Solution Brief

Intel® Select Solutions
High Performance Computing and AI
March 2021

Intel Select Solutions for HPC & AI Converged Clusters

Intel Select Solutions deliver the compute-intensive resources needed to run artificial-intelligence (AI) workloads on existing high-performance computing (HPC) clusters.

Enterprises are looking to simulation and modeling, artificial intelligence (AI), and big data analytics to help them achieve breakthrough discoveries and innovation. They understand these workloads benefit from a high-performance computing (HPC) infrastructure, yet they might still believe that separate HPC, AI, and big data clusters are the best choice for running these workloads.

Contributing to this belief are two challenges. The first challenge is a fundamental difference in how workloads request resources and how HPC systems allocate them. AI and analytics workloads request compute resources dynamically, an approach that isn't compatible with batch scheduling software used to allocate system resources in HPC clusters.

The second challenge is the pattern of using computing systems based on graphics processing units (GPUs) as dedicated solutions for AI workloads. Enterprises might not realize that adding these workloads to an existing HPC cluster is feasible without the use of GPUs.

Enterprises can deliver the compute infrastructure needed by AI workloads, with high levels of performance and cost-effectiveness, without adding the complexity of managing separate, dedicated systems. What they need is the ability to run HPC, big data analytics, and AI workloads within the same HPC infrastructure. They also need optimized resource scheduling that helps save time and reduce computing costs.

Creating a converged platform to run HPC, AI, and analytics workloads in a single cluster infrastructure supports breakthrough innovation. This innovation is made possible by the convergence of these workloads, while increasing the value and utilization of resources. Enter Intel Select Solutions for HPC & AI Converged Clusters. Intel Select Solutions are verified hardware and software stacks optimized across compute, storage, and networking resources for specific workloads. Built on Intel® Xeon® Scalable processors, Intel Select Solutions help ensure enterprises get the performance, agility, and security they require.

Intel Select Solutions for HPC & AI Converged Clusters

Core capabilities in Intel Select Solutions for HPC & AI Converged Clusters are delivered by a solution that runs AI workloads within an HPC environment. The architecture enables HPC batch schedulers to run all workloads—including simulation and modeling, big data analytics, and AI—on a common HPC infrastructure. It also enables partners to help customers build upon existing HPC investments to start running AI and big data workloads.

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 solution content 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™ 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

Intel Select Solutions for HPC & AI Converged Clusters feature Intel Xeon Gold processors.

Solution powered by: intel XEON GOLD

These solutions combine Intel Xeon Scalable processors, a high-performance parallel file system for storage, and Omni-Path Architecture (OPA) to deliver support for multiple types of workloads in the same infrastructure. This multi-workload support means:

- Customers can start their AI journeys on existing HPC infrastructures and potentially reduce the total cost of ownership (TCO) for HPC because Intel Xeon Scalable processor-based HPC environments do not require specialized hardware to run AI workloads
- Faster time to insights with improvements in AI inferencing
- No more burden of data transfer between multiple environments, reducing the time to results for data analytics and AI training runs
- Hybrid workflows supported in the same infrastructure, a capability that allows the solutions to make use of resources and improve efficiency across HPC, AI, and data-analytics workloads in a single environment

The Intel Select Solutions support advanced capabilities to run machine learning, deep-learning training models, and data analytics on the same HPC cluster. For example, the solutions help users to run Intel-optimized TensorFlow models on an HPC system. TensorFlow is a deep learning framework based on Python and designed for ease of use and extensibility on modern deep neural networks (DNNs), and it has been optimized for use on Intel Xeon processors. In addition, Apache Spark support in the solutions helps with machine learning and data analytics.

The solutions also provide a cohesive HPC and AI software stack with integrated open source tools for batch scheduling. This approach can reduce system complexity and licensing costs and can support hybrid workloads in the same HPC infrastructure.

Intel Select Solutions are verified solutions that combine Intel Xeon Scalable processors and other Intel technologies into a proven architecture based on the Intel HPC Platform Specification. This specification defines common industry practices and requirements for building Intel-based clusters. As an architectural foundation, the specification provides a consistent and stable platform, enabling development and deployment of a wide variety of compute-intensive and data-intensive workloads. Included in the foundation are the Intel software performance libraries and runtime environments that allow applications to experience optimized value from the underlying Intel processors and technologies. The Intel HPC Platform Specification enables organizations to achieve high performance with flexibility, scalability, balance, and portability.

The Intel Select Solutions for HPC & AI Converged Clusters simplify the challenge of building an HPC cluster and are designed to provide optimized performance for highly demanding hybrid workloads. In addition, the solutions are validated to ensure they:

- Include key components and technologies to deliver performance and scalability
- Comply with industry standards and best practices for Intel-based clusters, as defined in the Intel HPC Platform Specification
- Meet or exceed defined performance levels in targeted characteristics important to HPC applications

Hardware and software selections

Intel Select Solutions for HPC & AI Converged Clusters include several key hardware and software components. The solutions are built on top of Intel Select Solutions for Simulation & Modeling, with hardware that provides the right performance for converged HPC, AI, and big data analytics workloads.

Compute

These solutions use Intel Xeon Gold 6248R processors. Intel Xeon Scalable processors feature significant enhancements that can benefit HPC applications, including improvements in input/output (I/O), memory, fabric integration, and Intel Advanced Vector Extensions 512 (Intel AVX-512):[1]

- For HPC users adopting AI, the Intel Deep Learning Boost (Intel DL Boost) capability makes the configurations even more compelling because it accelerates AI workloads, increasing Int16 and Int8 peak operations/second. Intel DL Boost was designed to accelerate performance of AI deep learning (inference) workloads (for example, speech recognition, image recognition, object classification, machine translation, and others).
- Existing Intel AVX-512 fused-multiple add (FMA) instructions deliver significant performance for floating-point operations. However, with Intel DL Boost, the performance acceleration extends to integer operations and handles dense computations characteristic of convolutional neural network (CNN) and DNN workloads.

Intel Select Solutions for HPC & AI Converged Clusters use the following additional hardware:

- SSDs: Intel SSD DC S4610 Series
- Storage: HPC parallel file system
- Message fabric: Intel Omni-Path Host Fabric Interface Adapter 100 Series
- Management network switch: 10 gigabit Ethernet (GbE) switch

Fabric

OPA provides 100 gigabits per second (Gbps) bandwidth and a low-latency fabric for HPC clusters. OPA can also reduce cabling-related costs, power consumption, space requirements, and ongoing system-maintenance requirements.

Software

Software in the solutions includes the Slurm batch scheduler and the Magpie orchestration layer for Apache Spark and Alluxio. Magpie can also be used with other schedulers like Torque and LFS, and it provides a wide range of frameworks. As open source software, Magpie is less intrusive to the production software stack than its closed-source counterparts, and it supports multiple resource managers.

Additional software in the solution includes:

- The Linux operating system
- Intel Cluster Runtimes
- Intel Cluster Checker
- Intel HPC Platform RPM packages
- OpenHPC
- Intel Omni-Path Fabric software
- Intel Omni-Path Host Fabric Interface Driver
- Intel® OpenVINO™ toolkit
- Apache Spark
- TensorFlow
- Horovod

The latest release also includes the addition of Alluxio for improved file system I/O performance. HPC parallel file systems do not process small files efficiently, which can create a bottleneck with metadata-heavy big data analytics workloads. Alluxio efficiently buffers this metadata, and it is able to provide improved parallel file system I/O performance with big data analytics and AI workloads on HPC infrastructure.

Verified performance through benchmark testing

All Intel Select Solutions are verified to meet a specified minimum level of workload-optimized performance capabilities. Intel Select Solutions for HPC & AI Converged Clusters use the same performance watermarks as the Intel Select Solutions for Simulation & Modeling, which demonstrate optimized capabilities for HPC applications. These verified solutions meet or exceed design and testing standards across eight well-known industry benchmarks. These benchmarks cover important system aspects and indicate potential scale-up and scale-out performance for big data and AI workloads.

Intel Select Solutions for HPC & AI Converged Clusters also use the following benchmarks to verify performance: the TensorFlow ResNet-50 benchmark and the Spark-Bench suite of tests.

Configuration details

The Intel Select Solutions for HPC & AI Converged Clusters configuration is shown in Table 1.

Table 1. The Intel Select Solutions for HPC & AI Converged Clusters configuration

Ingredient	Intel Select Solutions for HPC & AI Converged Clusters **configuration details**
Application node	
CPU	2 x Intel Xeon Gold 6248R processor (or higher)
Memory	384 GB DRAM per node
4 x compute nodes	
CPU	2 x Intel Xeon Gold 6248R processor (or higher)
Memory	384 GB DRAM
Storage (boot)	240 GB Intel SSD DC S4610
Storage (capacity)	HPC parallel file system (470 megabits per second [Mbps] per client)
Network fabric	Omni-Path Host Fabric Interface Adapter 100 Series
Batch scheduler	Open source Magpie on SLURM
Software	CentOS Linux installation ISO (minimal or full) 7 build 2003 Intel Cluster Runtimes 2020.2 Intel Cluster Checker 2019.9 or higher Intel HPC Platform RPM packages for EL7 Software version 2018.0 OpenHPC1.3.9 or later Omni-Path Fabric Software (includes Omni-Path HFI Driver) Software version 10.10.3.1.1 or later Intel OpenVINO toolkit software version 2020.4 Intel Parallel Studio XE Cluster Edition 2020 Apache Spark version 2.4.6 Intel Optimization for TensorFlow v2.3.0 Horovod v0.19.0 Alluxio v2.3.0

Technology selections for Intel Select Solutions for HPC & AI Converged Clusters

In addition to the Intel hardware foundation used for these solutions, the following Intel technologies integrated in Intel Xeon Scalable processors deliver further performance and reliability gains:

- **Intel AVX-512:** Boosts performance for the most demanding computational workloads, with up to double the number of floating point operations per second (FLOPS) per clock cycle, compared to previous-generation Intel processors.[1]
- **Intel DL Boost:** The performance acceleration extends to integer operations and handles dense computations characteristic of CNN and DNN workloads. It accelerates AI workloads, increasing Int16 and Int8 peak operations/second. Intel DL Boost was designed to accelerate performance of AI deep learning (inference) workloads (for example, speech recognition, image recognition, object classification, machine translation, and others).
- **Intel Cluster Checker:** Inspects more than 100 characteristics related to cluster health. Intel Cluster Checker examines the system at both the node and cluster level, making sure all components work together to deliver optimal performance. It assesses firmware, kernel, storage, and network settings. It also conducts high-level tests of node and network performance using the Intel MPI Library benchmarks, STREAM, the High Performance LINPACK (HPL) benchmark, the High Performance Conjugate Gradients (HPCG) benchmark, and other benchmarks. Intel Cluster Checker can be extended with custom tests, and its functionality can be embedded into other software.
- **Intel Cluster Runtimes:** Supplies key software runtime elements that are required on each cluster to ensure optimal performance paths for applications. Intel runtime performance libraries, including Intel Math Kernel Library (Intel MKL) and Intel MPI Library, deliver excellent performance optimized for clusters based on Intel architecture.

- **Cluster Management Software Stack:** Provides a software stack that is required to deploy and manage Linux HPC clusters. The stack includes provisioning tools, resource management, I/O clients, development tools, and scientific libraries. Resource management tools such as Bright Cluster Manager, Warewulf, and xCAT support the software stack.
- **Converged parallel programming for Intel Xeon Scalable processors:** Enables the creation of a highly integrated portfolio of powerful technologies, software tools, and libraries. Intel Xeon Scalable processors offer an unparalleled flexible framework, based on a common programming model, that supports code-modernization initiatives across AI frameworks.

Part of the suite of Intel Select Solutions for HPC

Intel Select Solutions for HPC & AI Converged Clusters join the robust suite of Intel Select Solutions for HPC to address the most critical HPC workloads. As the foundation of Intel's HPC portfolio, Intel Select Solutions for Simulation & Modeling are Intel's most flexible solution for general HPC applications. To better visualize data, Intel Select Solutions for Professional Visualization build upon the functionality in Intel Select Solutions for Simulation & Modeling with optimizations to support simulation and visualization applications. Rounding out the HPC portfolio are the Intel Select Solutions for AI Inferencing and the Intel Select Solutions for Genomic Analytics.

Simplify deployments of AI workloads on HPC clusters

Intel Select Solutions for HPC & AI Converged Clusters combine Intel Xeon Scalable processors, OPA, and other Intel technologies with selected batch schedulers. They deliver optimized performance for running big data and AI workloads on HPC clusters with a single comprehensive, verified solution. Customers can begin their AI journeys today using existing, familiar infrastructure.

Visit intel.com/selectsolutions for more information on 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**

OPA: **intel.com/omnipath**

Intel HPC Platform Specification: **intel.com/content/www/us/en/high-performance-computing/hpc-platform-specification.html**

Intel HPC Application Catalog: **intel.com/content/www/us/en/high-performance-computing/hpc-application-catalog.html**

Magpie: **https://github.com/LLNL/magpie**

Where to buy: **intel.com/content/www/us/en/products/docs/select-solutions/where-to-buy.html**

[1]Intel Advanced Vector Extensions 512 (Intel AVX-512) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing Intel 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.

Printed in USA 0321/MM/PRW/PDF

Please Recycle 338948-003US