SOLUTION BRIEF

Intel® Select Solutions
Artificial Intelligence
April 2019

Intel® Select Solution for AI Inferencing

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

The resources needed to support inferencing on deep neural networks can be substantial. These operational needs typically drive organizations to update their hardware. However, investing in single-purpose hardware for inferencing can leave you exposed if your computational needs change before your expected refresh. High performance and speed for AI inferencing, coupled with the general-purpose flexibility of the Intel® hardware that your IT department is already familiar with, can meet your current needs and help protect your IT investments.

The Intel® Select Solution for AI Inferencing is a "turnkey platform" solution for low-latency, high-throughput inference performed on a CPU, not a separate accelerator card. It provides you with a jumpstart to deploying efficient AI inferencing algorithms on a solution composed of validated Intel architecture building blocks that you can innovate on and take to market. To do so, this solution makes use of a feature of 2nd Generation Intel® Xeon® Scalable processors, Intel® Deep Learning Boost (Intel® DL Boost), which accelerates AI inference by performing in one instruction inferencing calculations that previously took multiple instructions.

The Intel Select Solution for AI Inferencing also uses the Intel® Distribution of Open Visual Inference and Neural Network Optimization toolkit (Intel® Distribution of OpenVINO™ toolkit), a developer suite that accelerates high-performance deep learning (DL) inference deployments for computer-vision workloads. The toolkit takes models trained in different frameworks and optimizes them for different Intel hardware options in order to provide you with high flexibility for deployment. The toolkit also quantizes DL models, a process in which the toolkit transforms models from using large, high-precision 32-bit floating-point numbers, which are typically used for training, to using 8-bit integers. 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*, Apache MXNet*, and any framework supported by the Open Neural Network Exchange* (ONNX*) ecosystem. The Intel Select Solution for AI Inferencing also provides large amounts of memory to enable inferencing against larger—and more—models simultaneously.

Through the use of Intel DL Boost and the OpenVINO toolkit, in addition to other Intel hardware and software technologies, the Intel Select Solution for AI Inferencing can accelerate time to business insights from enterprise data and provide low latency and high end-to-end throughput for AI inferencing. And because the solution uses industry-standard, readily available IT components, it can help lower total cost of ownership (TCO) compared to using specialized hardware for inferencing. The Intel Select Solution for AI Inferencing uses Intel Xeon Scalable processors and Intel® Solid State Drives (SSDs) for high performance and reliability, which provides a verified, tested solution and simplifies deployment.

Intel Select Solution for AI Inferencing

The Intel Select Solution for AI Inferencing helps optimize price and performance while significantly reducing infrastructure evaluation time. Specifically, it combines the Intel Xeon Scalable processor platform, Intel SSDs, and Intel® Ethernet Network Adapters to empower enterprises to quickly harness a reliable, comprehensive solution that allows organizations to:

- **Prepare** AI infrastructure investments for the future with scalable compute and storage
- **Generate excellent TCO** with general-purpose hardware your IT organization is used to managing
- **Accelerate time to market** with a turnkey solution optimized for crucial software libraries and frameworks

Solution Components

The Intel Select Solution for AI Inferencing combines Intel compute, storage, and networking hardware to enable businesses to quickly deploy fast AI inferencing solutions on a performance-optimized infrastructure.

Intel® Xeon® Scalable Processor Family

2nd Generation Intel Xeon Gold processors provide the solution with an excellent performance-to-cost ratio and built-in technologies that enhance performance and efficiency for inferencing on AI models:

- **Intel® Advanced Vector Extensions 512 (Intel® AVX-512),** which provides 512-bit instructions that can accelerate performance for demanding workloads and usages like AI inferencing
- **Intel DL Boost,** which accelerates AI inference by doing in one instruction on the processor what previously took multiple instructions

Intel® SSD Data Center Family

Storage latency and size can be bottlenecks for AI inference. For this reason, the solution uses Intel® SSD DC P4610 drives for data storage. Based on Intel® 3D NAND technology, these enterprise data center SSDs provide a 3.2x lower annualized failure rate (AFR) than hard-disk drives (HDDs).[2]

For data caching, the solution uses the Intel® Optane™ SSD DC P4800X. These SSDs are based on Intel Optane technology and provide extremely low read latency to accelerate inferencing. Intel Optane SSD DC P4800X drives also economically accommodate larger cache sizes, which permits simultaneously caching more and larger DL models to improve inferencing performance.

Intel® Ethernet Connections and Intel® Ethernet Adapters

The 25Gb Intel® Ethernet 700 Series Network Adapters accelerate the performance of the Intel Select Solution for AI Inferencing. The Intel Ethernet 700 Series delivers validated performance ready to meet high-quality thresholds for data resiliency and service reliability with broad interoperability.[3] All Intel Ethernet products are backed by worldwide pre- and post-sales support and offer a limited lifetime warranty.

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 AI performance and help deliver AI readiness across the data center
- Provide hardware-enhanced platform protection and threat monitoring

The Intel Select Solution for AI Inferencing features 2nd Generation Intel Xeon Gold processors.

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 across the data center and on the network edge, Intel chose to measure and benchmark the number of images that can be inferenced per second (throughput) 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 and the OpenVINO toolkit.

Base Configuration

The Intel Select Solution for AI Inferencing is available in a "Base" configuration, as shown in Table 1. The Base configuration specifies the minimum required performance capability for the solution.

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

Table 1. The Base configuration for the Intel® Select Solution for AI Inferencing

INGREDIENT	INTEL® SELECT SOLUTION FOR AI INFERENCING **BASE CONFIGURATION**
SINGLE NODE	
PROCESSOR	2 x Intel® Xeon® Gold 6248 processor (3.40 GHz, 6 cores, 12 threads), or a higher number Intel Xeon Scalable processor
MEMORY	192 GB or higher (12 x 16 GB DDR4-2,666 MHz ECC RDIMM)
BOOT DRIVE	1 x 256 GB Intel® SSD DC P4101 (M.2 80 mm PCIe* 3.0 x4, 3D2, TLC) or higher
DATA TIER: DATA DRIVE	1.6 TB Intel SSD DC P4610 15 mm U.2 NVM Express* (NVMe*)
DATA TIER: CACHE DRIVE	375 GB Intel® Optane™ SSD DC P4800X U.2 NVMe
DATA NETWORK	25Gb Intel® Ethernet Converged Network Adapter XXV710-DA2 SFP28 DA Copper PCIe x 8 dual-port 10/25 gigabit Ethernet (GbE)
MANAGEMENT NETWORK PER NODE	Integrated 1 GbE port 0/RMM port
SOFTWARE	
LINUX* OS	CentOS* Linux release 7.5.1804/Red Hat* Enterprise Linux (RHEL*) 7
INTEL® MATH KERNEL LIBRARY (INTEL® MKL)	Intel MKL version 2018 Update 3
INTEL® MATH KERNEL LIBRARY FOR DEEP NEURAL NETWORKS (INTEL® MKL-DNN)	0.17 (included with the Intel® Distribution of OpenVINO™ toolkit)
INTEL DISTRIBUTION OF OPENVINO TOOLKIT	2018 R5
OPENVINO MODEL SERVER	0.3
TENSORFLOW*	1.12
PYTORCH*	1.0.0
APACHE MXNET*	1.3.1
INTEL® DISTRIBUTION FOR PYTHON*	2019 Update 1
APPLIES TO ALL NODES	
FIRMWARE AND SOFTWARE OPTIMIZATIONS	Intel® Volume Management Device (Intel® VMD) enabled** Intel® Boot Guard enabled** Intel® Hyper-Threading Technology (Intel® HT Technology) disabled Intel® Turbo Boost Technology enabled P-states enabled** C-states enabled** Power-management settings set to performance** Workload configuration set to balanced** Intel® Memory Latency Checker (Intel® MLC) streamer enabled** Intel MLC spatial prefetch enabled** Data Cache Unit (DCU) data prefetch enabled** DCU instruction prefetch enabled** Last-level cache (LLC) prefetch disabled** Uncore frequency scaling enabled**
MINIMUM PERFORMANCE STANDARDS	
Verified to meet or exceed the following minimum performance capabilities:	
IMAGENET DATA SET CLASSIFICATION USING RESNET-50 ON OPENVINO TOOLKIT	2,000 images per second (91 percent top-5 accuracy)[4]
IMAGENET DATA SET CLASSIFICATION USING RESNET-50 ON TENSORFLOW FRAMEWORK	1,300 images per second (91 percent top-5 accuracy)[5]

**Recommended, not required

Technology Selections for the Intel Select Solution for AI Inferencing

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

- **Intel Distribution of OpenVINO toolkit:** A command-line tool based on Python* that imports trained models from popular DL frameworks such as Caffe*, TensorFlow, and MXNet, in addition to any framework supported by ONNX. The OpenVINO toolkit performs analysis and adjustments for optimal inferencing on trained DL models on endpoint devices. Models serialized and adjusted by the OpenVINO toolkit are hardware agnostic and can run using smaller, faster, eight-bit integers (as opposed to 32-bit floating point numbers) with little loss of accuracy.
- **Intel® Math Kernel Library (Intel® MKL):** This library optimizes code for future generations of Intel processors with minimal effort. 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.
- **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.
- **AI frameworks optimized on Intel architecture:**
  - **TensorFlow:** This Python-based DL framework is designed for ease of use and extensibility on modern deep neural networks and has been optimized for use on Intel Xeon processors.
  - **Apache MXNet:** This open source, DL framework includes built-in support for Intel MKL and optimizations for Intel AVX-512 instructions.

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.

Deploy Optimized, Fast AI Inferencing on Industry-Standard, General-Purpose 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 the 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 solutions quickly and efficiently. Moreover, IT organizations that choose the Intel Select Solution for AI Inferencing get high-speed AI inferencing on general-purpose 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: **intel.com/selectsolutions**

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

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

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

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

Intel Deep Learning Boost: **intel.ai/intel-deep-learning-boost**

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

[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 initial product AFR of 0.66 percent vs. industry AFR average (2.11%). Source: Backblaze. "Hard Drive Stats for Q1 2017." May 2017. backblaze.com/blog/hard-drive-failure-rates-q1-2017/.

[3] 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.

[4] Testing conducted by Intel on February 26, 2019, on the Base hardware configuration detailed in Table 1.

[5] Testing conducted by Intel on March 7, 2019, on the Base hardware configuration detailed in Table 1.

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 does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Optimization Notice: 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. Notice Revision #20110804

Printed in USA 0419/LU/PRW/PDF Please Recycle 338879-001US