Dell HPC General Computing – Configuration and Supported Architecture with Dell PowerEdge 13th Generation Servers

A Dell Technical White Paper

Munira Hussain, Saeed Iqbal, Calvin Jacob

Dell HPC Engineering

December 2015| Rev 1.0

Contents

Figures

Introduction

High Performance Computing (HPC) Clustering is the concept of building cost-effective, computation intensive and performance optimized solutions built from commodity servers, storage, operating system, middleware libraries and high bandwidth low latency interconnects. It aims to provide intense computing power that can help perform data calculations and operations rapidly and scale effective with increasing number of nodes. From software perspective the design includes a hierarchical integrated architecture comprising of cluster configuration with middleware libraries, drivers tools that can be deployed, provisioned, monitored and managed on large number of nodes.

This white paper provides information on the latest Dell HPC General Research Computing Solution based on the Dell 13th Generation servers. The solution supports new generation Intel Xeon E5-2600 v3 based PowerEdge servers targeted to provide optimal performance and dense compute power.

Overview

This release includes support for the Bright Cluster Manager 7.1 with RedHat Enterprise Linux 6.6 errata kernel (2.6.32-504.16.2.el6.x86_64) on the following Dell Haswell based servers. The release includes support for updated Dell system management tools and utilities, network drivers and scripts, third party components and integration with Dell BIOS API.

Table 1. Supported Server Configurations

	Servers:	Interconnect support
Dell PowerEdge Rack series:	R730, R730XD,R630, R430, R930	PowerConnect 2800 Ethernet Force10 S , MXL and Z series Mellanox InfiniBand FDR/QDR
Dell PowerEdge M series:	M630, M830	
Dell PowerEdge C series:	C6320, C4130	
Dell PowerEdge FC series	FC830, FC630, FC430	
Dell PowerEdge VRTX series:	M630	

Reference Design and Targeted Platforms for various HPC needs

The supported servers in the above configuration cater to various general HPC compute needs. These are applicable based on the customer needs and application characteristics. For example some of the server selections are based on the infrastructure of the data center, power cooling, compute density, memory, network communication and I/O characteristics of the code.

Table 2. Characteristics of Servers in HPC Configuration

PowerEdge Servers	Number of Cores (Assumption is made by keeping Intel Xeon E5-2695 v3 @ 2.3GHz - 14C as baseline)	Number of cores in 1U/Rack	Total Memory/DIMM Slots	Support for GPGPU/Accelarator	I/O HD
R730	28	14	24	YES	8x3.5"/16x2.5"
R730XD	28	14	24	NO	16x3.5"/26x2.5"
R630	28	28	24	NO	10x3.5"/24x1.8"
R430	28	28	12	NO	4x3.5"/10x2.5"
C6320	28	56	16	No	6x2.5"
C4130	28	28	16	YES	4x2.5"
FC630	28	56	24	NO	2x2.5"
FC430	28	112	8	NO	2x18"
M630 (in M1000e)	28	~45	24	NO	2x2.5"SSD

GPU/Accelarator Nodes	High Density	Large Memory	Large I/O Capacity

The Table 2 above provides a snapshot of the advantages of the server configurations. The color coding refers to what category they fall into. For example a R730 or C4130 would be design choices for codes or algorithms that have potential performance improvement and can take advantage of GPU programming models. For HPC applications that are highly computation intensive, high density servers like C6320, FC630 and FC430 are applicable design choices.

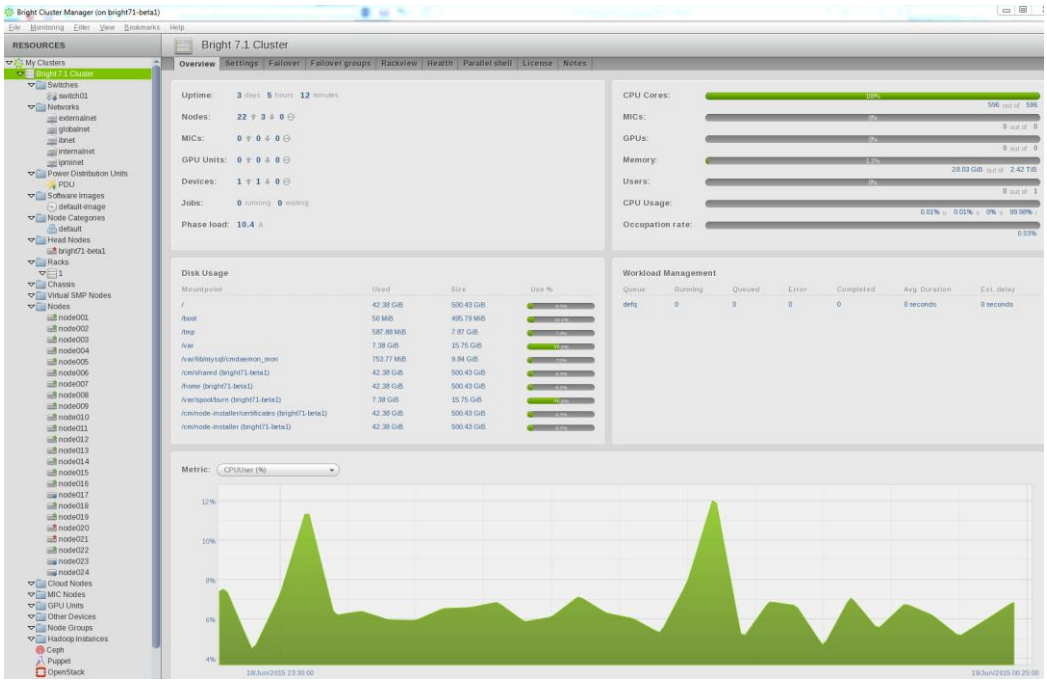
Details of Bright Cluster Manager 7.1

Bright Cluster Manager is a commercially supported software solution stack from Bright Computing. It aims to provide provisioning, deployment, installation, management and monitoring for High Performance Computing Clusters. Bright Cluster Manager 7.1 setups all the necessary services, ssh keys, passwords, kernel parameters, home directories that are needed for the configuration and performance of a cluster. It also integrates additional third party components eg: drivers and tools and middleware libraries.

Key features of Bright Cluster Manager 7.1 include:

- Bright Cluster Manager 7.1 Dell Edition is based on RHEL 6.6.z kernel.
- Dell BIOS Integration framework
- Cluster Management GUI and Shell
- Node Provisioning & Image Management
- Node Identification
- Software Update Management
- Cluster Monitoring
- User Management
- Parallel Shell
- High availability for the compute nodes
- Hardware RAID setup
- Scale to thousands of hosts / multiple clusters.

Figure 1. Bright Cluster Manager: Dashboard showing the Cluster Overview and Status

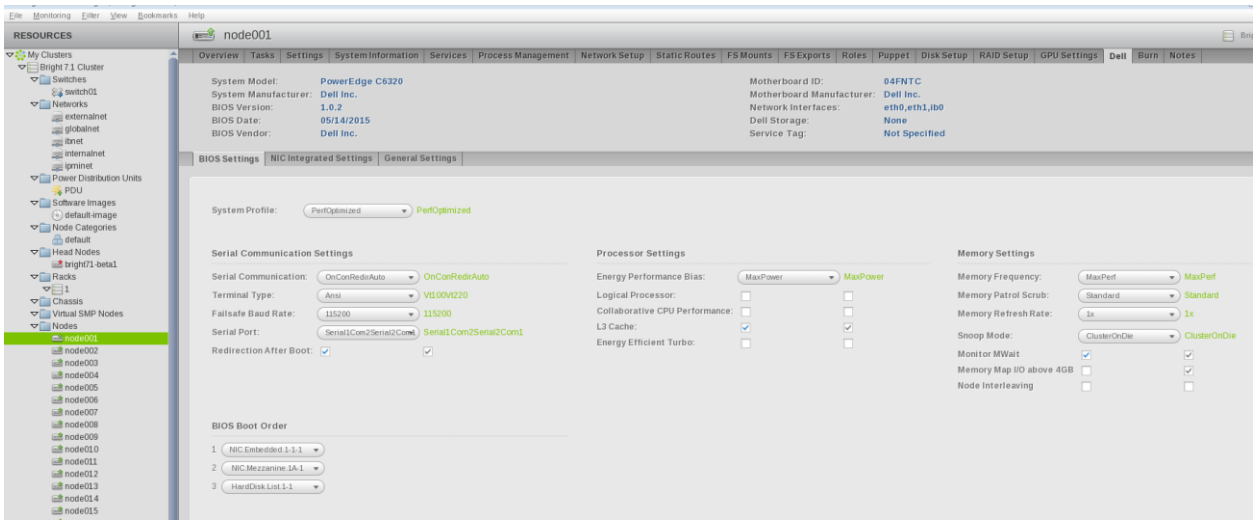


Additional Dell IP and tools are also integrated in the Dell iso that are available when customers download the respective iso from the Bright ftp website. The Dell version includes support for existing and latest Dell servers and is fully verified and validated across supported configuration. It also bundles the latest Dell System Management tools such as OpenManage and DTK that are deployed and configured at the time of cluster install. Furthermore, there are automated scripts available that set and tune BIOS settings on Dell servers for optimal performance. Additional integration includes add-on Dell network drivers and storage controller drivers that are generally not available in the base Operating System.

With this release operating and administrating the HPC has become seamless. Be it updating firmware on the compute nodes or setting the BIOS parameters, this could be performed with ease by accessing the nodes in the Bright CMGUI. Bright uses Dell Life Cycle Controller at the background to execute these tasks. Any anomaly or inconsistency can be seen in the Bright GUI and the necessary remedial steps can be taken.

Bright Cluster Manager 7.1 presents single GUI which reports details like health of a system, availability of resources, options to change BIOS settings, firmware update, etc. Bright Cluster Manager 7.1 uses Integrated Dell Remote Access Controller (iDRAC) for systems management, remote console redirection and firmware updates through Life Cycle Controller.

Figure 2: Dell BIOS Integration with 13G Servers



Third Party Software Drivers and Tools Integration

This release also includes support for high bandwidth low latency interconnect, ConnectX3 FDR/FDR10/QDR from Mellanox. Along with it provides GPU and Accelerator support and integration for the latest technology available. The current release includes the following versions of software stack, drivers and tools:

- Mellanox OFED 2.4, Mellanox OFED 2.3, Mellanox OFED 2.2 and Mellanox OFED 2.1 available as a drop down selection option
- nVidia CUDA 7.0 for GPGPU support
- Intel MPSS 3.5 for Intel Xeon Phi.
- No factory installation of cluster software.

Master /Compute Node Configurations

The section below goes into more details regarding the server configuration and specs. The specific nodes are design considerations for master and compute nodes. A master node is generally the node that manages home directories, security and stores repositories for the cluster nodes.

PowerEdge R730

Intel Haswell-EP 2S servers with up to 18 core processors and with minimum 4 GB DDR4 memory.

- Rack Mount 2U
- Intel® Haswell-EP up to 18 cores
- Processor support of 55W, 65W, 85W, 90W,105W, 120W, 135W,145W
- Two socket
- Two QPI links up to 9.6 GT/s
- 1333, 1600, 1866, 2133 MHz (3DPC) DDR4 DIMMs, RDIMM & LR-DIMM
- 4GB/8GB/16GB/32GB DIMM, maximum memory support upto 768
- Option for Broadcom/Intel/Qlogic NDCs (1Gb/s or 10Gb/s) (dual port or quad port)
- PCIe 3.0 x8 and x16 slots
- S130, H330, H730, H730P and H830 RAID controller support
- SAS, SATA, Near-line SAS, SSD, PCIe SSD support
- Maximum storage on R730 Up to 48TB via 8 x 3.5" BP
- Maximum storage on R730xd – Up to ~100TB via 16 X 3.5 HDD
- PCIe Switch Card controller for x4 PCIe SSD

PowerEdge R630

- Rack Mount 1U
- Intel® Haswell-EP up to 18 cores
- Processor support of 55W, 65W, 85W, 90W,105W, 120W, 135W,145W
- Two socket
- Two QPI links up to 9.6 GT/s
- 1333, 1600, 1866, 2133 MHz (3DPC) DDR4 DIMMs, RDIMM & LR-DIMM
- 4GB/8GB/16GB/32GB DIMM, maximum memory support up to 768GB
- Option for Broadcom/Intel/Qlogic NDCs (1Gb/s or 10Gb/s) (dual port or quad port)
- PCIe 3.0 x8 and x16 slots
- S130, H330, H730, H730P and H830 RAID controller support
- SAS, SATA, Near-line SAS, SSD, PCIe SSD support
- Maximum storage on R630 - Up to 18TB via 10 x 2.5" BP
- PCIe Switch Card controller for x4 PCIe SSD

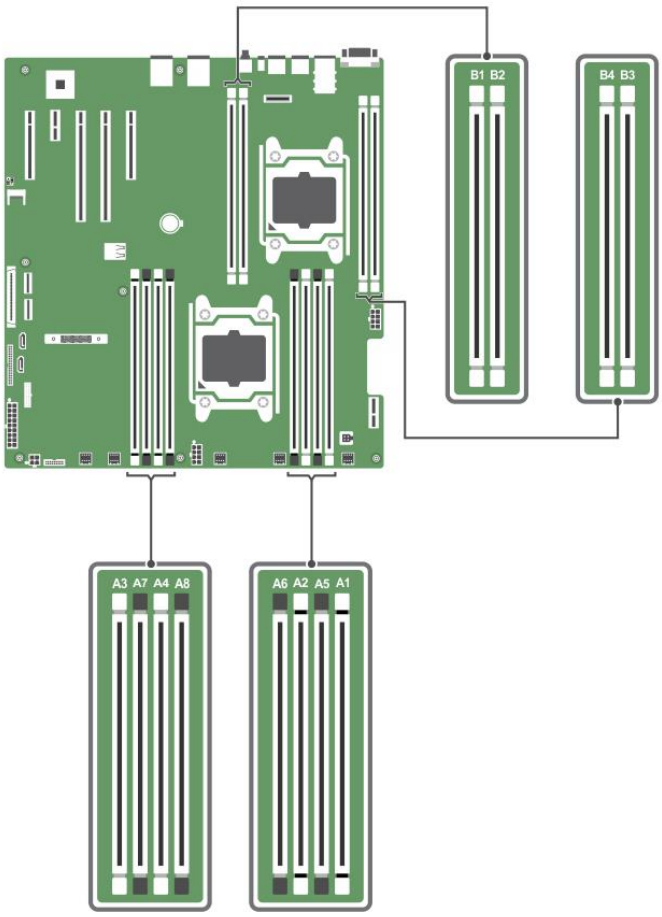
PowerEdge R430

- Rack Mount 1U
- Intel® Haswell-EP up to 16 cores
- Processor support of 55W, 65W, 85W, 90W,105W, 120W, 135W,145W
- Two socket
- Two QPI links up to 9.6 GT/s
- 1333, 1600, 1866, 2133 MHz (3DPC) DDR4 DIMMs, RDIMM & LR-DIMM
- 4GB/8GB/16GB/32GB DIMM, maximum memory support up to 384GB
- Option for Broadcom/Intel/Qlogic NDCs (1Gb/s or 10Gb/s) (dual port or quad port)

- Two PCIe 3.0 x8 slots
- S130, H330, H730, H730P and H830 RAID controller support
- SAS, SATA, Near-line SAS, SSD, PCIe SSD support
- Maximum storage on M630 - Up to 10 x 2.5" BP

Memory population for R430:
The R430 server supports DDR4 registered DIMMS (RDIMMS). Memory bus operating frequency can be 2133MT/s, 1866MT/s, 1600MT/s, or 1333MT/s.
The system contains 12 memory sockets split into four sets – two sets of 4 sockets and two sets of 2 sockets each. Each 4-socket set is organized into two channels and each 2-socket set is organized into one channel. In each channel of the 4-socket set, the release levers of the first socket are marked white and the second socket black. In the 2-socket set, each release lever is marked white.

Figure 3. R430 Memory Population



Compute Node Configurations:

PowerEdge C6320

This is a half-height sled that fits into a 2U chassis. The chassis can support upto four sleds.

- Intel® Xeon E5-2600 v3 up to 18 cores
- Processor support of 85W, 90W,105W, 120W, 135W, 145W
- Two socket
- Support for Mellanox ConnectX3 FDR mezz card
- Two QPI links up to 9.6 GT/s
- 1333, 1600, 1866, 2133 MHz (3DPC) DDR4 DIMMs, RDIMM & LR-DIMM
- 4GB/8GB/16GB/32GB/64GB DIMM, maximum memory support up to 256GB
- Option for Broadcom/Intel/Qlogic NDCs (1Gb/s or 10Gb/s) (dual port)
- One PCIe 3.0 x16 Riser slot
- Maximum storage up to 24 x 2.5" SAS, 12 x 3.5" SAS or 24 x 2.5" SDD
- iDRAC8 available for system Management
- Default upto six 2.5" SATA drives supported per sled

PowerEdge FC430

This is a quarter-width form factor and holds up to 8 nodes per PowerEdge FX2 enclosure

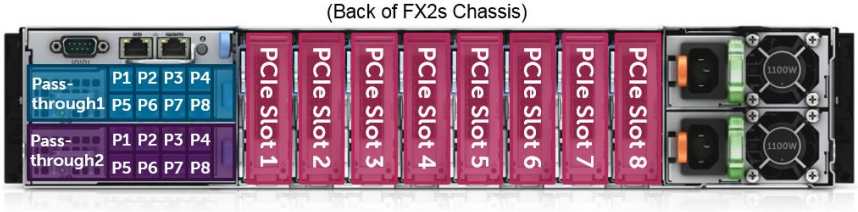
- Intel® Haswell-EP up to dual socket 14 cores /35MB cache support
- Support for front bay single port Mellanox ConnectX3 FDR Mezz with speeds of up to 56Gbp/s and latency of 0.7us. The front Mezz supports IB-pass through and is connected to CPU2
- Processor support of 55W, 65W, 85W, 90W,105W, 120W
- Two socket
- Two QPI links up to 9.6 GT/s
- Total 8 DIMM slots with upto 2133 MT/s DDR4 DIMMs
- 4GB/8GB/16GB/32GB DIMM, maximum memory support up to 256GB
- Option for Broadcom/Intel/Qlogic NDCs (1Gb/s or 10Gb/s) (dual port or quad port)
- One PCIe 3.0 x8 mezzanine slot and one PCIe x16 front cabled slot
- Maximum storage up to 2 x 1.8" SSD. However with Front Mezz IB only single storage 1.8" SSD available.
- CMC and iDRAC8 available for system Management

Figure 4. FC430 in a FX Chassis and Configuration

Ethernet and Port Mapping for Quarter-Width Sleds

The fixed mapping of which PCIe slots and Ethernet passthrough ports to quarter width server sleds

Location of Physical Ports
(Back of FX2s Chassis)



Mapping of PCIe Slots and Ethernet Ports to Server Sleds
(Front of FX2s Chassis)

Stream bandwidth for FC430 nodes with 8C and 12C with varying Snoop Modes and impact on Infiniband bandwidth and latency.

System info:

Dual socket Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz (8core) and Intel Xeon ® CPU E5-2680 v3 @ 2.50 GHz (12 core), 8 x 16GB DDR4

Bios: 1.0.4, perfoptimized, HT-disabled, nodeinterleave=disabled, ES=early snoop, HS=Homesnoop

Stream (FC430 -8Core Triad GB/s)		
	ES	HS
Full	89	89
como	45	45
com1	7	36

Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz (8core)					
			ES/HS	ES/HS	ES/HS
ib_write		core /memory	Bw (MB/s)	Bibw (MB/s)	Lat (us)
		como	5600/ 5800	10130/ 10522	0.8/ 0.8
		com1	5800/ 5820	10555/ 10550	0.77/ 0.78
		c1mo	5800/ 5800	11850/ 11840	0.66/ 0.66
		c1m1	5800/ 5820	11850/ 11850	0.66/ 0.66
		without numactl	5600/ 5800	10800/ 10800	0.76/ 0.76

Stream (FC430 -12c Triad GB/s)			
	ES	HS	COD
Full	115	116	116
como	58	58	30
com1	20	41	14

Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz (12core)					
			ES/HS/COD	ES/HS/COD	ES/HS/COD
ib_write		core /memory	bw (MB/s)	bibw (MB/s)	lat (us)
		como	5945 /5940 /5963	11800 /11820 /11493	0.69 /0.67 /0.69
		com1	5957/ 5955 / 5988	11800 /11850 / 11734	0.67 / 0.66 /0.65
		c1mo	6037 /6037 /5978	11840/11750 /11674	0.57/ 0.58 / 0.56
		c1m1	6039/ 6039 /6010	11860 /11850 /11837	0.57/ 0.57 /0.58
		without numactl	6060 /6050 / 6011	10800 /11850 / 11850	0.69 /0.66 /0.65

As seen in the performance numbers above, the 8core processor latencies are slightly higher by 0.1-0.2 us sec. This is because the memory bandwidth is impacted by the processor architecture which includes only a memory controller for 8core and lower processors. Also note that when there is no pinning the bandwidth and bi-directional bandwidth is slightly less by 100-200 MB/s for 8core servers. Additionally, early snoop (ES) has a greater impact in remote bandwidth as it falls to almost 7GB/s in memory bandwidth. Home snoop mode aims to provide a more consistent bw/bibw across numerous core/memory mappings for processors 8core and lower.

The Infiniband bandwidth and bi-bandwidth for 12core processor servers is fairly consistent across all snoop modes.

PowerEdge FC830

This server is a full-width form factor and holds up to 2 nodes per PowerEdge FX2 enclosure.

- Intel® Xeon E5-4600 v3 up to 18 cores
- Processor support of 55W, 65W, 85W, 90W,105W, 120W, 135W
- Four socket
- Four QPI links up to 9.6 GT/s
- 1333, 1600, 1866, 2133 MHz (3DPC) DDR4 DIMMs
- 4GB/8GB/16GB/32GB/64GB DIMM, maximum memory support up to 2048GB
- Option for Broadcom/Intel/Qlogic NDCs (1Gb/s or 10Gb/s) (dual port or quad port)
- Two PCIe 3.0 x16 mezzanine slots
- Maximum storage up to 8 x 2.5" SAS or 16 x 1.8" SSD
- CMC and iDRAC8 available for system Management

PowerEdge M630

This server is a half height blade that plugs into a 10U M1000e chassis. The chassis can support upto 16 M630 blades.

- Two socket
- Two QPI links up to 9.6 GT/s
- 1333, 1600, 1866, 2133 MHz (3DPC) DDR4 DIMMs, RDIMM & LR-DIMM
- 4GB/8GB/16GB/32GB DIMM,
- Option for Broadcom/Intel/Qlogic Mezz (1Gb/s or 10Gb/s)
- Two PCIe 3.0 x8 Mezz slots
- Support for dual port FDR /FDR10 and QDR Infiniband Mezz card
- SAS, SATA, Near-line SAS, support

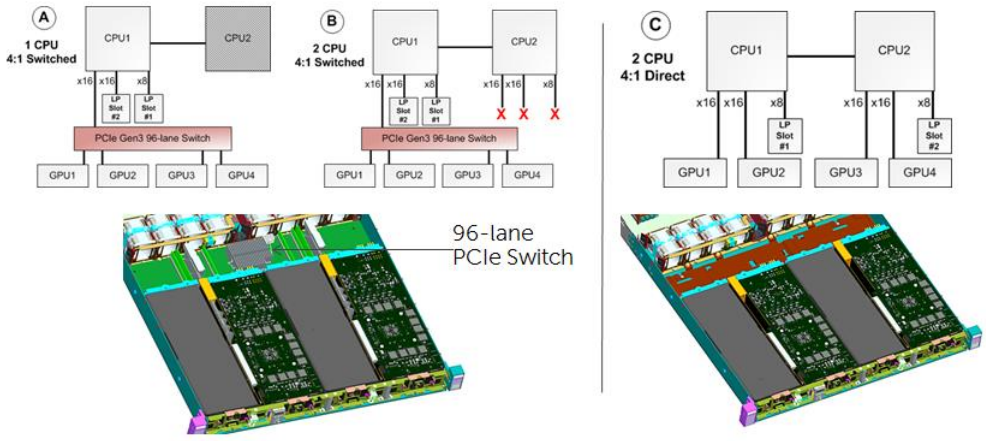
PowerEdge M830

This is a full-height form factor and holds up to 8 nodes per M1000e enclosure.

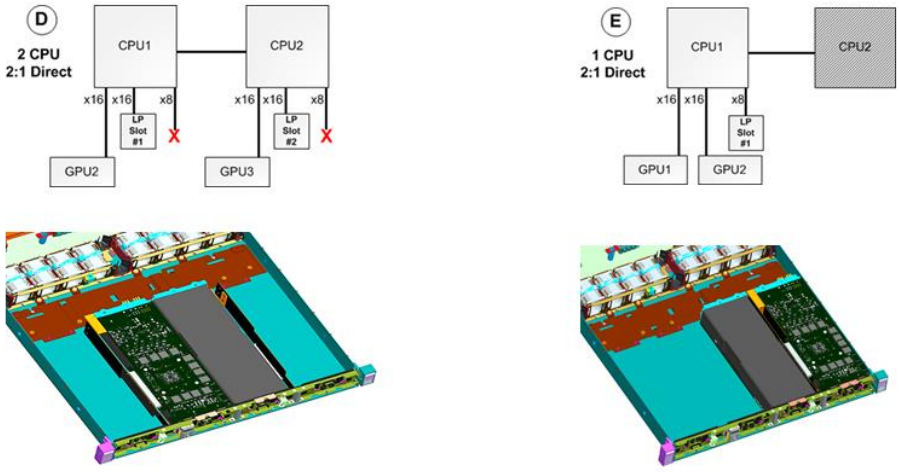
- Intel® Xeon E5-4600 v3 up to 18 cores
- Processor support of 55W, 65W, 85W, 90W,105W, 120W, 135W
- Four socket
- Support for Mellanox ConnectX3 FDR10
- Four QPI links up to 9.6 GT/s
- 1333, 1600, 1866, 2133 MHz (3DPC) DDR4 DIMMs, RDIMM & LR-DIMM
- 4GB/8GB/16GB/32GB/64GB DIMM, maximum memory support up to 2048GB
- Option for Broadcom/Intel/Qlogic NDCs (1Gb/s or 10Gb/s) (dual port or quad port)
- Four PCIe 3.0 x8 mezzanine slots
- Maximum storage up to 4 x 2.5" SAS or 12 x 1.8" SSD
- CMC and iDRAC8 available for system Management

PowerEdge C4130

The C4130 server is available in various flavors with 2GPUs/4GPUs 1CPU/2CPU. So the recommended configuration depends on the customer requirement, the GPU/accelerators used and the application being run on the servers.



- A is most economical for connecting 4 accelerators
- B is A with an additional CPU, both have SW
- C has no switch module, also C is the most balanced



- E most economical for connecting 2 accelerators
- D is the most balanced

PowerEdge R930

This is a 4S server generally used as a fat node for applications requiring large amount of memory. It supports 22nm Intel Xeon E7 4800/8800 v3 series processors also known as Haswell-EX. Intel Haswell-EX Processors have following features:

- Up to 18 execution cores per processor
- Each core supports two threads for up to 36 threads per processor
- 46-bit physical addressing and 48-bit virtual addressing
- 32 KB instruction and 32 KB data first-level cache (L1) for each core
- 256 KB shared instruction/data mid-level cache (L2) for each core

- Up to 37.5 MB last level cache (LLC): up to 2.5 MB per core instruction/data last level cache (LLC), shared among all cores
- Three QPI links up to 9.6 GT/s
- Four DMI2 lanes
- 32 PCIe Gen3 links capable of 8.0 GT/s
- No termination required for non-populated CPU (must populate CPU socket 1 first)
- Integrated 4 channel DDR4 memory controller
- 64 byte cache line size
- Execute disable bit
- Support for CPU turbo mode
- Intel 64 Technology

Conclusion or Summary

Deploying and managing a HPCC with Dell 13th Generation servers has never been so easy and seamless as it is now. With best of the breed servers and a choice of interconnects and memory options, Dell is able to clearly address the customer requirements in the most effective way. Management and maintenance of a cluster has been simplified in the Bright Cluster Manager CMGUI using intuitive methods and over all reporting enables the cluster admin to take remedial steps. With this release of HPCC Solution Stack from Dell, the maintenance downtime of the cluster is set to drop drastically and hence resulting in higher uptime of the cluster for meaningful jobs to be executed.