SUSE

White Paper
Mission Critical Computing

Powering Artificial Intelligence with High-Performance Computing

September 2020

SUSE.com

Your business might be starting to recognize that an HPC infrastructure is vital to supporting the AI and analytics applications you need to keep up with the huge volumes of data and processing to compete in today's world. Updating your current IT infrastructure to enable HPC to run AI workloads is vital to staying competitive in the current fast paced and changing environment. This paper will discuss the seven steps necessary to enable an HPC environment in your data center.

For this discussion, we're broadly defining AI as any system that is trying to solve the kind of problem—such as recognizing patterns—that human thinking does. AI encompasses machine learning (ML) and Deep Learning (DL) techniques. ML is a process in which curated examples are used to train a computer to recognize patterns. Deep learning is a branch of ML that uses digital neural networks to create systems that learn on their own.

AI applications are leveraged to accelerate competitive advantage in many industries:

- Management by exception
- Marketing
- Banking
- Finance
- Agriculture
- Healthcare
- Gaming
- Space Exploration
- Autonomous Vehicles
- Personal assistants (Chat bots)
- Social Media

Optimizing Your Environment to Enable HPC for AI Workloads

AI systems require scalable computing power. As the volumes of data grow, so do training times and computational requirements.

HPC environments today no longer imply bespoke supercomputers built for government and academic research institutions. In fact, enterprises can build high-performance infrastructures while dealing with far fewer vendors and far less customization than most realize.

Today's HPC systems are based on Linux clusters running on industry-standard x86 hardware—the same kind of Linux clusters enterprises are likely already using for their big data and cloud architectures.

The HPC community is making it easier than ever to implement a software-defined HPC infrastructure on the hardware you already have. The OpenHPC group and its members work to simplify the task of adopting HPC techniques and technologies and applying them to enterprise tasks.

How to Bring HPC to Your Organization

Here are seven things to keep in mind when considering how your organization is going to do AI with HPC.

PLAN YOUR WORKLOAD

Your obvious first step is to decide what you are going to run on your HPC cluster. Consider not only what you will start with, but also how you expect to scale the application or applications.

MPI vs Throughput Workloads - What type of workload will you run? HPC workloads can be broadly divided into two categories: parallel distributed (MPI) workloads and throughput workloads.

Message Passing Interface (MPI) applications are parallel distributed applications that consist of multiple concurrently running processes that communicate with each other with intensity.

Throughput workloads often require many tasks to be run to complete a specific job, with each task running on its own without communication between tasks. One example of a throughput workload would be the rendering the frames of digital content. Each frame can be computed on its own as well as in parallel.

There is a lot of information out there on what type of workload fits each environment and application.

Virtualized vs Bare Metal – Historically HPC workloads have been run on bare metal systems, but this is starting to change as virtualized environments are becoming more advanced and offer more value to the enterprise and HPC environments.

There are a number of benefits to running in a virtual environment these days around heterogeneity, security, and flexibility. Many workloads have been successfully virtualized to match the performance of bare-metal environments.

These are just some of the points that need consideration when starting to think about your HPC environment. Some others include resiliency, redundancy, and performance.

UPDATING YOUR HARDWARE

Your next step is to determine your hardware needs. As mentioned above, HPC can be implemented today as a software-defined infrastructure on top of a standard x86 hardware cluster.

Some of the base components you will need when assembling your HPC cluster are the following

- Management nodes
- Compute nodes
- Network Interconnects
- Storage

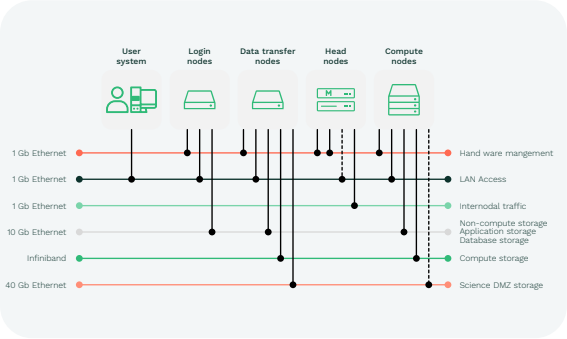
Networking and Storage will be covered a little later, but now we want to focus on the Management and compute nodes and what options you have. Your basic HPC cluster consists of at least one management or login node connected to a network of multiple compute nodes. There may be multiple management nodes used to run cluster-wide services, such as monitoring, workloads and storage usage. This depends on the size of the cluster. The login nodes are a key component to allow users to access the system. Users submit their jobs using these login nodes to the worker nodes via a workload scheduler. As technology has improved over the years, the components to building a cluster have changed. These days multicore Intel x86-architecture systems with varying amounts of cores and memory are commonly used for worker nodes.

Dell provides several great options for both management and compute nodes. Dell Technologies PowerEdge servers deliver high performance computing with the latest processors, accelerators, memory and NVMe storage. You can scale efficiently and predictably with a wide range of configuration and connectivity options. For AI applications, Dell offers several purpose-built platforms that feature high numbers of GPUs and dense compute environments:

- The PowerEdge C4140 Server is a two-processor server with up to four NVIDIA V100 GPUs in just 1U! This server has a patented interleaved GPU design to optimize both space and airflow for maximum compute performance. The PowerEdge C4140 is available with NVIDIA NVLINK™ direct GPU-to-GPU interconnect designed to speed communication between GPUs an order of magnitude faster than PCIe.
- The PowerEdge C6420 Server has up to four independent twoprocessor servers in 2U with up to 16 DIMMs, up to 24x 2.5" hard drives and M.2 boot storage. It's also available with direct contact liquid cooling (DCLC) to support higher-wattage processors for increased performance, energy efficiency and racklevel density.
- The DSS 8440 machine learning server is a two-processor 4U server with up to 16 GPU acceleratros and up to 10 local NVMe and SATA drives for optimized access to data. The DSS 8440 has an open architecture, based on industry standard PCIe fabric, allowing for customization of accelerators, storage options and network cards.

SEGMENTING YOUR NETWORK

If you already have a Linux cluster, your existing top-of rack switches, and Ethernet network may suffice for your HPC infrastructure. More advanced HPC interconnects can be utilized to solve specific throughput issues. The diagram below shows how you might segment your network traffic based on the network capacity of each type of interconnect.



Dell EMC switches

Dell EMC PowerSwitch S5200 ON Series Switches provide state of the art, high density open networking 25GbE top-of-rack and 100GbE spine/leaf switches to meet the growing demands of today's HPC/AI compute and storage traffic.

Dell EMC Networking Z9100 ON Series Switches are 10/25/40/50/100GbE fixed switches for high performance computing environments. With 32 ports of 100GbE, 64 ports of 50GbE, 32 ports of 40GbE, 128 ports of 25GbE or 128 ports 10GbE and two SFP+ ports of 10GbE/1GbE/100MbE, you can conserve rack space and simplifying migration to 100Gbps.

Mellanox InfiniBand

Mellanox® SB7800 series Switch IB 2™ InfiniBand® EDR 100Gb/s Switches deliver high bandwidth with low latency for the highest server efficiency and application productivity — ideal for HPC applications. You can get 36 ports at 100Gb/s per port, and can scale out to hundreds of nodes.

Gen Z Consortium

Dell is a founding member of the Gen Z Consortium, dedicated to creating a next generation interconnect that will bridge existing solutions while enabling unbounded innovation.

Utilize Innovative Storage Solutions

Unprecedented growth in the amount of data collected for analytics, artificial intelligence and other high-performance computing makes fast, scalable and resilient storage an imperative.

Direct-attached storage (DAS)

The Dell EMC PowerEdge R740xd Server has highly expandable memory (up to 3TB) and impressive I/O capability to match. Extraordinary storage capacity options make it well suited for data intensive applications that require greater storage while not sacrificing I/O performance.

The DSS 7000 series lowers your cost per gigabyte for storage while helping you meet the needs of an exascale future. It packs up to 90 hot serviceable 3.5 inch drives in 4U. Available with either one or two server nodes, the DSS 7000 can deliver up to 1.26PB of storage to tackle demanding storage environments.

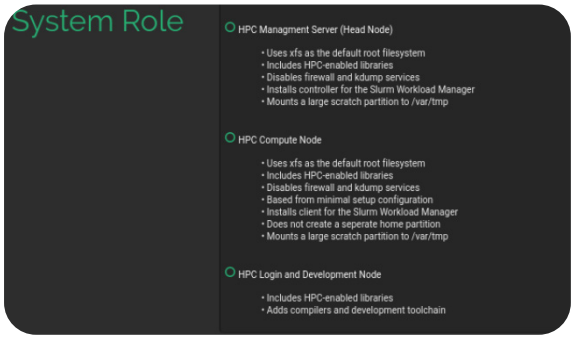
These combined with SUSE Enterprise Storage product will provide the most flexible and robust storage solution for your HPC environment running AI workloads. SUSE Enterprise Storage is a software-defined storage solution powered by Ceph designed to help enterprises manage ever-growing data sets.

SELECT THE RIGHT MIDDLEWARE

Much of the magic in HPC happens at the middleware layer, where you enable parallel computing through tools such as workload schedulers and Message Passing Interfaces (MPIs). Luckily, many of the tools that once came from various vendors are now available as open source.

All the tools you need are packaged together with SUSE® Linux Enterprise Server for High Performance Computing - a highly scalable, high performance open-source operating system designed to utilize the power of parallel computing.

SUSE Linux Enterprise for High-Performance Computing comes with several preconfigured system roles for HPC. These roles provide a set of preselected packages as well as an installation workflow that will configure the systems to make the best use of resources for the role selected.



HPC Management Server (Head Node)

This role includes the following features:

- Uses XFS as the default root file system
- Includes HPC-enabled libraries
- Disables firewall and Kdump services
- Installs the controller for the Slurm Workload Manager
- Mounts a large scratch partition to /var/tmp

HPC Compute Node

This role includes the following features:

- Uses XFS as the default root file system
- Includes HPC-enabled libraries
- Disables firewall and Kdump services
- Based on the minimal setup configuration
- Installs the client for the Slurm Workload Manager
- Does not create a separate home partition
- Mounts a large scratch partition to /var/tmp

HPC Development Node

This role includes the following features:

- Includes HPC-enabled libraries
- Adds compilers and a development toolchain

MANAGE YOUR CLUSTER

SUSE Linux Enterprise for HPC includes the Simple Linux Utility for Resource Management (SLURM) Workload Manager. SLURM is free to use, unifies some tasks previously distributed to discreet HPC software stacks, and is actively improved by the open source developer community. SLURM components include:

- **Cluster Manager:** Organizing management and compute nodes into clusters that distribute computational work.
- **Job Scheduler:** Computational work is submitted as jobs that utilize system resources such as CPU cores, memory, and time.
- **Cluster Workload Manager:** A service that manages access to resources, starts, executes, and monitors work, and manages a pending queue of work.

SLURM makes use of several software packages to provide the described facilities on workload manager server(s)

- **slurm:** Provides the "slurmctld" service and is the SLURM central management daemon. It monitors all other SLURM daemons and resources, accepts work (jobs), and allocates resources to those jobs.
- **slurm-slurmdbd:** Provides the "slurmdbd" service and an enterprise-wide interface to a SLURM database. This service uses a database to record job, user, and group accounting information. The daemon can support multiple clusters using a single database.
- **mariadb:** A MySQL compatible database that can be used for SLURM, locally or remotely.
- **munge:** An authentication service for creating and validating credentials containing the UID and GID of calling processes. Returned credentials can be passed to another process which can validate them using the unmunge program. This allows an unrelated and potentially remote process to ascertain the identity of the calling process. Munge is used to encode all inter-daemon authentications among SLURM daemons without the need for root privileges.

PLAN FOR EXPANSION

The faster and better your AI systems run, the more your enterprise will come to rely on them. At the same time, you can only expect data volumes to grow. It's likely you'll start with a small Proof of Concept (POC) that will need the capability to expand to accommodate ever more production workloads. Your architecture must be built from the ground up with this inevitability in mind.

Some Additional Considerations

The Cloud

If your organization is committed to the cloud—perhaps with a cloud-first strategy—you don't have to pass up the benefits of HPC. HPC-as-a-service offerings give you the benefits of HPC with the convenience of the cloud.

BUILDING A BUSINESS CASE

You will likely need to build a business case for your move to HPC. This should at minimum include two things: the value of your AI workloads to the business and the significant cost savings that you will gain by adding low-cost x86 servers or reusing existing hardware. This limiting of capital expenses will go a long way in helping to justify your infrastructure.

REDUCING RISK

One of the best things you can do to reduce the risks involved in transitioning your infrastructure to HPC is to work with trusted vendors. Dell is the worldwide industry leader in providing x86 servers, storage, and networking. SUSE has been a leader in enterprise Linux since its founding in 1992. SUSE Linux Enterprise Server currently runs on half of the world's top 50 supercomputers and SUSE is the commercial Linux leader in supercomputing's prestigious Top 500.

CHANGING THE WORLD THROUGH HIGH-PERFORMANCE COMPUTING AND ARTIFICIAL INTELLIGENCE

What could an HPC infrastructure mean for your AI goals and your business? It could help you join the ranks of companies changing the world. Behind the growth of innovative AI systems are HPC infrastructures and open source software. They have moved beyond the realm of government and academia to transform big data analytics and generate disruptive solutions that change our very way of life. The only question is whether your enterprise will be along for the ride.

Thank You

SUSE
Maxfeldstrasse
90409 Nuremberg
www.suse.com

For more information, contact SUSE at:
+1 800 796 3700 (U.S./Canada)
+49 (0)911-740 53-0 (Worldwide)