SOLUTION BRIEF

Intel® Select Solutions | Version 2
Artificial Intelligence
2nd Generation Intel® Xeon® Scalable Processors
November 2019

# Intel Select Solutions for AI Inferencing

**Accelerate artificial intelligence (AI) inferencing and deployment on an optimized, verified infrastructure based on industry-standard Intel® technology.**

Businesses increasingly look to artificial intelligence (AI) to increase revenue, drive efficiencies, and innovate their offerings. In particular, AI use cases powered by deep learning (DL) generate some of the most powerful and useful insights; some of these use cases can enable advances across numerous industries, for example:

- **Image classification**, which can be used for concept assignment like facial sentiment
- **Object detection,** which is utilized by autonomous vehicles for localization of objects
- **Image segmentation,** which provides the ability to outline organs in a patient's magnetic resonance imaging (MRI)
- **Natural language processing,** which enables textual analysis or translation
- **Recommender systems,** which can be used by online stores to predict customer preferences or suggest up-sell options

These use cases are only the beginning. As businesses incorporate AI into their operations, they discover new ways of applying AI. However, the business value of all AI use cases is dependent on how quickly answers can be inferenced from models trained by deep neural networks. The resources needed to support inferencing on DL models can be substantial, and they often require organizations to update their hardware to obtain the required performance and speed. However, many customers want to extend their existing infrastructures rather than purchase new single-purpose hardware. The flexibility of the Intel® hardware architecture that your IT department is already familiar with can help protect your IT investments. Intel Select Solutions for AI Inferencing are "turnkey platforms" that provide pre-bundled, verified, and optimized solutions for low-latency, high-throughput inference performed on a CPU, not on a separate accelerator card.

## Intel Select Solutions for AI Inferencing

Intel Select Solutions for AI Inferencing provide you with a jumpstart to deploying efficient AI inferencing algorithms on solutions built on validated Intel architecture so that you can innovate and go to market faster. To speed AI inferencing and time to market for applications built on AI, Intel Select Solutions for AI Inferencing combine several Intel and third-party software and hardware technologies.

### Software Selections

The software used in Intel Select Solutions for AI Inferencing includes developer and management tools to aid AI inferencing in production environments.

### Intel® Distribution of OpenVINO™ Toolkit

The Intel® Distribution of Open Visual Inference and Neural Network Optimization toolkit (Intel Distribution of OpenVINO toolkit) is a developer suite that accelerates high-performance AI and DL inference deployments. The toolkit takes models trained in different frameworks and optimizes them for multiple Intel hardware options in order to provide maximum performance for deployment. Using the toolkit's Deep Learning Workbench, models can be quantized to a lower precision, a process in which the toolkit transforms models from using large, high-precision 32-bit floating-point numbers, which are typically used for training and occupy more memory, to using 8-bit integers, which optimize memory usage and performance. Swapping out floating-point numbers for integers leads to significantly faster AI inference with almost identical accuracy.[1] The toolkit can convert and execute models built in a variety of frameworks, including TensorFlow, MXNet, PyTorch, Kaldi, and any framework supported by the Open Neural Network Exchange (ONNX) ecosystem. Additionally, pre-trained, public models are also available that can expedite development and improve image processing pipelines for Intel processors, without the need to search for or train your own models.

### Deep Learning Reference Stack

Intel Select Solutions for AI Inferencing come with the Deep Learning Reference Stack (DLRS), an integrated, high-performance open source software stack that is optimized for Intel Xeon Scalable processors and that is packaged into a convenient Docker container. The DLRS helps reduce the complexity associated with integrating multiple software components in production AI environments as it is a pre-validated, configured collection of required libraries and software components. The stack also includes highly tuned containers for the popular DL frameworks TensorFlow and PyTorch, along with the Intel Distribution of OpenVINO toolkit. This open source community release is part of an effort to ensure AI developers have easy access to all the features and functionalities of Intel platforms.

### Kubeflow and Seldon Core

As organizations gain experience with deploying models for inference in a production environment, a consensus has emerged around a set of best practices collectively referred to as "MLOps," which parallel "DevOps" software-development practices. To help teams apply MLOps, Intel Select Solutions for AI Inferencing use Kubeflow. With Kubeflow, teams can smoothly roll out new versions of their models with zero downtime. Kubeflow uses supported model-serving back ends like TensorFlow Serving to export trained models to Kubernetes. Model deployments can use canary testing or shadow deployments to validate new versions side by side with old ones. If teams detect problems, they can use model and data versioning, in addition to tracking, to ease root-cause analysis.

To maintain a responsive service as demand increases, load balancing in Intel Select Solutions for AI Inferencing automatically shards inferencing onto available instances of the servables across nodes. Multitenancy enables different models to be served, improving hardware utilization.

Finally, to speed up inferencing requests between servers on which AI inferencing runs and the endpoints where AI insights are needed, Intel Select Solutions for AI Inferencing can use Seldon Core to help manage inference pipelines. Kubeflow also integrates with Seldon Core to deploy DL models on Kubernetes, and it uses the Kubernetes API to manage containers deployed in inference pipelines.

## Hardware Selections

Intel Select Solutions for AI Inferencing combine 2nd Generation Intel Xeon Scalable processors, Intel® Optane™ DC Solid State Drives (SSDs), Intel® 3D NAND SSDs, and the Intel® Ethernet 700 Series, so your business can quickly deploy a production-grade AI infrastructure built on a performance-optimized platform that offers high-capacity memory for the most demanding applications and workloads.

### 2nd Generation Intel Xeon Scalable Processors

Intel Select Solutions for AI Inferencing feature the performance and capabilities of 2nd Generation Intel Xeon Scalable processors. For the "Base" configuration, the Intel Xeon Gold 6248 processor provides an optimized balance of price, performance, and built-in technologies that enhances performance and efficiency for inferencing on AI models. The Intel Xeon Platinum 8268 processor powers the "Plus" configuration, which is designed for even faster AI inferencing. Higher-number processors can also be used in either configuration. 2nd Generation Intel Xeon Scalable processors include Intel® Deep Learning Boost, a family of acceleration features that improves AI inferencing performance through the use of the specialized Vector Neural Network Instructions (VNNI) instruction set, which performs in a single instruction DL computations that formerly required three separate instructions.

### Intel Optane DC Technology

Intel Optane DC technology fills critical gaps in the storage and memory hierarchy, enabling data centers to accelerate their access to data. This technology also disrupts the memory and storage tier, delivering persistent memory, large memory pools, fast caching, and storage in a variety of products and solutions.

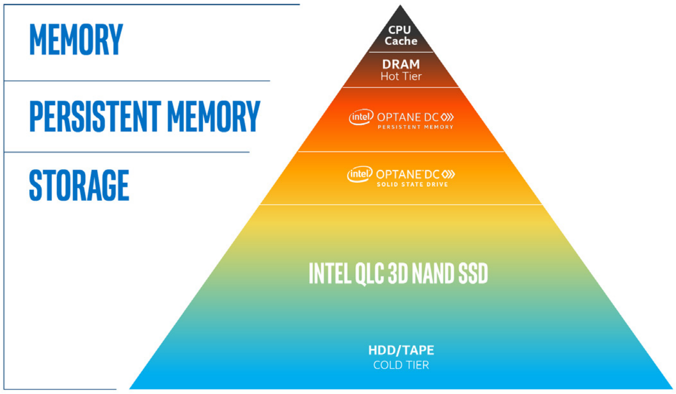

**Figure 1**. Intel Optane technology fills memory and storage performance gaps in the data center

### *Intel Optane DC SSDs and Intel 3D NAND SSDs*

AI inferencing performs best when the cache tier is on fast SSDs with low latency and high endurance. Workloads that require high performance can benefit from empowering the cache tier with the highest-performing SSDs rather than mainstream Serial ATA (SATA) SSDs. Intel Optane DC SSDs are used to power the cache tier in these Intel Select Solutions. Intel Optane DC SSDs offer high input/output (I/O) operations per second (IOPS) per dollar with low latency, coupled with 30 drive-writes-per-day endurance, so they are ideal for write-heavy cache functions.[2] The capacity tier is served by Intel 3D NAND SSDs, delivering optimized read performance with a combination of data integrity, performance consistency, and drive reliability.

### 25Gb Ethernet

The 25Gb Intel Ethernet 700 Series Network Adapters accelerate the performance of Intel Select Solutions for AI Inferencing. Paired with 2nd Generation Intel Xeon Platinum processors and the Intel SSD DC P4600, they provide up to 2.5x performance compared to 1Gb Ethernet (GbE) adapters and the Intel SSD DC S4500.[3,4] The Intel Ethernet 700 Series delivers validated performance ready to meet high-quality thresholds for data resiliency and service reliability with broad interoperability.[5] All Intel Ethernet products are backed by worldwide pre- and post-sales support and offer a limited lifetime warranty.

## Verified Performance through Benchmark Testing

All Intel Select Solutions are verified through benchmark testing to meet a pre-specified minimum capability level of workload-optimized performance. Because AI inferencing is an increasingly critical component of workloads in the data center, on the network edge, and in the cloud, Intel chose to measure and benchmark using a standard DL benchmarking approach and emulation of a real-world scenario.

For standard benchmarking, the number of images that can be processed per second (throughput) is measured on a pre-trained deep residual neural network (ResNet 50 v1) that is closely tied to broadly used DL use cases (image classification, localization, and detection) on TensorFlow, PyTorch, and the OpenVINO toolkit using synthetic data.

To emulate a real-world scenario, multiple clients are launched representing multiple request streams. These clients send images from the external client systems to the server for inferencing. On the server side, the incoming requests are load-balanced by Istio. The requests are then sent to multiple instances of a servable that contains a pipeline of pre-processing, prediction, and post-processing steps run through Seldon Core. The prediction is done using optimized DLRS container images of the OpenVINO toolkit model server. Once requests go through the pipeline, inferences are sent back to the requesting client. Throughput and latency are measured to help ensure that this test configuration can support the scale of inferencing in production environments.

## Base and Plus Configurations

Intel Select Solutions for AI Inferencing are available in two configurations, as shown in Table 1. The Base configuration specifies the minimum required performance capability for the solutions, and the Plus configuration provides an example of how system builders, system integrators, and solution and service providers can further optimize Intel Select Solutions for AI Inferencing to achieve higher performance and capabilities.

Customers can upgrade or expand on either of these configurations for additional capacity or performance. The Plus configuration utilizes higher performance 2nd Generation Intel Xeon Scalable processors and more memory to deliver up to 39 percent faster AI inferencing than the Base configuration.[6]

To refer to a solution as an Intel Select Solution for AI Inferencing, a solution provider must meet or exceed the defined minimum configuration ingredients and achieve the minimum benchmark-performance thresholds listed below.

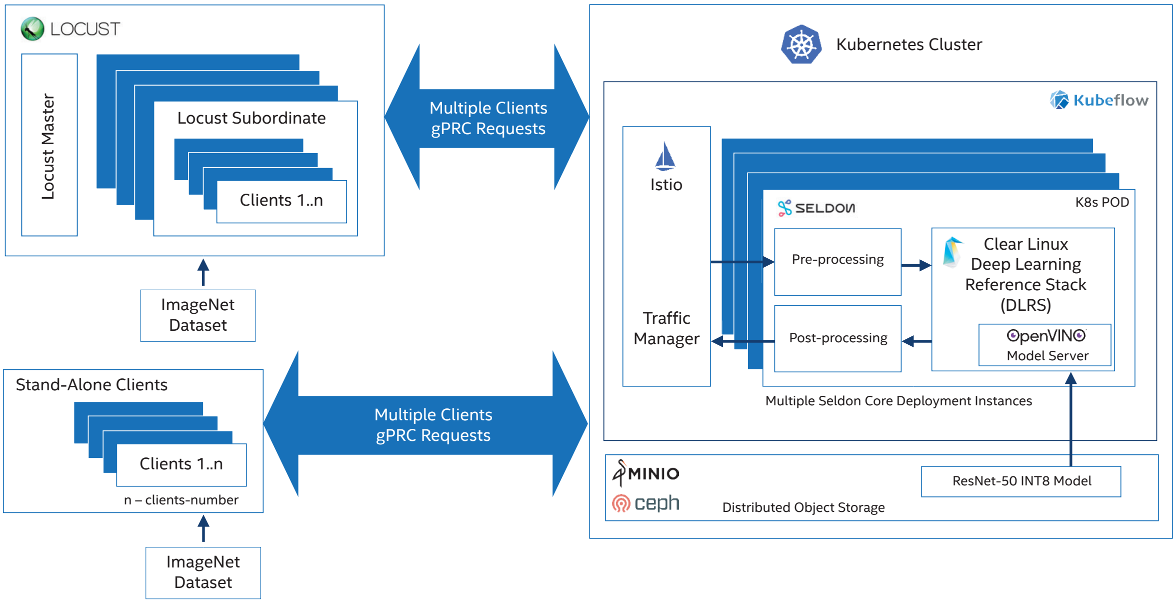

**Figure 2**. Architectural diagram of the real-world benchmark testing conducted on Intel Select Solutions for AI Inferencing

**Table 1.** The Base and Plus configurations for version 2 of the Intel Select Solutions for AI Inferencing

| INGREDIENT | INTEL SELECT SOLUTIONS FOR AI INFERENCING **BASE CONFIGURATION** | INTEL SELECT SOLUTIONS FOR AI INFERENCING **PLUS CONFIGURATION** |
|---|---|---|
| NUMBER OF NODES | Single-node configuration | Single-node configuration |
| PROCESSOR | 2 x Intel Xeon Gold 6248 processor (2.50 GHz, 20 cores, 40 threads), or a higher number Intel Xeon Scalable processor | 2 x Intel Xeon Platinum 8268 processor (2.90 GHz, 24 cores, 48 threads), or a higher number Intel Xeon Scalable processor |
| MEMORY | 192 GB or higher (12 x 16 GB 2,666 MHz DDR4 ECC RDIMM) | 384 GB (12 x 32 GB 2,934 MHz DDR4 ECC RDIMM) |
| BOOT DRIVE | 1 x 256 GB Intel SSD DC P4101 (M.2 80 mm PCIe 3.0 x4, 3D2, TLC) or higher | 1 x 256 GB Intel SSD DC P4101 (M.2 80 mm PCIe 3.0 x4, 3D2, TLC) or higher |
| STORAGE | **Data drive:** 1.6 TB NVM Express (NVMe) Intel SSD DC P4510<br>**Cache drive:** 375 GB Intel Optane SSD DC P4800X U.2 NVMe SSD | **Data drive:** 1.6 TB NVMe Intel SSD DC P4510<br>**Cache drive:** 375 GB Intel Optane SSD DC P4800X U.2 NVMe SSD |
| DATA NETWORK | 1 x Intel® Ethernet Converged Network Adapter XXV710-DA2 SFP28 DA Copper PCIe x 8 dual-port 25/10/1 GbE | 1 x Intel Ethernet Converged Network Adapter XXV710-DA2 SFP28 DA Copper PCIe x 8 dual-port 25/10/1 GbE |
| MANAGEMENT NETWORK | Integrated 1 GbE port 0/RMM port | Integrated 1 GbE port 0/RMM port |
| SOFTWARE | | |
| LINUX OS | CentOS Linux release 7.6.1810/Red Hat Enterprise Linux (RHEL) 7 | CentOS Linux release 7.6.1810/Red Hat Enterprise Linux (RHEL) 7 |
| INTEL MATH KERNEL LIBRARY (INTEL MKL) | Intel MKL version 2019 Update 4 | Intel MKL version 2019 Update 4 |
| INTEL DISTRIBUTION OF OPENVINO TOOLKIT | 2019 R1.0.1 | 2019 R1.0.1 |
| OPENVINO MODEL SERVER | 0.4 | 0.4 |
| TENSORFLOW | 1.14 | 1.14 |
| PYTORCH | 1.2.0 | 1.2.0 |
| MXNET | 1.3.1 | 1.3.1 |
| INTEL® DISTRIBUTION FOR PYTHON | 2019 Update 1 | 2019 Update 1 |
| INTEL® MATH KERNEL LIBRARY FOR DEEP NEURAL NETWORKS (INTEL® MKL-DNN) | 0.19 (implied the OpenVINO toolkit) | 0.19 (implied the OpenVINO toolkit) |
| DEEP LEARNING REFERENCE STACK (DLRS) | v4.0 | v4.0 |
| SOURCE-TO-IMAGE | 1.1.14 | 1.1.14 |
| DOCKER | 18.09 | 18.09 |
| KUBERNETES | v1.15.3 | v1.15.3 |
| KUBEFLOW | v0.6.1 | v0.6.1 |
| HELM | 2.14.3 | 2.14.3 |
| SELDON CORE | 0.3.2 | 0.3.2 |
| CEPH | v14.2.1 | v14.2.1 |
| MIN.IO (ROOK V1.0) | RELEASE.2019-04-23T23-50-36Z | RELEASE.2019-04-23T23-50-36Z |
| ROOK | 1.0.5 | 1.0.5 |
| OTHER | | |
| TRUSTED PLATFORM MODULE (TPM) | TPM 2.0 | TPM 2.0 |

MINIMUM PERFORMANCE STANDARDS

Verified to meet or exceed the following minimum performance capabilities:

| | | |
|---|---|---|
| CLASSIFICATION USING RESNET-50 ON OPENVINO TOOLKIT | 1,900 images per second (91 percent top-5 accuracy)[6] | 2,650 images per second (91 percent top-5 accuracy)[6] |
| SCALING IN EMULATED REAL-WORLD SCENARIO FROM 1 NODE TO 2 NODES | Up to 1.91x[7] | Up to 1.91x[8] |
| BUSINESS VALUE OF CHOOSING A PLUS CONFIGURATION OVER A BASE CONFIGURATION | The Plus configuration provides up to 39 percent faster inferencing performance.[6] | |

**Recommended, not required

## Technology Selections for Intel Select Solutions for AI Inferencing

In addition to the Intel hardware foundation used for Intel Select Solutions for AI Inferencing, Intel technologies deliver further performance and reliability gains:

- **Intel® Advanced Vector Extensions 512 (Intel® AVX-512):** A 512-bit instruction set that can accelerate performance for demanding workloads and usages like AI inferencing.
- **Intel Deep Learning Boost:** A group of acceleration features introduced in 2nd Generation Intel Xeon Scalable processors that provides significant performance increases to inference applications built using leading DL frameworks such as PyTorch, TensorFlow, MXNet, PaddlePaddle, and Caffe. The foundation of Intel Deep Learning boost is VNNI, a specialized instruction set that uses a single instruction for DL computations that formerly required three separate instructions.
- **Intel Distribution of OpenVINO toolkit:** A free software kit that helps developers and data scientists speed up AI workloads and streamline DL inferencing and deployments from the network edge to the cloud.
- **Intel Math Kernel Library (Intel MKL):** This library has implementations of popular mathematical operations that have been optimized for Intel hardware in a way that lets applications take full advantage of the Intel AVX-512 instruction set. It is compatible with a broad array of compilers, languages, operating systems, and linking and threading models.
- **Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN):** An open source, performance-enhancing library for accelerating DL frameworks on Intel hardware.

## WHAT ARE INTEL SELECT SOLUTIONS?

Intel Select Solutions are pre-defined, workload-optimized solutions designed to minimize the challenges of infrastructure evaluation and deployment. Solutions are validated by OEMs/ODMs, certified by ISVs, and verified by Intel. Intel develops these solutions in extensive collaboration with hardware, software, and operating system vendor partners and with the world's leading data center and service providers. Every Intel Select Solution is a tailored combination of Intel data center compute, memory, storage, and network technologies that delivers predictable, trusted, and compelling performance.

To refer to a solution as an Intel Select Solution, a vendor must:

1. Meet the software and hardware stack requirements outlined by the solution's reference-design specifications
2. Replicate or exceed established reference-benchmark test results
3. Publish a solution brief and a detailed implementation guide to facilitate customer deployment

Solution providers can also develop their own optimizations in order to give end customers a simpler, more consistent deployment experience.

## INTEL XEON SCALABLE PROCESSORS

2nd Generation Intel Xeon Scalable processors:

- Offer high scalability that is cost-efficient and flexible, from the multi-cloud to the intelligent edge
- Establish a seamless performance foundation to help accelerate data's transformative impact
- Support breakthrough Intel Optane DC persistent memory technology
- Accelerate artificial-intelligence (AI) performance and help deliver AI readiness across the data center
- Provide hardware-enhanced platform protection and threat monitoring

SOLUTION POWERED BY:

- **Intel Distribution for Python:** Accelerates AI-related Python libraries such as NumPy, SciPy, and scikit-learn with integrated Intel® Performance Libraries such as Intel MKL for faster AI inferencing.
- **Framework optimizations:** Intel has worked with Google on TensorFlow, with Apache on MXNet, with Baidu on PaddlePaddle, and on Caffe and PyTorch to enhance DL performance using software optimizations for Intel Xeon Scalable processors in the data center, and it continues to add frameworks from other industry leaders.

## Deploy Optimized, Fast AI Inferencing on Industry-Standard Hardware

Intel Select Solutions provide a fast path to data center transformation with workload-optimized configurations verified for Intel Xeon Scalable processors. When organizations choose Intel Select Solution for AI Inferencing, they get an optimized, pre-tuned, and tested configuration that is proven to scale so that IT can deploy AI inferencing in production environments quickly and efficiently. Moreover, IT organizations that choose Intel Select Solutions for AI Inferencing get high-speed AI inferencing on hardware that they are familiar with deploying and managing.

Visit intel.com/selectsolutions to learn more, and ask your infrastructure vendor for Intel Select Solutions.

## Learn More

Intel Select Solutions web page: **intel.com/content/www/us/en/architecture-and-technology/intel-select-solutions-overview.html**

Intel Xeon Scalable processors: **intel.com/xeonscalable**

Intel Optane DC technology: **intel.com/optane**

Intel SSD Data Center Family: **intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds.html**

Intel Ethernet products: **intel.com/content/www/us/en/products/network-io/ethernet.html**

Intel Ethernet 700 Series: **intel.com/ethernet**

Intel Distribution of OpenVINO toolkit: **https://software.intel.com/en-us/openvino-toolkit**

Intel Deep Learning Boost: **intel.ai/increasing-ai-performance-intel-dlboost**

Intel Framework Optimizations: **intel.ai/framework-optimizations**

Intel Deep Learning Reference Stack: **https://software.intel.com/en-us/blogs/2018/12/07/intel-introduces-the-deep-learning-reference-stack**

Intel Select Solutions are supported by Intel® Builders: **builders.intel.com**. Follow us on Twitter: **#IntelBuilders**

Kubeflow: **kubeflow.org**

Seldon Core: **seldon.io/tech/products/core/**

Seldon Deploy: **seldon.io/tech/products/deploy/**

[1] Intel. "Lower Numerical Precision Deep Learning Inference and Training." October 2018. https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training.

[2] Based on internal Intel testing. Source: Intel. "Product Brief: Intel Optane SSD DC P4800X Series." intel.com/content/www/us/en/solid-state-drives/optane-ssd-dc-p4800x-brief.html.

[3] Testing based on the 2nd Generation Intel Xeon Platinum 8260 processor and upgrading from a 1Gb to a 25Gb Intel® Ethernet Network Adapter XXV710 and from Serial ATA (SATA) drives to the NVM Express (NVMe)-based PCIe Intel SSD DC P4600.

[4] Performance results by HeadGear Strategic Communications are based on testing as of February 12, 2019. The comparative analysis in this document was done by HeadGear Strategic Communications and commissioned by Intel. Detailed configuration details: **Virtual Machine (VM) Host Server:** Intel Xeon Platinum 8160 processor, Intel Xeon Platinum 8160F processor (CPUID 50654, microcode revision 0x200004D), and Intel Xeon Platinum 8260 processor (CPUID 50656, microcode revision 04000014); Intel® Server Board S2600WFT (board model number H48104-850, BIOS ID SE5C620.86B.0D.01.0299.122420180146, baseboard management controller [BMC] version 1.88.7a4eac9e; Intel® Management Engine [Intel® ME] version 04.01.03.239; SDR package revision 1.88); 576 GB DDR4 2,133 MHz registered memory, 1 x Intel Ethernet Network Adapter XXV710-DA2, 1 x Intel Ethernet Converged Network Adapter X710-DA2; operating system drive configuration: 2 x Intel SSD DC S3500 in Intel® Rapid Storage Technology enterprise [Intel® RSTe] RAID1 configuration. Windows Server 2016 Datacenter edition 10.0.14393 build 14393, Hyper-V version 10.0.14393.0, Hyper-V scheduler type 0x3, installed updates KB4457131, KB4091664, KB1322316, KB3211320, and KB3192137. **E-mail Virtual-Machine Configuration:** Windows Server 2012 Datacenter edition 6.2.9200 build 9200; 4 x vCPU; 12 GB system memory, BIOS version/date: Hyper-V release v1.0, 2012, 11/26), SMBIOS version 2.4; Microsoft Exchange Server 2013, workload generation via VM clients running Microsoft Exchange Load Generator 2013, application version 15.00.0805.000). **Database Virtual-Machine Configuration:** Windows Server 2016 Datacenter edition 10.0.14393 build 14393, 2 x vCPU 7.5 GB system memory; BIOS version/date: Hyper-V release v1.0, 2012, 11/26), SMBIOS version 2.4, Microsoft SQL Server 2016 workload generation DVD Store application (dell.com/downloads/global/power/ps3q05-20050217-Jaffe-OE.pdf). **Storage Server:** Intel® Server System R2224WFTZS; Intel Server Board S2600WFT (board model number H48104-850, BIOS ID SE5C620.86B.00.01.0014.070920180847, BMC version 1.60.56383bef; Intel ME version 04.00.04.340; SDR package revision 1.60); 96 GB DDR4 2,666 MHz registered memory, 1 x Intel Ethernet Network Adapter XXV710-DA2, 1 x Intel Ethernet Converged Network Adapter X710-DA2; operating system drive configuration: 2 x Intel SSD DC S3500 in Intel RSTe RAID1 configuration. **Storage Configuration:** 8 x Intel SSD DC P4600 (2.0 TB) configured as RAID 5 volume using Intel® Virtual RAID on CPU (Intel® VROC), 8 x Intel SSD DC S4500 (480 GB) in RAID5 configuration using Intel® RAID Module RMSP3AD160F, 8 x Intel SSD DC P4510 in RAID 5 configuration using Intel VROC for VM operating system store, Windows Server 2016 Datacenter edition 10.0.14393 build 14393, Hyper-V version 10.0.14393.0, Hyper-V scheduler type 0x3, installed updates KB4457131, KB4091664, KB1322316, KB3211320, and KB3192137. **Windows Server 2016 Datacenter and Windows Server 2012 Datacenter Configured with Intel Xeon Platinum 8160 and Intel Xeon Platinum 8160F Processors:** Speculation control settings for CVE-2017-5715 (branch target injection)—hardware support for branch target injection mitigation is present: true; Windows operating system support for branch target injection mitigation is present: true; Windows operating system support for branch target injection mitigation is enabled: true; Windows operating system support for branch target injection mitigation is disabled by system policy: false; Windows operating system support for branch target injection mitigation is disabled by absence of hardware support: false. Speculation control settings for CVE-2017-5754 (rogue data cache load)—hardware requires kernel VA shadowing: true; Windows operating system support for kernel VA shadow is present: true; Windows operating system support for kernel VA shadow is enabled: true. Speculation control settings for CVE-2018-3639 (speculative store bypass)—hardware is vulnerable to speculative store bypass: true; hardware support for speculative store bypass disable is present: true; Windows operating system support for speculative store bypass disable is present: true; Windows operating system support for speculative store bypass disable is enabled system-wide: true. Speculation control settings for CVE-2018-3620 (L1 terminal fault)—hardware is vulnerable to L1 terminal fault: true; Windows operating system support for L1 terminal fault mitigation is present: true, Windows operating system support for L1 terminal fault mitigation is enabled: true. **Windows Server 2016 Datacenter and Windows Server 2012 Datacenter Configured with Intel Xeon Platinum 8160 and Intel Xeon 8160F Processors:** Speculation control settings for CVE-2017-5715 (branch target injection)—hardware support for branch target injection mitigation is present: true; Windows operating system support for branch target injection mitigation is present: true; Windows operating system support for branch target injection mitigation is enabled: true. Speculation control settings for CVE-2017-5754 (rogue data cache load)—hardware requires kernel VA shadowing: false. Speculation control settings for CVE-2018-3639 (speculative store bypass)—hardware is vulnerable to speculative store bypass: true; hardware support for speculative store bypass disable is present: true; Windows operating system support for speculative store bypass disable is present: true; Windows operating system support for speculative store bypass disable is enabled system-wide: true. Speculation control settings for CVE-2018-3620 (L1 terminal fault)—hardware is vulnerable to L1 terminal fault: false. **Network Switches:** 1/10GbE Supermicro SSE-X3348S, hardware version P4-01, firmware version 1.0.7.15; 10/25GbE Arista DCS-7160-48YC6, EOS 4.18.2-REV2-FX.

[5] The Intel Ethernet 700 Series includes extensively tested network adapters, accessories (optics and cables), hardware, and software, in addition to broad operating system support. A full list of the product portfolio's solutions is available at intel.com/ethernet. Hardware and software is thoroughly validated across Intel Xeon Scalable processors and the networking ecosystem. The products are optimized for Intel architecture and a broad operating system ecosystem: Windows, Linux kernel, FreeBSD, Red Hat Enterprise Linux (RHEL), SUSE, Ubuntu, Oracle Solaris, and VMware ESXi. Supported connections and media types for the Intel Ethernet 700 Series are: direct-attach copper and fiber SR/LR (QSFP+, SFP+, SFP28, XLPPI/CR4, 25G-CA/25G-SR/25G-LR), twisted-pair copper (1000BASE-T/10GBASE-T), backplane (XLAUI/XAUI/SFI/KR/KR4/KX/SGMII). Note that Intel is the only vendor offering the QSFP+ media type. The Intel Ethernet 700 Series supported speeds include 10GbE, 25GbE, 40GbE.

[6] 39.47 percent performance increase for the Plus configuration over the Base configuration derived from benchmark testing conducted by Intel using the ImageNet dataset classification with ResNet-50 on the OpenVINO toolkit on May 23, 2019. **Base configuration:** single node, 2 x Intel Xeon Gold 6248 processor (2.50 GHz, 20 cores, 40 threads), 12 x 16 GB 2,666 MHz DDR4 ECC RDIMM (192 GB total memory), boot drive: 1 x 256 GB Intel SSD DC P4101 (M.2 80 mm PCIe 3.0 x4, 3D2, TLC), data drive: 1.6 TB NVM Express (NVMe) Intel SSD P4510, cache drive: 375 GB Intel Optane SSD DC P4800X U.2 NVMe SSD, data network: 1 x 10Gb Intel Ethernet Converged Network Adapter XXV710-DA2, management network: integrated 1 gigabit Ethernet (GbE) port 0/RMM port. Software: CentOS Linux release 7.5.1804/Red Hat Enterprise Linux (RHEL) 7, Intel Math Kernel Library (Intel MKL) version 2018 update 3, Intel Distribution of OpenVINO toolkit 2019 R1 Runtime, OpenVINO Model Server 0.4, TensorFlow 1.14, PyTorch 1.01, MXNet 1.31, Intel Distribution for Python 2019 update 1, Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) 0.18 (implied by OpenVINO), Deep Learning Reference Stack (DLRS) v4.0, Kubernetes v1.15.1, Kubeflow v0.6.1, Seldon Core, Ceph v14.2.1, Min.io (Rook v1.0) RELEASE.2019-04-23T23-50-36Z. ImageNet Dataset Classification using ResNet-50 on OpenVINO toolkit: 1,880 images per second (91 percent top-5 accuracy). **Plus configuration:** single node, 2 x Intel Xeon Platinum 8268 processor (2.90 GHz, 24 cores, 48 threads), 12 x 32 GB 2,934 MHz DDR4 ECC RDIMM (384 GB total memory), boot drive: 1 x 256 GB Intel SSD DC P4101 (M.2 80 mm PCIe 3.0 x4, 3D2, TLC), data drive: 1.6 TB NVMe Intel SSD P4510, cache drive: 375 GB Intel Optane SSD DC P4800X U.2 NVMe SSD, data network: 1 x 10Gb Intel Ethernet Converged Network Adapter XXV710-DA2, management network: integrated 1 GbE port 0/RMM port. Software: CentOS Linux release 7.5.1804/RHEL 7, Intel MKL version 2018 update 3, Intel Distribution of OpenVINO toolkit 2019 R1 Runtime, OpenVINO Model Server 0.4, TensorFlow 1.14, PyTorch 1.01, MXNet 1.31, Intel Distribution for Python 2019 update 1, Intel MKL-DNN 0.18 (implied by OpenVINO), DLRS v4.0, Kubernetes v1.15.1, Kubeflow v0.6.1, Seldon Core, Ceph v14.2.1, Min.io (Rook v1.0) RELEASE.2019-04-23T23-50-36Z. ImageNet Dataset Classification using ResNet-50 on OpenVINO toolkit: 2,650 images per second (91 percent top-5 accuracy).

[7] Testing conducted by Intel on October 9, 2019. Test configuration: two nodes, 2 x Intel Xeon Gold 6248 processor (2.50 GHz, 20 cores, 40 threads), 12 x 16 GB 2,666 MHz DDR4 ECC RDIMM (192 GB total memory), boot drive: 1 x 256 GB Intel SSD DC P4101 (M.2 80 mm PCIe 3.0 x4, 3D2, TLC), data drive: 1.6 TB NVM Express (NVMe) Intel SSD P4510, cache drive: 375 GB Intel Optane SSD DC P4800X U.2 NVMe SSD, data network: 1 x 10Gb Intel Ethernet Converged Network Adapter X722, management network: integrated 1 gigabit Ethernet (GbE) port 0/RMM port. Software: CentOS Linux release 7.6.1810/Red Hat Enterprise Linux (RHEL) 7, Intel Math Kernel Library (Intel MKL) version 2019 update 4, Intel Distribution of OpenVINO toolkit 2019 R1.0.1, OpenVINO Model Server 0.4, Intel Distribution for Python 2019 update 1, Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) 0.19, Deep Learning Reference Stack (DLRS) v4.0, Docker v18.09, Helm v2.14.3, Kubernetes v1.15.3, Kubeflow v0.6.1, Seldon Core v0.3.2, Rook v1.0.5, Ceph v14.2.1, Min.io (Rook v1.0) RELEASE.2019-04-23T23-50-36Z. Scaling in emulated real-world scenario—throughput test: normalized performance 1 (Intel® Hyper-Threading Technology: off).

[8] Testing conducted by Intel on October 9, 2019. Test configuration: two nodes, 2 x Intel Xeon Platinum 8268 processor (2.90 GHz, 24 cores, 48 threads), 12 x 16 GB 2,666 MHz DDR4 ECC RDIMM (192 GB total memory), boot drive: 1 x 256 GB Intel SSD DC P4101 (M.2 80 mm PCIe 3.0 x4, 3D2, TLC), data drive: 1.6 TB NVM Express (NVMe) Intel SSD P4510, cache drive: 375 GB Intel Optane SSD DC P4800X U.2 NVMe SSD, data network: 1 x 10Gb Intel Ethernet Converged Network Adapter X722, management network: integrated 1 gigabit Ethernet (GbE) port 0/RMM port. Software: CentOS Linux release 7.6.1810/Red Hat Enterprise Linux (RHEL) 7, Intel Math Kernel Library (Intel MKL) version 2019 update 4, Intel Distribution of OpenVINO toolkit 2019 R1.0.1, OpenVINO Model Server 0.4, Intel Distribution for Python 2019 update 1, Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) 0.19, Deep Learning Reference Stack (DLRS) v4.0, Docker v18.09, Helm v2.14.3, Kubernetes v1.15.3, Kubeflow v0.6.1, Seldon Core v0.3.2, Rook v1.0.5, Ceph v14.2.1, Min.io (Rook v1.0) RELEASE.2019-04-23T23-50-36Z. Scaling in emulated real-world scenario—throughput test: normalized performance 1.91 (Intel Hyper-Threading Technology: off).

Performance results are based on testing as of the date set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

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.

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

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.

Printed in USA 1119/MM/PRW/PDF Please Recycle 341602-001US