DELL EMC HPC Solution for Life Sciences v1.1: Deployment Best Practices

PowerEdge C6320 compute subsystem with Intel® Omni-Path fabric

Dell Engineering
October 2016

A Dell EMC Deployment and Configuration Guide

Revisions

Date	Description
October 2016	Initial release

Table of contents

1 Introduction

Dell HPC Solution for Life Sciences is a pre-integrated, tested, tuned and purpose-built platform, leveraging the most relevant of Dell's High Performance Computing line of products and best-in-class partner products due to the high diversity in life sciences applications. It encompasses all the hardware resources required for various life sciences data analysis while providing an optimal balance of compute density, energy efficiency, and performance from Enterprise server line-up of Dell.

Dell HPC solution for Life Sciences v1.1 provides higher flexibility for the solutions. A platform is available in five variants, depending on the cluster interconnects selected, which can be either 10 Gigabit Ethernet (GbE), Intel® Omni-Path (OPA), InfiniBand® (IB) EDR or IB FDR. In this version, the following options are available:

- PowerEdge C6320 compute subsystem with Intel® OPA fabric
- PowerEdge C6320 compute subsystem with IB EDR fabric
- PowerEdge C6320 compute subsystem with 10 GigE fabric
- PowerEdge FX2 compute subsystem with IB FDR fabric
- PowerEdge FX2 compute subsystem with 10 GigE fabric

The solutions are nearly identical for IB, Intel® OPA and 10 GigE versions, except for a couple of changes in the switching infrastructure and network adapters. These differences are outlined in Network Components section. The solution ships in a deep 48U rack enclosure, which was chosen because of its ease of mounting PDUs and for effortless cable management. This rack houses the compute, storage, and networking modules of the solution. Also, there are software modules which deploy, manage, and maintain the cluster.

This deployment guide describes the PowerEdge C6320 compute subsystem with Intel® OPA solution for diverse life sciences applications including molecular dynamics simulation solution into the flexible architecture as well as improving the performance of genomics data analysis platform.

2 Audience

This deployment guide describes the Dell EMC HPC Solution for Life Sciences and its configuration on the PowerEdge C6320 with Intel® Omni-Path interconnect. It assumes the reader is familiar with Dell PowerEdge products and switches, HPC cluster deployments, Bright Cluster Manager and standard HPC validation. It focuses on the special aspects of the Dell EMC HPC Solution for Life Sciences, and the genomics applications, molecular dynamics simulation applications, installation and benchmarking.

3 Solution Overview

The Dell HPC Solution for Life Sciences with PowerEdge C6320 compute subsystem with Intel® OPA fabric consists of 24 nodes of PowerEdge C6320 in one 48U rack. This solution also includes two master nodes, two login nodes, one CIFS gateway, Dell EMC HPC NFS Storage Solution - High Availability (NSS7.0-HA), and Dell EMC HPC Lustre Storage.

The configuration used for solution validation and performance benchmarking is shown here. NSS7.0-HA serves as a primary storage shared among master, login and compute nodes while Dell EMC HPC Lustre Storage is used for the performance and storage capacity needs.

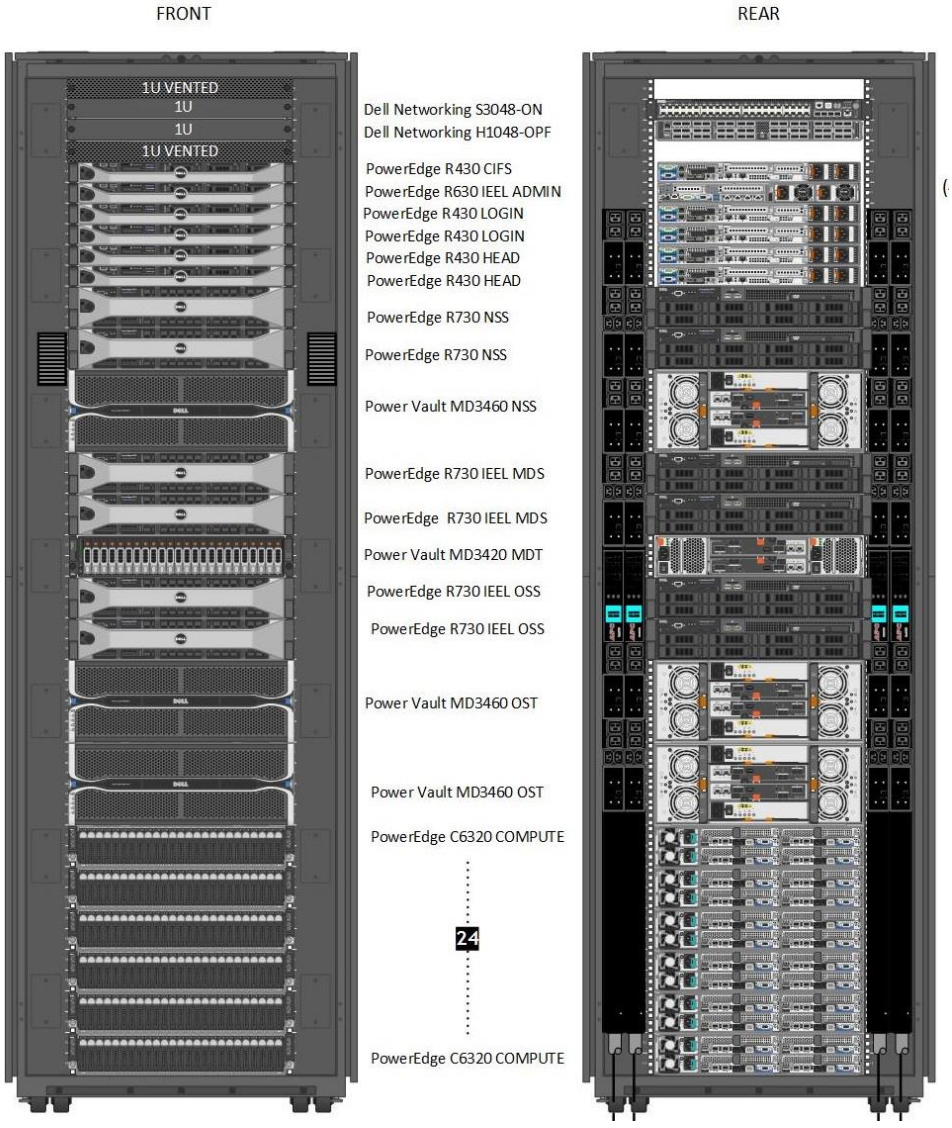

Figure 1 Dell HPC Solution for Life Sciences with PowerEdge C6320 rack servers and Intel® OPA fabric

3.1 Component details

3.1.1 Master node

The PowerEdge R430 is the choice of master nodes, and high level component details are listed in Table 1.

Table 1 PowerEdge R430 configuration

Component	Information
2x PowerEdge R430	
RAID controller	Dell PowerEdge RAID Controller (PERC) H730
Intel Omni-Path	Intel Omni-Path Host Fabric Adapter 100 Series 1 Port PCIe x16, Low Profile
1GbE	On-Board Broadcom 5720 Quad Port 1Gb LOM
Processors	2x Intel Xeon E5-2680 v4 processors
Memory	8 x 8GB RDIMM, 2400MT/s, Dual Rank
Disks	2 x 500GB 7.2K RPM NLSAS 6Gbps 2.5in Hot-plug Hard Drive,3.5inHYB CARR,13G
Internal optical drive	DVD+/-RW, SATA

Each PowerEdge R430 has two disks. It is recommended these to be configured in RAID-1 for the operating system.

3.1.2 Login node

Login nodes are where you login for access to the compute nodes. These nodes will be used for all access, compilation and job submission needs. Login node and high availability are optional. Power Edge R430 is the recommended server, and its configuration please refer to section 3.1.1.

3.1.3 Compute node

The Dell PowerEdge C6320 is an ultra-dense 2U server that can support up to four independent two socket (2S) servers. Each independent server features dual Intel Xeon E5-2600v3 or Intel Xeon E5-2600v4 series processors with up to 22 cores, C612 chipset for I/O connectivity, DDR4 memory, dual-port embedded 10 Gigabit Ethernet controllers (SFP+), and integrated iDRAC8 systems management with a dedicated RJ45 connection. The detailed configuration is listed in Table 2.

Table 2 PowerEdge C6320 configuration

Component	Information
24x PowerEdge C6320	
Intel Ommi-Path	Intel Omni-Path Host Fabric Adapter 100 Series 1 Port PCIe x16, Low Profile
1GbE	SFP+ to RJ45 transceiver in LOM1 of Intel x520 SPF+ (management network)
Processors	2x Intel Xeon E5-2697 v4 processors
Memory	8 x 16GB RDIMM, 2400MT/s, Dual Rank
Disks	1x 1TB 7.2K RPM SATA 6Gbps 2.5in Cabled Hard Drive per blade.

3.1.4 Common Internet File System (CIFS) gateway

Table 3 PowerEdge R430 configuration as CIFS gateway

Component	Information
PowerEdge R430	
Processors	2x Intel Xeon E5-2680 v4 processors
Memory	6 x 8GB RDIMM, 2400 MT/s, Dual Rank
Disk	2 x 500GB 7.2K RPM NLSAS 6Gbps 2.5in Hot-plug Hard Drive,3.5inHYB CARR,13G
RAID controller	Dell PowerEdge RAID Controller (PERC) H730
Intel Omni-Path	Intel Omni-Path Host Fabric Adapter 100 Series 1 Port PCIe x16, Low Profile
1GbE	I SFP+ to RJ45 transceiver in LOM1 of Intel X710 Dual Port 10Gb Direct Attach, SFP+
10GbE	Intel X710 Dual Port 10Gb Direct Attach, SFP+

3.2 Storage configuration

Both storages, NSS7.0-HA and Dell EMC HPC Lustre Storage are managed and operated through Dell Networking S3048-ON and Dell Networking H1048-OPF switch as described in 3.3.

3.2.1 NSS7.0-HA

A pair of Dell PowerEdge R730 servers were configured as an active-passive HA pair and function as a NFS server for the HPC compute cluster. Both NFS servers were connected to a shared Dell PowerVault MD3460 storage enclosure extended with one Dell PowerVault MD3060e storage enclosure (a 480 TB solution with the two PowerVault MD storage arrays) at the back-end. The user data is stored on an XFS file system created on this storage. The XFS file system was exported to the clients by using NFS. For the HA functionality of the NFS servers, a private 1 Gigabit Ethernet network was configured to monitor server health and heartbeat, and to provide a route for the fencing operations by using the Dell Networking 3048-ON switch.

3.2.2 Dell EMC HPC Lustre Storage

This solution continues to use the Dell PowerEdge R630 server as the Intel Management Server, while the Object Storage Servers and Metadata Servers in the configuration will be based on the Dell PowerEdge R730. The PowerEdge R730 is a 2U dual socket server with 7 PCIe expansion slots available enabling future expansion. Also new to this solution is support of the new Intel Omni-Path Host Fabric Interface (HFI) adapter. The storage backend of the solution will utilize the PowerVault MD3460 fully populated with your choice of 4TB, 6TB or 8TB SAS hard disk drives. For the Lustre software, the solution supports version 3.0 of the Intel Enterprise Edition for Lustre software. The solution uses two Dell PowerEdge R730 servers as Object Storage Servers (OSS) in an active-active configuration. 120 drive system configuration (60 drives per PowerVault MD3460) is used. The number of OSTs is 12, and each OST consists of 10x 4TB drives. Total raw storage size is 480TB.

3.3 Network configuration

This solution comprises of two network switches, Dell Networking S3048-ON and Dell Networking H1048-OPF switch, for management network and high-speed interconnects respectively.

The port assignment of the Dell Networking S3048-ON switch for the Intel® OPA or IB versions of the solution is as follows.

- Ports 01-04 and 27–52 are assigned to the cluster's private management network to be used by Bright Cluster Manager® connecting master, login, CIFS gateway and compute nodes. The PowerEdge C6320 server's ethernet and iDRAC constitute a majority of these ports.
- Ports 06–09 are used for the private network associated with NSS7.0-HA.
- The rest of the port 05 and ports 12–26 are allocated to the Lustre solution for its private management network
- Port 10 and 11 are used for the PDUs.

Figure 2 Dell Networking S3048-ON switch

Note: It is required to install four SFP+ to RJ45 transceivers in Dell Networking S3048-ON switch ports 49-52.

High-speed interconnect among master, login and compute nodes is configured through the Dell Networking H1048-OPF with 1:1 non-blocking topology. One port is assigned to each master, login and CIFS gateway and compute nodes. NSS7.0-HA needs two ports while DELL EMC HPC Lustre Storage requires five ports.

Figure 3 Dell Networking H1048-OPF switch

4 Cluster installation

4.1 Installing the head node

It is recommended that all the nodes are connected beforehand so that how things are connected is known.

1. The BIOS of the head node should have the local time set.
2. The head node should be booted from a Bright Cluster Manager (BCM) DVD or a flash. We created a bootable USB drive with a BCM ISO image containing BCM version 7.2 and Red Hat Enterprise Linux (RHEL) version 7.2.
3. Install Bright Cluster Manager should be selected in the text boot menu. This brings up the GUI installation Welcome screen.

Figure 4 Welcome screen

4. At the Welcome screen, Continue should be clicked. By default, this continues with a Normal (recommended) installation mode.
5. At the License screens: At the Bright Computing Software License screen, the acceptance checkbox should be ticked. Continue should then be ticked.
6. At the Linux base distribution screen, the acceptance checkbox should be ticked. Continue should then be clicked.
7. Continue at the Kernel Module screen.
8. At the Hardware Information screen, all the relevant hardware is detected. Continue should be clicked.
9. At the Nodes screen:
   - The number of racks and compute nodes are specified
   - The base name for the compute nodes is set. Accepting the default of node means nodes names are prefixed with node.
   - The number of digits to append to the base name is set. For example, accepting the default of 3 means nodes from node001 to node999 are possible names.

- The correct hardware manufacturer is selected

Continue is then clicked.

10. At the Network Topology screen, the default network layout is chosen. Click continue
11. At the Additional Network Configuration screen, add the Intel® OPA network and 1GbE network and configure the use of IPMI/iLO BMCs on the nodes. Adding an IPMI/iLO network is needed to configure IPMI/iLO interfaces in a different IP subnet, and is recommended. When done, Continue should be clicked.
12. At the Networks screen, the network parameters for the head node should be entered for the interface facing the network named externalnet. Uncheked DHCP checkbox, and add static values. Then OK button should be clicked.
13. At the Nameservers screen, add proper DNS search domains and external DNA name servers.
14. At the Network Interface screen, review IP addresses assigned to the network interfaces. Continue should be clicked.
15. If Intel® OPA network is properly enabled, the Subnet Managers screen is displayed. At this screen, nodes that are to run the subnet manager for the Intel® OPA network should be selected. Continue should then be clicked.
16. At the Installation source screen, the USB drive containing the BCM/RHEL should be selected, click Continue. Clicking on the Continue button starts the media integrity check.
17. At the Workload Management setup screen, select Torque/Maui from the dropdown menu. The recommended scheduler is Torque/Maui. The master node should be configured to run jobs. Slurm has known issues with next generation sequencing data analysis pipeline especially when some sub-processes are concatenated through piping.

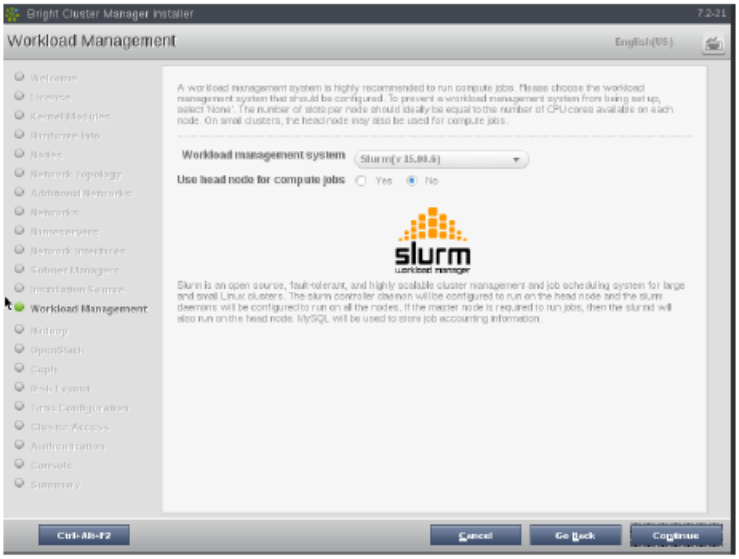

Figure 5 Workload Management Setup

18. At the Disk Partitioning and Layouts screen, two 500GB RAID 1 volume on the head node should be selected. The installation will be done onto this volume, overwriting all its previous content. Use master-standard.xml as the layout template.

19. At the Time Configuration screen, a time-zone should be selected, and optionally, NTP time-servers should be added. Continue should be clicked.
20. At the Cluster Access screen, accept the defaults and click Continue.
21. At the Authentication screen, a hostname should be entered for the head node. Also a password should be entered for use in system administration. Continue.
22. At the Console screen, a text or graphical console can be configured for the nodes in the cluster.
23. At the Summary screen, the network summary should be reviewed. The Start button then starts the installation. Yes should be clicked to confirm that the data on the listed volume may be erased.
24. Installation Progress screen should eventually complete. Clicking on Reboot and then clicking Yes to confirm, reboots the head node.