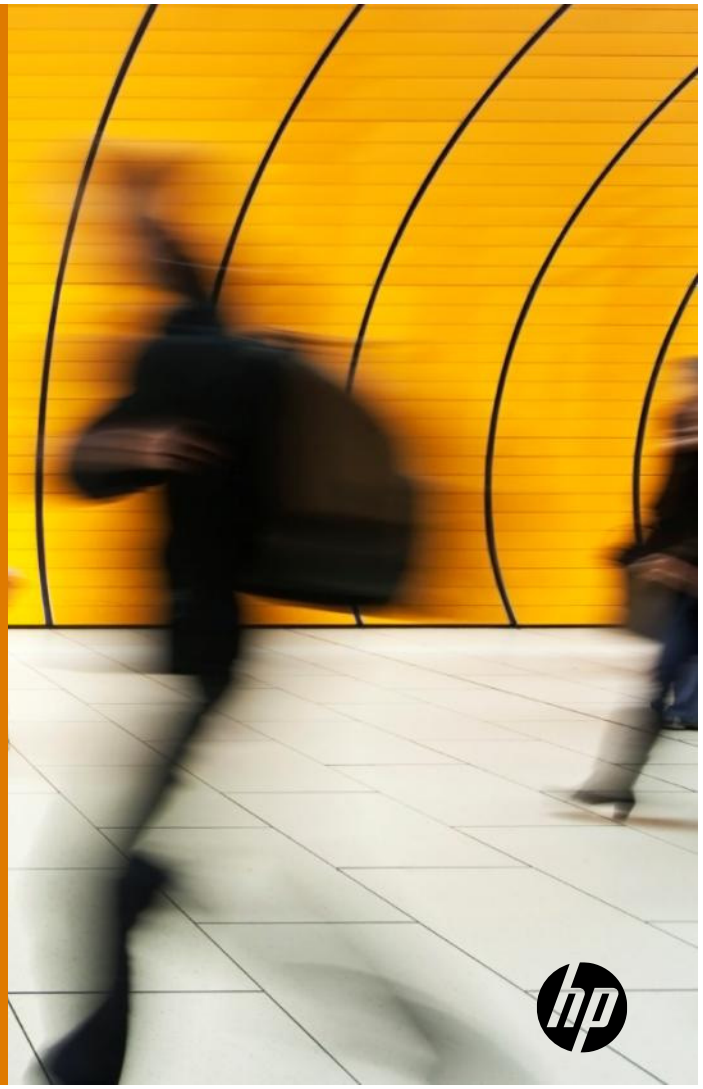
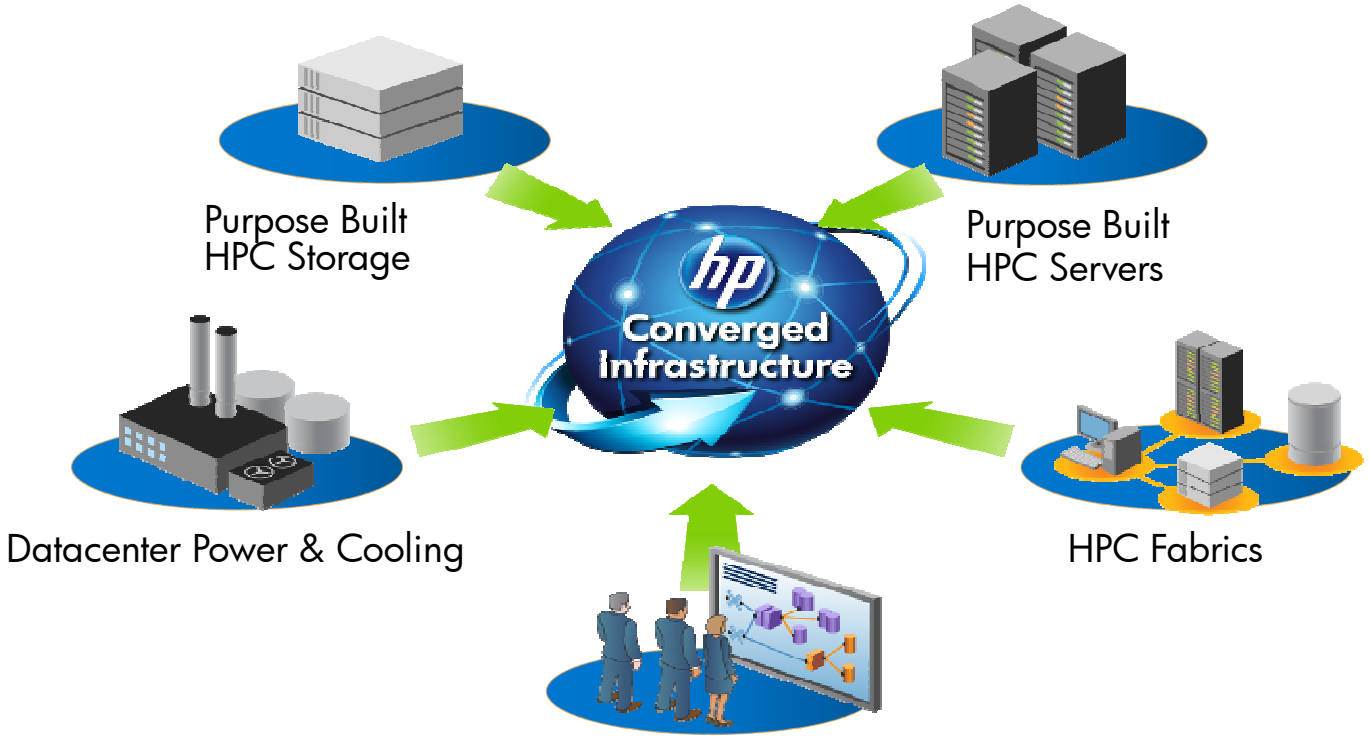
TAITO CLUSTER

Cheran Sorin Cristian
EMEA HPC Competency Center

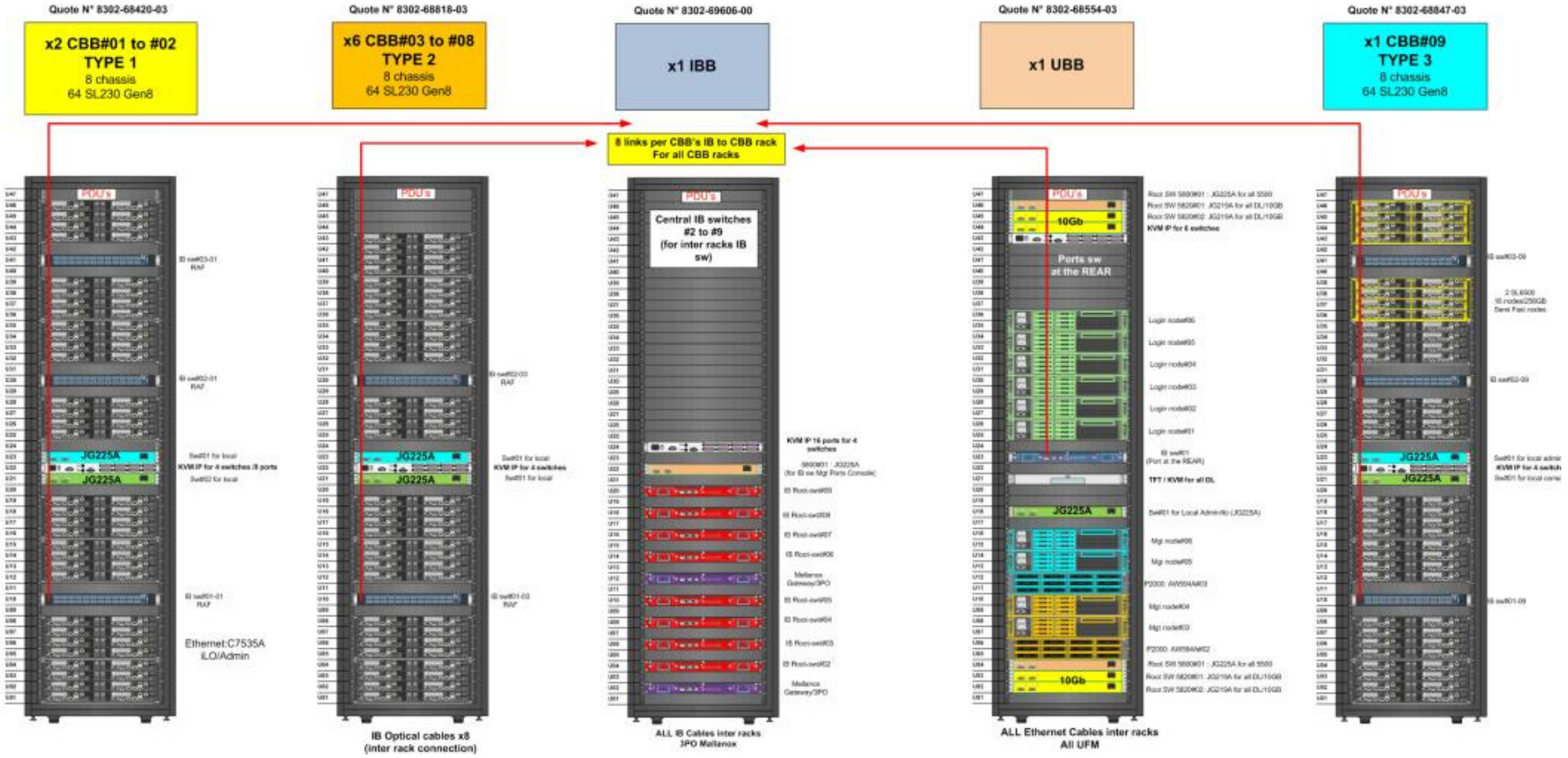


HPC typical setup



Taito Cluster

Taito Cluster



Taito Cluster

- 560 Compute nodes SL 230 Gen8
  - 2x 8 cores E5-2670 at 2.6 GHz
  - 64GB
- 16 Compute nodes SL230 Gen8 Fat nodes (256GB)
- 4 Service nodes DL 380 Gen8
- 6 Login nodes Dl 380 Gen8
- Infiniband Mellanox 4XFDR
- Cluster Management Utlity
- Slurm
- Unified Fabric Manager
- Intel Cluster Suite XE

SERVERS IN TAITO

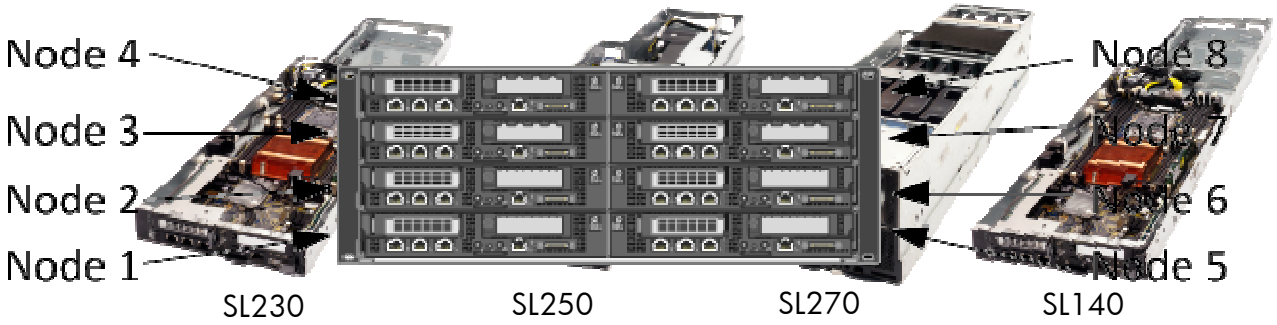
3 types of nodes

Login nodes – 6 DL 380 Gen8

Service nodes – 4 DL 380 Gen8

Compute nodes –
560 SL 230 Gen8
16 SL230 Gen 8 Fat nodes

SL6500 Chassis



- Shared power & fans for reduced component quantity and increased power efficiency
- Ability to mix and match SL half-width nodes
- Front cabling for increased rear air-flow and ease of serviceability
- Individually serviceable nodes

- **SL230** –Socket-R, ultra-dense server for virtualization and HPC applications (1U)
- **SL250** –Socket-R, hybrid-compute node for GPU computing and data base applications in HPC (2U)
- **SL270** –Socket-R, high-performance GPU solution, optimized for extreme GPU density (4U)

- **SL140** – Socket-B, cost-effective, power-efficient and ultra-dense solution (1U)

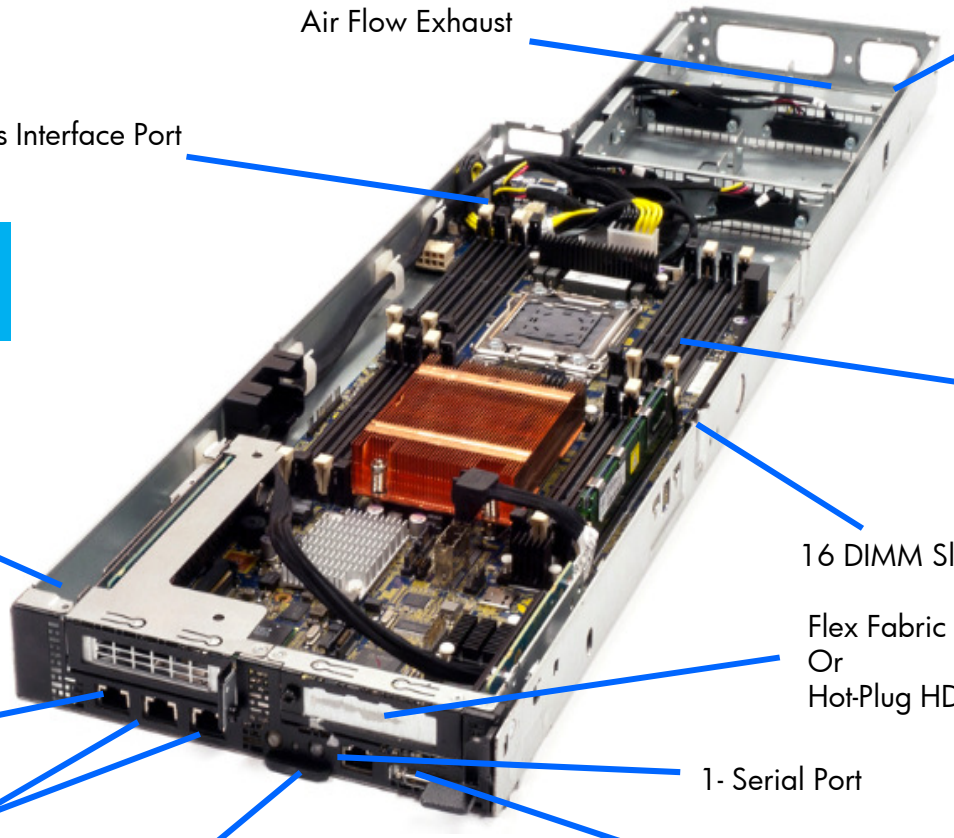
*Needs1200mm deep racks

	SL140s Gen8	SL230s Gen8	SL250s Gen8	SL270s Gen8
Processor	E5-2400 - 4/6/8 Cores	E5-2600 - 4/6/8 Cores		
Chipset	Intel® C600			
Memory	12xDR3, RDIMM/UDIMM, up to 1333MHz –ECC	16xDDR3, RDIMM/UDIMM up to 1600MHz-ECC		
Max Memory	256GB	512GB		
Internal Storage	2 LFF NHP 4 SFF NHP Opt: 2 SFF HP	2 LFF NHP 4 SFF NHP Opt: 2 SFF HP	4 SFF HP 2 LFF NHP	8 SFF HP
Max Internal Storage	4TB 3.5" SAS; 1.2TB 2.5" SAS; 6TB SATA; 480GB 2.5" SSD	4TB 3.5" SAS; 1.2TB 2.5" SAS; 6TB 3.5" SATA; 2TB 2.5" SATA; 480GB 2.5" SSD	2TB 2.5" hot plug SAS; 1.2TB 2.5" non-hot plug SAS; 2TB 2.5" hot plug SATA; 2TB 2.5" SATA; 480GB 2.5" SSD	4TB SAS; 4TB SATA; 960GB SSD
Networking	1x Integrated NC366i Dual Port Gigabit Server Adapter	1x Integrated NC366i Dual Port Gbe 1xDual Port networking daughter card: QDR IB, 10GbE		
I/O Slots	1xPCIe Gen3: 1x16 HL/LP	1xPCIe Gen3: 1x16 HL/LP	4xPCIe Gen3: 1x8 HL/LP; 3x16 HL/LP	9xPCIe Gen3: 1x8 HL/LP; 8x16 HL/LP
Integrated Management	HP iLO Mgt Engine, SIM, IRS		Opt: HP Insight Control, iLO Adv	
Form Factor	1U HW - 8 trays per s6500 (4U)	1U HW – 8 trays per s6500 (4U)	2U HW – 4 trays per s6500 (4U)	4U HW – 2 trays per s6500 (4U)

HP ProLiant SL230s Gen8 1U Half Width Tray

Item	SL230 Gen8
Processor	**Two Intel® Xeon® E5-2600 Series** **4/6/8 Cores,**
Chipset	**Intel® Sandy Bridge EP Socket-R**
Memory	(512 GB) 16 sockets DDR3 up to 1600MHz ECC
Max Memory	512 GB
Internal Storage	Two LFF non-hot plug SAS, SATA bays or Four SFF non-hot plug SAS, SATA, SSD bays Two Hot Plug SFF Drives (Option)
Max Internal Storage	**8TB**
Networking	Dual port 1GbE NIC/ Single 10G Nic
I/O Slots	One PCIe Gen3 x16 LP slot 1Gb and 10Gb Ethernet, IB, and FlexF abric options
Ports	Front: (1) Management, (2) 1GbE, (1) Serial, (1) S.U.V port, (2) PCIe, and Internal Micro SD card & Active Health
Power Supplies	750, 1200W (92% or 94%), high power chassis
Integrated Management	iLO4 hardware-based power capping via SL Advanced Power Manager
Additional Features	Shared Power & Cooling and up to 8 nodes per 4U chassis, single GPU support, Fusion I/O support
Form Factor	16P/8GPUs/4U chassis

HP ProLiant SL230s Gen8 1U Half Width Tray



Air Flow Exhaust

Rear Kit Options
- 2 – 3.5 HDDS
- 4 – 2.5 HDDs
- 4 – SDD's
- 1- GPU

Chassis Interface Port

8 Nodes per 4U chassis
16 CPUs per 4U chassis
8 GPUs per 4U chassis

2 Socket-R CPUs

PCIe Expansion Slot
Or
Hot-Plug HDD
Or
Fusion I/O Accelerator

16 DIMM Slots

Flex Fabric Slot
Or
Hot-Plug HDD

Management
Port – iLO4

1- Serial Port

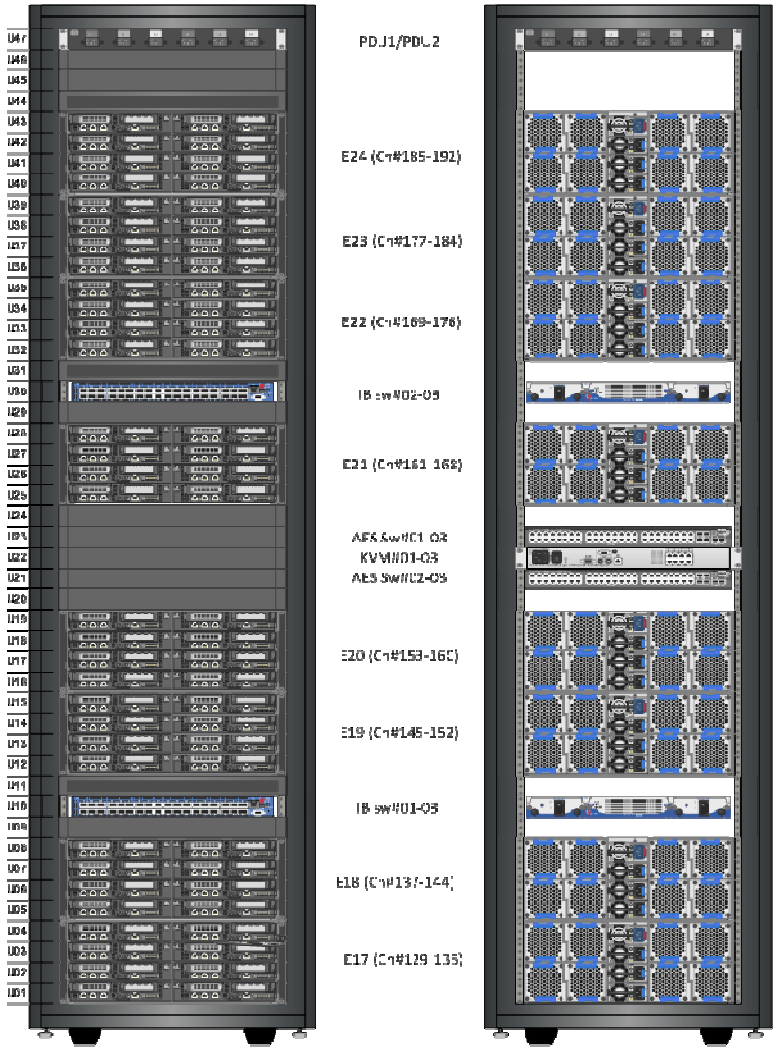
2- 1GbE Ports

SUV Port – Serial/USB/Video

SID Tab – System Identification

CBB Rack – 9 racks

- Compute Build Blocks contain:
- 8 SL6500 with 8 Sl230 Gen8 each
- 2 IB switches 36 Ports each
  - racks 2,5,9 have 3 IB switches
- 2 Ethernet Switches
- KVM
- 2 PDUs



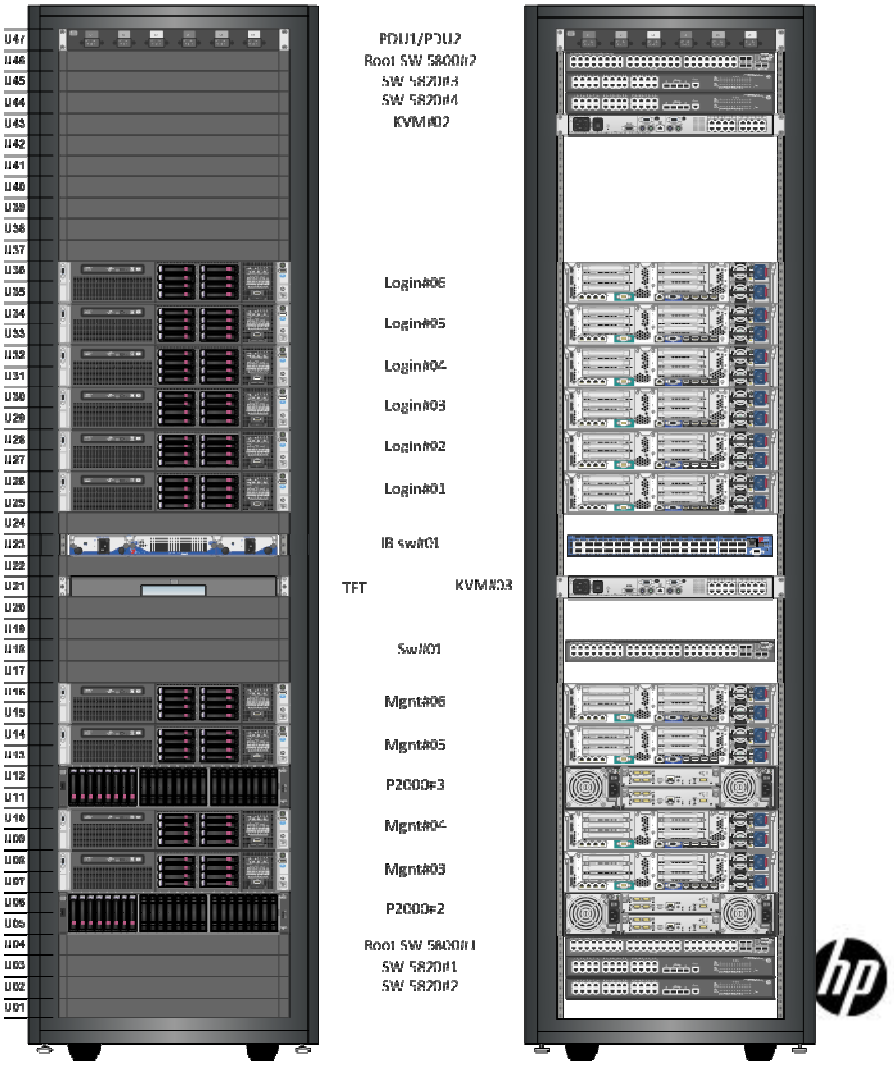
ProLiant DL380p Gen8

Key features and benefits

Key feature	Customer benefit
Intel® Xeon E5-2600 series with dual QPI	Up to 80% more performance (1)
4 Channels Per DIMM and 24 Memory DIMM slots	200% more memory capacity than G7 (768G) (2) Max Memory
Up to 16 SFF or 8 LFF hard drives	33% more internal storage capacity (3)
Flexible Network options (Flex-LOM)	Flexibility of choice with 4x1G or 2x10G Ethernet; or 2x10G Flex Fabric
Up to 6 PCIe Gen3 slots	200% the I/O capacities with PCIe- Gen3 (4)
HP Smart Storage Solution	Up to 200% more performance with HP Smart Drives, Smart Array (5)
iLO Management Engine	4th generation of iLO manageability
Active Health	Always on Diagnostics, 5x faster diagnose root cause (6)

Login and Service Nodes

- The Utility Building Block contains:
  - 4 service nodes
  - 6 login nodes
  - 2 Storage arrays
  - 1 IB switch
  - Ethernet Switches
  - KVM
  - TFT



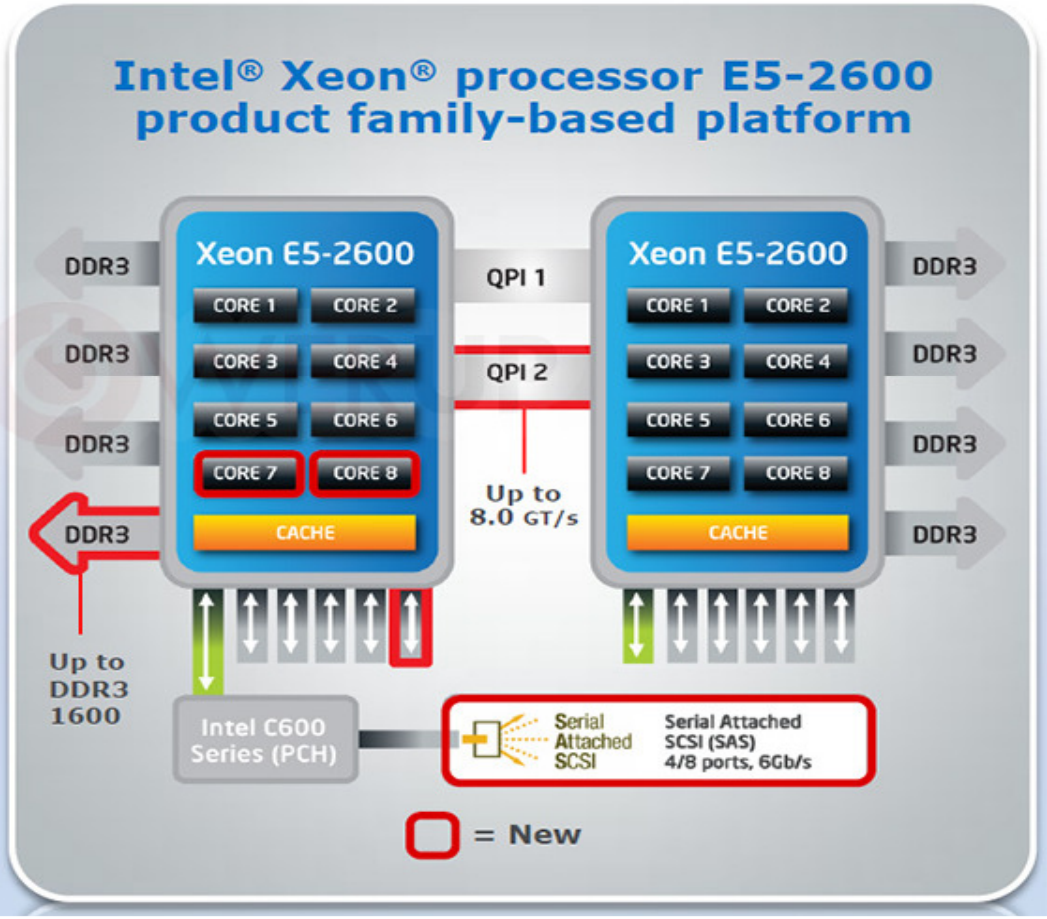
Intel® Westmere-EP vs. Intel® Sandy Bridge-EP-EN

Feature	Westmere-EP	E5-2600 (EP)	E5-2400 (EN)
Cores	Up to 6 cores / 12 threads	Up to 8 cores / 16 threads	
Cache Size	12 MB	Up to 20 MB	
Max Memory Channels per Socket	3	4	3
Max Memory Speed	1333 MHz	1600 MHz	
New Instructions	AES-NI	Adds AVX	
QPI frequency	6.4 GT/s	Up to 8.0 GT/s	
Inter-Socket QPI Links	1	2	1
PCI Express	• 36 Lanes PCIe2* on Chipset	40 Lanes/Socket Integrated PCIe3	24 Lanes/Socket Integrated PCIe3
Server/Workstation Power TDP	Server/Workstation: 130W, 95W, 80W, LV (Low Power)	150 (Workstation Only) 130, 115, 95, 80, 70, 60 (Low power)	95, 80, 70, 60, 50 (Low Power)

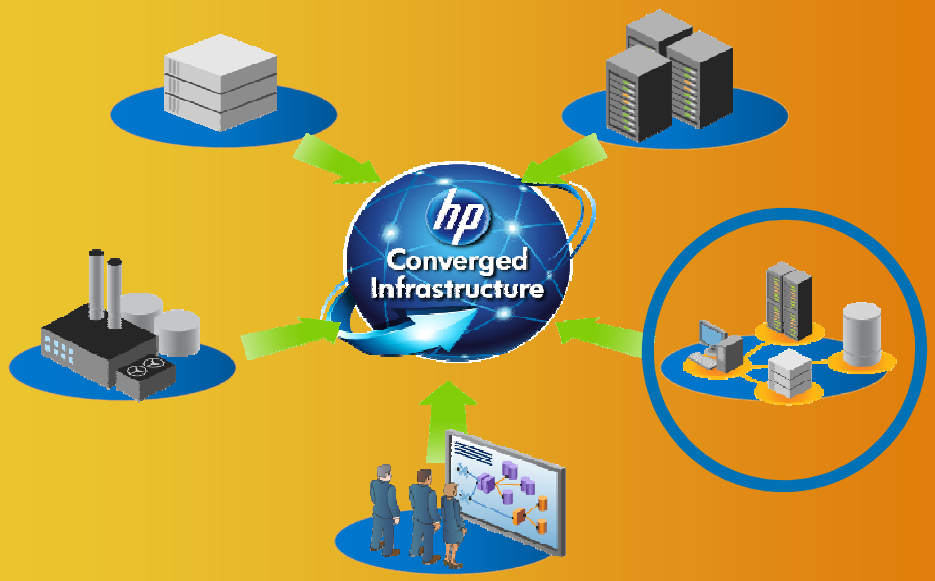
Intel® Xeon® Processor E5-2600 Product Family

More Capabilities for a Next-Gen Data Center

	Intel® Xeon® processor 5600 series-based platform	Intel® Xeon® processor E5 product family-based platform
Processor / Protocol	QPI 6.4 GT/s 6C/12 T/12MB cache Turbo 1.0	QPI 8.0GT/s, 2 QPI links 8C / 16 T/20MB cache Turbo 2.0 AVX
Memory	3 Channels Up to 1333Mhz Up to 18 DIMMs Up to 288GB	4 channels Up to 1600Mhz Up to 24 DIMMs Up to 768GB LRDIMMs
I/O	Two-chip IOH/ICH Support for up to 32 lanes of PCIe 2.0	Intel® Integrated I/O Support for up to 80 lanes of PCIe 3.0 DDIO
Power Management	NM 1.5	NM 2.0



INTERCONNECTS IN TAITO



IB in TAITO

- Link Speed = Link Width * Signal Rate
- The following bandwidth can be reached

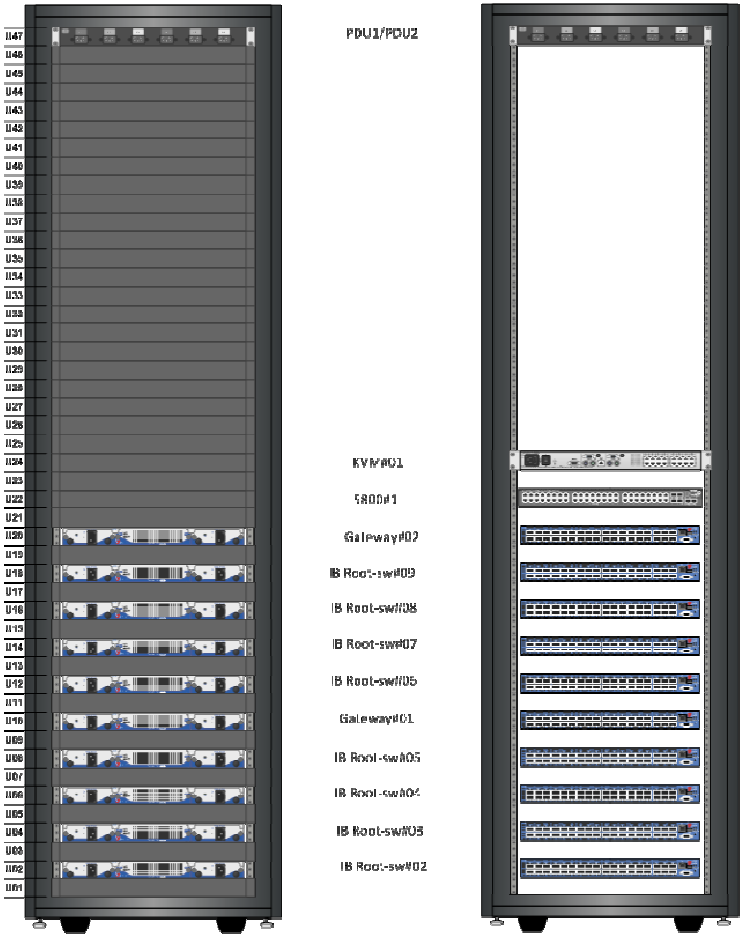
	Single (SDR)	Double (DDR)	Quad (QDR)	Fourteen (FDR)	Enhanced (EDR)
1X	2 Gbit/s	4 Gbit/s	8 Gbit/s	14 Gbit/s	25 Gbit/s
4X	8 Gbit/s	16 Gbit/s	32 Gbit/s	56 Gbit/s	100 Gbit/s
12X	24 Gbit/s	48 Gbit/s	96 Gbit/s	168 Gbit/s	300 Gbit/s

- The CURRENT Technology is 4xFDR.

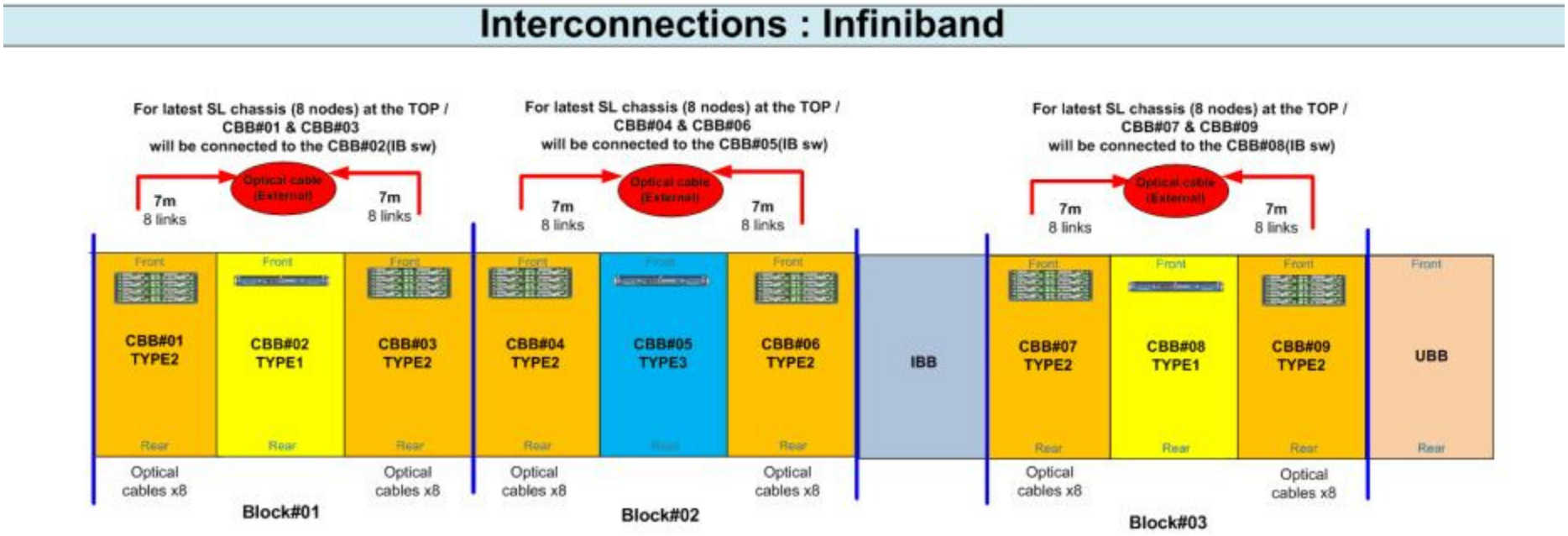
IB in TAITO

- Each CBB contains 2 Mellanox switches 36 ports with the exception of 2,5,9 (3 swtiches each) = 21 Mellanox 36 ports 4X FDR.

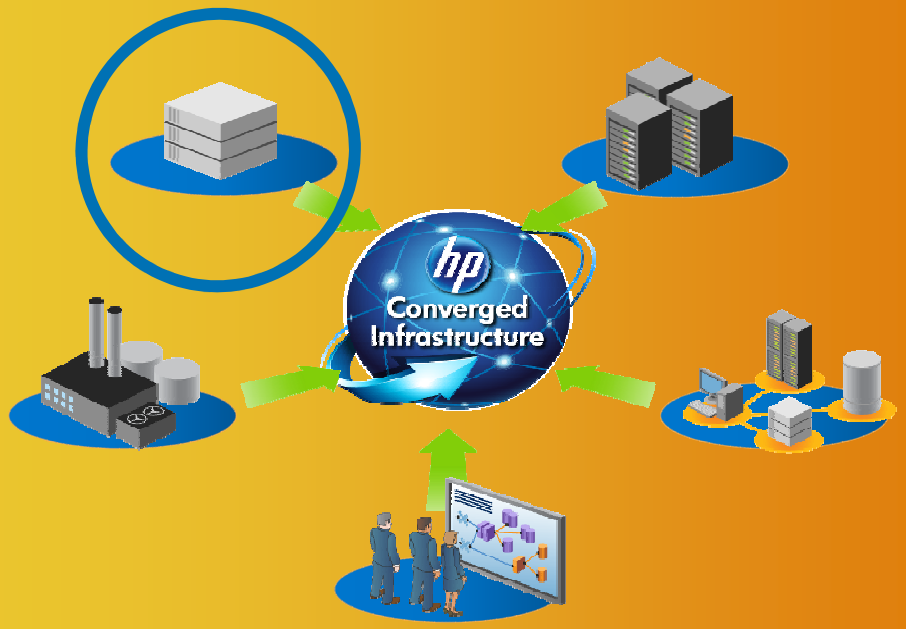
- Inifiniband Building Block contains:
  - 9 switches
  - 2 gateways.



IB in TAITO


Interconnections : Infiniband
For latest SL chassis (8 nodes) at the TOP / CBB#01 & CBB#03 will be connected to the CBB#02(IB sw)
For latest SL chassis (8 nodes) at the TOP / CBB#04 & CBB#06 will be connected to the CBB#05(IB sw)
For latest SL chassis (8 nodes) at the TOP / CBB#07 & CBB#09 will be connected to the CBB#08(IB sw)
Optical cable (External)
7m
8 links
Front
Rear
CBB#01 TYPE2
CBB#02 TYPE1
CBB#03 TYPE2
CBB#04 TYPE2
CBB#05 TYPE3
CBB#06 TYPE2
IBB
CBB#07 TYPE2
CBB#08 TYPE1
CBB#09 TYPE2
UBB
Optical cables x8
Block#01
Block#02
Block#03

STORAGE IN TAITO

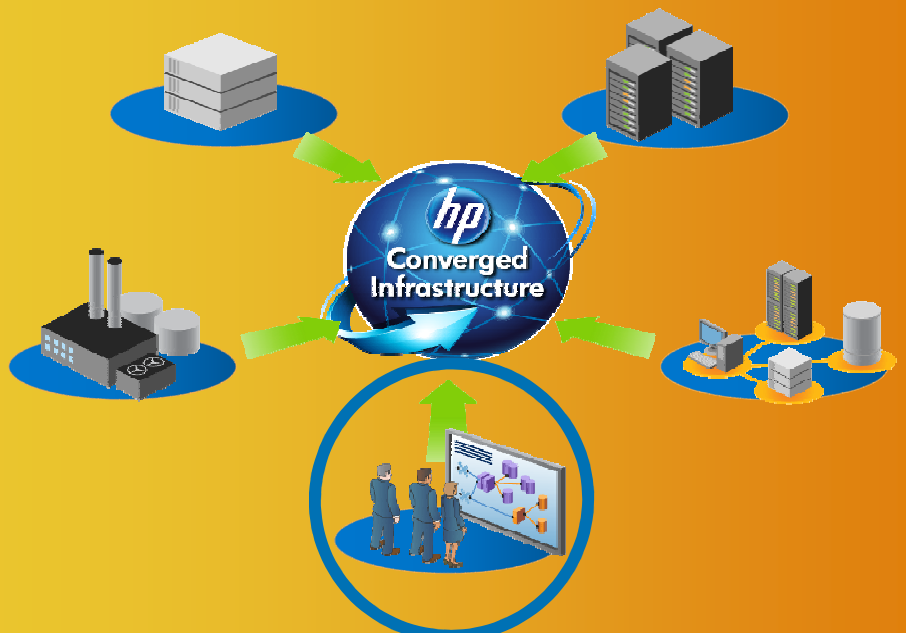


P2000 array used for the HA clustering

- The P2000 array was used to configure the HA clustering.
- IT holds the configuration files for
  - CMU - CLuster management utility
  - UFM – Unified FAbric Manager
  - SLURM – Simple Linux Utility for Resource Manager
- Contains 8 disks of 300 GB configured as follows:

lun	1	2	3
vdisk	1	2	2
size (GB)	256	512	64
file system type	ext4	ext4	ext4
used for	UFM	CMU	Slurm

SOFTWARE IN TAITO



Software Architecture for TAITO

Operating System

- Linux: CentOs 6.2

Workload manager

- Slurm

Cluster Management tools

- HP : Cluster Management Utility
- IB : Unified Fabric Manager

Applications

....

....

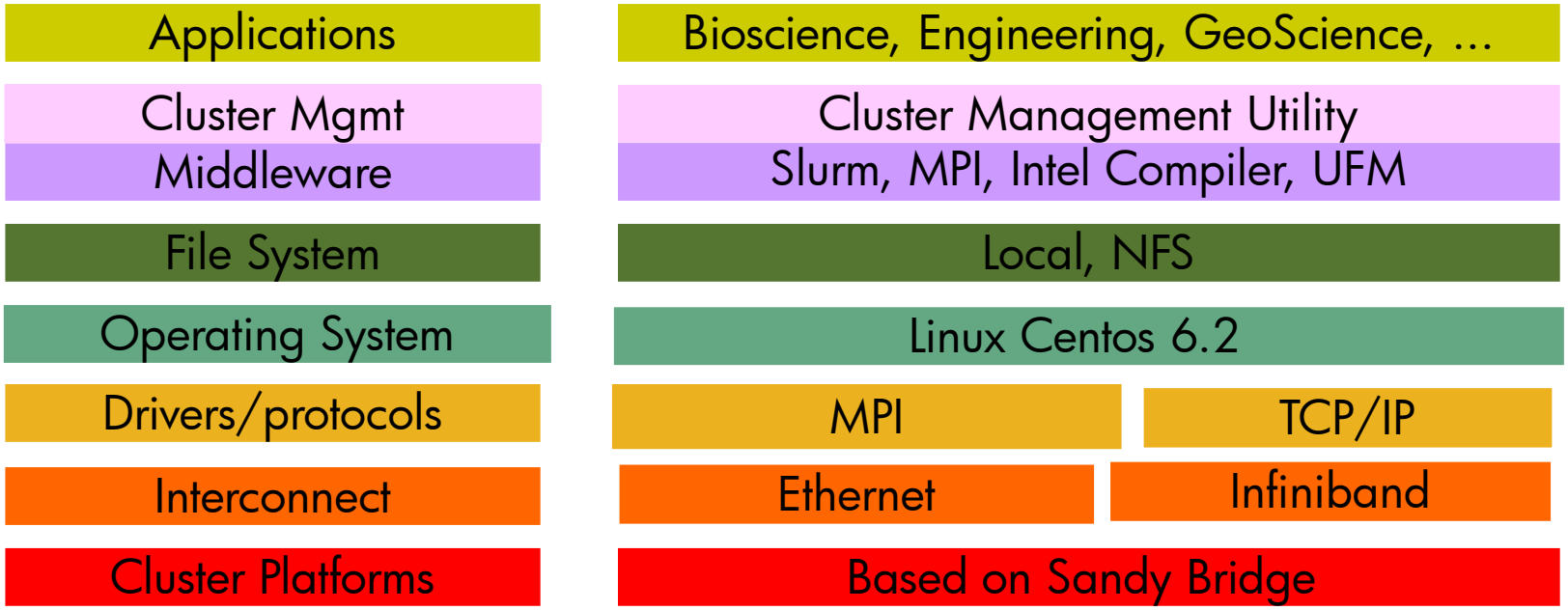
....

Libraries used by applications

- MPI : Intel MPI
- Intel Compiler

Cluster Solution Design Choices

Top level view of SW and HW components of TAITO



HP Insight CMU

Hyperscale cluster lifecycle management software

Proven

- 10 years+ in deployment, Top500 sites included with 1000's of nodes

Built for Linux, with support for multiple Linux distributions

- Including Hybrid support w/Windows

Provision

- Simplified discovery, firmware audits
- Fast and scalable cloning

Monitor

- 'At a glance' view of entire system; zoom to component
- Customizable
- Lightweight

Control

- GUI and CLI options
- Easy, friction-less control of remote servers

Worldwide CMU Deployments

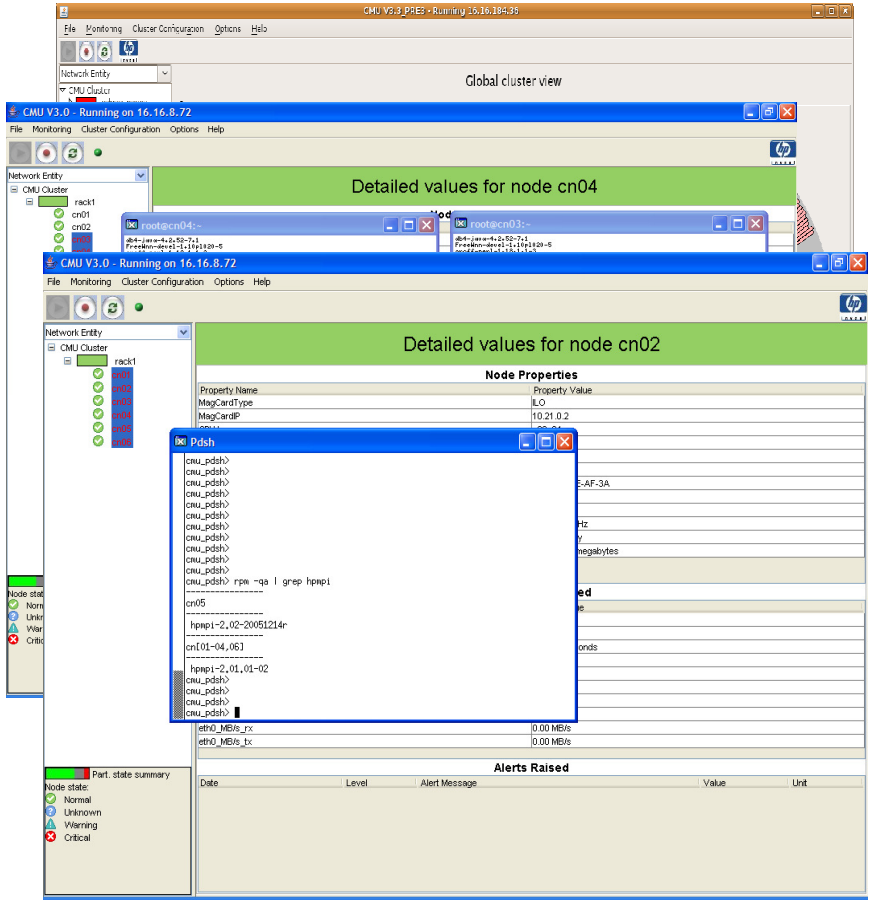
HP ships 2 CMU clusters per week WW

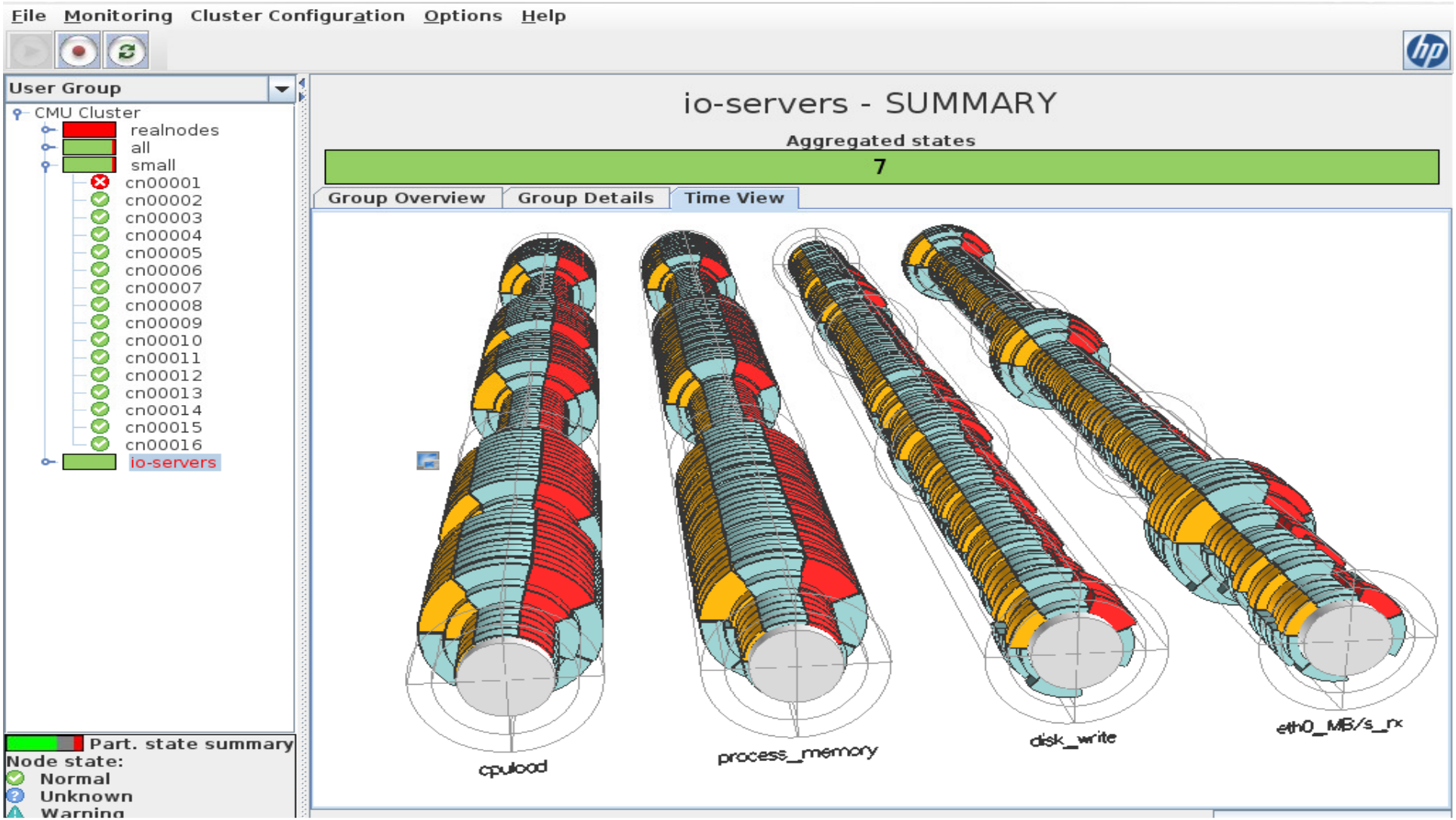
CMU main functionalities

Deployment
- Imaging (cloning)
- Autoinstall (kickstart | autoyast | preseed)
- Diskless

Scalable live monitoring
- Scalable non intrusive monitoring engine (+collectl)
- Monitoring GUI / monitoring API

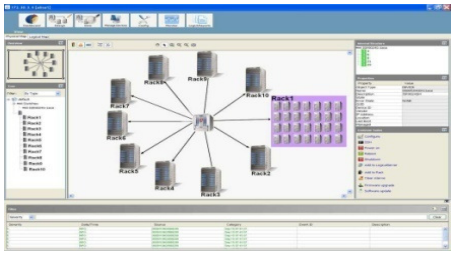
Day to day administration
- interactive cli (+ cmu_* linux commands)
- **cmudiff**, command broadcast
- multiple window broadcast (one window per host)
- single window PDSH, one command on all the hosts
- GUI (JAVA based for the desktop)




File Monitoring Cluster Configuration Options Help
User Group
CMU Cluster
realnodes
all
small
cn00001
cn00002
cn00003
cn00004
cn00005
cn00006
cn00007
cn00008
cn00009
cn00010
cn00011
cn00012
cn00013
cn00014
cn00015
cn00016
io-servers
Part. state summary
Node state:
Normal
Unknown
Warning
io-servers - SUMMARY
Aggregated states
7
Group Overview
Group Details
Time View
cpuload
process_memory
disk_write
eth0_MB/s_rx

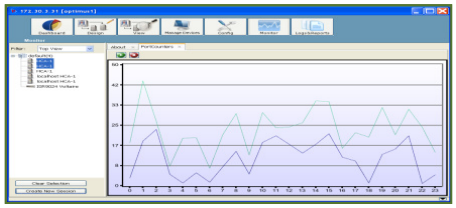
Mellanox Unified Fabric Manager (UFM)

- Monitor and troubleshoot
  - Monitor and analyze traffic behavior and fabric utilization
  - Visualize events and correlate to racks and applications
  - Detect and report problems, identify inefficiencies
- Optimize performance and utilization
  - Apply optimal routing based on application requirements, fabric topology, and load
  - Optimize performance via congestion, QoS and fabric partitioning configuration
- Provision and automate
  - Provide fabric and I/O partitioning
  - Expose the entire functionality via an extensible API, used for 3rd party integration or for automation and scripting



Central views correlated to ports, racks and apps

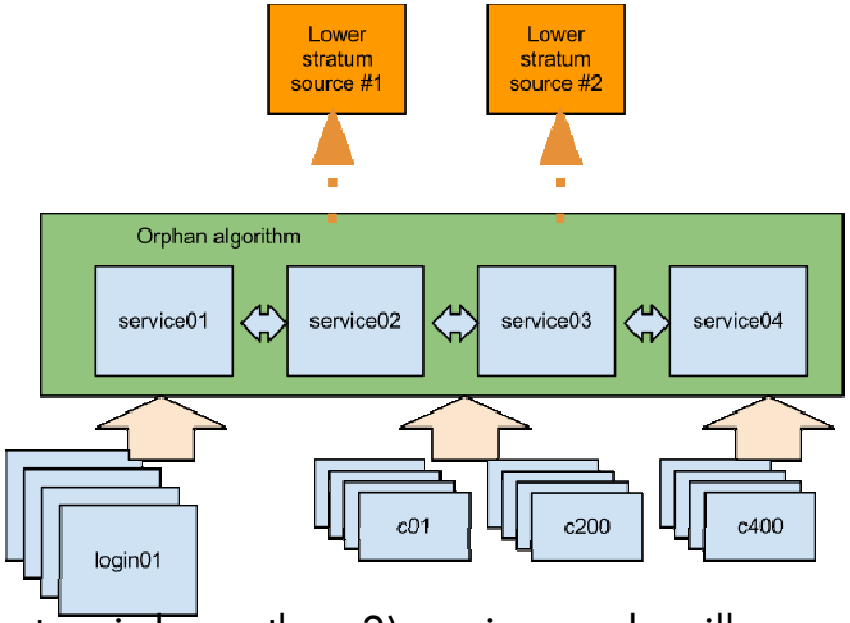
Dashboard of fabric and device utilization



Granular, real-time monitoring

OS configuration – NTP

- No NTP stratum 0 clock available
- Needed to synchronise reliabily the compute nodes together.
- The orphan algorithm for that. It works as follows:



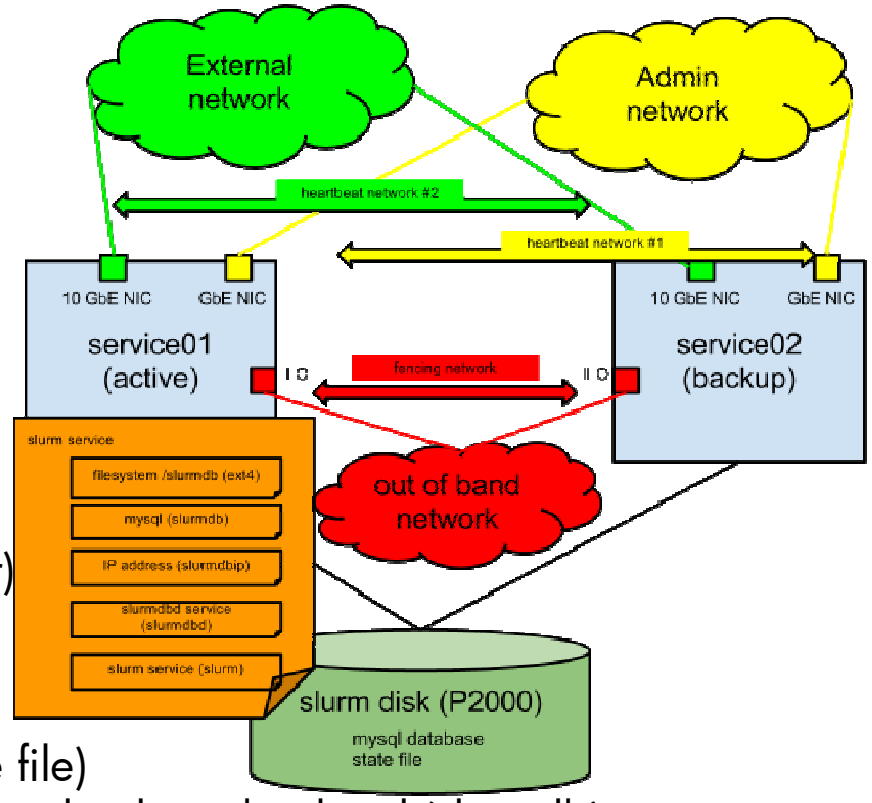
A. when a server with a low stratum is available (stratum is lower than 3) services node will use this time source to synchronise themself. (and also amongst them)

B. when no low stratum source is available service will use the orphan algorithm to “average” their clock to provide a somewhat reliable time source.

In all case compute node and login will use the services nodes as source for their time.

HA Configuration Slurm/CMU

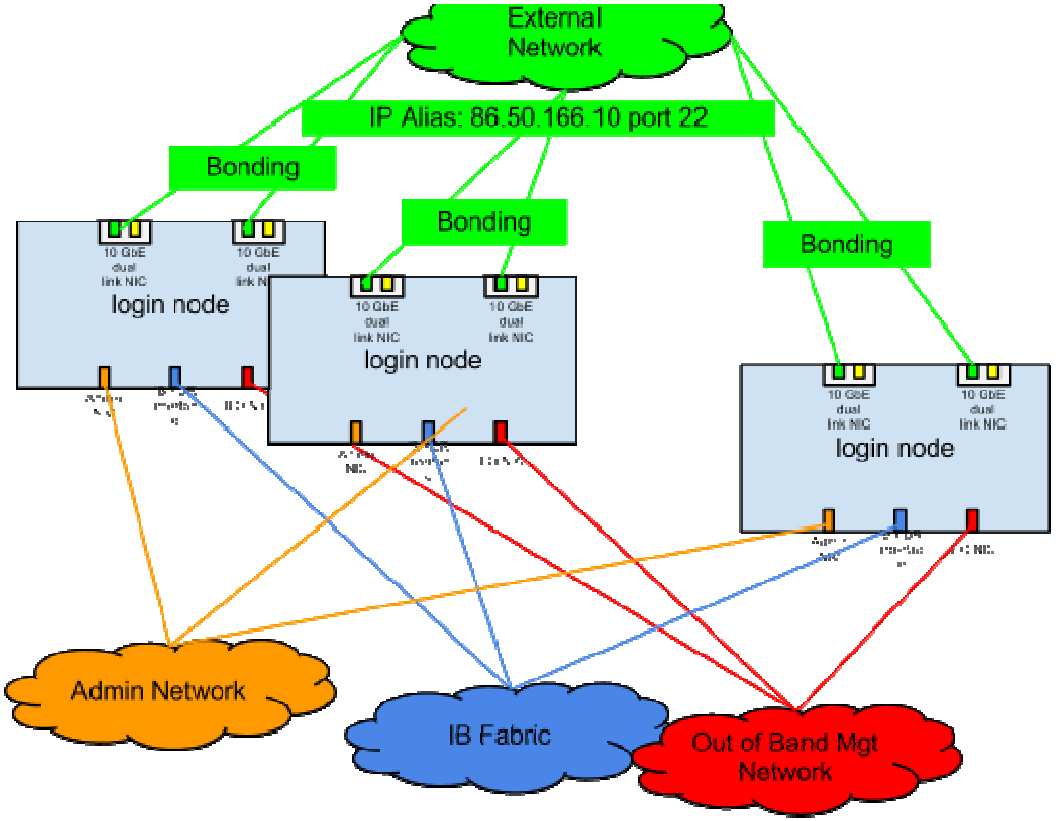
- Software used: pacemake/corosync
- SLurm on service nodes 01 – 02
- CMu on service nodes 03 – 04
- One uses:
  - Fencing (using ILO to shutdown/startup partner)
  - Database (using mysql to store slurm account information)



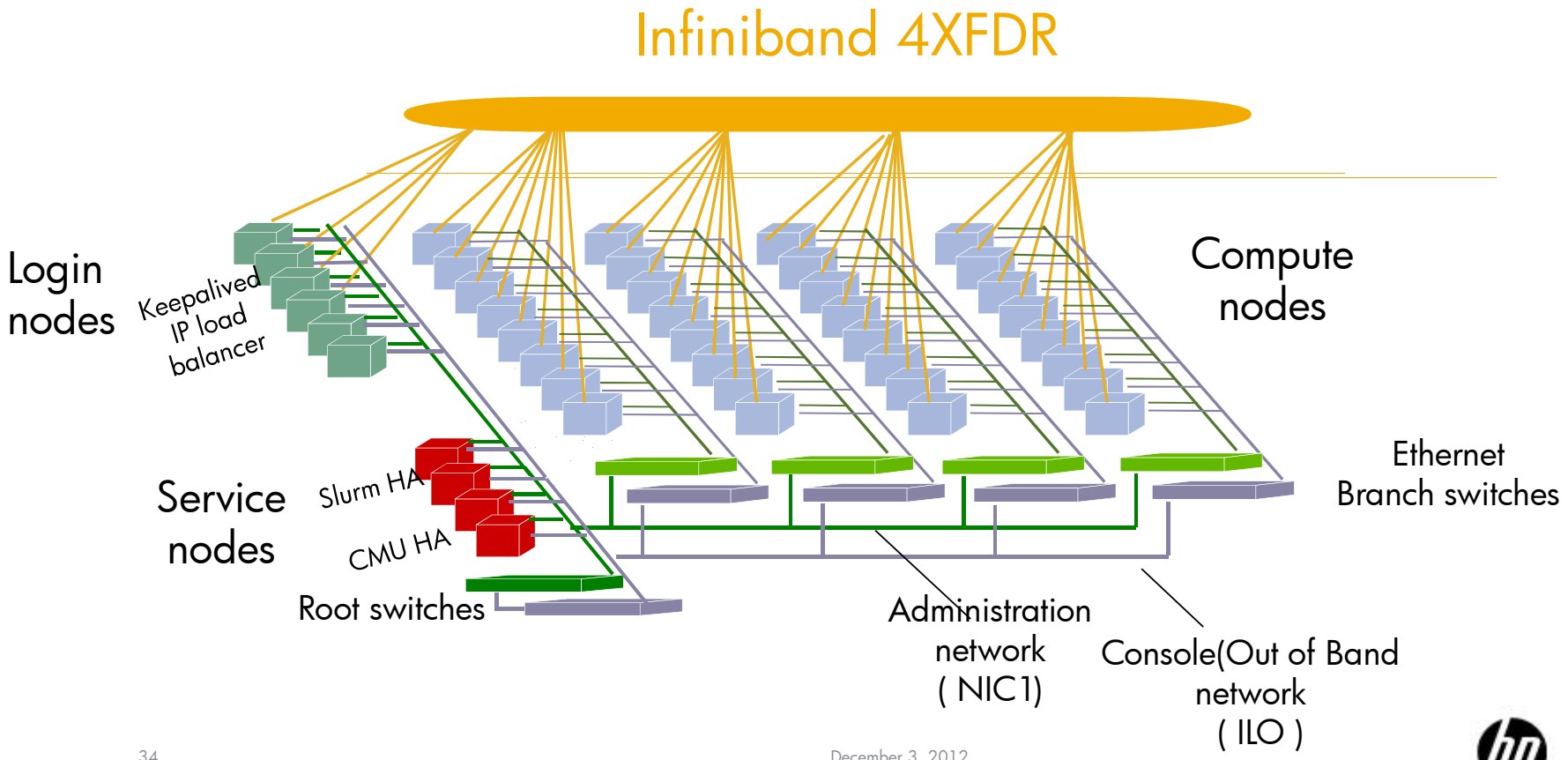
- Shared Storage (needed for database and state file)
- Shared IP (allows the slurm daemon to access the database backend (slurmdb)

Login node IP Balancer

- An ip load balancer is implemented on the login node to provide seamless access to the cluster by the user



HPC Cluster : high level view



THANK YOU