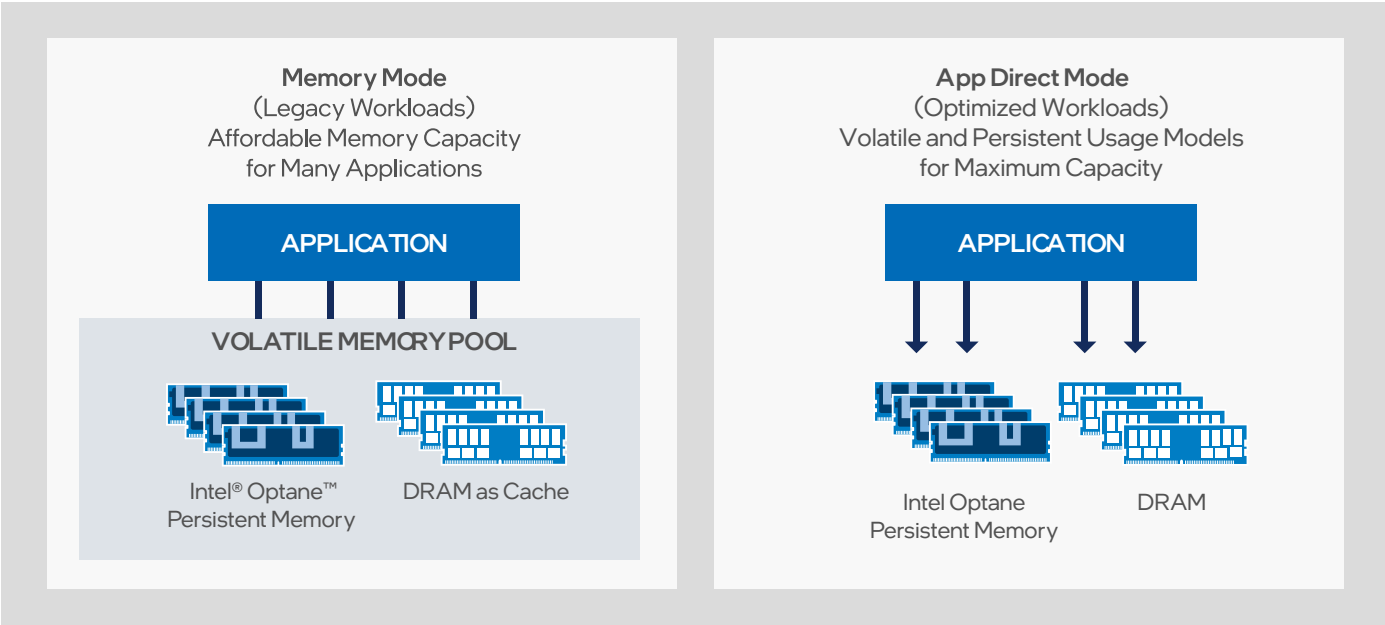

Figure 2. Intel® Optane™ persistent memory operating modes.

Maximizing the Advantage of On-Premise Cloud

From a software perspective, this solution consists of VMware Cloud Foundation 4.0.1, which is based on several main VMware components: VMware vSphere with Tanzu, VMware Tanzu Kubernetes Grid (TKG) Service for vSphere, VMware vSAN, VMware vRealize Suite, VMware NSX-T Data Center, and VMware SDDC Manager to provide IaaS capabilities.

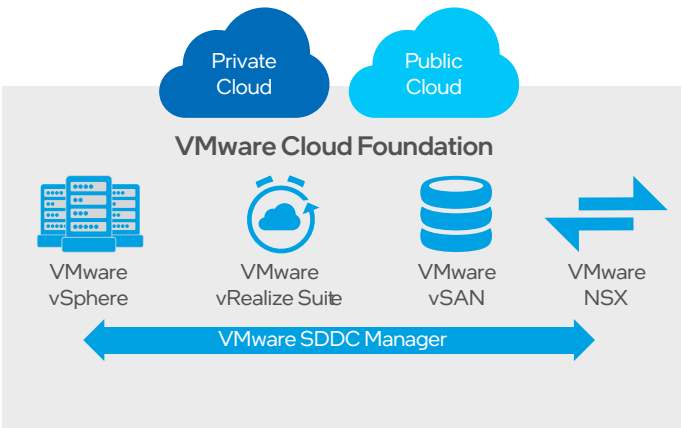

Figure 3. VMware Cloud Foundation is a cloud solution that is managed through VMware SDDC Manager and built on VMware vSphere, vRealize Suite, vSAN, and NSX-T Data Center

The solution is built from a number of key components:

- **VMware vSphere with Tanzu Workload Management:** VMware vSphere extends virtualization to storage and network services and adds automated, policy-based provisioning and management. vSphere is the starting point for building an SDDC platform. VMware vSphere with Tanzu enables streamlined development, agile operations, and accelerated innovation for all enterprise applications. It consists of two core components: ESXi and vCenter Server. ESXi is the virtualization platform used to create and run VMs and appliances, while vCenter Server manages multiple ESXi hosts as clusters, using shared pool resources. VMware vSphere with Tanzu workload management enables the deployment and operation of compute, networking, and storage infrastructure for vSphere with Tanzu. It makes it possible to use vSphere as a platform for running Kubernetes workloads natively on the hypervisor layer. Kubernetes workloads may be run directly on ESXi hosts and upstream Kubernetes clusters can be created within dedicated resource pools by using the TKG Service.
- **VMware Tanzu Kubernetes Grid (TKG) Service for vSphere:** TKG is used to provision and manage the lifecycle

of Tanzu Kubernetes clusters, which are proprietary installations of Kubernetes open-source software, built and supported by VMware.

- **VMware vRealize Suite:** VMware vRealize Suite provides IT organizations with a modern platform for infrastructure automation, consistent operations, and governance based on DevOps and machine-learning principles.
- **VMware vSAN:** VMware vSAN joins all storage devices across a vSphere cluster into a shared data pool and eliminates the need for external shared storage. Two vSAN cluster configurations are possible:
  - A hybrid vSAN cluster uses two types of storage devices: flash devices for the cache tier and magnetic drives for the capacity tier.
  - An all-flash vSAN cluster uses flash devices for both the cache and capacity tiers.

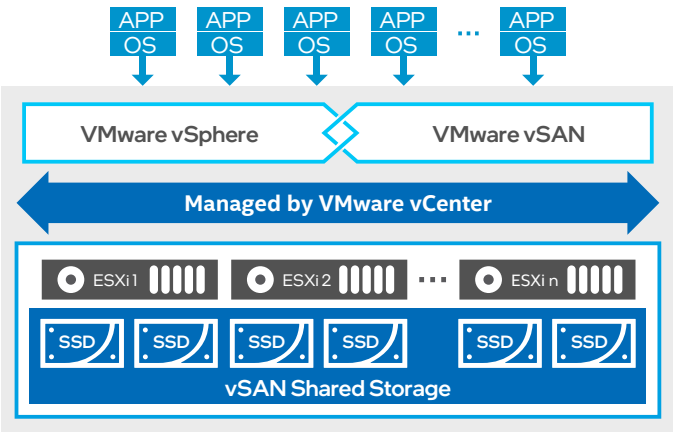

Figure 4. VMware vSAN overview

- **VMware NSX-T Data Center:** NSX-T Data Center (formerly NSX-T) enables a virtual cloud network with a software-defined approach. Working like a network hypervisor, it reproduces a complete set of Layer 2 through Layer 7 networking services: routing, switching, access control, firewalls, QoS, and DHCP in software. All these components can be used in any combination to create isolated virtual networks on demand. The services can then be extended to a variety of endpoints within and across clouds.
- **VMware SDDC Manager:** Software-Defined Data Center (SDDC) Manager manages the bring-up of the VMware Cloud Foundation system, creates and manages workload domains, and performs lifecycle management to keep the software components up to date. SDDC Manager also monitors the logical and physical resources of VMware Cloud Foundation.
- **Hybrid Cloud Extension to AWS:** VMware Cloud on AWS is used as the destination for the hybrid architecture. VMware Hybrid Cloud Extension (HCX) enables VM migration, workload rebalancing, and protection between on-premises and cloud. In addition to business continuity, it provides network extension for multi-tier applications without changing the VM properties.
- **VMware Cloud on AWS:** The solution takes advantage of existing tools, processes, and familiar VMware technologies, along with native integration with AWS. This makes it easy to adopt, greatly reduces service disruption associated with migrating critical services to the cloud and eliminates the need for rearchitecting the environment to suit a public cloud infrastructure.
- **VMware Cloud Bare-Metal Types:** The latest addition to VMware Cloud on AWS bare-metal infrastructure, i3en.metal bare-metal instances, address a variety of workloads, including data- or storage-intensive workloads requiring high random I/O access. i3en.metal instances extend the security capabilities of VMware Cloud on AWS by providing in-transit hardware-level encryption between instances within the SDDC boundaries.

The Technology Behind the Building Blocks

Data Warehousing Building Blocks

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. Numerous methods can be used to organize a data warehouse. Hardware, software, and data resources are the main components of this architecture, and VMware Cloud Foundation provides an excellent platform on which to deploy data warehousing solutions (see Figure 5).

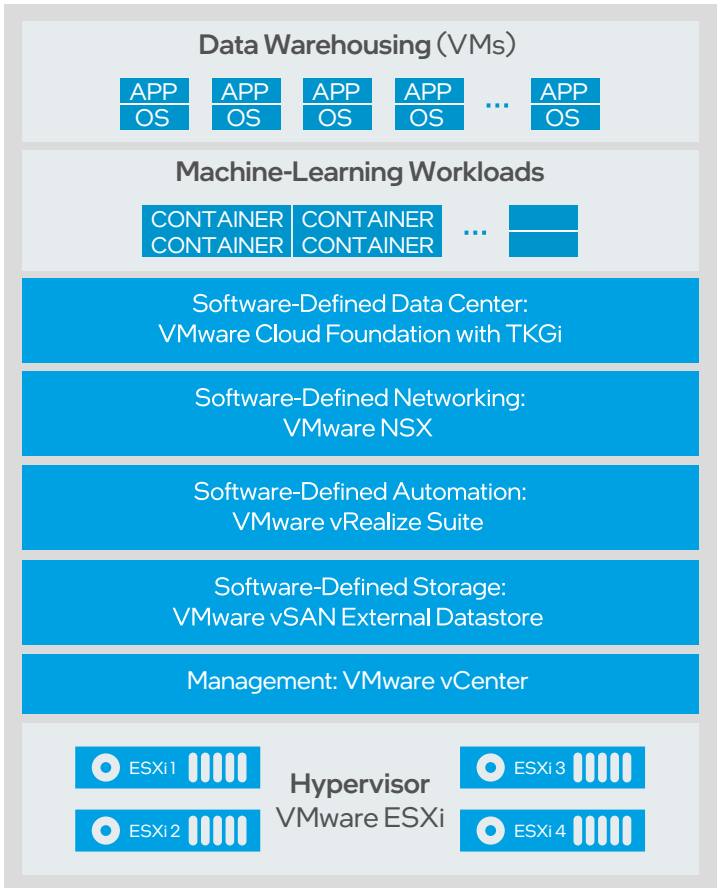

Figure 5. VMware Cloud Foundation can be used for all data analytics, AI, and machine-learning workloads

The VMware hybrid/multicloud platform supports data warehousing, including industry-proven solutions based on Microsoft SQL Server 2019 or Oracle Database 19c. The entire platform runs on vSAN, which provides additional storage policy configuration options in terms of data redundancy (multiple redundancy levels are available). vSAN can be used by both platform administrators and end users (such as when processing persistent volume claims on Kubernetes deployments) to obtain the maximum usage of the entire platform storage system.

Analytics and AI Building Blocks
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 hybrid/multi-cloud 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. This reference architecture demonstrates how to train a machine-learning model and then how it can be deployed on a hybrid/multi-cloud cluster. It also shows how Intel architecture-optimized deep learning libraries can be used to boost inference performance.

Intel® oneAPI as a Building Block for Innovative Solutions
Modern workload diversity necessitates the need for architectural diversity; no single architecture is best for every workload. XPUs, including CPUs, GPUs, FPGAs, and other accelerators, are required to extract high performance. Intel® oneAPI products deliver the tools needed to deploy applications and solutions across these architectures. Its set of complementary toolkits—a base kit and specialty add-ons—simplify programming and help developers improve efficiency and innovation.

DataRobot as a Building Block for Optimized Machine Learning
This solution also demonstrates DataRobot, a popular automated machine-learning platform that takes advantage of optimizations for Intel architecture. Organizations worldwide use DataRobot to empower the teams they already have in place to rapidly build and deploy machine-learning models and create advanced AI applications. With a library of hundreds of powerful open-source machine-learning algorithms, the DataRobot platform encapsulates many best practices and helps to accelerate and scale data science capabilities while increasing transparency, accuracy, and collaboration. DataRobot makes it easy to explain AI through human-friendly visual insights and automated model documentation with blueprints that describe each step in the modeling process and the algorithms used. Enterprises can evaluate any model using several tools. All DataRobot models are ready for production and can be deployed with a single click to make AI fully operational. Enterprises can monitor models using a centralized dashboard to view service health and usage in real time. They can also manage model accuracy to easily understand which features have drifted and deploy updates with no service interruption.

Driving Analytics/AI Workload Performance Advantage with Intel

VMware Cloud Foundation 4.0.1 introduced Intel® DL Boost with VNNI to VMs. Intel & VMware ran experiments to show the improvement of inference performance with an Intel architecture-optimized container stack that uses the new VNNI instruction set.

Based on tests conducted by Intel-VMWare:[1]

- Framework optimizations from Intel Optimization for TensorFlow can provide 2.33X improvement for the Base configuration and 2.61X performance improvement for the Plus configuration.
- Upgrading from an older version of VMWare Cloud Foundation (such as 3.9) to version 4.x provides up to 1.64x improvement for the Plus configuration.
- Using Intel DL Boost and VNNI with int8 precision quadruples performance over fp32 precision.

Why Tech Mahindra for Digital Transformation?

Tech Mahindra along with esteemed partners like Intel & VMware, helps enterprises accelerate their digital transformation with a comprehensive suite of offerings that covers the entire IT infrastructure stack. With platform-enabled solutions powered by AI and Automation, Tech Mahindra helps enterprises accelerate their journey to a digital future focusing on helping organizations develop a future ready infrastructure aligned to their core business needs.

Tech Mahindra fosters customer success through a 3-pillar strategy:

Imagine	with enterprises to define how their industries will change in the future and how they will proactively transform to be ready for that eventuality.
Build	systems and processes for enterprises so that they can handle revolutionary change.
Run	the business value of clients by helping them differentiate themselves in an increasingly crowded market.

Leadership across verticals

Communications	5G Slice to Price - Communications Value Chain Produce to Play - Media & Entertainment Value Chain	Retail	Design to Delight - Retail & Consumer Goods Value Chain
BFSI	Cradle to Grave - Banking, Financial Services Value Chain	Energy & Utilities	Wells to Wheels - Oil & Gas Value Chain Generation to Consumption utilities Value Chain
Manufacturing	Concept to Customer Delight -Manufacturing Value Chain	Healthcare	Delivering Future of Care Health & Lifesciences Value Chain

Learn more at https://www.techmahindra.com/en-in/

Conclusion

With the need for high-performance data analytics and AI on the rise, enterprises seek flexible solutions that can run workloads on-premises or in the public cloud. The proposed hybrid/multi-cloud platform combines the best of Tech Mahindra, Intel, and VMware solutions. This reference architecture has been validated to meet expected key performance indicators in demanding customer workload scenarios for VMware Cloud Foundation 4.0, which can take advantage of Intel® DL Boost and VNNI.

These results strongly illustrate that the right combination of software and hardware and software that is optimized for the hardware can substantially increase application performance. And the faster applications run, the faster business can run.

intel®

Source: For detailed deployment of guide, refer https://www.intel.com/content/www/us/en/cloud-computing/modernize-data-center-for-hybrid-cloud.html

Testing by Intel, August 2020. See endnote 3 for hardware configuration. 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=4.18.0-193.14.2.el8_2.x86_64; other SW: VMware Cloud Foundation 4.0.1, VMware ESXi hypervisor 7.0b 16386292.Plus Fat VM configuration: 96 vCPUs; otherwise, identical to the Base Fat VM configuration.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex. Performance results are based on testing as of the dates shown in configurations and may not reflect all publicly available updates. No product or component can be absolutely secure. Your costs and results may vary.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

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

No product or component can be absolutely secure.

Your costs and results may vary.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/benchmarks.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Intel Advanced Vector Extensions (Intel AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at intel.com/go/turbo.