UNIFYING AI, ANALYTICS, AND HPC ON A SINGLE CLUSTER

Maximizing Efficiency and Lowering Costs for Tomorrow's Enterprise

Allene Bhasker and Keith Mannthey, Solution Architects, Data Center Group, Intel Corporation

The next few years will be remembered as the time when artificial intelligence (AI)—including machine and deep learning—became mainstream all over the enterprise. One study shows that more than 60% of enterprises are currently putting AI solutions in place, with predictive analytics as the most common application.

Hosting Strategy for AI, Analytics, and HPC Workloads

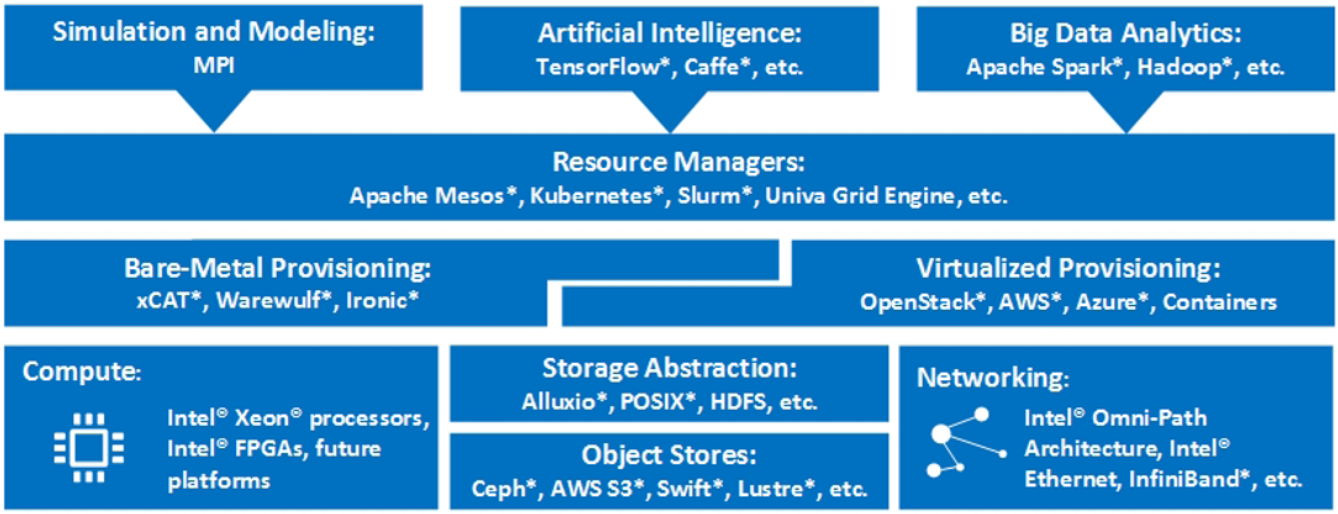
As IT organizations decide where to host AI workloads and AI-driven analytics, many consider purpose-built servers equipped with specialized accelerators and GPUs. Those who are forward-looking may think in terms of clusters of these servers to handle the expected growth of AI's role in their day-to-day operations. A broader perspective still notes that AI, analytics, and HPC workloads all run well on similar cluster hardware, based on robust individual cores and high-speed interconnects.

At this point, a common question arises: "What would it take to run AI, analytics, and HPC workloads together on the same cluster?" This approach is particularly attractive if you consider a business process that uses all three types of processing. For example, running simulation and modeling, data cleaning, and AI-based inference steps all on the same cluster is far more efficient than maintaining separate clusters.

Convergence onto a single cluster also has obvious cost benefits. Server utilization is higher with a single cluster, so you can buy fewer servers. A simpler environment is less expensive to configure and maintain—and lets you avoid expensive requirements to move and stage data among multiple clusters. Integrating these workloads onto a single environment also helps reduce latency, something that gets more important every year as real-time requirements emerge.

Build AI and Analytics Capabilities on the Existing HPC Platform

The approach to convergence focuses on adding AI and analytics capabilities on top of an HPC cluster. The Intel HPC Platform Specification defines requirements for a base cluster solution that includes common industry standards and practices for Intel-based solutions (**Figure 1**). This provides a common and consistent interface for HPC applications, and many commercial HPC software vendors have validated application support of solutions compliant with this platform specification.



1 **Generalized Intel solution stack for converged clusters**

Beyond the base definition, additional requirements for specific capabilities and functionalities are described in distinct sections. Compliant solutions are composed by meeting the requirements of the base solution plus the desired capability layers. This streamlines introduction or expansion of capabilities while still maintaining the interoperability with applications targeting the platform. The path to a converged platform involves adding new sections to the Intel HPC Platform Specification that describe the requirements for AI and analytics capabilities.

Combining Solutions Built for Different Customer Environments

Intel is developing a series of solution architectures to help define requirements that converge AI, analytics, and HPC workloads into a single, unified cluster. Making multiple resource managers work together smoothly is a daunting challenge. And the solutions implement various approaches to integrating capabilities such as maintaining job queues and scheduling jobs in a centralized way for all types of workloads.

- **Solution 1: Extend HPC Batch Schedulers.** This approach extends batch schedulers using wrapper scripts that submit jobs on behalf of AI and analytics workloads. This simple approach has almost no systems overhead.
- **Solution 2: Univa* Grid Engine and Resource Broker.** For shops that are already using Univa Grid Engine*, this solution uses Univa Resource Broker* to integrate Apache Mesos*-compatible AI and analytics software.
- **Solution 3: Apache Mesos and Batch Schedulers.** This forthcoming solution architecture integrates Apache Mesos and batch schedulers to work together seamlessly across HPC, AI, and analytics workloads.

The solution architectures are flexible in terms of supporting different ways of provisioning, whether on bare metal or with virtual machines and containers on hybrid clouds. They also include storage abstraction to unify data across object stores, providing a single source of data to be used in place, without large-scale data movement. Intel is involved with enablement activities across the software ecosystem, including open-source contributions and co-engineering with technology providers. This optimization work is key to making sure that all three types of workloads benefit from the full range of Intel® platform features for performance and security.

To make it easier to deploy converged cluster solutions, Intel makes pre-optimized, integrated infrastructure available through participating OEMs as **Intel® Select Solutions**. Because these architectures are validated in advance, mainstream enterprises now have a clear path to the efficiency and cost benefits of converged clusters for tomorrow's AI, analytics, and HPC workloads.