intel experience what's inside™

AI ON HPC INFRASTRUCTURE: GETTING STARTED IN THREE STEPS

Tips on Convergence

A truly converged high-performance computing (HPC) and artificial intelligence (AI) infrastructure cannot be achieved overnight. It's a complex journey of managing various stakeholders through multiple steps, and balancing strategic and technical requirements.

This document outlines three steps for any organization to get started with AI applications using its current HPC infrastructure.

Click on each section to find out more.

STEP 1 Think holistically about your requirements

- Needs
- Expectations
- Limitations

STEP 2 Think holistically about your AI solution

- Software Selection
- Software Application and Development
- Frameworks

STEP 3 Validate and scale your HPC technology

- Support
- Scaling Out

STEP 1: NEEDS

How are you planning to use AI to meet your mission or business needs?

If your organization currently uses HPC for simulation and modeling, start by evaluating where and how AI can benefit your mission or business needs. Also look at what data you will need to perform the desired AI application.

It is important to ensure the data strategy you develop involves secure data collection, storage, and retrieval. Also, in order to provide your stakeholders with an optimal AI-enabled HPC environment, your software, hardware, and human skills will all weigh into the equation. For environments where HPC systems need to support multiple users with unique workloads, such as in academia and government, ensuring system flexibility is key.

INSIGHTS:

Start by exploring some of the AI use cases that others have successfully deployed to gain inspiration.

Click on each section to find out more.

STEP 1 Think holistically about your requirements

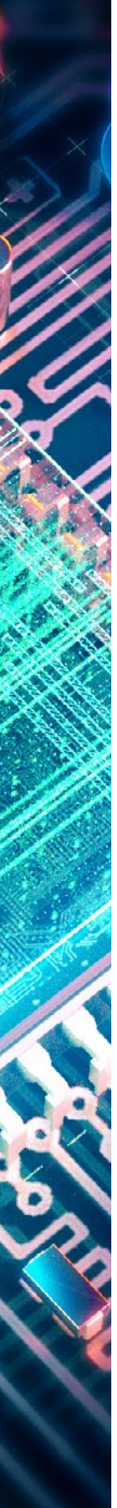
- Needs
- Expectations
- Limitations

STEP 2 Think holistically about your AI solution

- Software Selection
- Software Application and Development
- Frameworks

STEP 3 Validate and scale your HPC technology

- Support
- Scaling Out



STEP 1: EXPECTATIONS

What infrastructure do you need to start running AI applications?

It is important to keep in mind that AI is not a single, monolithic workload, nor is its future clearly defined. It is an evolving field where new algorithms and ideas are continuously shaping implementation.

To help future-proof your investments, look first at optimizing algorithms on your existing Intel® Xeon® processor-based HPC systems until you have clearly determined whether you need an accelerator for your model training needs.

INSIGHTS:

HPC systems based on Intel® Xeon® Scalable processors enable you to exploit the fundamental principle of data parallelism in deep learning training and allow you to benefit from large memory support on the server platform.

For organizations wanting to optimize their existing HPC infrastructure for specific workloads Intel® Select Solutions offer easy and quick-to-deploy infrastructure optimized for analytics clusters and HPC applications across simulation and modeling, professional visualization, genomics analytics and more. You can also ask for Intel Select Solutions from participating OEM partners.

Click on each section to find out more.

STEP 1 Think holistically about your requirements

- Needs
- **Expectations**
- Limitations

STEP 2 Think holistically about your AI solution

- Software Selection
- Software Application and Development
- Frameworks

STEP 3 Validate and scale your HPC technology

- Support
- Scaling Out

STEP 1: LIMITATIONS

What are the current limitations to running HPC, High Performance Data Analytics (HPDA), and AI applications on the same infrastructure?

The convergence of HPC and AI has begun. However, most HPC software stacks are not fully compatible with cloud-centric AI workloads and the developer skill sets needed for each are different. This creates a potential barrier for enterprises, government organizations and academia looking to integrate AI solutions into their HPC infrastructure.

To minimize current and future barriers while helping future-proof your investments, focus on using a familiar Intel® Xeon® processor-based HPC infrastructure that is optimized for a diverse and demanding set of applications, including HPC, HPDA, and AI.

INSIGHTS:

Intel is actively working with the broader AI ecosystem on bridging the gaps in this mixed workload environment through hardware tuning, system architecture, and software ecosystem engagements, as can be seen in *Figure 1: Unifying the three pillars – HPC, HPDA, and AI.*

Click here to see Figure 1.

Click on each section to find out more.

STEP 1: LIMITATIONS

Intel ecosystem engagements across HPC, HPDA, and AI.

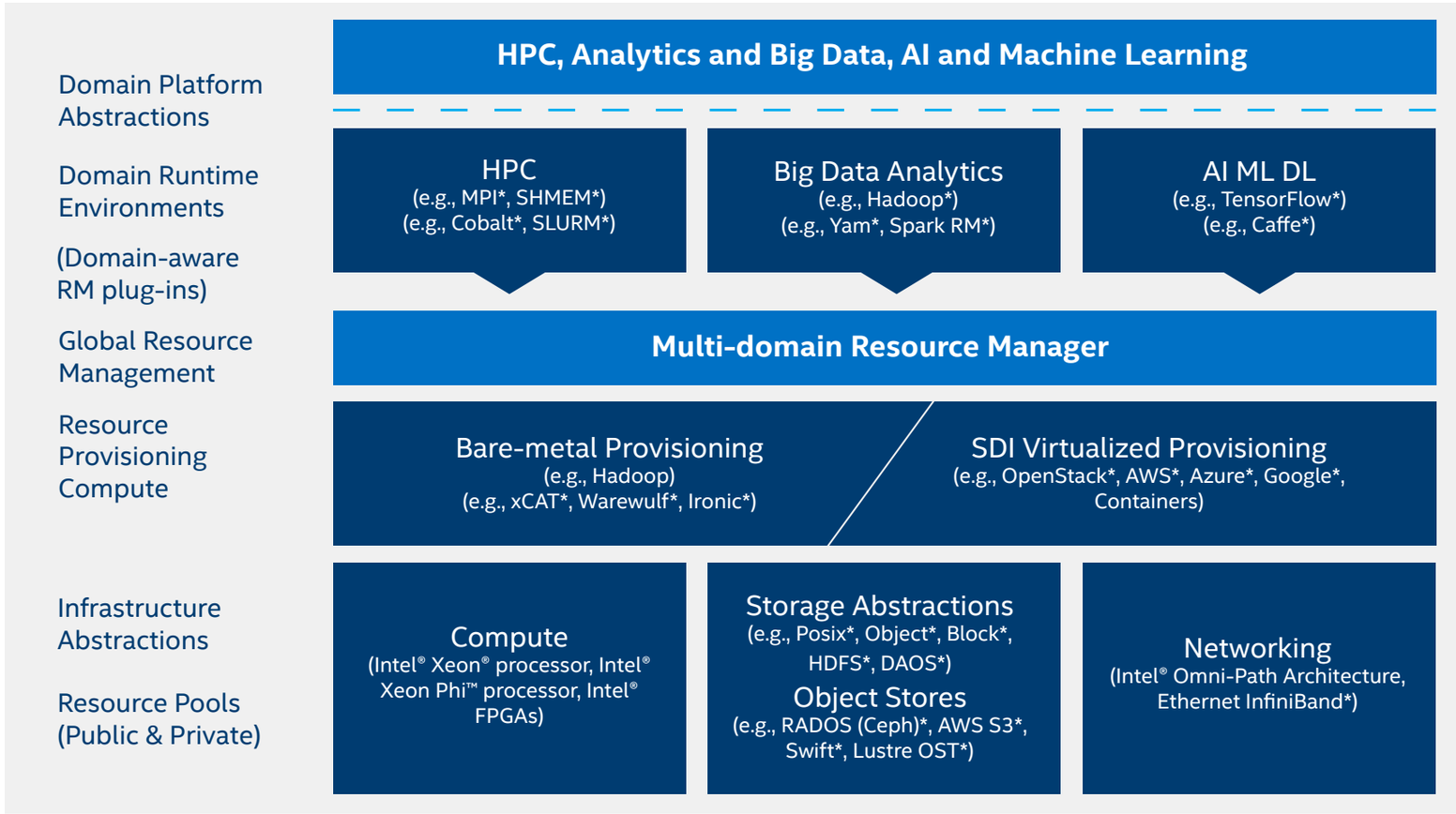


Figure 1: Unifying the three pillars – HPC, HPDA, and AI.

Click on each section to find out more.

STEP 1 Think holistically about your requirements

- Needs
- Expectations
- **Limitations**

STEP 2 Think holistically about your AI solution

- Software Selection
- Software Application and Development
- Frameworks

STEP 3 Validate and scale your HPC technology

- Support
- Scaling Out

STEP 2: SOFTWARE SELECTION

What are the software solutions that best serve your workloads?

Choosing the software necessary for your intended AI workflows will allow for easier planning and optimization of the physical HPC infrastructure needed to support it.

Modern AI applications are based on open source frameworks such as TensorFlow*, PyTorch*, and MXNet*, with codes written in common languages including Python*, Java*, and C++*. Choose an AI framework that supports the application(s) you need to implement. It's important to get a clear understanding of how much of the AI workloads you have identified will be training and how much will be inference.

INSIGHTS:

HPC systems enabling research through AI, visualization, simulation, and modeling workflows benefit from software offered by Intel, the open source community, and independent software vendors (ISVs). To help accelerate the development process Intel's interoperable HPC frameworks support all the languages for modern AI applications.

Click on each section to find out more.

STEP 1 Think holistically about your requirements

- Needs
- Expectations
- Limitations

STEP 2 Think holistically about your AI solution

- **Software Selection**
- Software Application and Development
- Frameworks

STEP 3 Validate and scale your HPC technology

- Support
- Scaling Out

STEP 2: SOFTWARE APPLICATION AND DEVELOPMENT

Are ready-made applications adequate for your unique usage scenarios?

If available applications cannot address your unique usage scenarios your developers can create or modify existing software. While the HPC community offers libraries to assist in this endeavor, developers coding applications for HPC and AI may require specialized skills such as optimization for parallel computing.

Intel's HPC interoperable framework assists developers with tools to modernize applications for advanced workloads and support for development languages including Python, C++, and Fortran*.

INSIGHTS:

Consider AI frameworks that are optimized for your current HPC infrastructure. If you are developing your own algorithms, focus on optimizing your algorithms on your existing Intel® Xeon® processor-based infrastructure before considering additional investments in GPUs or accelerators.

Click on each section to find out more.

STEP 1 Think holistically about your requirements

- Needs
- Expectations
- Limitations

STEP 2 Think holistically about your AI solution

- Software Selection
- **Software Application and Development**
- Frameworks

STEP 3 Validate and scale your HPC technology

- Support
- Scaling Out

STEP 2: FRAMEWORKS

What are the options for using deep learning frameworks to maximize the potential of AI on HPC systems?

Unoptimized versions of current deep learning frameworks such as TensorFlow* and PyTorch do not take full advantage of CPU cores during the execution of deep learning approaches like convolutional neural networks (CNNs). This is because the user-controllable parameters do not provide sufficient micro-architectural information on the underlying non-uniform memory access (NUMA) configuration to achieve optimal performance on multi-socket Intel® Xeon® processor-based platforms.

INSIGHTS:

Intel's optimizations for popular deep learning frameworks have significantly increased processor-level performance. In addition, system-level optimizations can further increase the performance of CNN workloads on Intel® Xeon® Scalable processors used in deep learning and HPC applications.

These optimizations and Best Known Methods (BKMs) can be found in the white paper, "Boosting Deep Learning Training and Inference Performance on Intel® Xeon® and Intel® Xeon Phi™ processors".

Click on each section to find out more.

STEP 1 Think holistically about your requirements

- Needs
- Expectations
- Limitations

STEP 2 Think holistically about your AI solution

- Software Selection
- Software Application and Development
- **Frameworks**

STEP 3 Validate and scale your HPC technology

- Support
- Scaling Out

STEP 3: SUPPORT

Do you have the dedicated human resources necessary to grow your HPC environment?

Just as the infrastructure needs for HPC and AI deployments vary, so do the skills and processes needed to deploy and run them. Some supercomputers are supported by many internal staff members with the expertise necessary to manage and grow their HPC environment while others may need external support.

Organizations lacking dedicated human resources and expertise should consider working with Intel or original equipment manufacturers (OEMs) to accelerate system deployment.

INSIGHTS:

Companies keen to build internal AI capabilities can hone their AI skills with courses on algorithms, machine learning, and deep learning from the Intel® AI Academy.

Click on each section to find out more.

STEP 1 Think holistically about your requirements

- Needs
- Expectations
- Limitations

STEP 2 Think holistically about your AI solution

- Software Selection
- Software Application and Development
- Frameworks

STEP 3 Validate and scale your HPC technology

- **Support**
- Scaling Out

STEP 3: SCALING OUT

What do you need to scale out the system?

Once you've built your business case and refined your algorithms with fresh, accurate data from the trial, you're ready to wade deeper and scale the AI application.

Intel has worked closely with key customers to scale their TensorFlow-based AI workloads running on an existing Intel® Xeon® Scalable processor-based HPC infrastructure across thousands of nodes at high efficiencies and performance. To help achieve similar results, ensure you are using the most recent and optimized AI frameworks, optimized Intel libraries, and the latest Intel Xeon Scalable processors in your HPC infrastructure.

INSIGHTS:

To find out more about how to optimize HPC architectures for AI convergence read this solution brief from Intel.

For more information about the convergence of HPC and AI please visit: intel.com/HPC

Click on each section to find out more.

1018/CAT/JO/PDF

338070-001EN