Santos Dumont Architecture

10-10-2016

Bull
atos technologies

Outline

- Project
- Santos Dumont General Information
- Mobull
- Cooling systems
- Compute nodes
- Service nodes
- Storage
- Software
- Top500
- Questions

Project

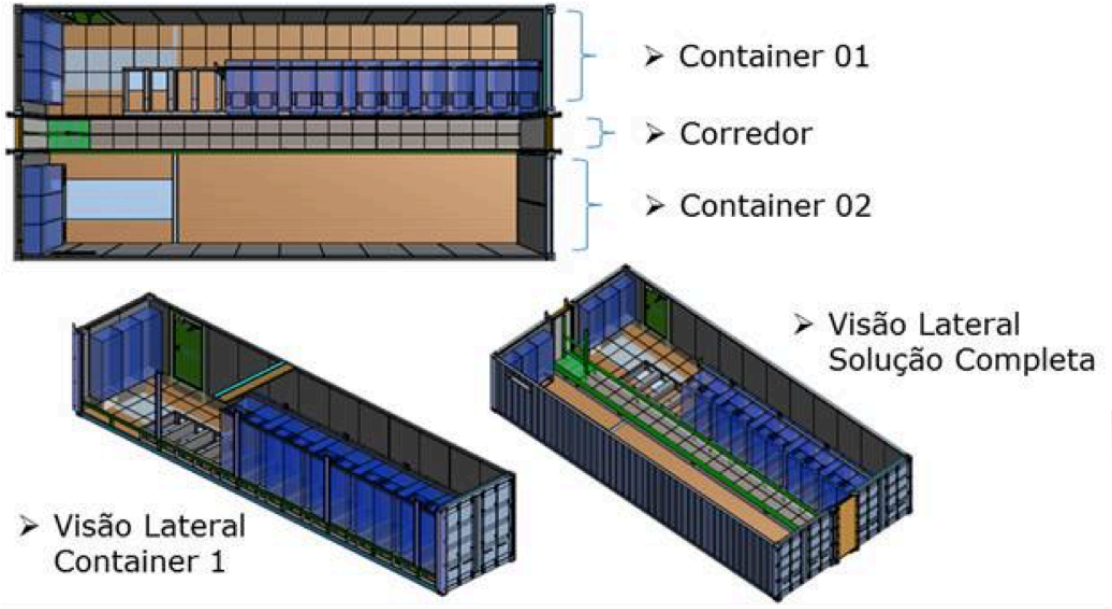
- Signed on September 2014
- Started on January 2015
- Manufactured & Build at Bull's France Factory in Angers
- Acceptance test executed on April 2015
- One training week at Grenoble France
- System administration training at headquarters, Paris, France
- Transported by cargo ship
- Arrived at LNCC in July 2015
- First tests on November 2015

Santos Dumont General view

- Santos Dumont is a supercomputer based on the HPC cluster concepts
- The solution includes from the Datacenter to software training
- The main technologies used are:
  - Mobull
  - Water Cooling systems
  - Bullx DLC Hardware based on Intel E5-2600v2 processors, nVidia K40, Xeon Phi KNC 7120P
  - Bullx S technology for FAT-nodes
  - Infiniband FDR network, GbE network, 10 GbE network
  - Seagate Lustre
  - RedHat Enterprise Linux 6.4 operating system
  - Bullx SCS AE 4
  - SLURM resource manager
  - Intel Parallel Studio, GNU compilers
  - Bullx MPI, Intel MPI

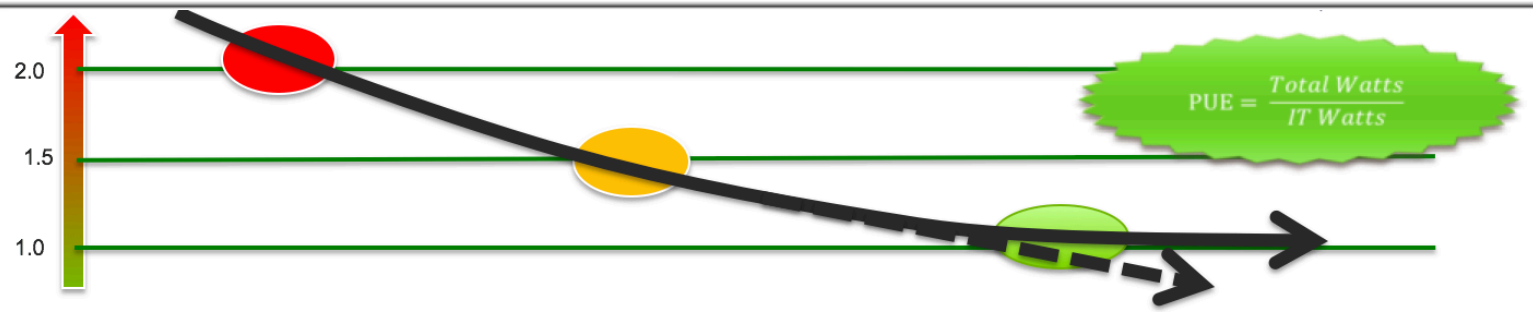
Mobull

- Mobull is a container based solution for Datacenters
- Designed for plug & boot
- The LNCC configuration is based on 2 Mobull and a corridor
- Each Mobull can host up to 14 standard 42” racks
- The corridor was specially designed for LNCC solution



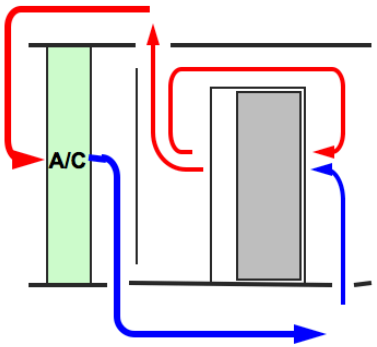
Mobull

Cooling & Power Usage Efectiveness



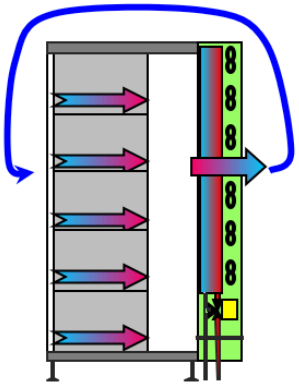
Air-cooled
10(-20) kW/rack
Room 20° C
A/C water 7-12° C

PUE ≥ 1.7



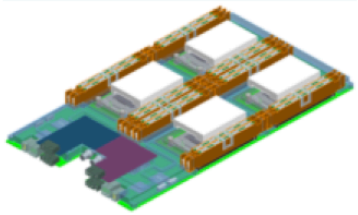
Water-cooled doors
40 kW/rack
Room 20°C
Water 7-12°C

PUE ~ 1.4



Direct-Liquid-cooling
80 kW/rack
Room up to 27°C
Water Up to 30 °C

PUE < 1.1

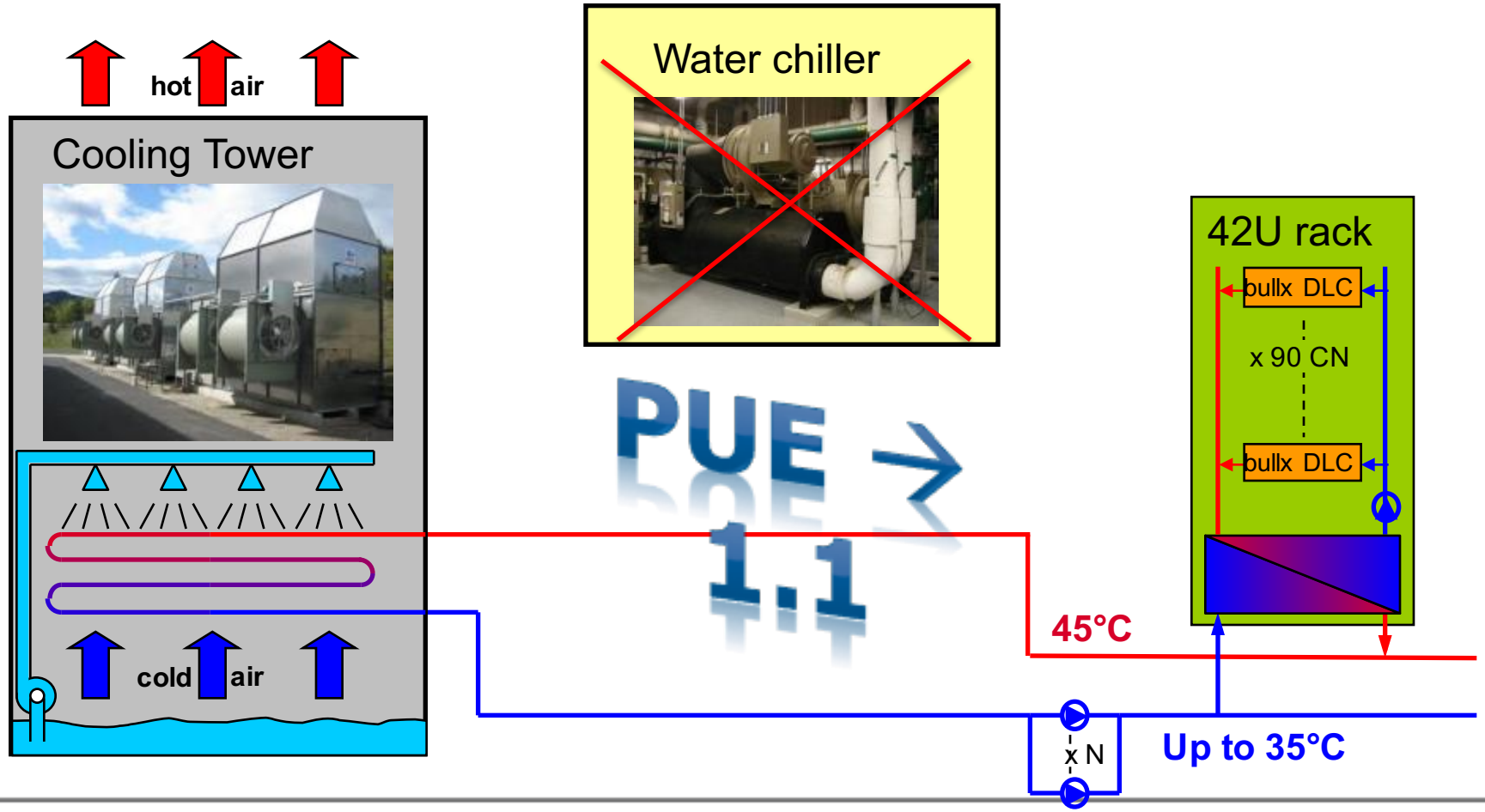


Co-generation

Bull atos technologies

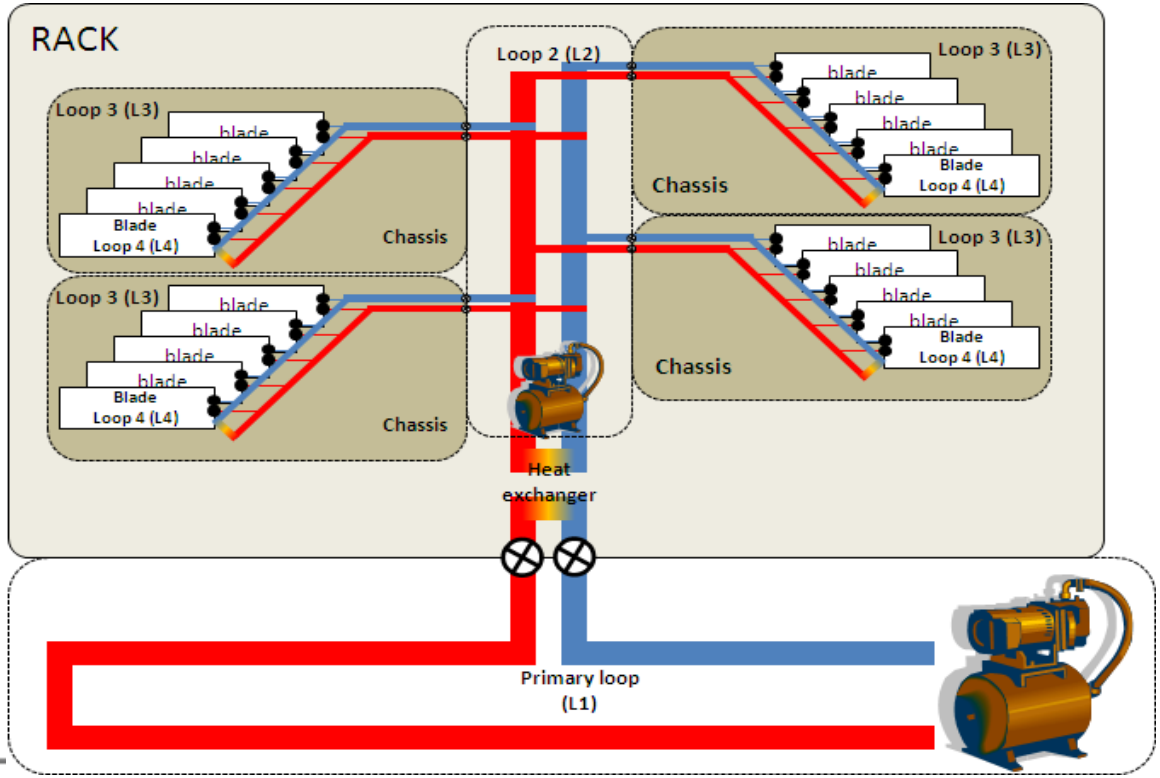
Direct Liquid Cooling - DLC

- With hot/warm cooled servers, water chillers are not used.



Direct Liquid Cooling - DLC

- DLC technology applies to the compute nodes racks (14 of 20).
- There are 3 water loop.
- 2 Heat exchanger (HYC) are at the bottom of the rack to exchange heat from the primary loop and the second loop.

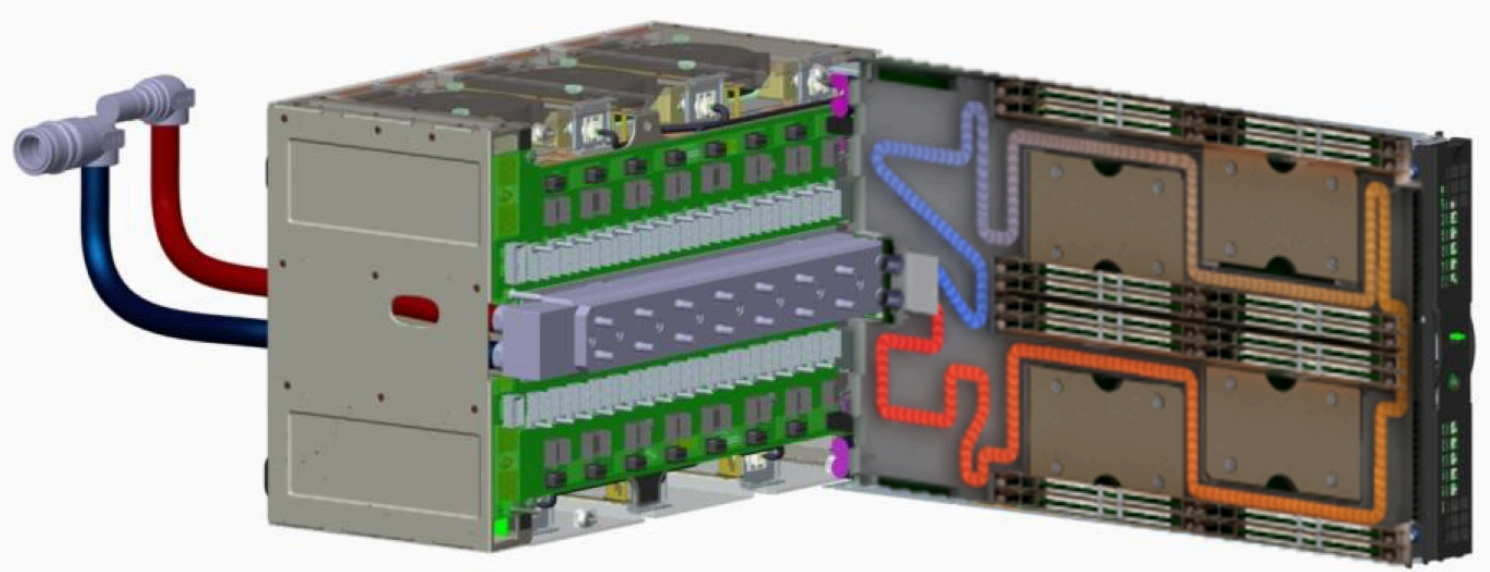


Cooling systems

- The LNCC cooling system is based on two types of water pipes (warm and cold).
- There are 2x inlet (warm and cold) and 4 outlet water pipes.

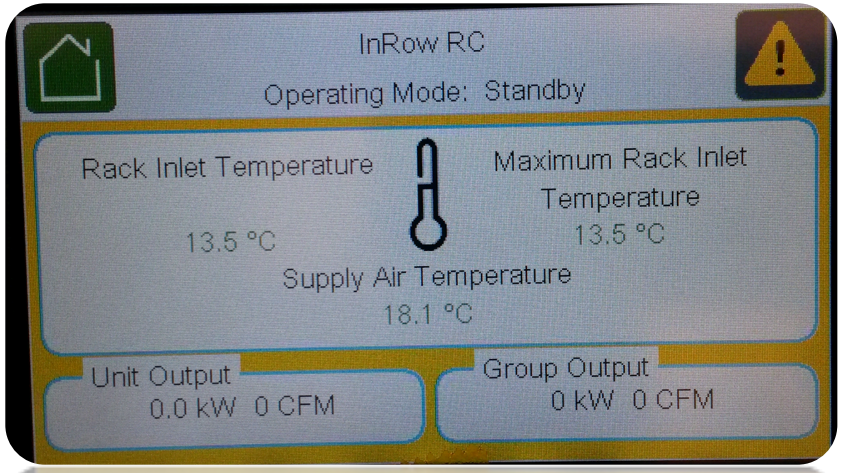
Direct Liquid Cooling - DLC

- The compute nodes has a cold plate with an internal water flow.



InRows

- For the air cooled components, like service nodes, switches, storage, etc., the InRows technology is used.
- InRows have the same height of a standard rack and are located between them.
- They pull the air from the hot corridor and push it to the cold one.
- In that process they cool the air using cold water
- They work with cold water (7-12 C) degrees.



Santos Dumont cluster design

- The Santos Dumont was designed to be a general purpose cluster.
- In that way, the state of art technologies were used for the compute nodes:
  - Thin nodes (cpu & memory).
  - Accelerators or hybrid nodes (cpu & accelerator & memory).
  - Fat-node (huge number of cores & memory).
- Thin & hybrid nodes has the same CPU & memory configuration.
- All these nodes interconnected with a high performance network: Infiniband FDR.
- A Lustre based solution was configured as the production parallel filesystem of the cluster.

Compute nodes

- The compute nodes are based on the bullx DLC B700 technology
  - B510 for thin nodes
  - B515 for accelerators nodes
- The rack can host 4x 7U chassis
- Each chassis has capacity for 9 blades
- B510 blade host 2 compute nodes
- B515 blade host 1 compute node with 2 accelerators

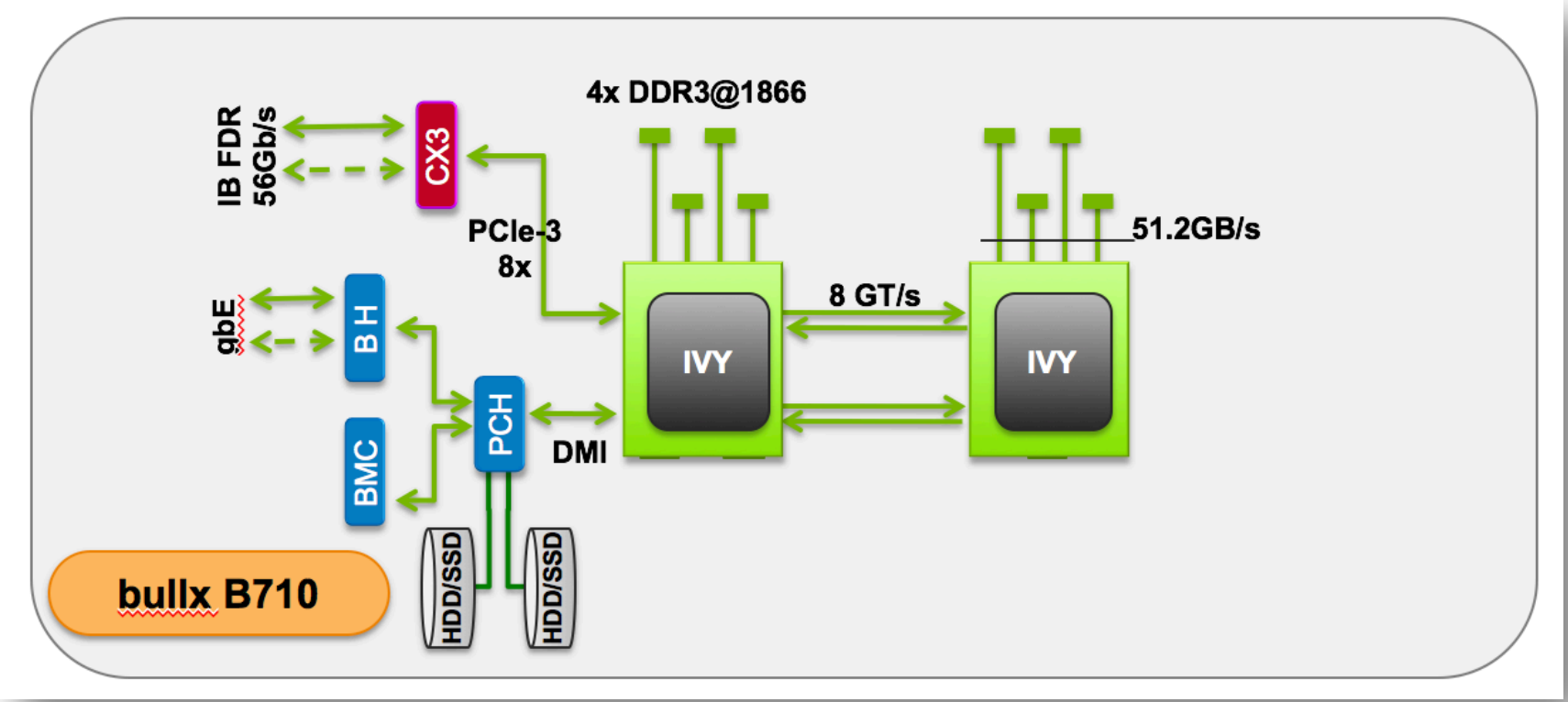
B510 – thin nodes

- The cluster has 7x full racks of B510 nodes for a total of **504** compute nodes

- Each compute node has the following configuration:
  - 2x Intel Xeon E5-2695v2 (12c, 2.4 Ghz)
  - 64 GB DDR3 RAM (8x 8GB DIMM)
  - 1x 120GB SSD disk
  - 1x Infiniband FDR ConnectX-3
  - 1x GbE
- Each compute node provides 460.8 Gflops
- Each chassis has 18 compute nodes
- Each rack has 72 compute nodes

B510 – thin nodes

Block diagram



B515 – Accelerators nodes

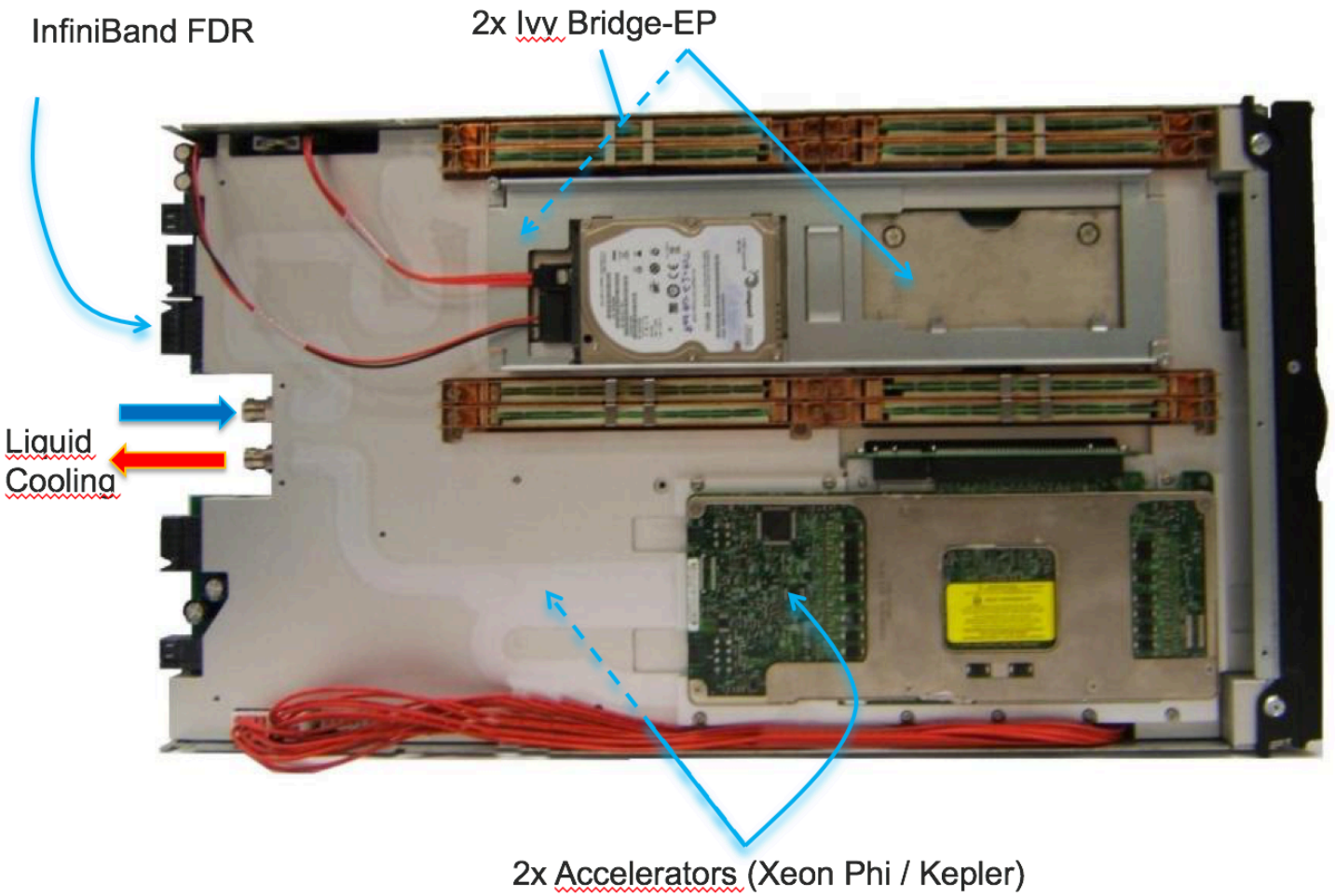
nVidia nodes

- The cluster has 5x and a half racks of B515 nodes with 2x K40 nVidia boards for a total of **198** nodes and **396** nVidia K40

- Each compute node has the following configuration
  - 2x Intel Xeon E5-2695v2 (12c, 2.4 Ghz)
  - 64 GB DDR3 RAM (8x 8GB DIMM)
  - 1x 120GB SSD disk
  - 1x Infiniband FDR ConnectX-3
  - 1x GbE
  - 2x nVidia K40 12GB GDDR5
- Each compute node provides 3320.8 Gflops
- Each chassis has 9 compute nodes
- Each rack has 36 compute nodes

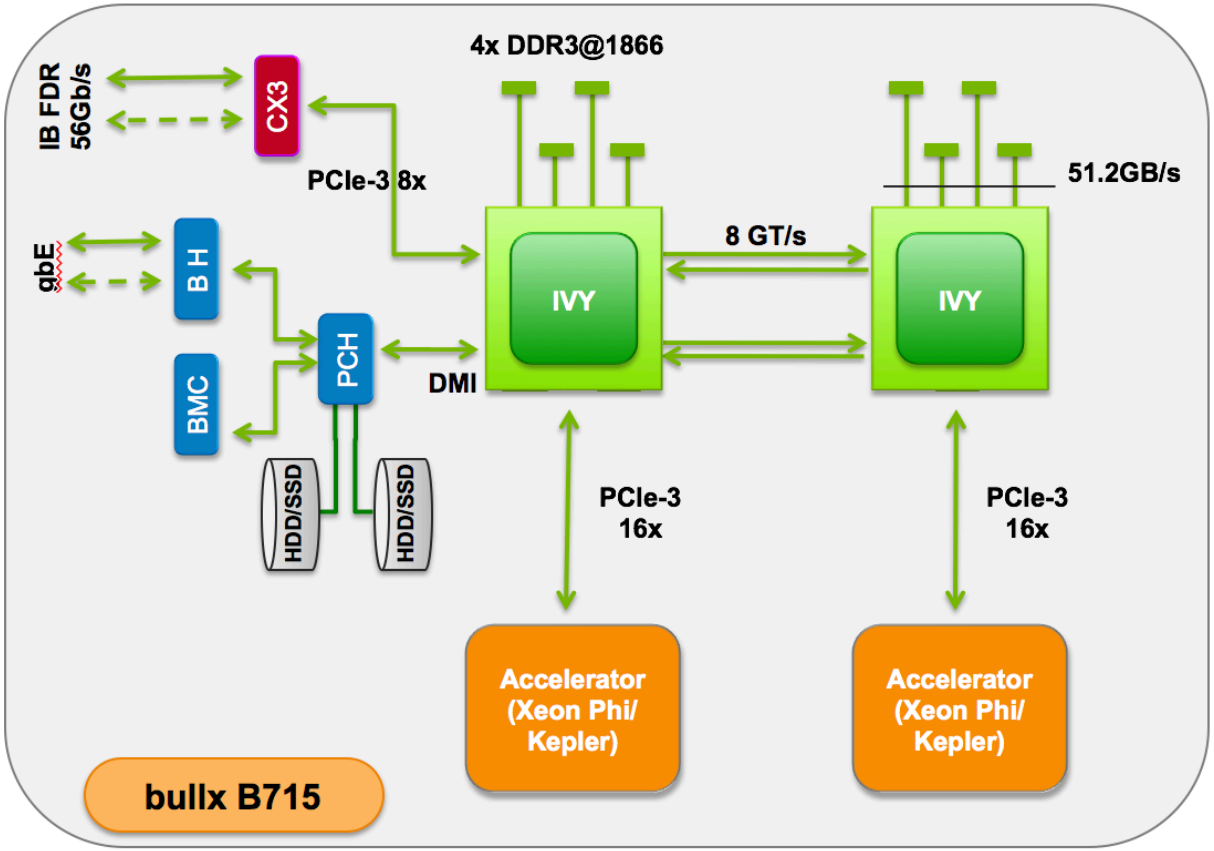
B515 – Accelerators nodes

nVidia nodes



B515 – block diagram

Block diagram



B515 – Accelerators nodes

nVidia nodes

- nVidia K40 specification:
  - 2.880 cuda cores
  - 12GB GDDR5
  - 1,43 Tflops dual-precision
  - 288 Gbytes/sec Memory bandwidth ECC off

B515 – Accelerators nodes

Intel Phi KNC nodes

- The cluster has 1x and a half racks of B515 nodes with 2x Intel Phi 7120P boards for a total of **54** nodes and **108** Intel Phi 7120P

- Each compute node has the following configuration
  - 2x Intel Xeon E5-2695v2 (12c, 2.4 Ghz)
  - 64 GB DDR3 RAM (8x 8GB DIMM)
  - 1x 120GB SSD disk
  - 1x Infiniband FDR ConnectX-3
  - 1x GbE
  - 2x Intel Phi 7120P 16GB GDDR5
- Each compute node provides 2876.8 Gflops
- Each chassis has 9 compute nodes
- Each rack has 36 compute nodes

B515 – Accelerators nodes

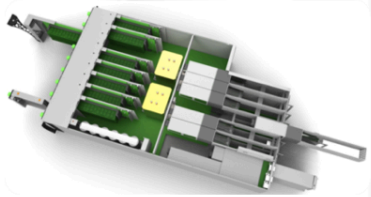
Intel Phi nodes

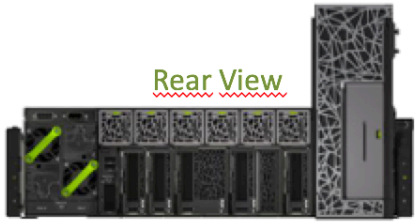
- Intel Phi 7120P specification:
  - 61 cores
  - 16 GB GDDR5
  - 1,208 Tflops dual-precision
  - 352 Gbytes/sec Memory bandwidth

S6130 – Fat node

BCS2 technology

bullx S



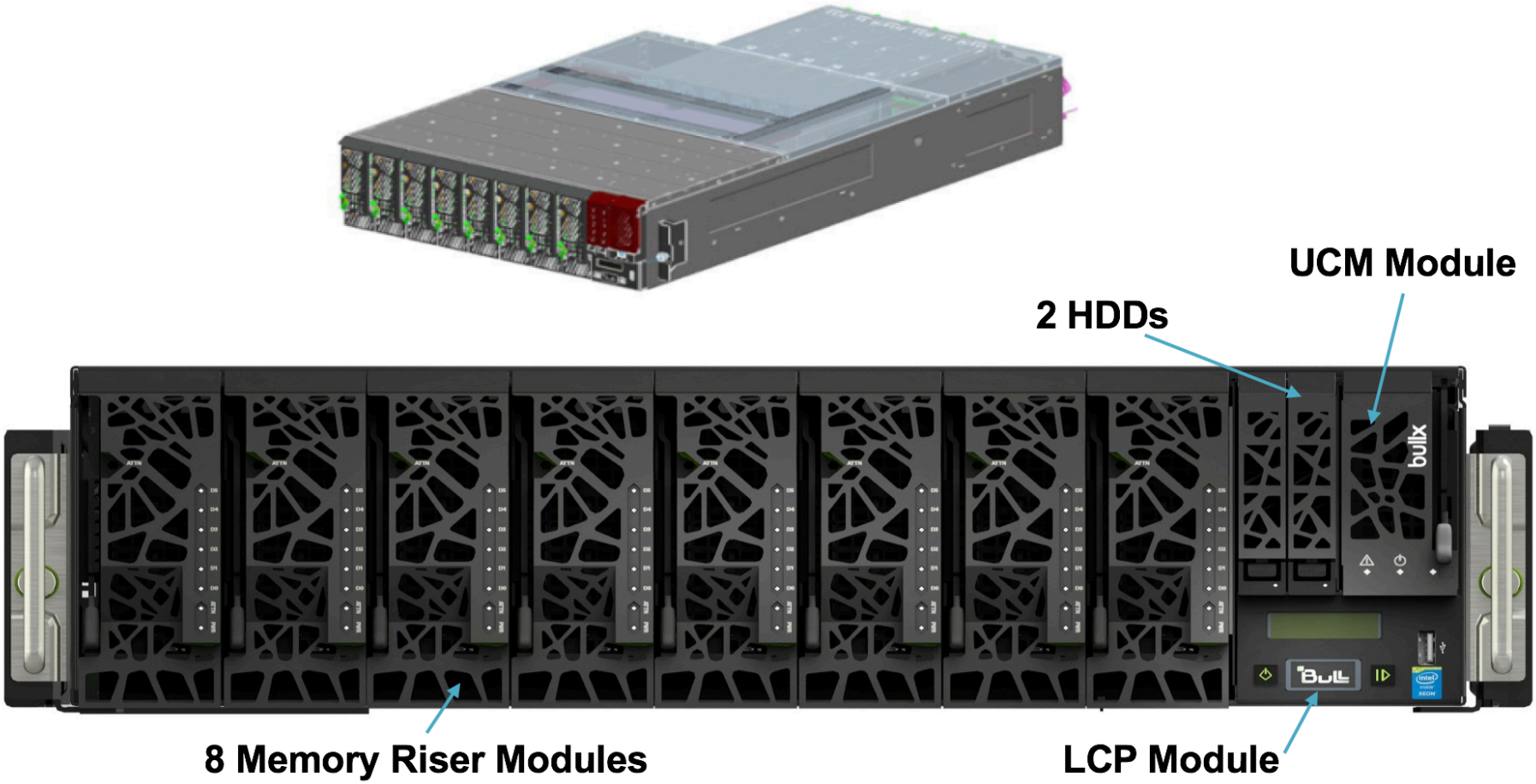


2 Haswell EX CPUs	4 Haswell EX CPUs	8 Haswell EX CPUs	16 Haswell EX CPUs
3U Form Factor	6U Form Factor	12U Form Factor	24U Form Factor
Up to 36 cores	Up to 72 cores	Up to 144 cores	Up to 288 cores
Up to 3TB RAM	Up to 6TB RAM	Up to 12TB RAM	Up to 24TB RAM
7 PCI-e Gen3	14 PCI-e Gen3	28 PCI-e Gen3	56 PCI-e Gen3
Active/Passive PS	Active/Passive PS	Active/Passive PS	Active/Passive PS
	HW Partitioning	HW Partitioning	HW Partitioning

Bull atos technologies

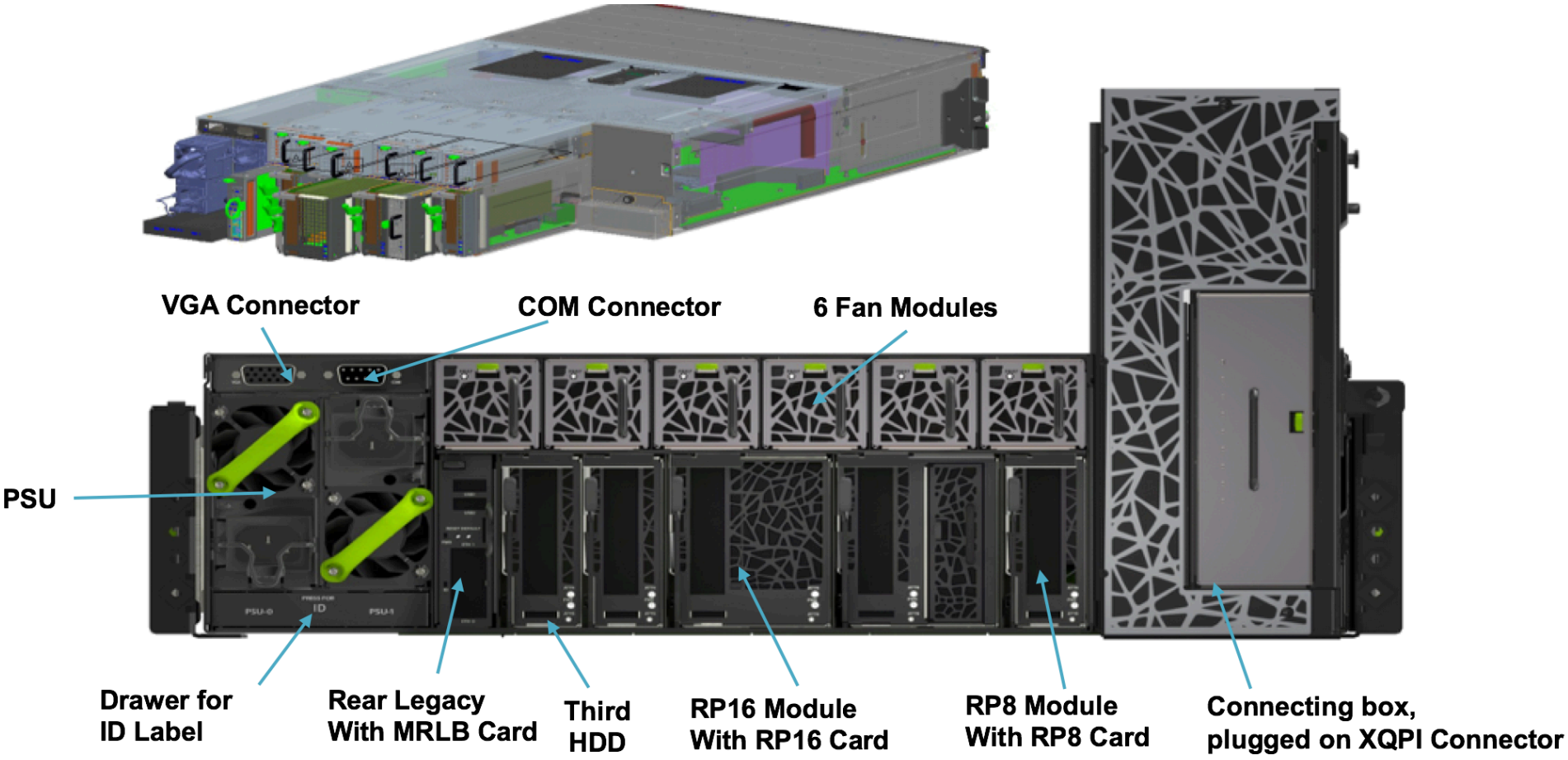
S6130 – Fat node

Front view



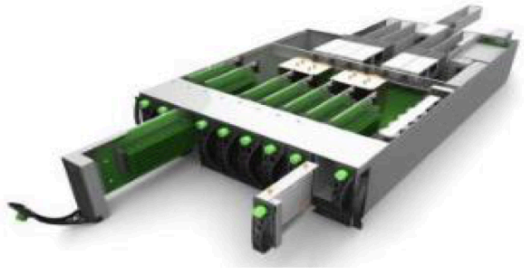
S6130 – Fat node

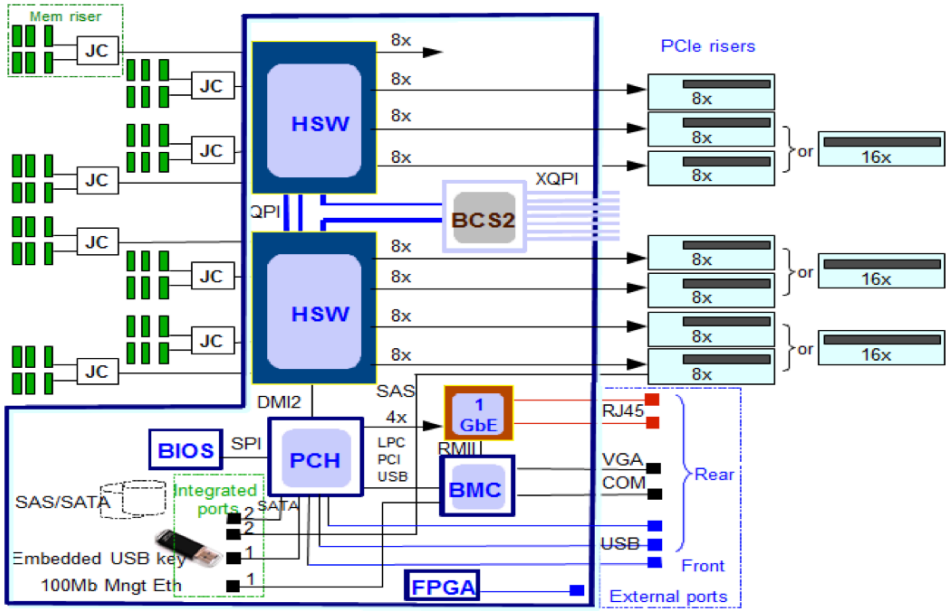
Rear view



S6130 – Fat node

Rear view



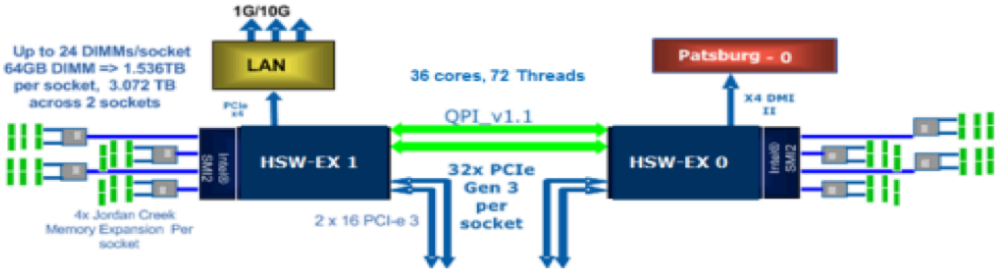


- 3U, rack-able within a standard 19” rack
- Motherboard
  - 2-socket module or node (2S)
    - CPU: Haswell-EX and then Broadwell-EX
  - Maximum memory slots: 24 DIMMs per socket DDR3 and DDR4
  - Maximum use of available PCIe lanes offering low profile slots:
    - seven (7) 8x PCIe Gen3 slots
    - or three (3) 16x PCIe Gen3 slots and one (1) 8x PCIe Gen3 slot
    - or intermediate configurations where 2 (two) 8x slots are replaced by a 16x slot
- BCS2, PCH, BMC, Gigabit Ethernet controller
- Hot-plug PSU: 2N for 2 power rails (2 x 1400W total)
- Hot-plug Fans: 5+1
- Hot-plug Disks: 1+1 (Raid1)
- LCP
- Hot-plug UCM
- Hot-plug Memory riser
- Hot-plug PCIe riser

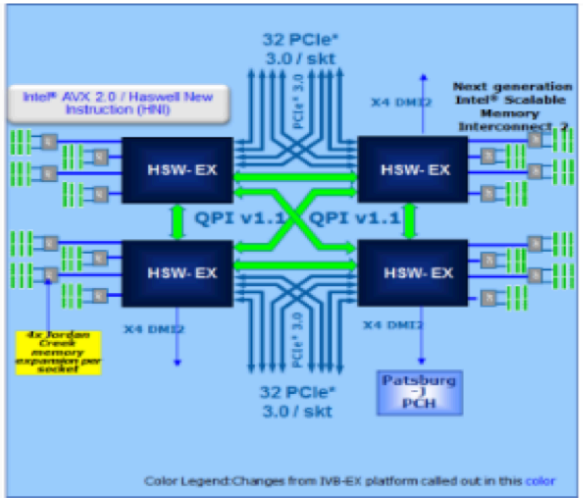
S6130 – Fat node

Glueless server used by competitors

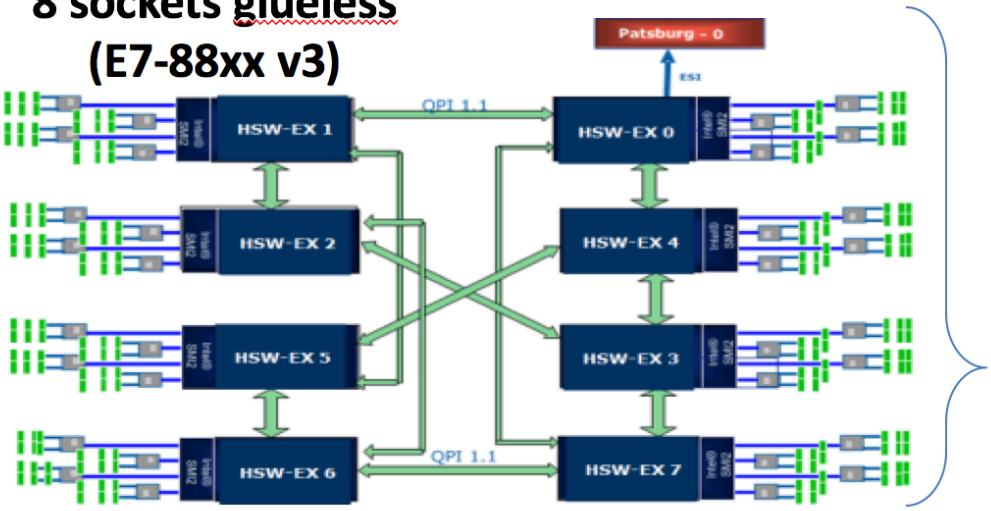
2 sockets (E7-28xx v3)

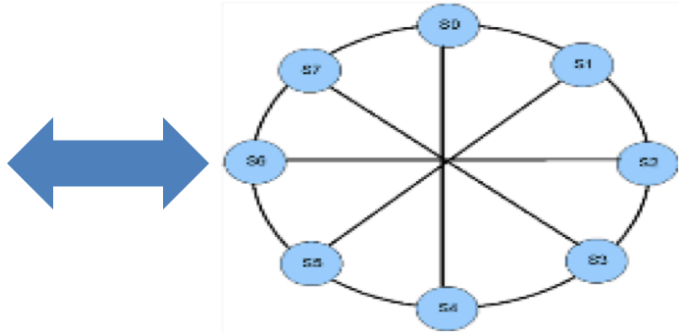


4 sockets (E7-48xx v3)



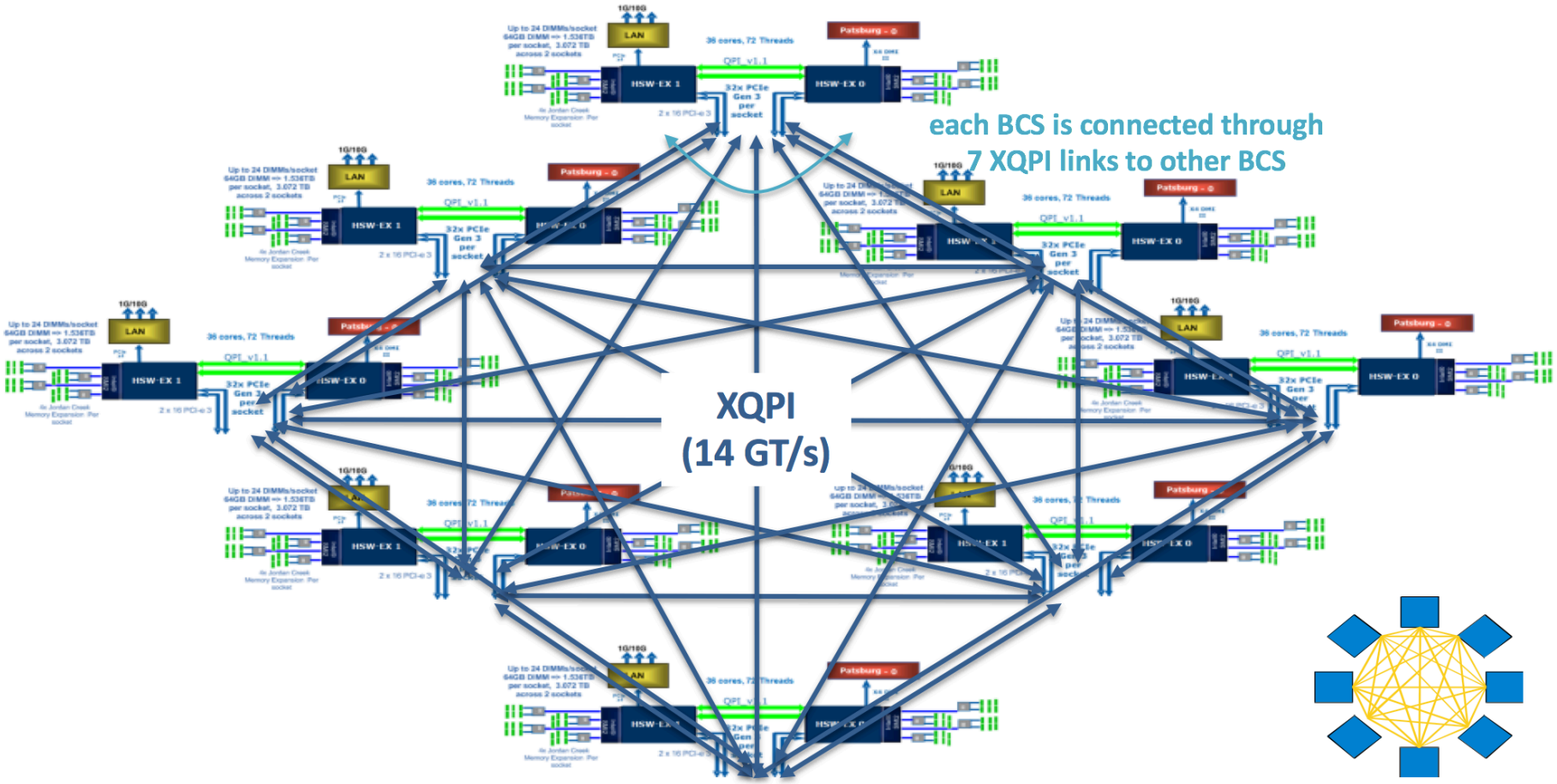
8 sockets glueless (E7-88xx v3)





S6130 – Fat node

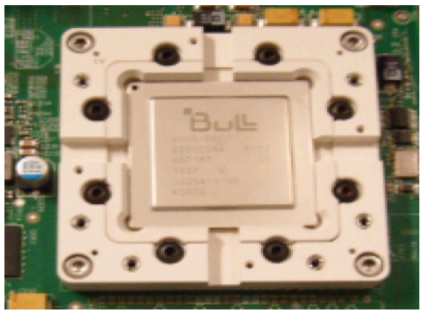
Glue technology used by Bull



S6130 – Fat node

XQPI links

- BCS2 provides 7 XQPI links to connect up to 7 others modules in order to build a 16-socket system (maximum)



- Bandwidth:
  - 1 XQPI link: 14 GT/s each direction
  - 1 Transfer = 2 bytes => 14 GT/s = 224 Gb/s
  - Transfer rate between 2 modules:
    - 4 sockets (4 XQPI links): equivalent to ~88x 10 GigE ports
    - 8 sockets (2 XQPI links): equivalent to ~44x 10 GigE ports
    - 16 sockets (1 XQPI link): equivalent to ~22x 10 GigE ports

S6130 – Fat node

XQPI links

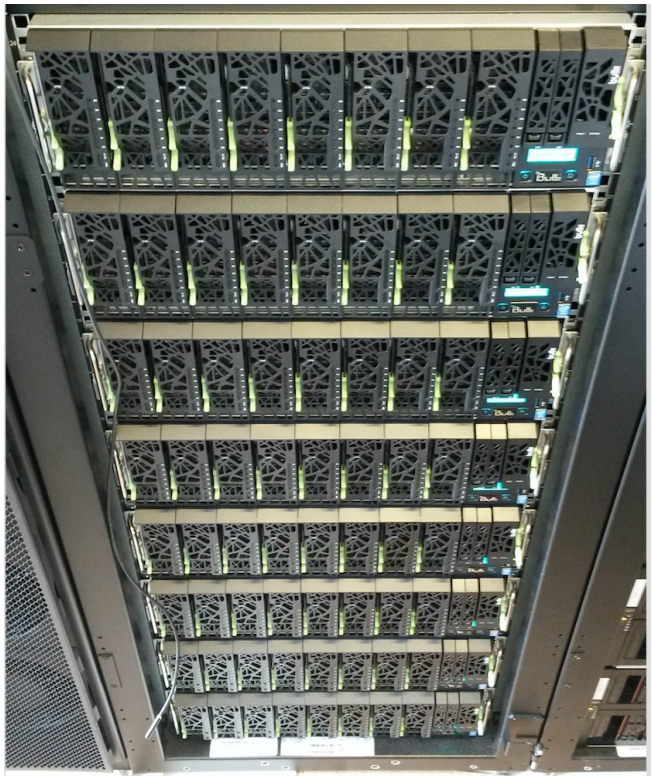
	Ivy-Bridge-EX	Haswell-EX	Broadwell-EX
Availability	2015	H1-2016	H2-2016
Max cores/socket	15 cores	18 cores	24 cores
Max cores/system	240 cores	288 cores	384 cores
Memory type	DDR3	DDR3 & DDR4	DDR3 & DDR4
Memory slots	384	384	384
Max memory capacity	24 TB	24 TB	24 TB (48 ? TB)
Max Memory BW / socket	78 GB/s	91 GB/s	91 GB/s
Max Memory BW total	1248 GB/s	1455 GB/s	1455 GB/s
SPECint_rate2006	9120	11600	
SPECfp_rate2006	6520		

Bull
atos technologies

S6130 – Fat node

LNCC configuration

- The cluster has 1x bullx S6130 with the following configuration:
  - 16x Intel Ivy-Bridge E7-2870v2 15c 2.3Ghz
  - 6TB DDR3 RAM
  - 1x 120GB SSD disk
  - 1x Infiniband FDR ConnectX-3
  - 1x GbE
- **4.4 Tflops R_{peak}**



S6130 – Fat node

Rear view

Service nodes

Administration, MWS & Login

- Service nodes includes:
  - Administration nodes
    - Administration tasks
    - Accessed by sysadmins
  - Managed WorkStation (MWS) nodes
    - Monitoring and remote management
    - Accessed by sysadmins
  - Login nodes
    - User access, home for compilation and job management
    - Accessed by users and sysadmins

Service nodes

Administration nodes

- The Administration nodes have all the administration tools for the sysadmin manage the cluster
- The server are based on the bullx R423-E3 products (2U chassis node)
- There are 2x administration nodes configured in High Availability with the following configuration:
  - 2x E5-2695v2 12c, 2.4GHz
  - 128 GB DDR3@1600 RAM
  - 2x 500GB 7.2 krpm SATA2 RAID1
  - 1x IB FDR network port
  - 1x GbE network port
  - 1x Ethernet BMC network port
  - 2x 10GbE network ports
- NetApp CDE2680-24
  - 12x 300GB disks
  - 2x 300 GB Global Hot-Spare disk

Service nodes

MWS nodes

- The MWS nodes are configured in a High-Availability cluster
- Their resource are VMs to monitor and remote access the compute nodes
- The server are based on the bullx R423-E3 products (2U chassis node)
- Hereafter the configuration of each node:
  - 2x E5-2695v2 12c, 2.4GHz
  - 64 GB DDR3@1600 RAM
  - 2x 500GB 7.2 krpm SATA2 RAID1
  - 1x GbE network port
  - 1x IB FDR network port
  - 1x Ethernet BMC network port

Service nodes

Login nodes

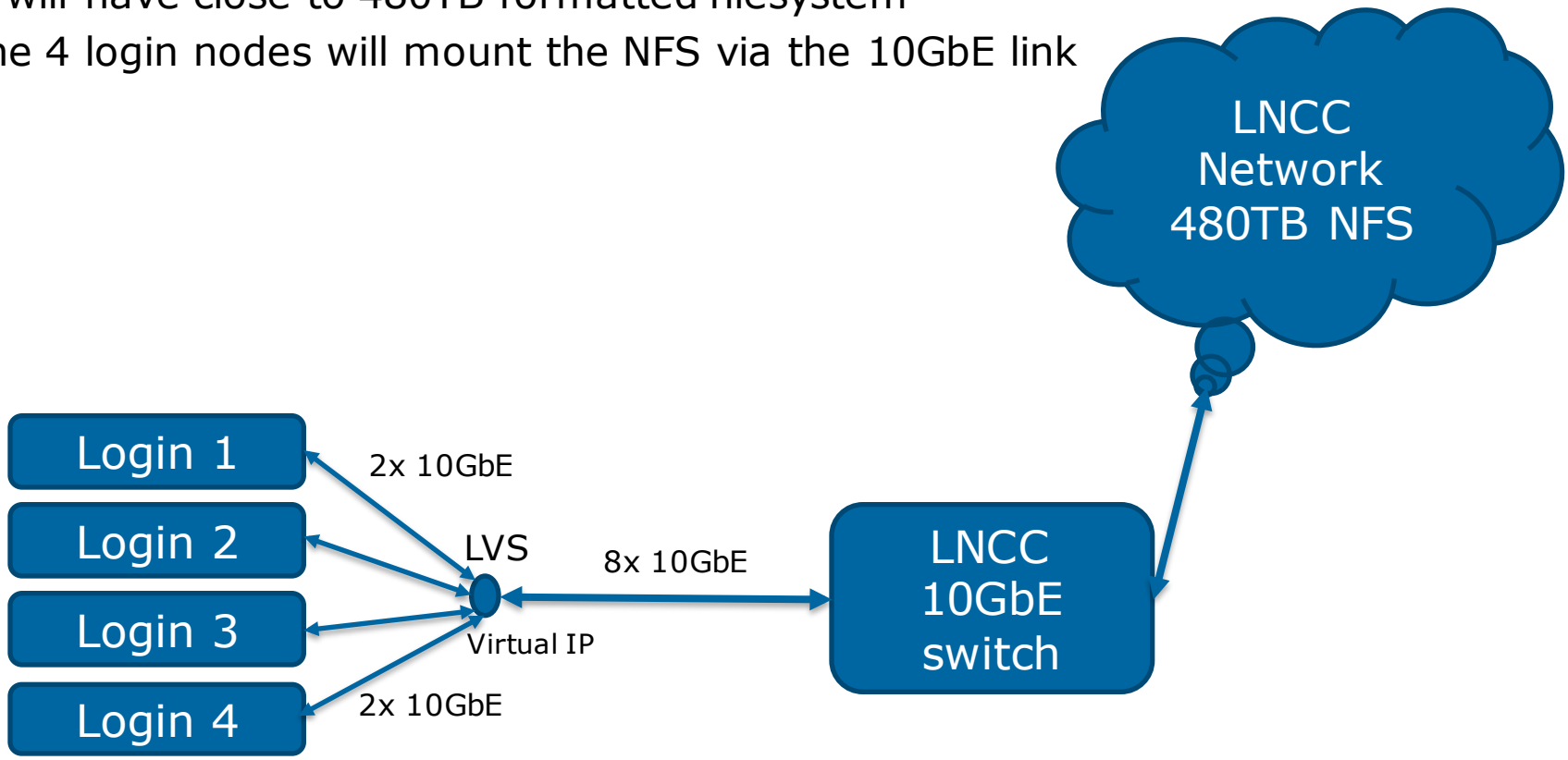
- The are configured 4 login nodes
- There is configured an Linux Virtual Server (LVS) on the 4 nodes
- The server are based on the bullx R423-E3 products (2U chassis node)
- The configuration of each node is the following:
  - 2x E5-2695v2 12c, 2.4GHz
  - 128 GB DDR3@1866RAM
  - 2x 500GB 7.2 krpm SATA2 RAID1
  - 1x GbE network port
  - 1x IB FDR network port
  - 1x Ethernet BMC network port
  - 2x 10GbE network ports

- The 2x 10GbE Ethernet links are configured in aggregated bonding

Service nodes

Login nodes - home

- The home of the user will be mounted on a LNCC NFS storage
- It will have close to 480TB formatted filesystem
- The 4 login nodes will mount the NFS via the 10GbE link

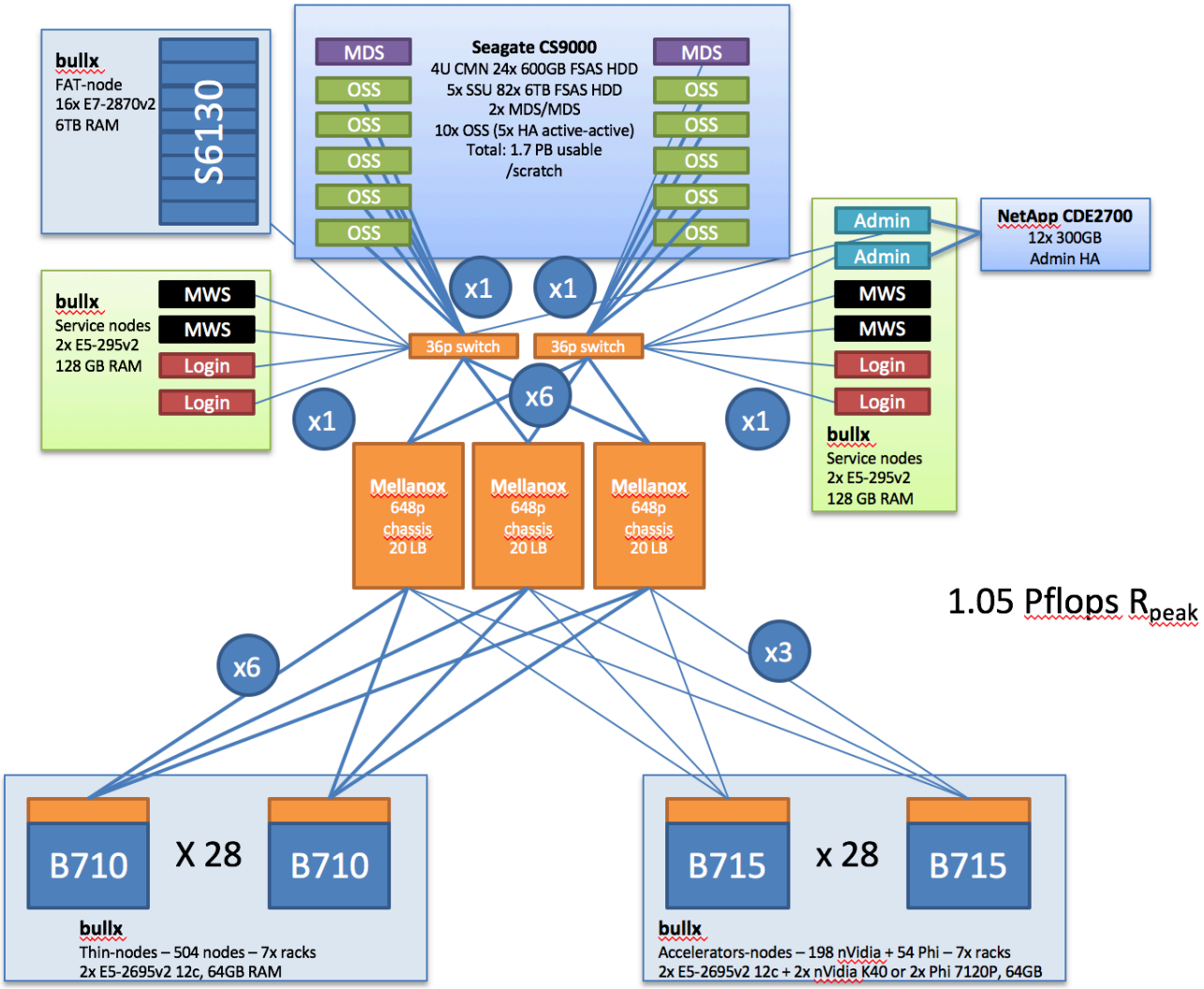


Infiniband network

Application & storage network

- The cluster has a high bandwidth-low latency network implemented with Infiniband FDR technology
- The topology is a fat-tree full-nonblocking
- The network is shared for the application and to access the central storage
- Each compute node has 1x FDR interface for application and storage
- Each administration node has 1x FDR interface for administration
- Each login node has 1x FDR interface to access the storage
- Each B700 chassis has an embedded Infiniband network with 36 ports (18 for local compute nodes and 18 to connect on the IB switches)
- There are 3x Mellanox up to 648port switches with 20 Leaf Boards on each

Infiniband network



GbE network

Administration network

- The cluster has a GbE network used for administration
- Each node has a BMC to manage the node (start/stop, access to console, virtual media, etc.)
- The GbE network is only used for administration purpose like:
  - Power on nodes
  - Access to nodes
  - Deploy nodes
  - Check status
  - SLURM communications
  - etc.

10GbE network

Login access network

- The 4x login nodes and the 2x administration nodes have a dual port 10GbE interface on each
- These 12x links are connected to the LNCC 10GbE switches
- The aim is to provide a high-bandwidth access to the login nodes from the LNCC network

Lustre – Distributed File System

Login access network

- The Santos Dumont has a 1.7 PB formatted Lustre Distributed File System
- The solution is based on the Seagate ClusterStor 9000
- It is composed by:
  - Cluster Management Unit
    - 2U 4x nodes (MDS/MGS/Management nodes)
  - 5 Scalable Storage Unit (SSU)
    - Each has 2x HA configured OSS nodes
    - 82x 3.5” 6TB disks FSAS
    - 2x SSD 100GB
    - ClusterStor GridRAID

Lustre – Distributed File System

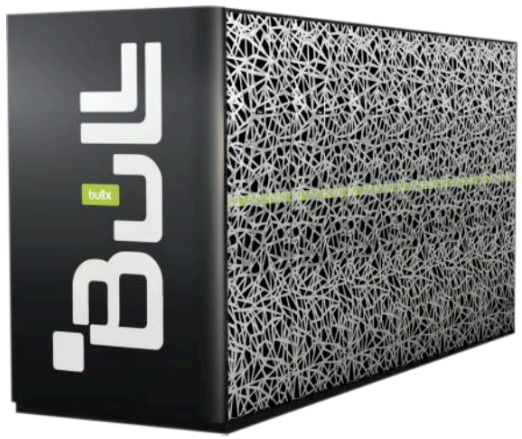
Seagate solution

- Each SSU has 2x OSS configured in HA

Software

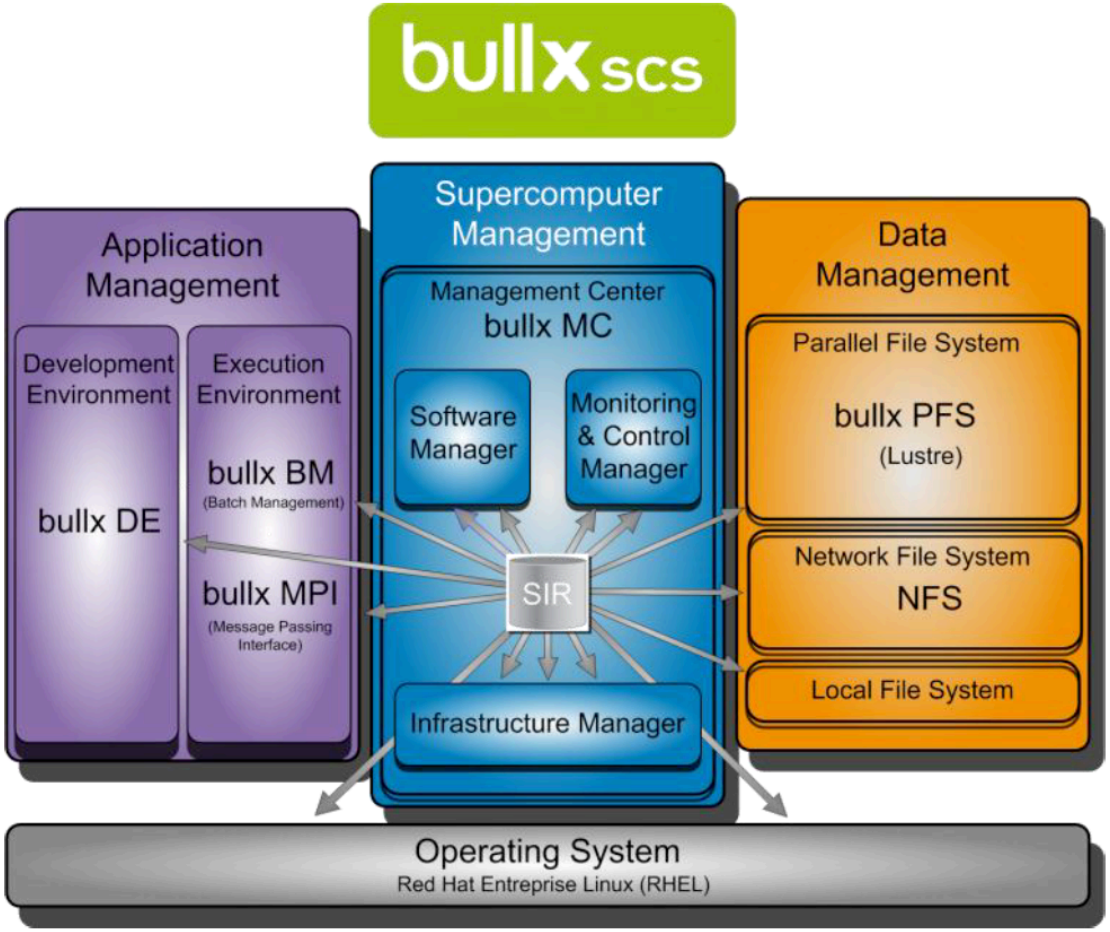
bullx Supercomputing Cluster Suite Advanced Edition 4 update 4

- The cluster has the bullx Supercomputing Cluster Suite Advanced Edition 4 update 4 (SCS AE 4u4) software stack
- bullx supercomputer suite is a comprehensive, powerful and robust software solution that meets the requirements of even the most challenging high performance computing needs
- It is the result of Bull's long experience in deploying HPC software to the strictest specifications and major investments and continued efforts in Research & Development



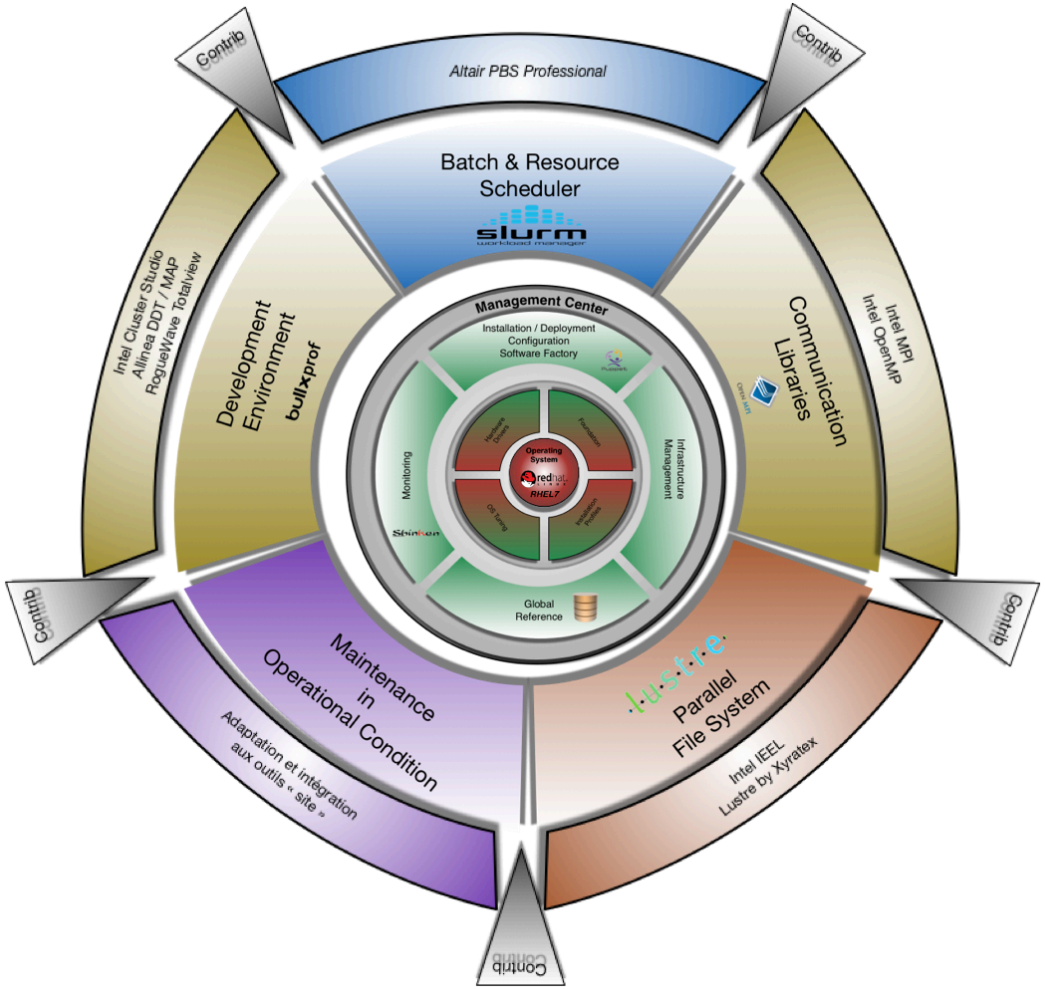
Software

bullx Supercomputing Cluster Suite Advanced Edition 4 update 4



Software

bullx Supercomputing Cluster Suite Advanced Edition 4 update 4



Software

bullx Supercomputing Cluster Suite Advanced Edition 4 update 4

- Languages: Fortran, C, C++, Python, CUDA
  - Fortran: Intel Fortran ifort (IFORT) 15.0.2 20150121, GNU Fortran (GCC) 4.4.7 20120313 (Red Hat 4.4.7-11)
  - C: Intel C icc (ICC) 15.0.2 20150121, GNU gcc (GCC) 4.4.7 20120313 (Red Hat 4.4.7-11).
  - C++: Intel C++ icpc (ICC) 15.0.2 20150121, GNU g++ (GCC) 4.4.7 20120313 (Red Hat 4.4.7-11)
  - Python: python-2.6.6-52
  - CUDA: Cuda compilation tools, release 6.0, V6.0.1
- MPI: Bullx MPI, Intel MPI
  - Bullx MPI: v1.2.8.4
  - Intel MPI: v5.0.3.048
  - Open MPI: To be installed
- SLURM – Resoure Manager
  - Version 2.6

Santos Dumont summary

- 7x Rack DLC B710 Thin nodes (CPU)
  - 504 compute nodes
    - 2x Intel E5-2695v2 (12c, 2.4Ghz)
    - 64 GB DDR3 RAM
- 5,5x Rack DLC B715 Hybrid nodes (CPU + GPU)
  - 198 compute nodes (396 GPU boards)
    - 2x Intel E5-2695v2 (12c, 2.4Ghz)
    - 64 GB DDR3 RAM
    - 2x nVidia K40
- 1,5x Rack DLC B715 Hybrid nodes (CPU + Phi)
  - 54 compute nodes (108 Phi Boards)
    - 2x Intel E5-2695v2 (12c, 2.4Ghz)
    - 64 GB DDR3 RAM
    - 2x Phi 7120

Santos Dumont summary

- 1x Rack FAT-node + Ethernet Network
  - 1 compute node
    - 16x Intel Ivy (15c, 2.3Ghz)
    - 6 TB RAM
  - 5x Cisco switches
- 3x Rack Infiniband Network
  - 3x Mellanox 648 ports
  - 2x Administration nodes
  - 4x Login nodes
  - 4x MWS nodes
- 2x Rack Seagate Storage Lustre
  - 1.7 PB usable space
  - Seagate v1.5 ClusterStor

Santos Dumont summary

- IB FDR full-nonblocking fat-tree
- Lustre v2.1 for /scratch
- Software stack:
  - RedHat Linux 6.4
  - Supercomputing Cluster suite AE 4 u4
  - Intel Compilers
- Cooling system:
  - DLC Direct Liquid Cooling. PUE <= 1.1
  - InRow Cold water. PUE <=1.4

Santos Dumont benchmarks

HPL Linpack

- Top 500 June 2015
  - **456,8 Tflops** Hybrid Nvidia K40, #146 June 2015

146	Laboratório Nacional de Computação Científica Brazil	**Santos Dumont GPU** - Bullx B710, Intel Xeon E5-2695v2 12C 2.4GHz, Infiniband FDR, Nvidia K40 Bull, Atos Group	10,692	456.8	657.5

  - **363,2 Tflops** Hybrid Intel Xeon Phi, #178 June 2015

178	Laboratório Nacional de Computação Científica Brazil	**Santos Dumont Hybrid** - Bullx B710, Intel Xeon E5-2695v2 12C 2.4GHz, Infiniband FDR, Intel Xeon Phi 7120P Bull, Atos Group	24,732	363.2	478.8

  - **321,2 Tflops** Intel Xeon, #208 June 2015

208	Laboratório Nacional de Computação Científica Brazil	**Santos Dumont CPU** - Bullx B71x, Intel Xeon E5-2695v2 12C 2.4GHz, Infiniband FDR Bull, Atos Group	18,144	321.2	348.4

Questions?

Thanks

Gerardo Ares
gerardo.ares@atos.net

10-10-2016

Bull
atos technologies