Design Principles for HPC

Design principles for scalable, multi-rack HPC systems

Dell EMC HPC Engineering

January 2018

A Dell EMC Technical White Paper

Revisions

Date	Description
January 2018	Initial release

The information in this publication is provided "as is." Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Dell believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Table of contents

Executive summary

HPC system configuration can be a complex task, especially at scale, requiring a balance between user requirements, performance targets, data center power, cooling and space constraints and pricing. Furthermore, many choices among competing technologies complicates configuration decisions. This document describes a modular approach to HPC system design at scale where sub-systems are broken down into modules that integrate well together. The design and options for each module are backed by measured results, including application performance. Data center considerations are taking into account during design phase. Enterprise class services, including deployment, management and support, are available for the whole HPC system.

1 Introduction

A high performance computing (HPC) system comprises many different components, both hardware and software, that are selected, configured and tuned to optimize performance, energy efficiency and interoperability. The number of different components, and choice in the selection of each component, can make HPC design complex and intractable, especially as systems scale to support growing performance requirements. This white paper describes the principles for HPC design used at Dell EMC. Rigor during architecture design and engineering lead to a system that is well-configured and well-tuned for maximum performance, while being easy-to-use, reliable, ideal for the specific environment in which it will be hosted, and backed by enterprise support. The HPC design process encompasses not just Dell EMC teams in engineering, services, sales and support, but also other partners who develop hardware, software and applications.

This document presents design choices and recommendations for a Dell EMC HPC system. It discusses different aspects of an HPC system and how these come together in a simple, complete system. Designing a system for high performance computing (HPC) requires more than just sizing the system for its use case. Sizing for the use case is a critical step to ensure the system will be able to deliver to its stated goals – performance, application support, user access and so on. However, in addition to the use case, it is equally important to consider the environment in which the system will operate. A perfectly designed system that is limited by data center constraints will be disappointing to its users. Environmental concerns include the data center location, power delivery options, cooling methods, weight restrictions and so on.

This section describes the end-to-end view of HPC systems. That is followed by the different compute, storage, networking, software and datacenter considerations for HPC systems. Expanding a system over time for growth is an important aspect, and scalability is an important aspect of the initial system design, which is also discussed. Services and support options for the full system are described towards the end.

1.1 A complete view of HPC systems

HPC systems are comprised of individual components connected together. Given that all components are standard-based technology, it’s tempting to think of an HPC system as a loose collection of parts that simply work together. While such a system might work, it is unlikely to work well or to deliver on the full promise of the technology it includes.

At Dell EMC the architectural design of an HPC system includes many technology evaluations and decision points, optimizing for interoperability as well as performance, value (performance per dollar), and energy efficiency (performance per watt). HPC system designs are validated in house in the HPC Innovation Lab with a range of ISVs, from operating systems, to cluster managers, to workload application performance. A global HPC Solution Architect team is available to map individual customer requirements and considerations to known good best practices for system design, all while taking into account the software environment (shown in Figure 1) that empowers the System Administrator and end user to accelerate their workload.

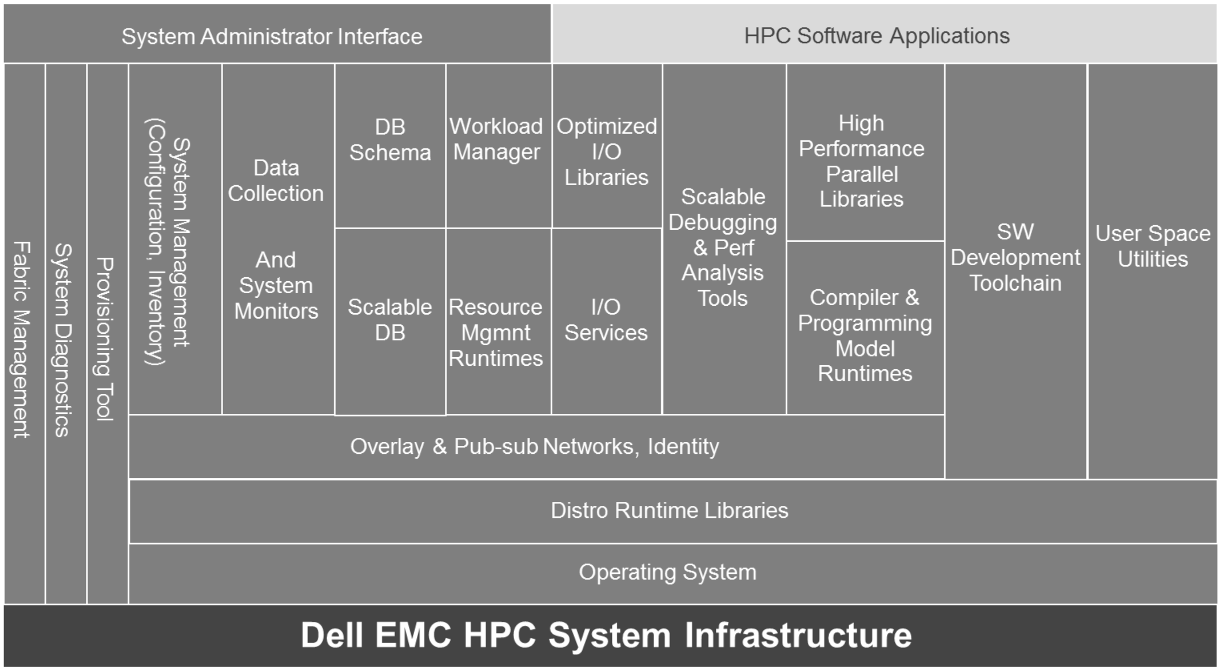

System Administrator Interface
HPC Software Applications
Fabric Management
System Diagnostics
Provisioning Tool
System Management (Configuration, Inventory)
Data Collection
And System Monitors
DB Schema
Scalable DB
Workload Manager
Resource Mgmnt Runtimes
Optimized I/O Libraries
I/O Services
Scalable Debugging & Perf Analysis Tools
High Performance Parallel Libraries
Compiler & Programming Model Runtimes
SW Development Toolchain
User Space Utilities
Overlay & Pub-sub Networks, Identity
Distro Runtime Libraries
Operating System
Dell EMC HPC System Infrastructure

Figure 1 – End-to-end HPC system view

2 Hardware Components

The use case of the HPC system is the starting point. The requirements of the applications that the system is expected to support, number of users, data storage and processing requirements, size, power, cooling and weight restrictions of the datacenter will determine the size of the HPC system. This document does not discuss how to size an HPC system for a particular set of applications, but instead focuses on the common aspects across all HPC systems. Components are selected from a leading HPC portfolio that is built on flexible and scalable design principles, and leverages our broad end-to-end enterprise portfolio. A summary of the hardware components is in Figure 2 and described in the sections below.

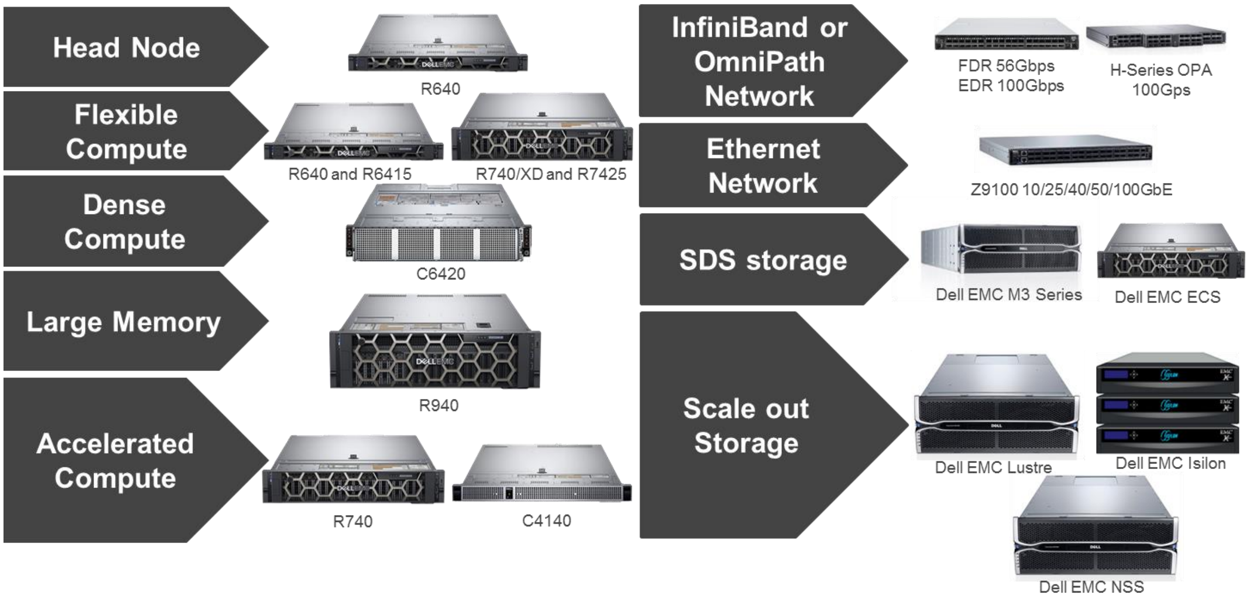


Figure 2 – HPC hardware options

2.1 Infrastructure Unit

The servers that provide management functionality for the HPC system fall into the category of infrastructure servers. These roles include the cluster master node(s), login node(s) and compiler node(s).

The master node manages the software image(s) for the cluster, handles cluster deployment, management and monitoring. It also runs the resource scheduler and job manager for the cluster like Slurm. The master node can be run in high availability mode with an active-passive server pair (plus shared storage for a unified cluster view) to protect this important cluster function and minimize downtime. Cluster management software like Bright Cluster Manager would run on the master node(s).

The login nodes service user login requests and compiler nodes are set up for with application build environments. These roles can be combined on a single server or distributed across multiple servers depending on the size of the cluster. The infrastructure servers can also be set up to run in a high-availability mode providing an extra layer of resiliency for these key roles.

Infrastructure servers are key to keeping the HPC system up and accessible to users. The Dell EMC PowerEdge R640 or R740 Server is recommended for this role. This is a 1U or 2U server with rich configuration choices in memory and local storage, which also provides

Integrated Dell Remote Access Controller (iDRAC) enterprise manageability and systems security. The PowerEdge R740 can support more local storage and peripherals in its 2U form factor and is a good choice for HPC systems where additional storage is needed on the master node, or when the master node needs to host GPUs as well.

2.2 Compute Unit

The compute unit of the HPC system is the main computational work horse. This can include CPUs, GPUs or FPGAs – any unit that provides computational power. The size of the compute unit i.e. the number of compute nodes depends on the use case of the HPC system, expected performance, budget, space and data center requirements. Decisions include the type/model of compute server as well as the configuration of the compute server itself.

Compute and networking must be designed together in HPC since the building block of the compute unit will determined by the switch radix of the interconnect between the compute nodes. For example, when using a fabric with 48-port switches, 24 compute nodes per switch (with 24 switch ports available for uplinks) would make a good building block. A 36-port switch would be associated with 18 compute nodes per switch (with 18 switch ports available for uplinks). This is expanded upon in Section 2.4 on the networking unit.

There are many choices for the compute server itself, and density, floor weight, rack power and cooling capability can help guide the selection. This is discussed further in Section 4 on data center considerations.

Studies comparing different processor and GPU options, memory and network configurations across multiple workloads are available at www.hpcatdell.com to help guide individual server configuration.

Looking at the servers available today, the Dell EMC PowerEdge C6420 Server is designed for HPC and provides a dense compute platform with four 2S server sleds in a 2U chassis. The Dell EMC PowerEdge C4140 Server is a GPU-optimized two socket platform and can support up to four NVIDIA GPUs, including the PCI-e and SMX2 versions. It also has support for PCIe FPGA cards. The PowerEdge C4140 provides a high density accelerator solution in 1U of rack space. The PowerEdge C6320p is an accelerator platform for the Intel® Xeon Phi™ line of CPUs and ideal for highly parallel, vectorized workloads, balancing performance with power efficiency.

The Dell EMC portfolio also includes several other compute platforms for environments where server density is not the key factor, like the Dell EMC PowerEdge R640 and PowerEdge R740 servers. These servers support two Intel CPU sockets with multiple PCI-e, disk and network options in 1U and 2U of rack space respectively. AMD recently announced its re-entry into the server CPU space with its EPYC™ line of processors. The Dell EMC PowerEdge R6415, R7415 and R7425 servers are AMD-based platforms with all the advantaged of the EPYC architecture – extra memory capacity, support for GPUs and multiple NVMe devices.

For use cases that need a large memory system or many cores, the four socket Dell EMC PowerEdge R940 Server provides up to 1.5TB of memory at the time of writing and 112 physical CPU cores.

These servers are part of Dell EMC's 14th generation server line-up and include next generation systems management with Integrated Dell Remote Access Controller 9 (iDRAC9), expanded interconnect and disk options including Non-Volatile Memory Express (NVMe) support, and support for the Intel Xeon® Scalable Processor Family or the AMD EPYC processors.

Due to its density and feature set, the PowerEdge C6420 Server is the most popular HPC compute platform and described in more detail in Appendix A.

2.3 Storage Unit

The storage component stores user home directories, application data, results, provides fast scratch space and can include archives and backup. Depending on the use case and the size of the HPC system, the storage unit can range from just disks on the master node that are exported to the compute unit, to separate file systems for user data, application data, a parallel file system, fast scratch backup recovery and tape archives.

To satisfy these use cases, the technologies used in the storage solution could include SAS/SATA disks, SSDs, NVMe and any combination of these. The choice of storage solution would depend on performance, capacity and cost. The Dell EMC options described here are designed to satisfy the various storage use-cases and incorporate the best technologies for the use-case.

The Dell EMC Ready Bundle for HPC NFS Storage is a tuned NFS storage solution (NSS) that provides reliable and efficient performance. The NSS solution is ideal for home directories and application data for small to mid-size clusters. It cannot be used as a parallel file system, but can support long term data storage. The performance of the NSS has been extensively characterized to right-size the best fit for this storage solution. This solution can be delivered as a fully tuned appliance-like configuration, or as a build your own.

The Dell EMC Ready Bundle for HPC Lustre Storage provides a fully supported, high performance, parallel file system based on commodity hardware that can scale to petabytes of storage. The key here is best performance at a reasonable price point. The performance characteristics of the Lustre building blocks are well characterized in the HPC Innovation Lab to assist in selecting the most appropriate Lustre storage configuration in terms of performance and capacity.

Dell EMC Isilon is a proven scale-out network attached storage (NAS) solution that can handle the unstructured data prevalent in many different workflows. The Isilon storage architecture automatically aligns application needs with performance, capacity, and economics. As performance and capacity demands increase, both can be scaled simply and non-disruptively, allowing applications and users to continue working.

All these storage solutions are individually scalable for capacity and performance independent of the compute unit. Typically the storage solution will be hosted in independent racks from the compute solution allowing both to grow and scale with capacity and need. The storage solution and compute units will be interconnected together at the data path and this is explained in the next section.

2.4 Networking Options

Most HPC systems have at least two network fabrics. One is used for administration and management and the second for inter process communication (IPC) and storage traffic. Depending on the use case, separate networks for management and storage traffic may also be configured.

The choices for the network range from Ethernet – 1 GbE, 10GbE, 25 GbE, 40 GbE, 100 GbE to Mellanox InfiniBand and Dell EMC H-Series based on Intel® Omni-Path Architecture (OPA). For Omni-Path, there are two further choices – a PCI-e adapter, or an Intel CPU that has integrated fabric plus a carrier card that exposes the network ports. From a functionality and network design point of view, both these Omni-Path options would be identical.

Depending on the server adapter and switch choices, this network could be built with optical or copper wires. A typical HPC system will use copper cabling within a rack and fiber cables between racks, balancing ease of cable management with cost.

Administration fabric – 1GbE is available on all platforms and suffices for deploying the software on a cluster, management and monitoring of system health. This single network can be used for Intelligent Platform Management Interface (IPMI) communication via the onboard iDRAC systems management controller.

Two 48-port GbE switches will allow management of the entire rack and can be uplinked to the core switches outside the rack. For example, the Dell Networking S3048-ON is a 1GbE top-of-rack open networking switch. The S3048-ON design provides (48) 1000BASE-T ports that support 10MbE/100MbE/1GbE and four 10GbE SFP+ uplinks. Each 10GbE interface can be used as uplinks to the network spine/core, as stack ports to connect up to six units in a stacked configuration, or a combination of both, depending on network architecture.

For larger HPC systems, it can be useful for the master node to have a 10GbE connection at the Ethernet core instead of a 1GbE connection. This helps bring nodes back up quickly in the event of downtime.

IPC and storage fabric – Typically InfiniBand or Omni-Path is used as the high performance fabric providing low latency and high bandwidth connectivity between compute platforms for IPC and access to the storage. Depending on the system use case, 10GbE or 25GbE can be a good alternatives too.

There is no single design for this IPC fabric. A common network architecture is a fat-tree. Depending on the use case, some systems might need a complete non-blocking fabric that allows for simulation communication between nodes in the cluster with zero network contention. This implies that the number of downlinks from a switch should be equal to the number of uplinks. This level of performance requires a large number of switches and cables and the number of components increases significantly with the number of nodes in the system. Another option is a blocking-fabric with some reasonable blocking factor – say 2:1 over subscription where the number of downlinks to the servers is 2x the number of uplinks. Other blocking factors are similarly possible. Other network architectures for the IPC fabric include 2D mesh, 3D mesh, torus, dragonfly, etc. Since the Dell EMC HPC systems are built from direct attached components, any of these network topologies is a viable option for the final system and there is great flexibility in network design choices.

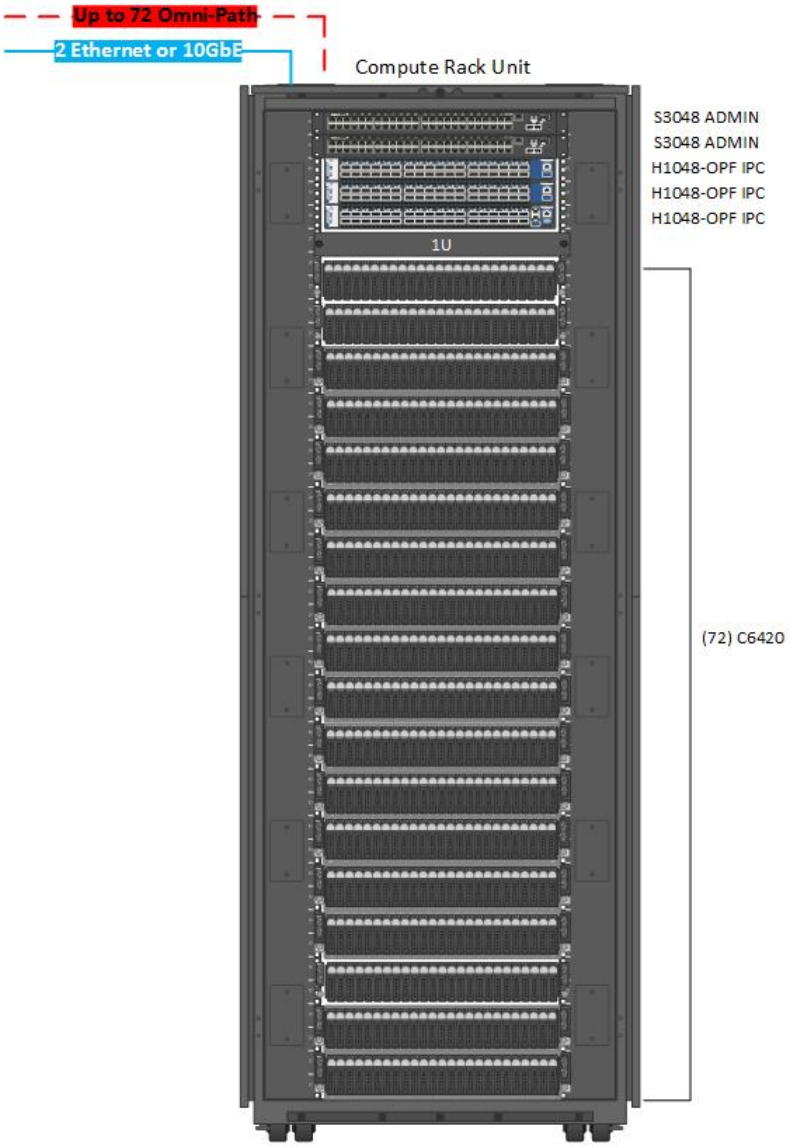


Figure 3 - Example rack configuration with 72 compute nodes, GbE and OPA switches. The OPA switches shown here could alternately be distributed between the servers to reduce intra-rack cable lengths and reduce cable congestion at the top.

A popular network design for HPC uses top-of-rack (TOR) or leaf switches within the rack, and core switches that connect to the leaf switches. Section 2.2 on compute units mentioned a building block unit that includes servers and a switch. Using 48-port OPA switches, 24 servers can be connected to one switch for a 1:1 non-blocking network. With 36-port IB switches, 18 servers can be connected to a single switch to start building a 1:1 non-blocking network. Using the PowerEdge C6420 systems as an example configuration, each rack can host 72 compute nodes as shown in Figure 3. Three Omni-Path or four InfiniBand switches can provide connectivity to all the 72 compute nodes within the racks, and two 48-port Ethernet switches are sufficient for the administration fabric network. The number of uplinks to the core switch can vary depending on the network topology and blocking factor desired for the network. For simplicity the core switches for the system are assumed to be in a separate rack, say along with the infrastructure nodes. Multiple such racks can be configured, and the uplinks from all these racks can be combined at the core switches as shown in Figure 4 and Figure 5. Both these schematics use the 48-port Omni-Path switch as an example, but the same principles can be applied to InfiniBand with the 36-port switch as well. Note that for larger clusters or more complex networks a custom network topology can be designed for the specific use case.

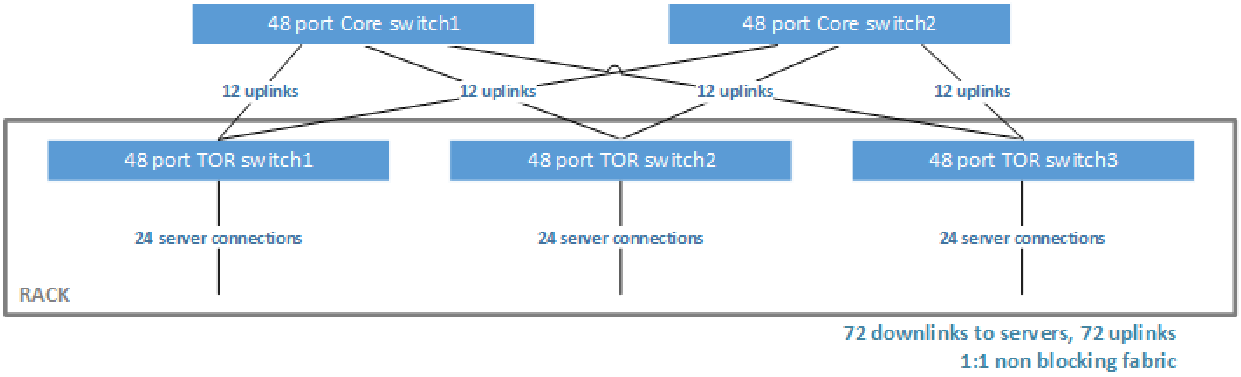


Figure 4 – 72-node compute rack with TOR switches and core switches

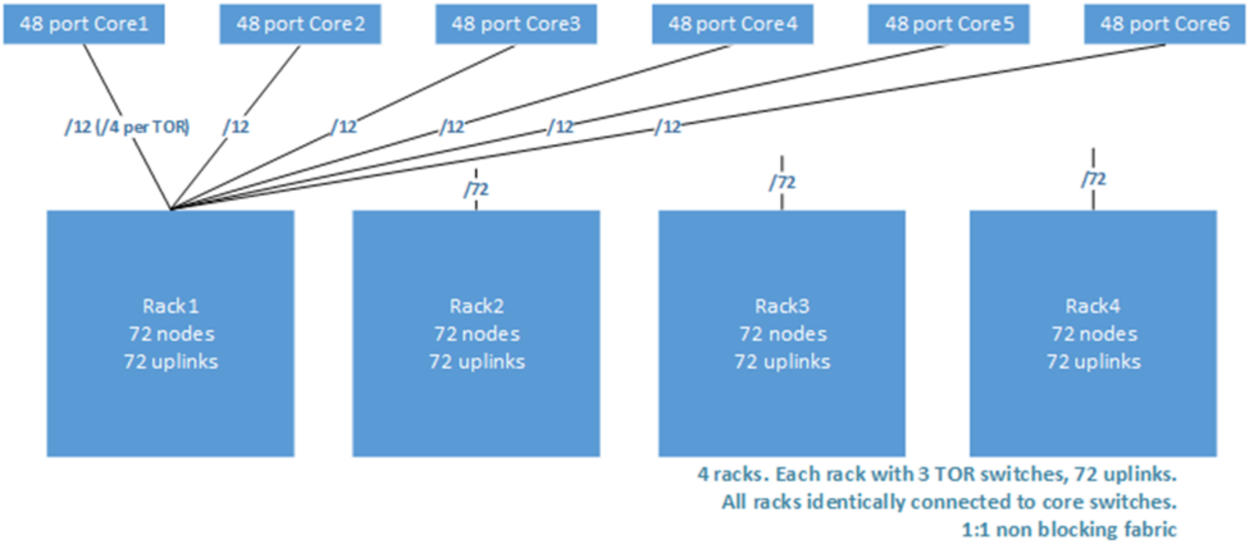


Figure 5 – 288-node compute unit with four compute racks and core switches

The storage unit will also have uplinks to the core switch from any storage unit edge switch(es), and the number of ports needed on the core will depend on the size of the storage sub system. Using storage edge switches to connect to the core versus connecting the storage directly to the core helps avoid storage traffic hotspots on the core that could cause traffic congestion in large environments.

The number of compute, infrastructure and storage switch ports and required blocking factor of the fabric will determine the total number of switches needed for a specific configuration.

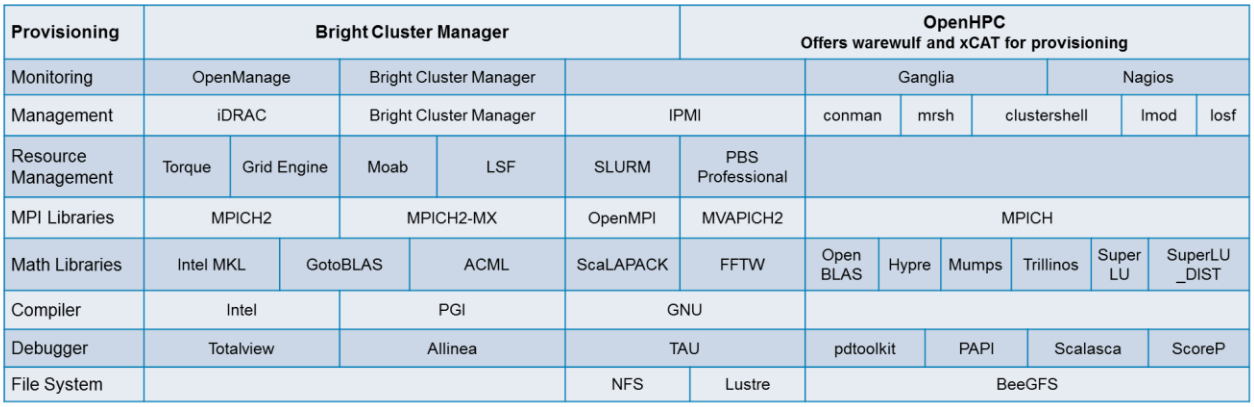
The PowerEdge line of server platforms allows not just multiple interconnect choices but also support for multiple generations of interconnects. For example, EDR InfiniBand is an option today, but when HDR InfiniBand is available in the market, the same server platforms will support HDR as well. A future expansion to the HPC system can therefore simply and easily incorporate the latest technology at the time while retaining the ease of use and manageability of the entire system.

3 Software Components

This section describes some of the software components that make up the HPC system. As with the hardware, the software is assembled in a modular manner. There are multiple choices for individual components (like resource manager, MPI, compilers etc.) and support for drivers based on the hardware selections. The software components are validated on in-house HPC systems along with the hardware as mentioned in Section 1.1.

3.1 Cluster Software Stack

HPC cluster software is a suite of tools that are packaged together to make it easy to deploy, administer and maintain the HPC system. Right from easy deployment out of the box, to user support and managing the life cycle of the HPC system, this software is the administrator’s main console to the HPC system. Bright Cluster Manager and OpenHPC are cluster software packages that have been validated on Dell EMC HPC systems and include PowerEdge-specific drivers, systems management packages, recommended operating systems tuning options, resource managers, MPI libraries and compilers. This is shown in Figure 6.



<table>
<tr><td><b>Provisioning</b></td><td colspan="7"><b>Bright Cluster Manager</b></td><td colspan="15"><b>OpenHPC</b>
<b>Offers warewulf and xCAT for provisioning</b></td></tr>
<tr><td>Monitoring</td><td colspan="3">OpenManage</td><td colspan="3">Bright Cluster Manager</td><td colspan="3"></td><td colspan="8">Ganglia</td><td colspan="5">Nagios</td></tr>
<tr><td>Management</td><td colspan="3">iDRAC</td><td colspan="3">Bright Cluster Manager</td><td colspan="3">IPMI</td><td colspan="2">conman</td><td colspan="3">mrsh</td><td colspan="5">clustershell</td><td colspan="2">lmod</td><td>losf</td></tr>
<tr><td>Resource Management</td><td>Torque</td><td colspan="2">Grid Engine</td><td colspan="2">Moab</td><td>LSF</td><td>SLURM</td><td colspan="2">PBS Professional</td><td colspan="13"></td></tr>
<tr><td>MPI Libraries</td><td colspan="3">MPICH2</td><td colspan="3">MPICH2-MX</td><td>OpenMPI</td><td colspan="2">MVAPICH2</td><td colspan="13">MPICH</td></tr>
<tr><td>Math Libraries</td><td colspan="2">Intel MKL</td><td colspan="2">GotoBLAS</td><td colspan="2">ACML</td><td>ScaLAPACK</td><td colspan="2">FFTW</td><td>Open BLAS</td><td colspan="3">Hypre</td><td colspan="2">Mumps</td><td colspan="3">Trillinos</td><td colspan="2">Super LU</td><td colspan="2">SuperLU _DIST</td></tr>
<tr><td>Compiler</td><td colspan="3">Intel</td><td colspan="3">PGI</td><td colspan="3">GNU</td><td colspan="13"></td></tr>
<tr><td>Debugger</td><td colspan="3">Totalview</td><td colspan="3">Allinea</td><td colspan="3">TAU</td><td colspan="3">pdtoolkit</td><td colspan="4">PAPI</td><td colspan="4">Scalasca</td><td colspan="2">ScoreP</td></tr>
<tr><td>File System</td><td colspan="6"></td><td colspan="2">NFS</td><td>Lustre</td><td colspan="13">BeeGFS</td></tr>
</table>

Figure 6 - Cluster software stack

3.2 Systems Management Software

The cluster software packages allow the administrator to manage the entire HPC system as one unit. In turn, the cluster software relies on systems management software to manage the compute units, storage units and the network. Systems management tools based on iDRAC and IPMI for server and storage management and switch management utilities from the switch vendors are pre-integrated into the cluster software package and any additional tools can be easily added on. This makes it simple to identify hardware errors and impending failures across the multiple racks of the HPC system. The Dell EMC OpenManage suite provides enterprise-class systems management functionality for systems monitoring.

3.3 Application Performance

The prime motivation for any HPC system is to support its users and their applications. To this end, the HPC and AI Innovation Lab has partnerships with many ISVs and runs application benchmarks in house. Application performance across system generations, across multiple types of architectures, BIOS tuning results are available for many categories of applications including digital manufacturing, life sciences and research. Work is also being done on containers and deep learning. Tests include performance as well as power monitoring to provide effective energy efficiency recommendations.

The in-house expertise and partnerships with ISVs lead to detailed best practices and HPC systems can be then configured taking into account each customer’s specific use case and priorities.

4 Data Center Considerations

The final configuration of the HPC system is a series of trade-offs. An existing datacenter provides many of the bounds for the system – the physical limitations of space and "per rack" limitations of power, cooling, and weight. These topics are discussed below.

4.1 Power Configuration

A trade-off with high density servers is in the input power requirements to each rack. As server platform densities increase, and each platform supports richer and more power hungry configurations (CPU TDP, more memory DIMMs, more disks), the input power needed to drive a single rack increases significantly. For example, a 42U rack filled with a high end PowerEdge C6420 HPC configuration and 72 servers (Figure 4) would require ~39-43kW of power to accommodate peak load. A less dense rack with 36 1U R640 servers would require 20-22kW.

Power Distribution Unit (PDU) vendors should be consulted for detailed PDU capacities. Countries differ in terms of specific voltage supplied, but in general, the distribution voltages center around 100, 200, and 400 volts. With single phase distribution, the capacity is simply the voltage times the current limit. With all PDUs, North American versions are derated by 20%. A 50A PDU available to Europe might be the same unit but limited to 40A in North America. Three phase distribution is becoming much more common. A 208V 3-phase input PDU splits out to three equal lines of 120V power rated to the amperage of the PDU. There are very few available 400-415V 3-phase input PDUs, but they will become more commonplace with increasing rack densities. These break out into 3 equal lines of 230-240V power, each with the capacity of voltage multiplied by amperage.

Higher densities will necessitate deeper racks, simply due to needing more PDUs to provide the total rack power capacity. Wide racks may also be required. The higher the capacity PDU is, the deeper it is in general. Very dense servers can also require C19/C20 plugs that have longer back-shells on their cords. Without a wide rack, a straight power cord can possibly interfere with service access to removable items from the rear of the server.

Liquid cooling will also become more common, especially in very dense servers. Liquid cooling is often accomplished by running water through a distribution manifold up and down the rack with connections to each server. It is very common for water and power distribution to be at the rear of the server so water manifolds may be in space competition with PDUs. This will be another driver toward adoption of 400V power.

The subject of 480V distribution comes up occasionally. Three phase 480V breaks out into three feeds of 277V which, while a common distribution in some industries is not common at all to IT equipment. For the foreseeable future, there will be very few servers offering 277V-capable power supplies.

The Dell Enterprise Infrastructure Planning Tool (EIPT) is an easy-use utility to determine high-level server level power and weight estimates. .

4.2 Rack Cooling

Delivering adequate power to a rack for maximum performance without throttling is one factor that determines the density of infrastructure, but the other aspect is dissipating the heat generated to continue operations within required ambient temperature limits.

Focusing first on the server itself, the hardware configuration of the platform determines the ambient temperature needs of the server. Choices of chassis backplane, number of disks, CPU wattage, additional controllers, copper or fiber network cables determine the heat produced and impact the air flow through the platform and thus the maximum ambient temperature required for the platform to operate safely and effectively.

Using the Dell EMC PowerEdge C6420 Server as an example, ambient temperature requirements range from 21 C to 35 C. A typical HPC configuration with a mid-tier processor, optimal memory, a network adapter with copper wires and one disk per server would require only 35 C in the data centers. More disks would raise the minimum requirement to 30 C.

Within the server, there is a choice of an (only) air-cooled or liquid-cooled hybrid option with approximately 30-40% of the server heat requiring an air cooled facility solution of some kind. Before discussing liquid cooling, air cooled options should first be discussed and ruled out. This is to ensure that the more cost-effective and flexible solutions are evaluated first.

Many conventionally cooled data centers cannot effectively deliver the amount of air volume commensurate with the heat loads mentioned in section 4.1. Raised floor cooling comes with uncertainty in the delivery volume to any specific location. This can be enhanced with containment into pod groupings, but there can still be natural discrepancies between two identically configured pods. Row cooling removes most of that uncertainty, but still requires balancing at the pod level; and the row coolers take up rack space.

Rear door heat exchangers add additional depth and add capital cost but are easier to design around specific rack level capacities and can be a more cost-effective solution than liquid cooling. With cold enough facility water, rear doors can deliver very high cooling capacities. There is some loss of flexibility since cooling capacity is devoted to a specific rack.

If these options are not viable, or if the area of interest includes facility cooling efficiency, or if the requirement is for a richer configuration (more drives and high wattage CPUs) that is not supported by air cooling, then the cold-plate cooling solution is another option.

The PowerEdge C6420 liquid cooled option employs direct liquid cooling solutions from CoolIT that include CPU cold plates, rack manifolds and a heat exchanger unit. The CPU cold plates replace the heat sinks in air cooled systems. Coolant tubes come out of each sled and connect to a manifold. The rack manifold circulates and distributes water based coolant to servers within a rack. The heat exchanger can be one per rack or a larger one for multiple racks. Either version connects to the rack manifold(s) to pump coolant to the servers and into the cold plates that are in direct contact with components. The heated water is then run through the heat exchanger that transfers the heat either into the facility water or data center air.

Figure 7 shows a liquid cooled PowerEdge C6420 Server with the cold plate and coolant tubes. Figure 8 shows the rack manifold and an example of a liquid-to-liquid heat exchanger. The liquid ports on C6420 servers are in the rear, so rack manifolds would typically be mounted vertically in the PDU area. For smaller deployments, smaller rack manifolds could also be mounted horizontally within the U-space. Depending on the power capacity that must be managed, a heat exchanger could be 2U or 4U mounted in the rack, or a standalone unit in its own rack.

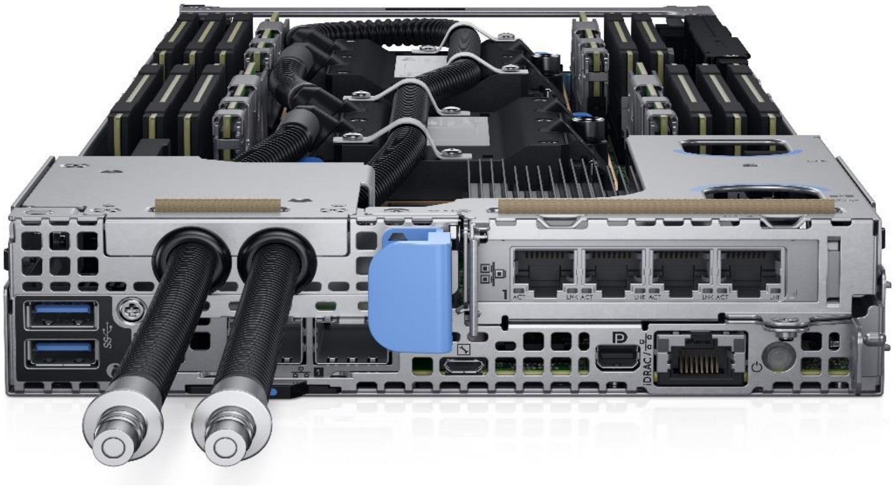


Figure 7 - PowerEdge C6420 liquid cooled server sled

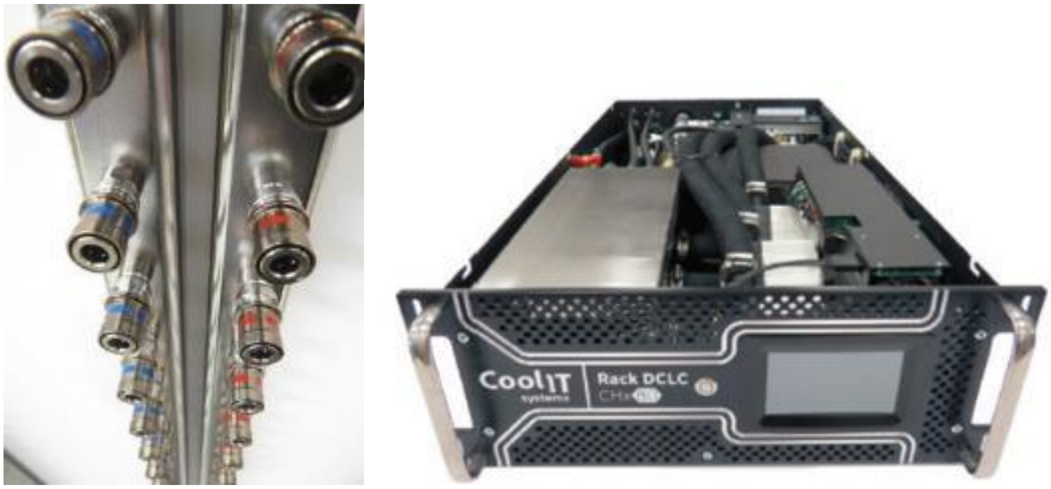


Figure 8 – Rack manifold and Heat Exchanger

4.3 Rack Weight

A compute rack, as described in Figure 3 with 72 servers and five top-of-rack switches, will weigh ~1800 lbs. and consume up to 43kW of power. This configuration maximizes the space in the rack, but could be hard to accommodate in older datacenters due to the large concentration of power needed for one rack and the weight of the rack. In these cases, the trade-off will be to reduce server density to acceptable power and weight limits.

For example, assuming a datacenter weight limit of 1500 lbs. per rack and 25kW of power per rack, a reasonable configuration would be 10 chassis for 40 PowerEdge C6420 servers in the rack with two top of rack Omni-Path switches and one Ethernet switch. This rack would contain 23U of equipment and satisfy both weight and maximum power limits.

4.4 Rack Cabling

As is clear by now, a dense HPC system with multiple PDUs, network switches, power cables, network fabrics and additional liquid cooling manifolds can make simply assembling the rack a messy proposition. A well-installed rack is key, but so are manageability (removing a server for maintenance for example), troubleshooting (replacing a failed cable) and even performance (thick cable bundles blocking air flow can throttle the platform). It is typical to use copper cables within the rack and fiber based cables for inter-rack links striking a balance between manageability and cable cost.

5 Scaling the Design

HPC systems are scale out by definition. As needs grow, additional compute power can be added to a system with more CPUs, GPUs or FPGA-based servers, additional storage capacity or throughput can be added with more storage arrays. The network must support this expansion. Ideally, the architecture of the system should accommodate this need to scale from the design phase.

The design principles presented in this paper make scaling easy. The compute building block comprehends and is built around the network and switch radix as described in Section 2.2. A compute unit is built with multiple compute rack, each with its top-of-rack switches (Figure 3). The storage solution is hosted in a rack(s) of its own and can include NFS-based storage, parallel file systems, archives, etc. The infrastructure servers and core switch network will be in the final rack(s). The core switch network is based on the size of the total system with room to grow as the system scales.

An example schematic is shown in Figure 9 where each rack is built off the units described in Section 2.

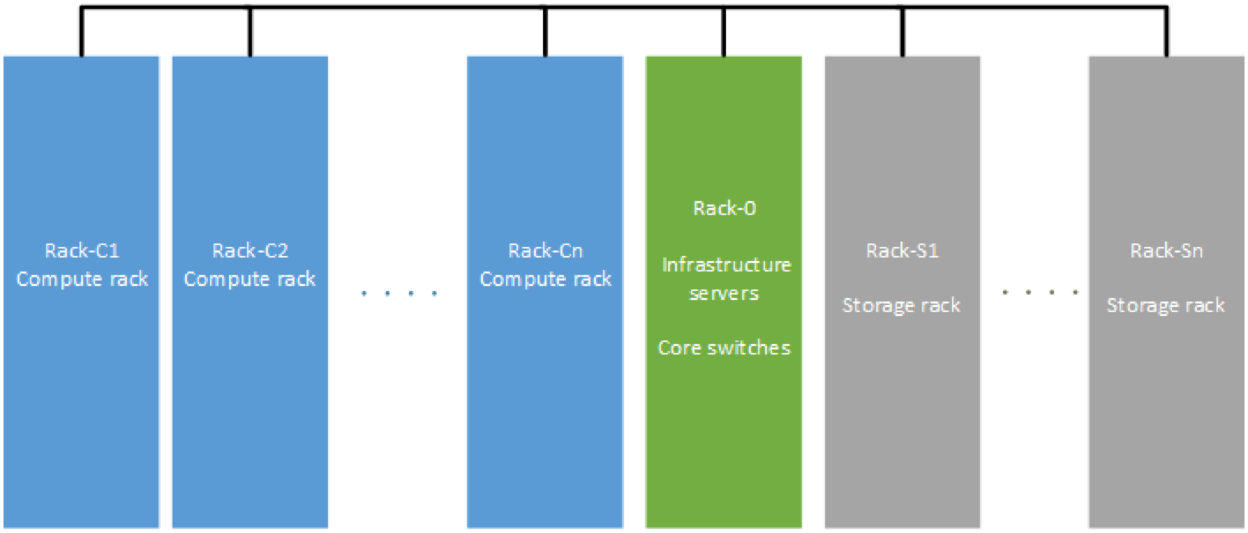
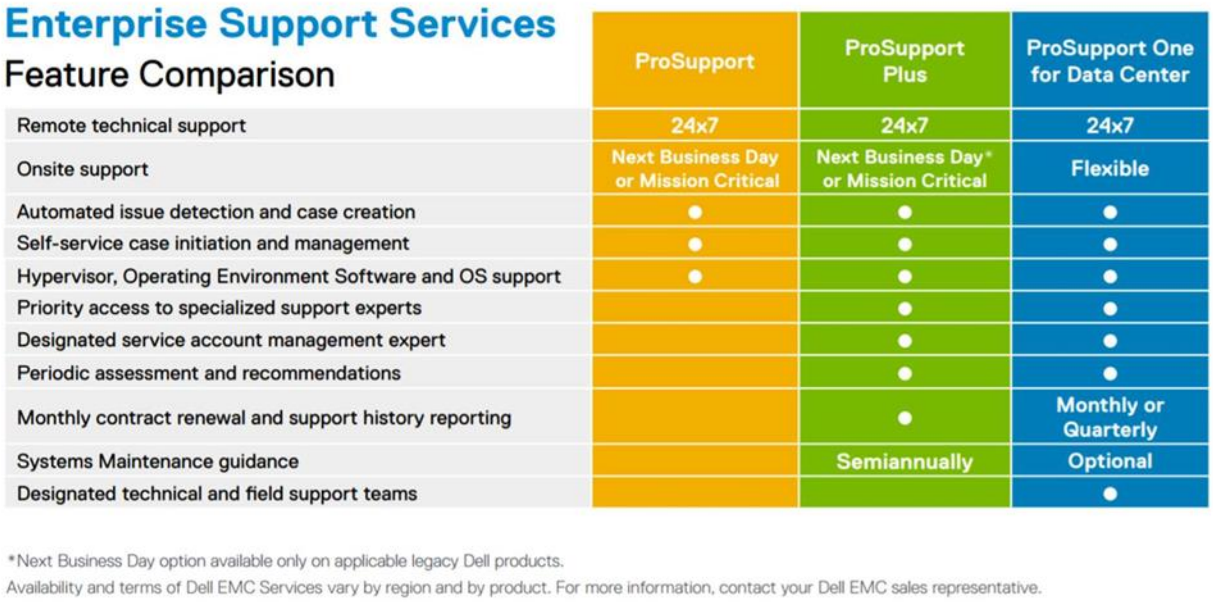


Figure 9 - Scalable design for HPC system

While this design will work for most HPC systems and of different sizes, there will always be a need for custom architectures based on unique use cases or data center needs. A team of HPC solution architects work alongside the customer to convert specific requirements into designs that are built on best practices.

6 Enterprise Support and Services

A good design and a well-configured HPC system is the starting point for the end users, equally important is services and support for the entire life cycle of the system. HPC trained services team can rack and stack, install and complete the software configuration of the system if desired. Different levels of support are offered, as described in Figure 10, and as illustrated, HPC support is handled via standard Dell ProSupport channels leading to an enterprise-class experience.



Enterprise Support Services
Feature Comparison

	ProSupport	ProSupport Plus	ProSupport One for Data Center
Remote technical support	24x7	24x7	24x7
Onsite support	Next Business Day or Mission Critical	Next Business Day* or Mission Critical	Flexible
Automated issue detection and case creation	●	●	●
Self-service case initiation and management	●	●	●
Hypervisor, Operating Environment Software and OS support	●	●	●
Priority access to specialized support experts		●	●
Designated service account management expert		●	●
Periodic assessment and recommendations		●	●
Monthly contract renewal and support history reporting		●	Monthly or Quarterly
Systems Maintenance guidance		Semiannually	Optional
Designated technical and field support teams			●

*Next Business Day option available only on applicable legacy Dell products.
Availability and terms of Dell EMC Services vary by region and by product. For more information, contact your Dell EMC sales representative.

Figure 10 - Enterprise support options

7 Conclusion

This document presents Dell EMC's HPC design principles. HPC systems are configured based on careful design that includes all aspects from hardware, software, application performance, data center considerations, power and cooling. Extensive evaluation in the HPC and AI Innovation Lab leads to best practice recommendations including performance characteristics across multiple HPC domains. This allows added insights into ideal configurations that balance performance, value (performance per dollar), and energy efficiency (performance per watt).

A complete view of the HPC system includes the environment that hosts the systems. Datacenter analysis, including aspects of power, cooling and space requirements, is incorporated into the HPC system design.

Enterprise-class services and support help ensure the HPC system continues to operate optimally, freeing up the system administrators to focus on the users and their research.

A Appendix – Dell EMC PowerEdge C6420 Server

The PowerEdge C6420 Server is the most popular HPC compute platform and is described in more detail here. Four server sleds are hosted in a 2U chassis as shown in Figure 11. The chassis provides shared infrastructure for power and cooling. Each individual sled is a standard two socket server platform with individual networking as seen in Figure 12.

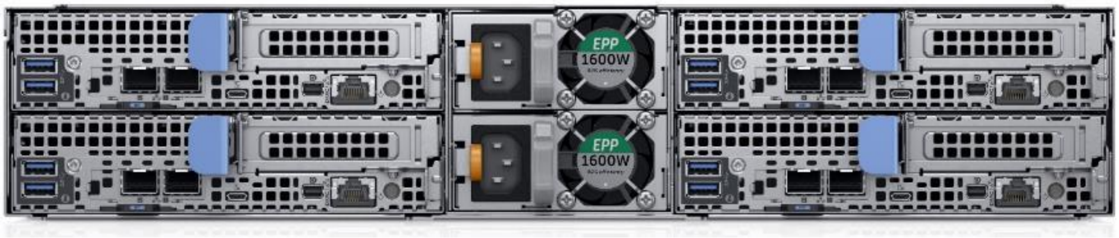


Figure 11 - PowerEdge C6400 chassis with four PowerEdge C6420 server sleds (rear view)

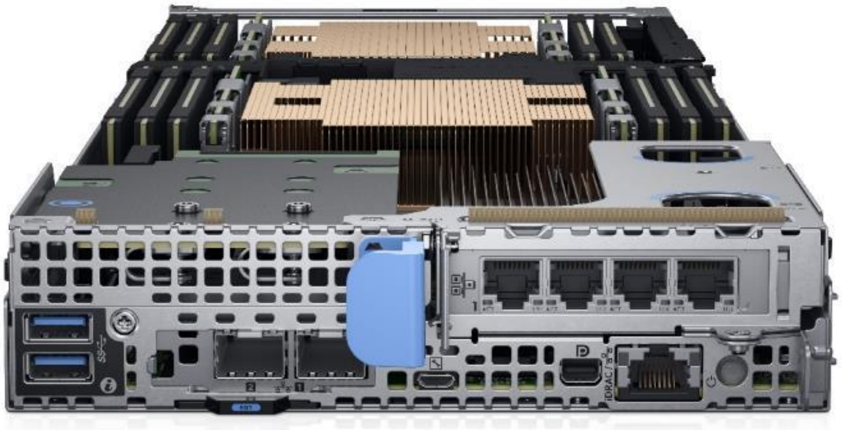


Figure 12 - PowerEdge C6420 server sled. Dual socket, 16 DIMM slots with support for Intel 3D Xpoint memory. Shown here with dual port 10GbE SFP+ mezzanine card and quad port 10GbE PCI-e card

The sleds support M.2 boot devices and, additionally, the chassis can host up to 24 2.5” disks for the four servers as shown in Figure 13. The unit is also available in a diskless configuration for use cases where compute nodes boot off the network and do not need local disks. The diskless configuration has no hard drive backplane leading to better air flow and cooling and fewer thermal restrictions. Using the PowerEdge C6420 Server as a component for the compute portion of the HPC system allows the HPC system to scale in units of four servers and, thus, provides flexibility for a range of compute sizes.

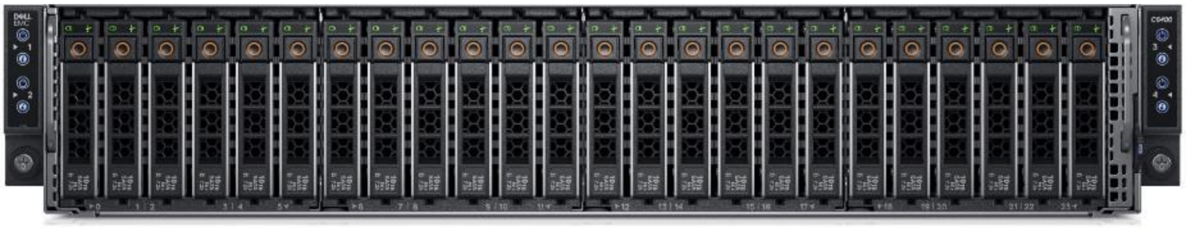


Figure 13 - PowerEdge C6400 chassis with 24 2.5" disks (front view)

Some key features of this platform are provided in Table 1 and Table 2.

Table 1 – PowerEdge C6400 chassis

PowerEdge C6420 chassis	
Chassis	PowerEdge C6400, 2U chassis
Chassis options for disks and backplane	• 3.5” disks, up to 12 disk drives in chassis • 2.5” disks, up to 24 disk drives in chassis • 2.5” disks with NVMe, up to 8 NVMe drives and 16 SAS drives in chassis • No backplane for diskless option
Disks	Multiple choices of NVMe drives, SAS and SATA SSD drives, 10K SAS, 15K SAS, 7.2K NLSAS and SATA drives.
Power Supplies	2 * 1600W, 2000W or 2400W
Servers	Up to four PowerEdge C6420 server sleds

Table 2 – PowerEdge C6420 server sled

PowerEdge C6420 server sled	
Processor	Dual socket supporting Intel Xeon Scalable Processor Family (CPUs with architecture code named “Skylake”)
Memory	16 DIMM slots Recommended configuration for HPC: • 12 DDR4 memory modules • 4 remaining memory slots can be used for 3D XPoint memory when available.
Internal drives	• M.2 boot drive, 120GB capacity available at the time of writing • Internal SD card • Option for NVMe internal drives using the PCI-e slot

Disk Controller	PERC 9 (H330, H730P), HBA330 and chipset RAID controller
Hard disk drives	• With 4 server sleds in a chassis - up to 6 2.5” drives per server with up to 2 as NVMe; or up to 3 3.5” drives per server • With 2 server sleds in chassis - up to 12 2.5” drives per server
PCI-e slots	• 1 x16 PCI-e slot • 1 x8 mezzanine slot for internal storage controller • 1 x16 mezzanine slot for network cards • 1 x16 buried riser for M.2 device only
Network options	• On board 1GbE RJ45 connector PCI-e cards: • Mellanox EDR InfiniBand HCA • Intel Omni-Path Host Fabric Adapter • 1 GbE, 10GbE, 25 GbE, 40 GbE, 100 GbE adapters Mezzanine cards: • Intel Omni-Path carrier for Xeon-F CPUs that have integrated fabric • 10GbE card
Systems Management	iDRAC 9

A typical HPC configuration for compute-centric workloads that balances performance, value and energy efficiency would include a PowerEdge C6420 Server with two Intel Xeon Gold 6148 processors, 12 x 16GB 2667 MT/s memory modules, one SAS drive or an M.2 drive, and one Mellanox EDR InfiniBand or Intel Omni-Path adapter.

A configuration for workloads that need more memory capacity and more local disk storage would include a PowerEdge C6420 Server with two Intel Xeon Gold 6148 processors, 12 x 32GB 2667 MT/s memory modules, six SAS drive and an M.2 drive, and one Mellanox EDR InfiniBand or Intel Omni-Path adapter.

The Dell EMC server portfolio includes power supplies ranging in capacity from 495W to 2400W. The ordering webpage helps “right-size” a power supply to the specific configuration of the platform taking into account CPU, memory and disk power requirements to ensure the server is not starved for power.

Focusing on the PowerEdge C6420 specifically, the shared chassis has two power supply slots and can hold 1600W, 2000W or 2400W power supplies. The Dell Enterprise Infrastructure Planning Tool (EIPT) is an easy-to-use utility to configure different platforms and determine power, as well as weight requirements. For a typical HPC PowerEdge C6420 configuration with a mid-bin processor (150W+), 12 memory DIMMs, one disk, and an InfiniBand or Omni-Path card, the recommended platform power supply is two 2400W.