Architecting High Performance Computing Systems for Fault Tolerance and Reliability

Blake T. Gonzales
HPC Computer Scientist
Dell Advanced Systems Group
blake_gonzales@dell.com

Agenda

- HPC Fault Tolerance and Reliability
- Architecture Design Techniques
- Dell HPC Solutions: Purpose Built Reliability
- Questions

HPC Fault Tolerance and Reliability

HPC Fault Tolerance and Reliability

- Complex nature of HPC systems can have a detrimental effect on their ability to reliably complete the tasks at hand.

- Research performed by HPC systems is important!

- Reliability and fault tolerance is of utmost concern in HPC.

$$\mathrm{MTTF} = \int_0^\infty R(t).dt$$

$$\mathrm{MTTF} = \int_0^\infty e^{-\lambda t}.dt$$

$$A = \frac{\mathrm{MTTF}}{\mathrm{MTTR} + \mathrm{MTTF}}$$

Single Points of Failure

- Shared-memory multiprocessor (SMP) systems are generally prone to system wide failures due to single errors in memory, CPU or disk.

- With the ubiquitous use of clustered HPC technology in the last decade, the risk of system wide failures due to single points of failure can be minimized!

- Clustered solutions must be designed correctly.

HPC Subsystem Design

- Cluster solutions have many “moving parts”

- It is important to design each subsystem with an eye to how it relates to the other subsystems.

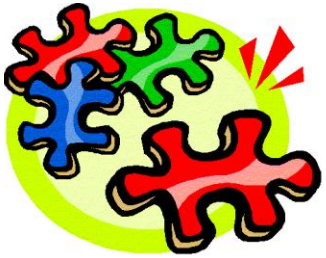
- Find key components that are likely to cause system wide failures, and implement architecture design techniques to prevent such failures.

Architecture Design Techniques

Component Classification

- Failure has little effect on overall reliability
  - Compute Nodes
  - Out-of-band management

- Failure has major effect on overall reliability
  - Head / Admin Node(s)
  - Job Scheduler
  - Storage
  - Power
  - Cooling
  - Cabling
  - Network
  - The list goes on…



Compute Nodes

- Irony: The workhorse subsystem of an HPC cluster, is the same subsystem that requires the least amount of built-in fault tolerance.
- High fault tolerance to an occasional failed job
- Generally does not require added fault tolerant subsystems

Compute Nodes

- Common to have several compute nodes inoperable on large systems

- When a single compute node fails, typically only one job is effected, or subset of jobs, on the system.

Login Nodes

- Customer Facing, Outage Perception
- Provide Multiple Identical Nodes
- Publish Entry Points

Head / Administrative Nodes

- Consider Separating these Functions
  - Provisioning
  - Image Management
  - Job Scheduling (Multiple Nodes!)
  - Network Boot for Compute Nodes

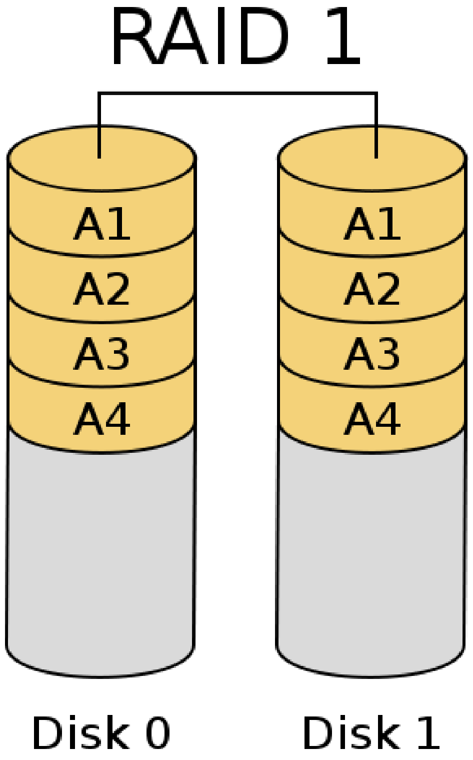
Non-Compute Node Disk Protection

- Mirroring - RAID 1 (this doesn’t protect against data corruption though)

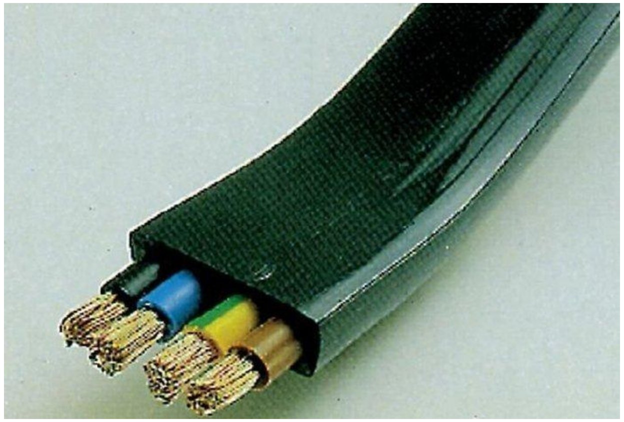
- Hot Spares

- Backups (software stack, compute node images)

- Disk Cloning (weekly, multiple copies)



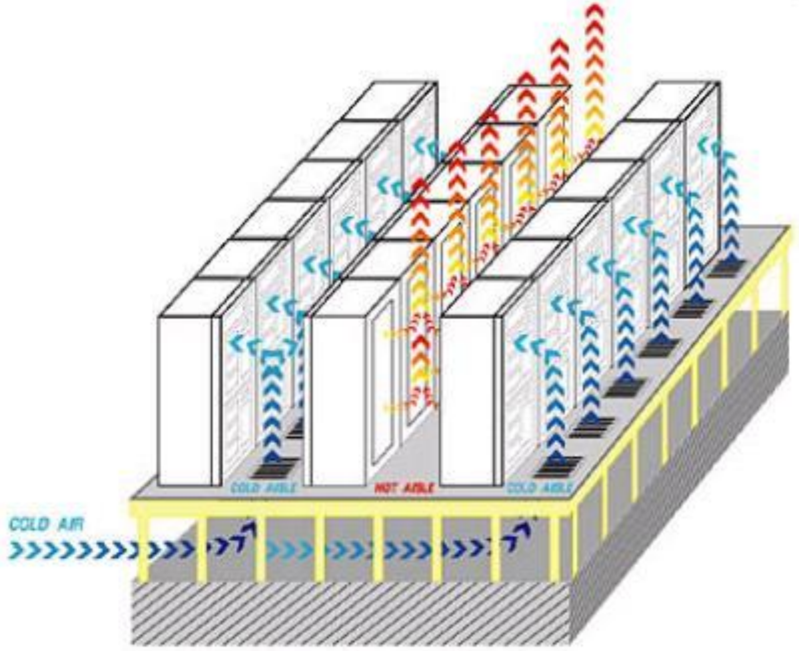
Power Distribution



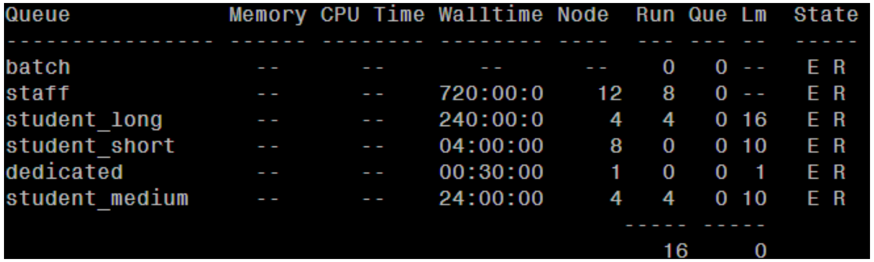
- Continuous Power Feed
  - Generators
  - Battery Backup or UPS
- Multiple Data Center or Rack based PDUs
- Deliver Power Feeds to Multiple Power Supplies
- Hot Swapable Power Supplies
- Labeling is critical

Cooling

- For every watt consumed in a component, there is a cooling power component that must be consumed as well!
- Air Handlers, Chillers, Fans
- Hot spots correlate with compute nodes
- Hot/Cold Aisles
- Chilled Doors, In-row cooling



Job Scheduling



```
Queue            Memory CPU Time Walltime Node  Run Que Lm  State
---------------- ------ -------- -------- ----  --- --- --  -----
batch              --        --       --      --    0   0 --   E R
staff              --        --  720:00:0   12    8   0 --   E R
student_long       --        --  240:00:0    4    4   0 16   E R
student_short      --        --  04:00:00    8    0   0 10   E R
dedicated          --        --  00:30:00    1    0   0  1   E R
student_medium     --        --  24:00:00    4    4   0 10   E R
                                              ----- -----
                                                 16     0
```

- “Job Scheduling State” database

- Multiple paths to the database

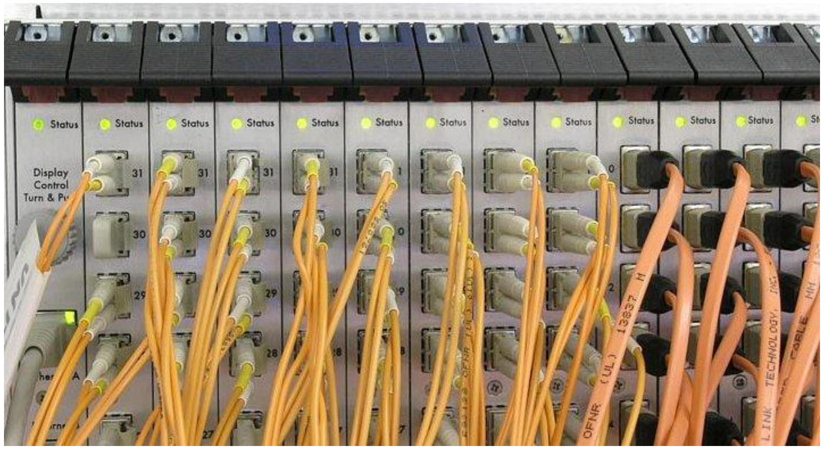
- Multiple failover daemons on separate nodes

- Jobs may continue to run on compute node infrastructure, even if daemon nodes fail. But you need a “map” of the activity.

- Checkpoint / Restart

Networking and Interconnect

- Little redundancy needed in admin network
- Out-of-band management is crucial
- Interconnect hubs require redundancy
  - Ethernet & Infiniband
  - MPI / Storage
  - Power
  - Uplinks/Downlinks



Cable Management

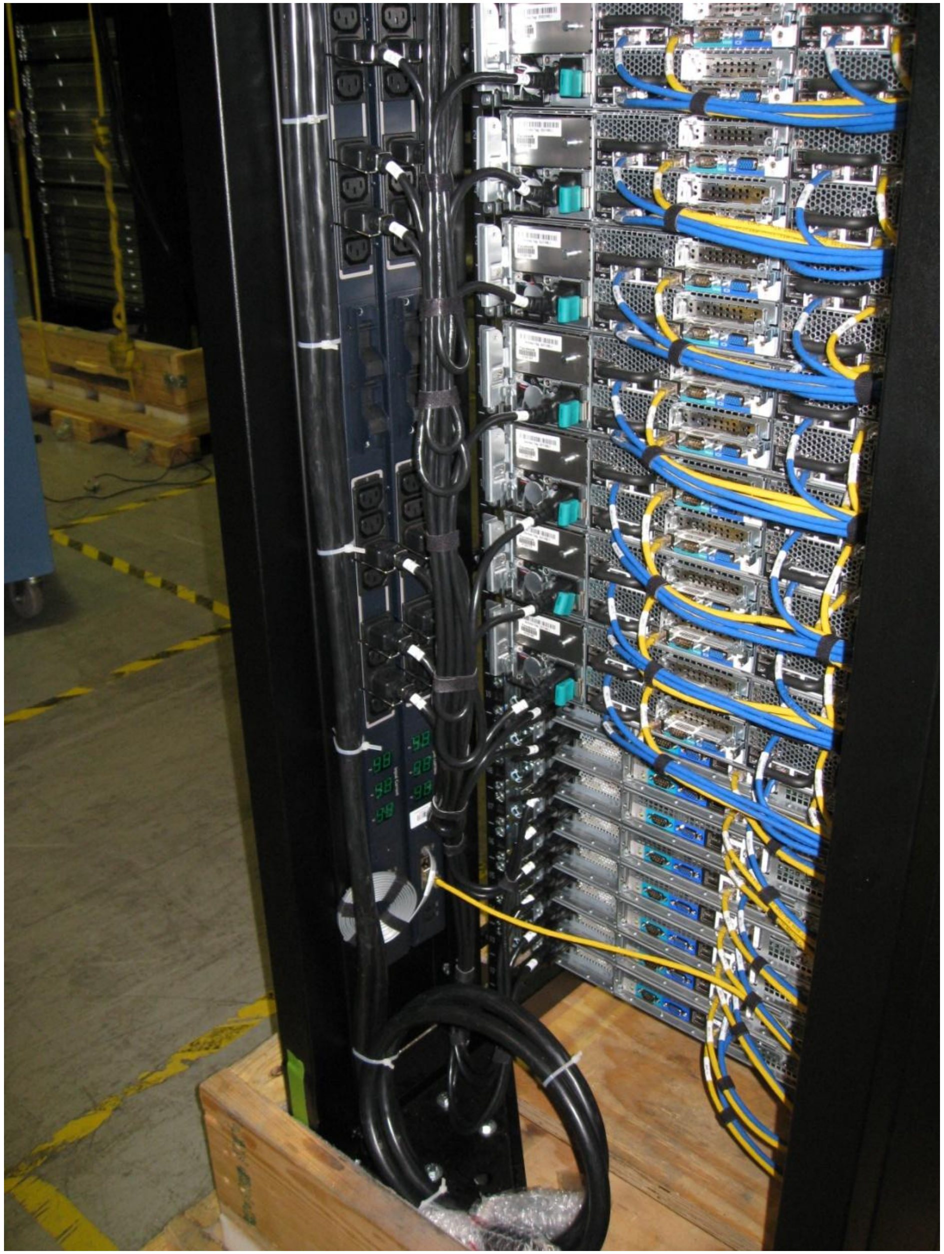
- Reliability, Really?

- Heavy IB Cables

- Labeling is Critical

- Organize the chaos

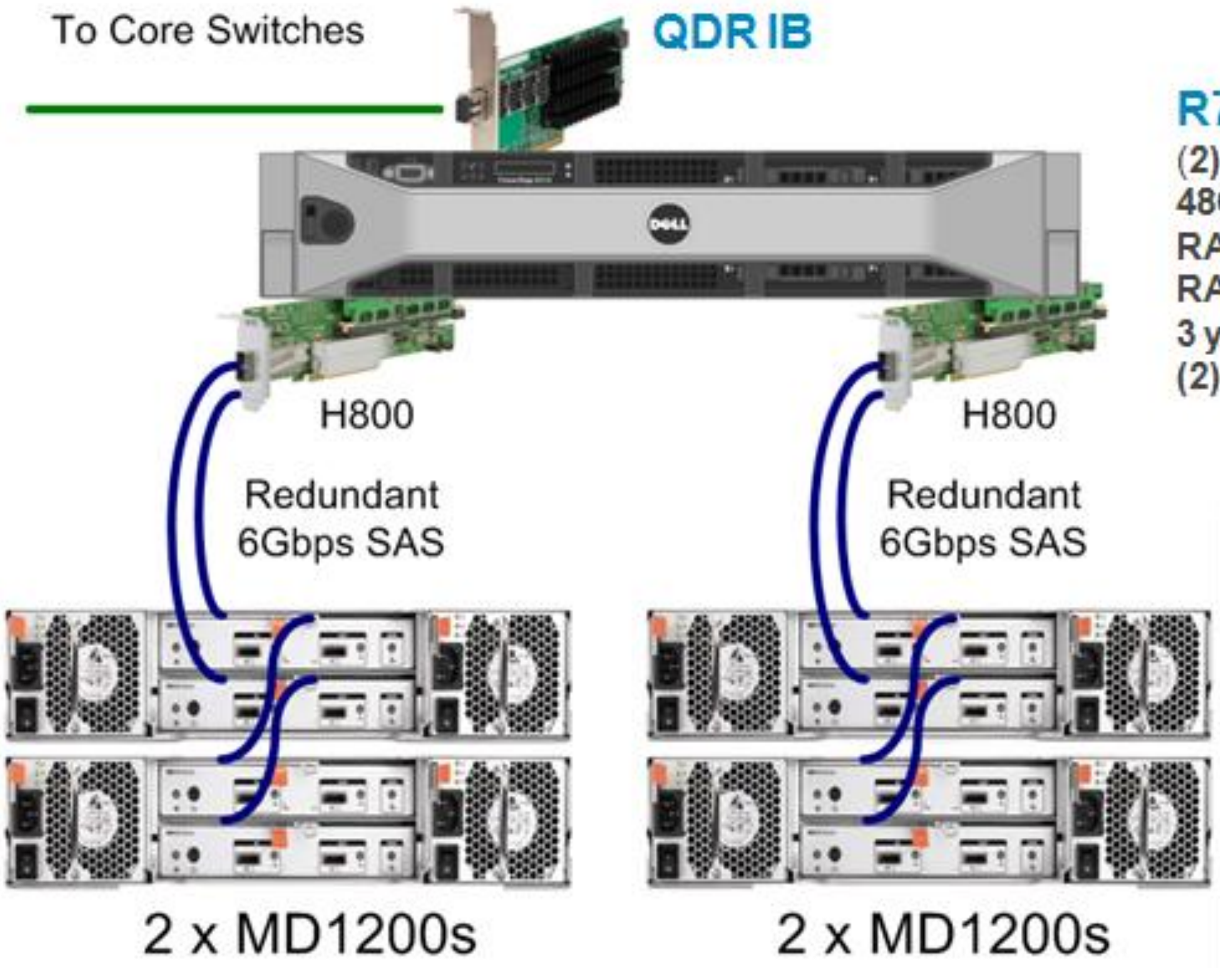
- You’ll be glad next time you have a outage and time is of the essence!



Continual Testing!

Dell HPC Solutions: Purpose Built Reliability

The Dell HPC NFS Storage Solution (NSS)



To Core Switches

QDR IB

H800

Redundant 6Gbps SAS

H800

Redundant 6Gbps SAS

2 x MD1200s

2 x MD1200s

R710: (NFS Gateway)

(2) 2.4GHz Westmere (4c)
48GB memory
RAID-1 OS w/ host-spare
RAID-0 swap (2 drives)
3 years support including FS
(2) PERC H800 w/ RAID-60 and LVM

Summary

48-96TB

RAID-60 and LVM

- *RAID-6 within each MD1200*
- *RAID-0 across MD1200's*
- *LVM to combine LUNS*

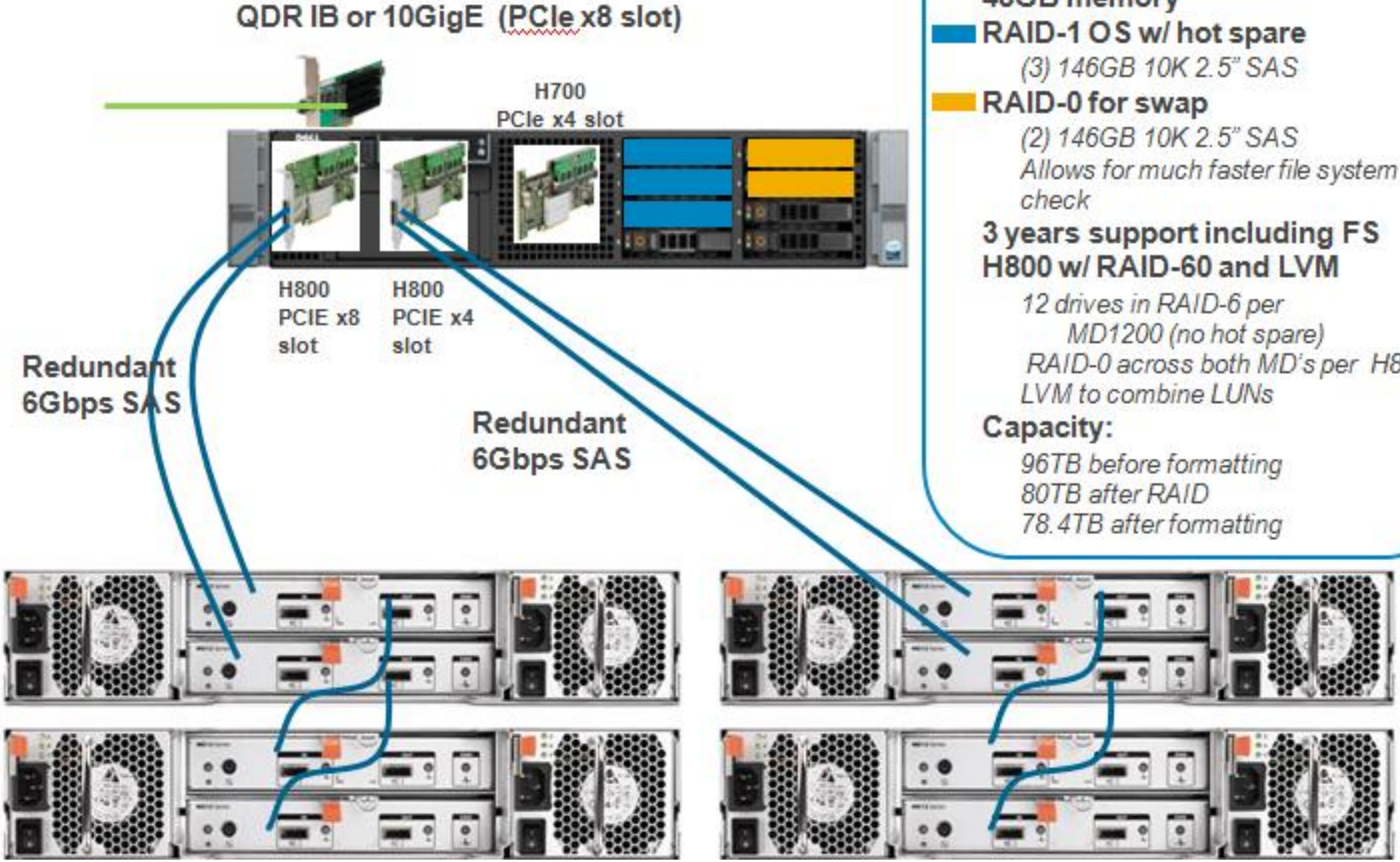
10GigE NFS Performance

- *Sequential Read: 855 MB/s*
- *Sequential Write: 1,180 MB/s*

InfiniBand NFS Performance

- *Sequential Read: 1,350 MB/s*
- *Sequential Write: 1,470 MB/s*

Behind the Bezels (BTB)



R710: (NFS Gateway)

(2) 2.4 GHz Westmere (4c)

48GB memory

🟦 **RAID-1 OS w/ hot spare**

(3) 146GB 10K 2.5" SAS

🟧 **RAID-0 for swap**

(2) 146GB 10K 2.5" SAS

Allows for much faster file system check

3 years support including FS

H800 w/ RAID-60 and LVM

12 drives in RAID-6 per
MD1200 (no hot spare)

RAID-0 across both MD's per H800

LVM to combine LUNs

Capacity:

96TB before formatting

80TB after RAID

78.4TB after formatting

The Dell Terascala HPC Storage Solution (HSS)

- Full Lustre solution, fully configured, tested, tuned, and deployed
  - On-site installation with client deployment and training included

- Redundant, highly available solution

- Simple, linear scalability

- Full management system with easy to use GUI

MetaData Server (MDS)

Base OSS

Dell PowerEdge C6100

- **Four 2-Socket Nodes in 2U**
  - Intel Westmere-EP
- **Each Node:**
  - 12 DIMMs each
  - 2 GigE (Intel)
  - 1 Daughter Card (PCIe x8)
    - 10GigE
    - QDR IB
  - One PCIe x16 (half-length, half-height)
  - Optional SAS controller (in-place of IB)
- **Chassis Design:**
  - Hot Plug, Individually Serviceable System Boards / Nodes
  - Up to 12 x 3.5” drives (3 per node)
  - Up to 24 x 2.5” drives (6 per node)
- **N+1 Power supplies (1100W or 1400W)**
- **NVIDIA HIC certified**
- **DDR and QDR IB PCIe card certified**

Dell PowerEdge C410x

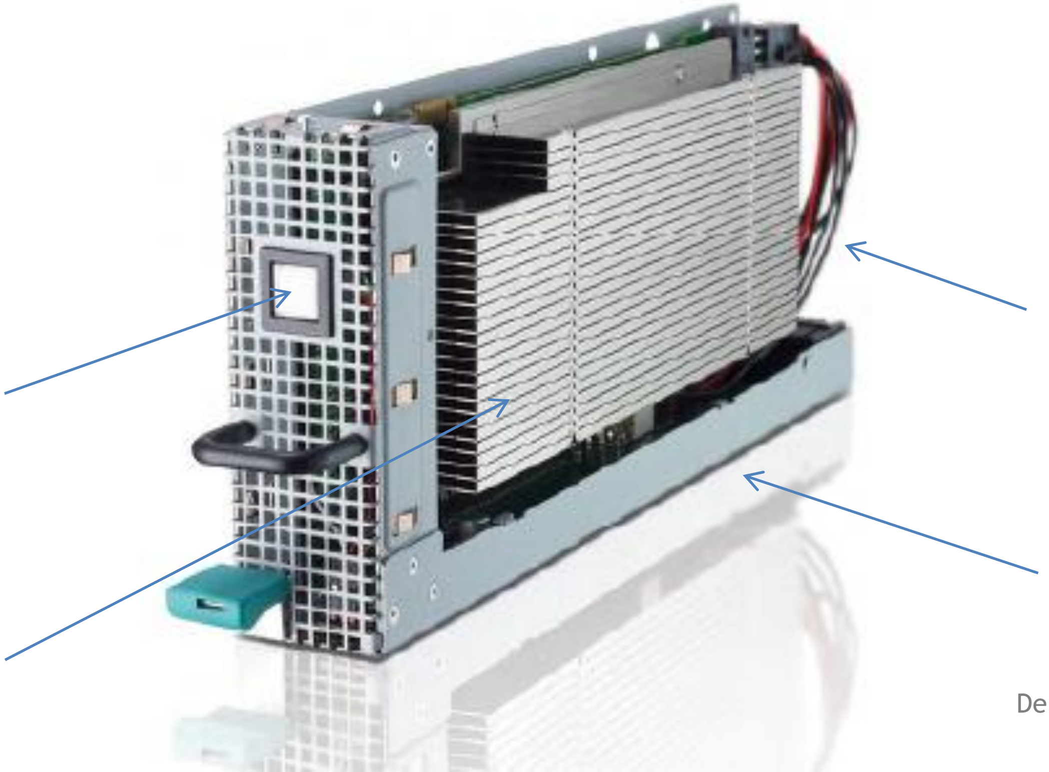
- 3U chassis (external)
  - “Room-and-Board” for PCIe Gen-2 x16 devices
  - Up to 8 hosts

- Sixteen (16) x16 Gen-2 Devices
  - Initial Target = GPGPUs
  - Support for any FH/HL or HH/HL device
  - Each slot Double-Wide
  - Individually Serviceable

- N+1 Power (3+1)
  - Gold (90%)
- N+1 Cooling (7+1)

Dell PowerEdge C410x

- Sixteen (16) x16 Gen-2 Modules
  - PCIe Gen-2 x16 compliant
  - Independently serviceable



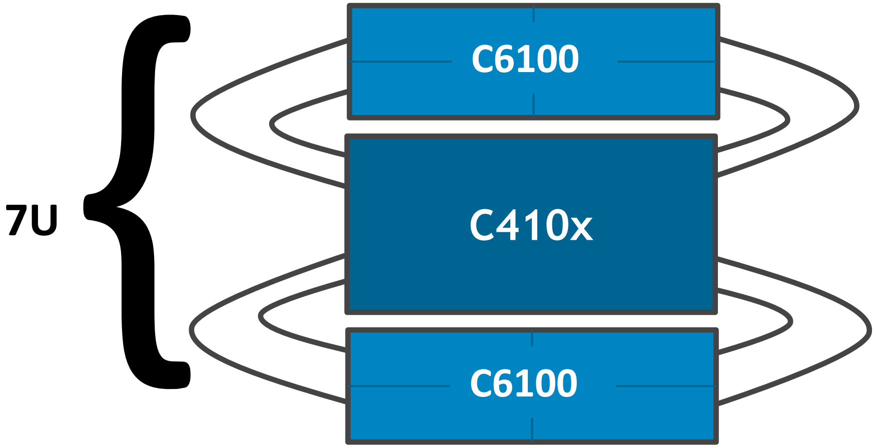
Power connector for GPGPU card

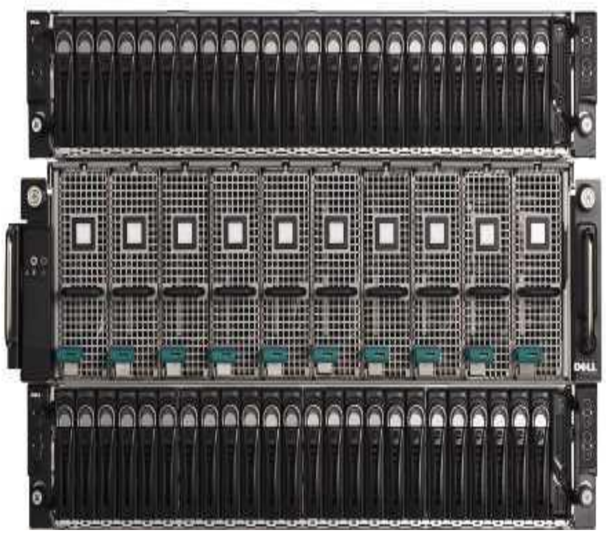
LED and On/Off

Board-to-board connector for X16 Gen 2 PCIe signals and power

GPU card

C410x "Sandwich"





8-Card C410x Sandwich				16-Card C410x Sandwich
2 x C6100				2 x C6100
8 GPUs				16 GPUs
1 – QDR IB daughtercard				1 – QDR IB daughtercard
7U total				7U total
8 GPUs total				16 GPUs total
8 nodes total				8 nodes total
8/7 nodes / U				8/7 node / U
8/7 GPUs per U				16/7 GPUs per U
1 GPU per PCIe x16				2 GPUs per PCIe x16

Dell PowerEdge C410x

- Increased density (more GPUs per RackU)
- Introduced “flexibility”
    - GPU/Host ratio = 1:1, 2:1, 3:1, 4:1, …, (8:1), …, (16:1)
- Purposely Separate the Host from the GPUs
- Purpose-built to power, cool and manage PCI-e devices
    - (N+1) Power (3+1 “Gold” power supplies)
    - (N+1) Cooling (7+1 fans)
    - Onboard BMC Web interface to monitor, manage & configure
    - Each PCI-e Module is individually serviceable

-- no un-cabling

-- no un-racking

- - no opening of compute nodes

- - no bumped DIMMS

- - no disturbed dust

- - vertical insertion

HPC at Dell

Blake T. Gonzales
HPC Computer Scientist
blake_gonzales@dell.com

Jim Gutowski
HPC Business Development Manager
james_gutowski@dell.com

HPC Blogs, Case Studies, Guides, Tips and Tricks
http://HPCatDell.com/

Twitter
@HPCatDell