PetaScale I/O Challenges: Hyperion and Sequoia

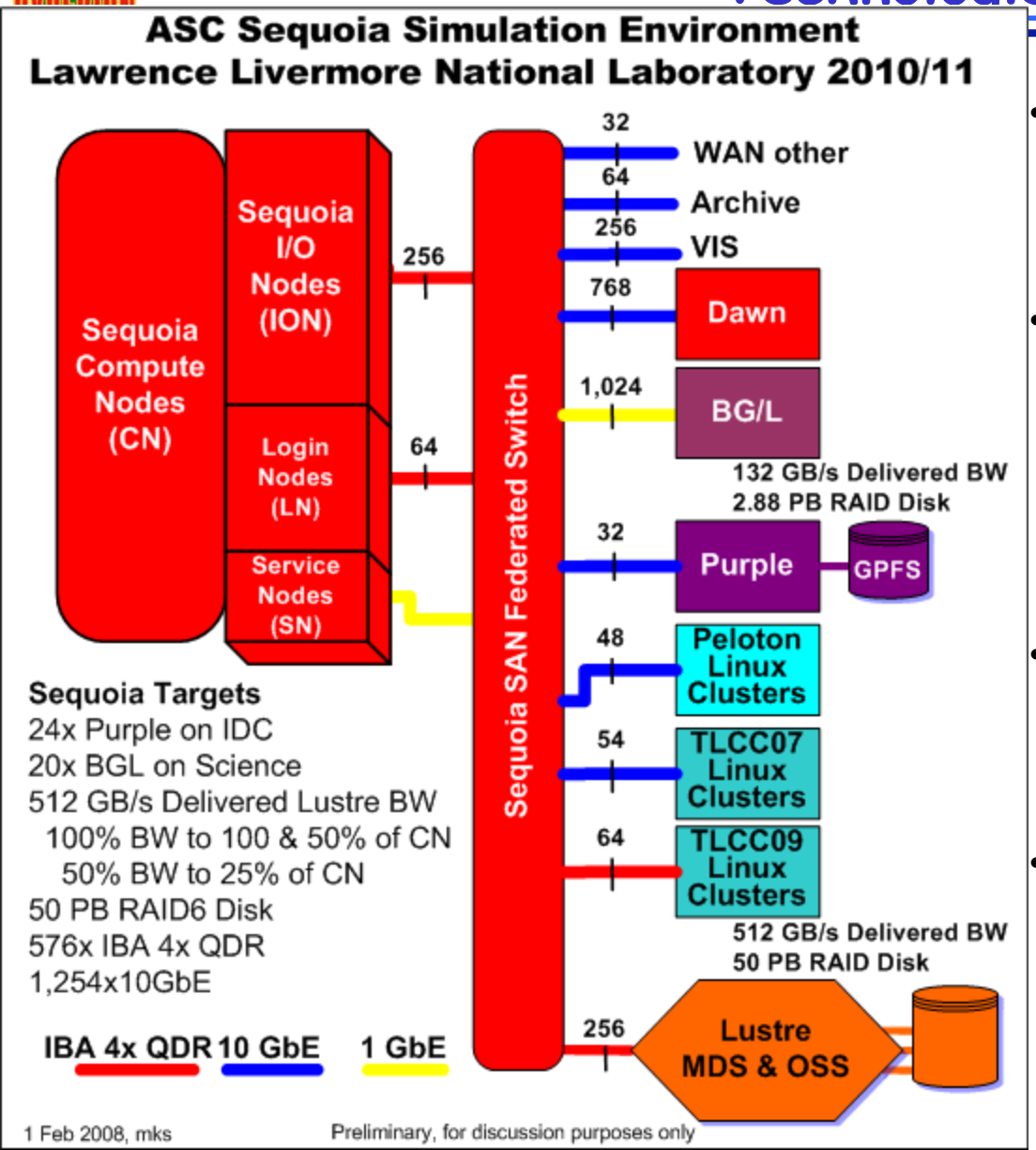
Matt Leininger
LLNL

OpenFabrics Workshop
March 23, 2009

Overview

- ASC Sequoia
- Hyperion Advanced Technology Testbed
- Petascale I/O challenges
- Future directions

Sequoia Peta-Scale Simulation Environment Requires the Development of Critical Enabling Technologies



- ASC Sequoia requires a leap forward in simulation environment capability
- Deploying simulation environments for ASC Purple & BlueGene/L required developing and testing at unprecedented scale
- A new test environment is required for Sequoia and future Linux clusters
- Hyperion will be a cluster testbed for breakthrough hardware and software technology development

Hyperion Project Goals 1

- Development and testing environment for critical enabling Linux Cluster Technologies
  - InfiniBand Open Source Software - OFED
  - Lustre & other Open Source Parallel File System Testing and Scaling
  - Open Source Tri-Lab Linux software stack development & testbed
  - “Intel Cluster Ready” process to push technology out to wider HPC community

Hyperion Project Goals 2

- Evaluation testbed for new hardware & software technologies
  - Petascale I/O technology scaling for Sequoia and future capacity systems
  - Processor, memory, networking, storage, visualization, etc.
  - Designed for future technology refresh, expansion, and upgrades

Hyperion Project Goals 3

- Innovative approach for forming long term collaborations
  - OpenFabrics Alliance
  - Lustre Center of Excellence
  - Other Gov't agencies, alliances, and computing centers
  - End customers (e.g. Financial services, Oil & Gas, Pharma, etc.)

Hyperion Collaborations Allow Partners to Build A Resource None Could Afford Alone

Founding Members

- **Intel, Dell, and Supermicro**
  - Processors, Nodes, Racks and integration
- **QLogic, Cisco and Mellanox**
  - IBA switches & HCAs
  - Ethernet switches & NICs
  - IBA⇔Ethernet routers
- **DDN, Sun, LSI**
  - Storage Hardware
- **RedHat**
  - Linux testing and sys admin
- **Sun and RedHat**
  - Parallel File Systems

National Nuclear Security Administration

Hyperion Collaborations Allow Partners to Build A Resource None Could Afford Alone

Benefits to Members & Wider Community

- Software scalability & performance testing enables wider market adoption of cluster technologies
- Enables large-scale cluster deployments to become more common & routine
- HPC engagement between Hyperion members
- Large scale testbed for critical HPC technologies
- HPC engagement with critical vendors influencing larger general HPC markets
- Early product evaluation at large-scale

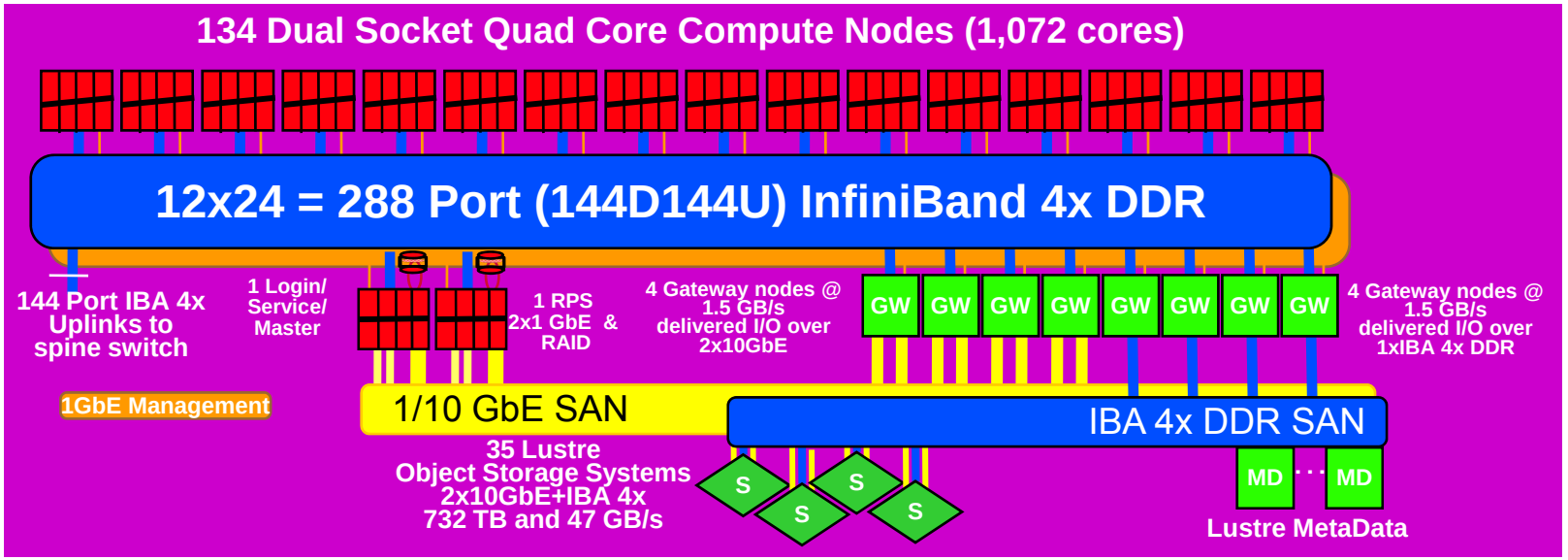
Sustainable Plan for Hyperion Budget

- LLNL Investment
  - $4.25M for Hyperion base
    - $3.2M for 8 SU base system
    - $0.75M for SSU & SAN
    - $0.15M for siting & install
    - $0.15M for RedHat on-site
  - $0.5M/yr ongoing for tech refresh
  - Yearly power & cooling costs
  - System Admin 0.75 FTE
- Funding Sources
  - ~$5.4M collaborator contributions
  - FY08
    - $2.8M
    - System admin, power & cooling
  - FY09 and beyond
    - $2.0M
    - System admin, power & cooling

Hyperion Governance Model

- Collaboration structure
  - Legal documents are LLNS subcontracts and loan agreements
  - Have non-binding MOU for broad perspective
- Executive Board
  - Members set policy, collaboration interaction models
  - Allocations determined by member investments
  - Deals with strategic issues, priorities
  - Meets quarterly
- Operations Team
  - Sets day-to-day schedule of activities and platform usage
  - Meets weekly or bi-weekly
  - Interacts with community with instant messaging, email, & wiki

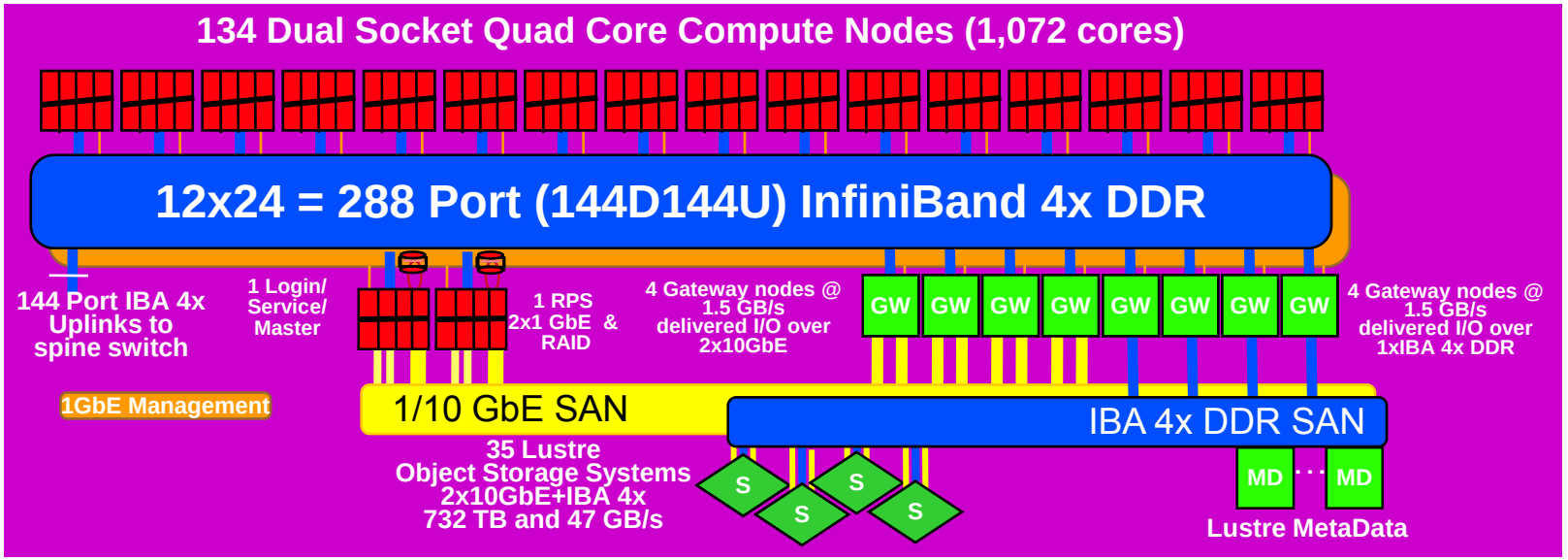
Hyperion Phase 1
11.5 TF/s Scalable Unit



Hyperion Phase 1 Deployment is an 4 SU 46 TF/s Cluster with full IBA

- 576 Total nodes and 4,608 cores, 12.1 TB/s memory bandwidth, 4.6 TB capacity
- IBA is expandable to 1,728 IB ports single plane and can double to dual plane
- 80.0 GF/s dual socket 2.5 GHz quad-core Intel LV Harpertown nodes
  - 8 GB from 4 channels FB-DIMM 667 RAM @ 21.6 GB/s
- Nodes utilize PCI-Express generation 2 I/O which provides an upgrade path to IBA 4x QDR
- Storage Scalable Units (SSU) from DDN, LSI and Sun yielding >47GB/s and 732TB disk
- Full system will require 400 kW of power, 112 tons of cooling. Sited at LLNL

Hyperion Phase 2
11.1 TF/s Scalable Unit



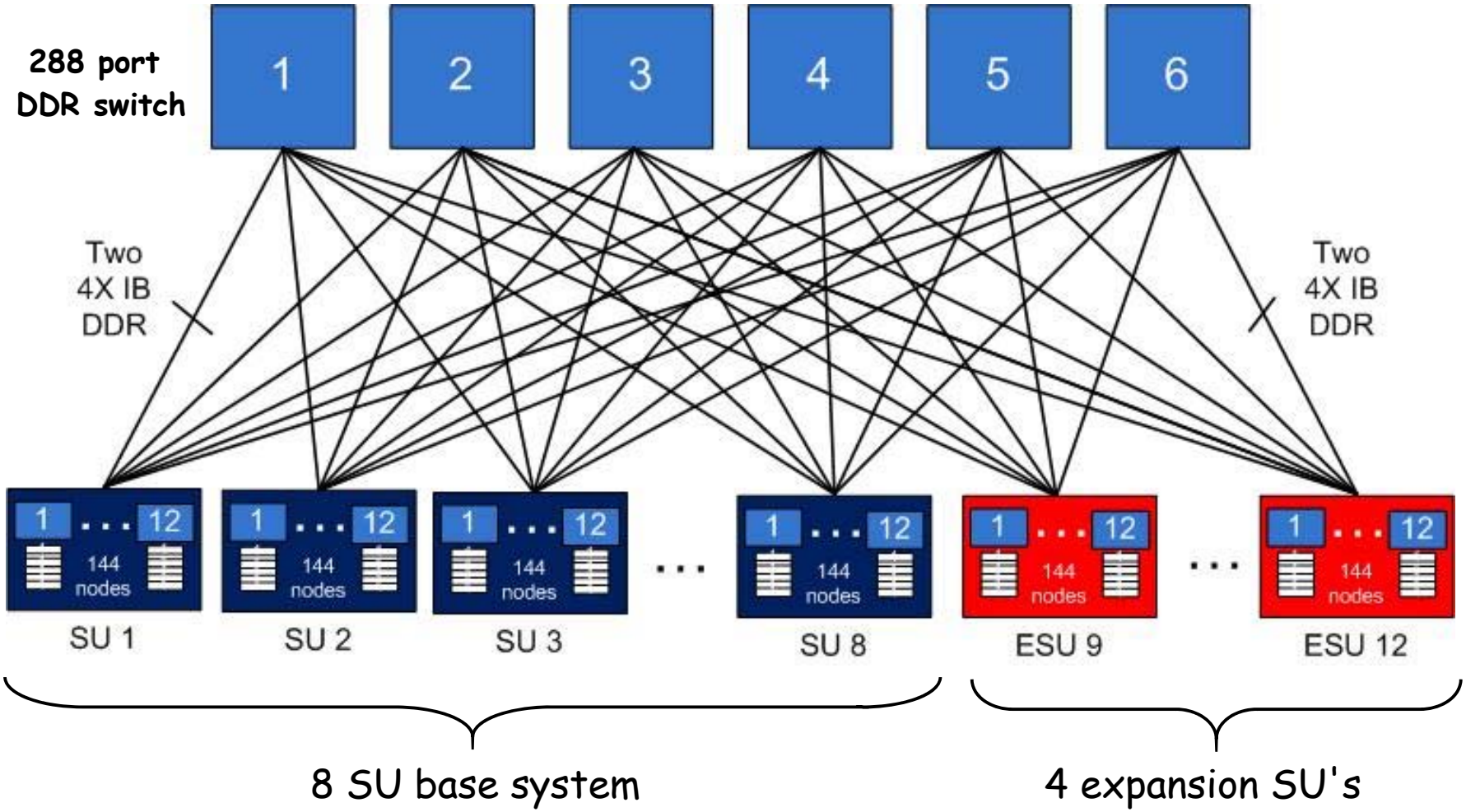
Hyperion Phase 2 Deployment is an 4 SU 44.4 TF/s Cluster with full IBA

- 576 Total nodes and 4,608 cores, 36.9 TB/s memory bandwidth, 4.6 TB capacity
- IBA is expandable to 1,728 IB ports single plane and can double to dual plane
- 76.8 GF/s dual socket 2.4 GHz quad-core Intel Xeon Nehalem nodes
  - 12 GB from 6 channels DDR3 1067 DRAM @ 64 GB/s
- Nodes utilize PCI-Express generation 2 I/O which provides an upgrade path to IBA 4x QDR
- Storage Scalable Units (SSU) from DDN, LSI and Sun yielding >47GB/s and 732TB disk
- Full system will require 400 kW of power, 112 tons of cooling. Sited at LLNL

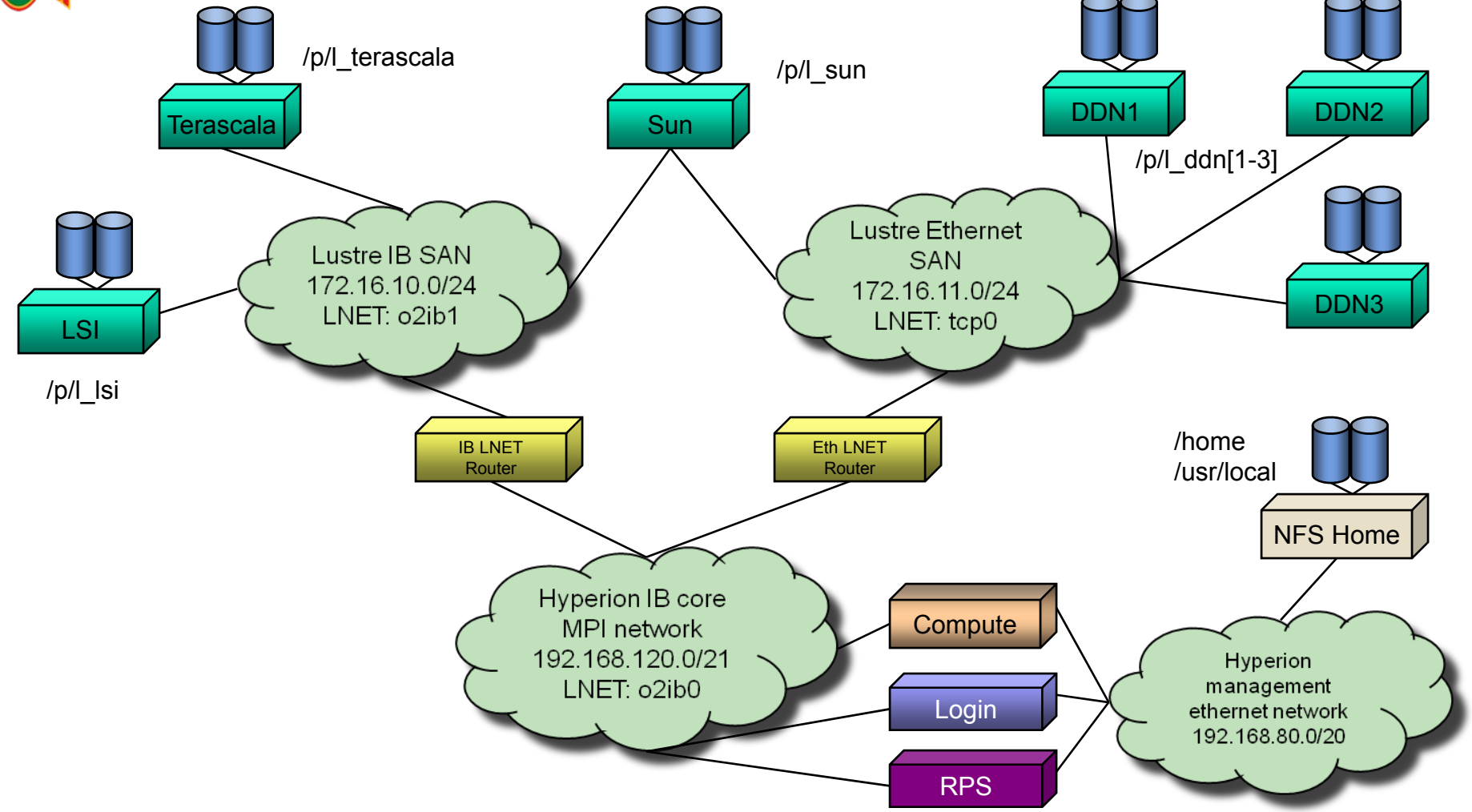
Hyperion Storage

- Data Direct Networks
  - S2A 9550 IBA 4X SDR 512x250 GB SATA
  - S2A 9900 IBA 4X DDR 280x500 GB SATA
  - S2A 9900 IBA 4X DDR 160x147 GB SAS
  - 233 TB total RAID6 (8+2) storage capacity
- LSI/Engenio
  - Four Engenio 7000 (XBB2)
  - IBA 4X DDR 192x500GB SATA
  - 307 TB total RAID6 (8+2) storage capacity
- Sun
  - Eight “Thor” Sun Fire X4540
  - 48x500 GB SATA, IBA 4X DDR/10GbE
  - 192 TB raw total storage capacity

Hyperion 8 Scalable Units with Network for 50% Expansion



Hyperion File Systems and Networks

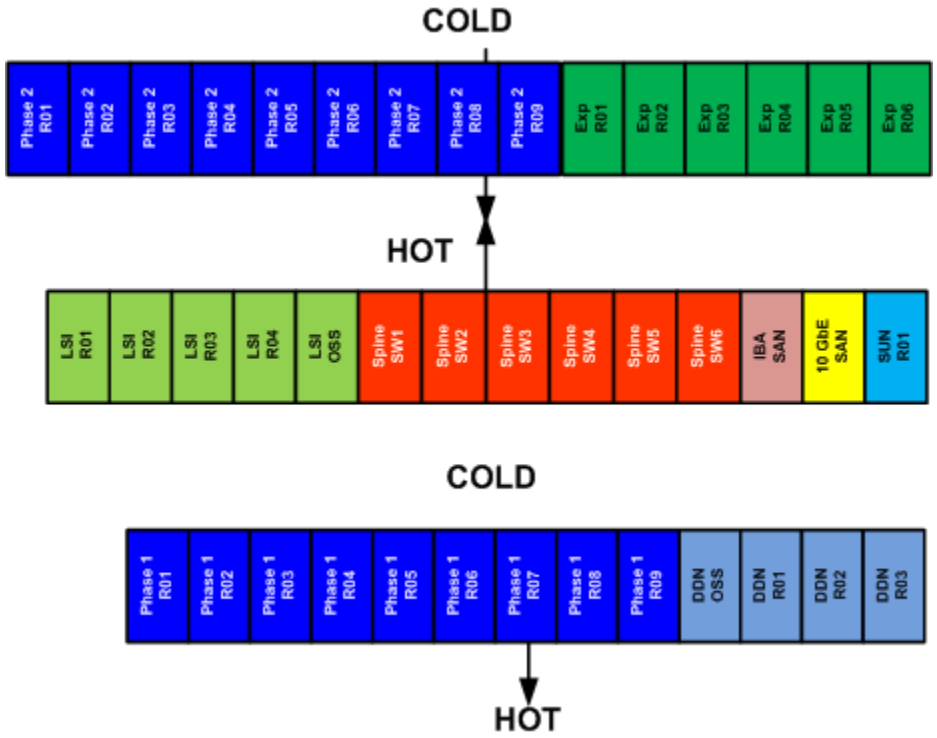


Hyperion Phase 1
46 TF/s 4SU in 9 racks

HyperionPhase1_R01	HyperionPhase1_R02	HyperionPhase1_R03	HyperionPhase1_R04	HyperionPhase1_R05	HyperionPhase1_R06	HyperionPhase1_R07	HyperionPhase1_R08	HyperionPhase1_R09
GW03 GW04 CN67 CN68	GW03 GW04 CN67 CN68	GW03 GW04 CN67 CN68			GW03 GW04 CN67 CN68	GW03 GW04 CN67 CN68	GW03 GW04 CN67 CN68	GW03 GW04 CN67 CN68
GW01 GW02 CN65 CN66	GW01 GW02 CN65 CN66	GW01 GW02 CN65 CN66			GW01 GW02 CN65 CN66	GW01 GW02 CN65 CN66	GW01 GW02 CN65 CN66	GW01 GW02 CN65 CN66
IB SW1	IB SW1	IB SW1	IB SW1		IB SW1	IB SW1	IB SW1	IB SW1
CN61 CN62 CN63 CN64	CN61 CN62 CN63 CN64	CN61 CN62 CN63 CN64	GW03 GW04 CN59 CN60		CN61 CN62 CN63 CN64	CN61 CN62 CN63 CN64	CN61 CN62 CN63 CN64	CN61 CN62 CN63 CN64
Terminal Server 1	Terminal Server 1	Terminal Server 1	Terminal Server 1		Terminal Server 1	Terminal Server 1	Terminal Server 1	Terminal Server 1
Mgmt Ethernet Switch 1	Mgmt Ethernet Switch 1	Mgmt Ethernet Switch 1	Mgmt Ethernet Switch 1		Mgmt Ethernet Switch 1	Mgmt Ethernet Switch 1	Mgmt Ethernet Switch 1	Mgmt Ethernet Switch 1
CN57 CN58 CN59 CN60	CN57 CN58 CN59 CN60	CN57 CN58 CN59 CN60	GW01 GW02 CN57 CN58		CN57 CN58 CN59 CN60	CN57 CN58 CN59 CN60	CN57 CN58 CN59 CN60	CN57 CN58 CN59 CN60
CN53 CN54 CN55 CN56	CN53 CN54 CN55 CN56	CN53 CN54 CN55 CN56	CN53 CN54 CN55 CN56	LSM01	CN53 CN54 CN55 CN56	CN53 CN54 CN55 CN56	CN53 CN54 CN55 CN56	CN53 CN54 CN55 CN56
IB SW2	IB SW2	IB SW2	IB SW2	LSM02	IB SW2	IB SW2	IB SW2	IB SW2
CN49 CN50 CN51 CN52	CN49 CN50 CN51 CN52	CN49 CN50 CN51 CN52	CN49 CN50 CN51 CN52	LSM03	CN49 CN50 CN51 CN52	CN49 CN50 CN51 CN52	CN49 CN50 CN51 CN52	CN49 CN50 CN51 CN52
CN45 CN46 CN47 CN48	CN45 CN46 CN47 CN48	CN45 CN46 CN47 CN48	CN45 CN46 CN47 CN48	LSM04	CN45 CN46 CN47 CN48	CN45 CN46 CN47 CN48	CN45 CN46 CN47 CN48	CN45 CN46 CN47 CN48
CN41 CN42 CN43 CN44	CN41 CN42 CN43 CN44	CN41 CN42 CN43 CN44	CN41 CN42 CN43 CN44	RAID01	CN41 CN42 CN43 CN44	CN41 CN42 CN43 CN44	CN41 CN42 CN43 CN44	CN41 CN42 CN43 CN44
IB SW3	IB SW3	IB SW3	IB SW3		IB SW3	IB SW3	IB SW3	IB SW3
CN37 CN38 CN39 CN40	CN37 CN38 CN39 CN40	CN37 CN38 CN39 CN40	CN37 CN38 CN39 CN40	RPS01	CN37 CN38 CN39 CN40	CN37 CN38 CN39 CN40	CN37 CN38 CN39 CN40	CN37 CN38 CN39 CN40
CN33 CN34 CN35 CN36	CN33 CN34 CN35 CN36	CN33 CN34 CN35 CN36	CN33 CN34 CN35 CN36		CN33 CN34 CN35 CN36	CN33 CN34 CN35 CN36	CN33 CN34 CN35 CN36	CN33 CN34 CN35 CN36
CN29 CN30 CN31 CN32	CN29 CN30 CN31 CN32	CN29 CN30 CN31 CN32	CN29 CN30 CN31 CN32	RAID02	CN29 CN30 CN31 CN32	CN29 CN30 CN31 CN32	CN29 CN30 CN31 CN32	CN29 CN30 CN31 CN32
IB SW4	IB SW4	IB SW4	IB SW4		IB SW4	IB SW4	IB SW4	IB SW4
CN25 CN26 CN27 CN28	CN25 CN26 CN27 CN28	CN25 CN26 CN27 CN28	CN25 CN26 CN27 CN28	RPS02	CN25 CN26 CN27 CN28	CN25 CN26 CN27 CN28	CN25 CN26 CN27 CN28	CN25 CN26 CN27 CN28
Terminal Server 2	Terminal Server 2	Terminal Server 2	Terminal Server 2		Terminal Server 2	Terminal Server 2	Terminal Server 2	Terminal Server 2
Mgmt Ethernet Switch 2	Mgmt Ethernet Switch 2	Mgmt Ethernet Switch 2	Mgmt Ethernet Switch 2	RAID03	Mgmt Ethernet Switch 2	Mgmt Ethernet Switch 2	Mgmt Ethernet Switch 2	Mgmt Ethernet Switch 2
CN21 CN22 CN23 CN24	CN21 CN22 CN23 CN24	CN21 CN22 CN23 CN24	CN21 CN22 CN23 CN24		CN21 CN22 CN23 CN24	CN21 CN22 CN23 CN24	CN21 CN22 CN23 CN24	CN21 CN22 CN23 CN24
CN17 CN18 CN19 CN20	CN17 CN18 CN19 CN20	CN17 CN18 CN19 CN20	CN17 CN18 CN19 CN20	RPS03	CN17 CN18 CN19 CN20	CN17 CN18 CN19 CN20	CN17 CN18 CN19 CN20	CN17 CN18 CN19 CN20
IB SW5	IB SW5	IB SW5	IB SW5		IB SW5	IB SW5	IB SW5	IB SW5
CN13 CN14 CN15 CN16	CN13 CN14 CN15 CN16	CN13 CN14 CN15 CN16	CN13 CN14 CN15 CN16		CN13 CN14 CN15 CN16	CN13 CN14 CN15 CN16	CN13 CN14 CN15 CN16	CN13 CN14 CN15 CN16
CN09 CN10 CN11 CN12	CN09 CN10 CN11 CN12	CN09 CN10 CN11 CN12	CN09 CN10 CN11 CN12	RAID04	CN09 CN10 CN11 CN12	CN09 CN10 CN11 CN12	CN09 CN10 CN11 CN12	CN09 CN10 CN11 CN12
CN05 CN06 CN07 CN08	CN05 CN06 CN07 CN08	CN05 CN06 CN07 CN08	CN05 CN06 CN07 CN08		CN05 CN06 CN07 CN08	CN05 CN06 CN07 CN08	CN05 CN06 CN07 CN08	CN05 CN06 CN07 CN08
IB SW6	IB SW6	IB SW6	IB SW6	RPS04	IB SW6	IB SW6	IB SW6	IB SW6
CN01 CN02 CN03 CN04	CN01 CN02 CN03 CN04	CN01 CN02 CN03 CN04	CN01 CN02 CN03 CN04		CN01 CN02 CN03 CN04	CN01 CN02 CN03 CN04	CN01 CN02 CN03 CN04	CN01 CN02 CN03 CN04

Twice the density of standard 1U nodes with 72 nodes per rack

Layout for Hyperion Phase 1 and 2 with IBA Switches, Storage, and Expansion

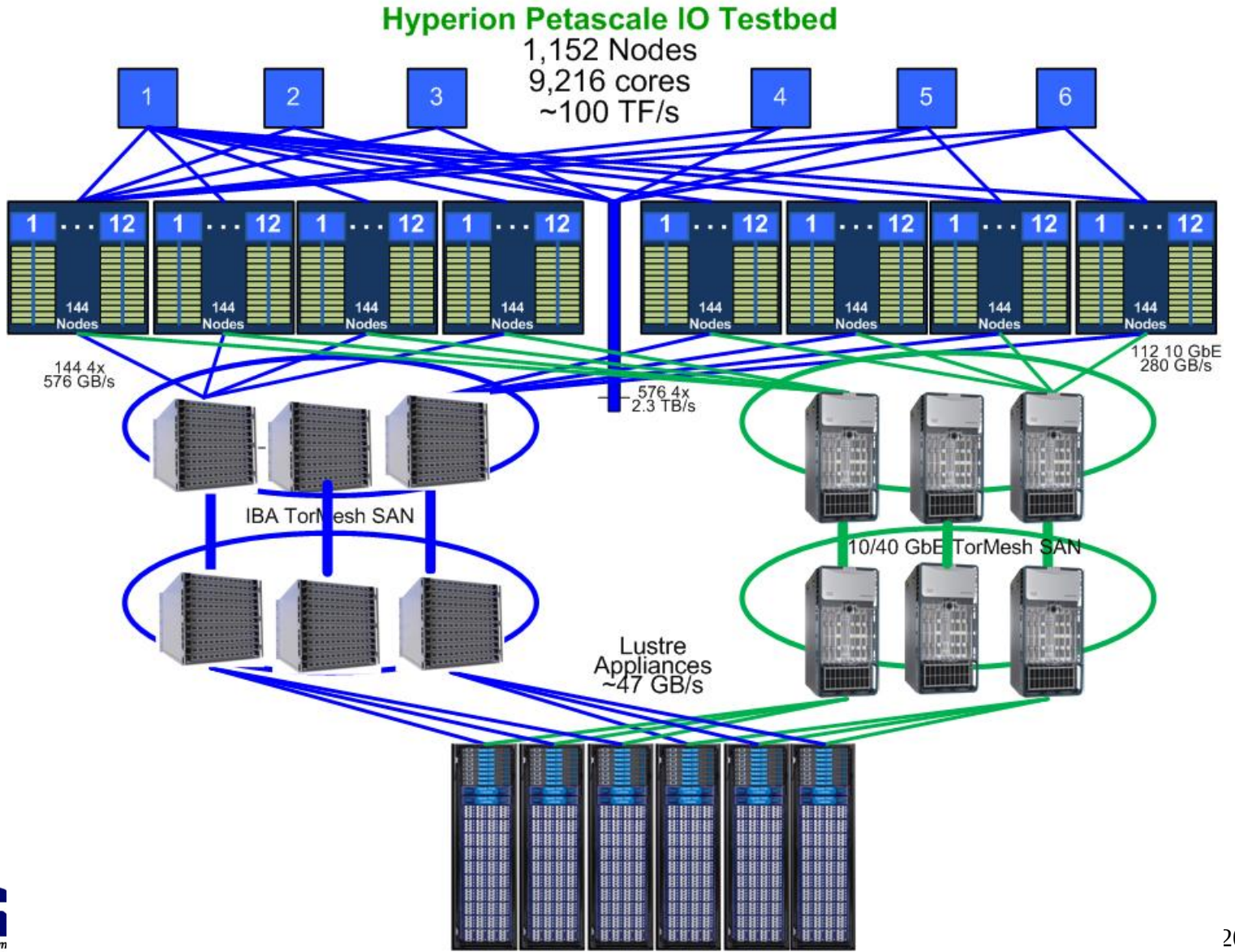


Siting at LLNL on the open collaboration network

Ongoing Hyperion Testing

- File System
  - Sun File system testing Lustre 1.8 pre-releases
  - Finding scalability issues early in the devel cycle
  - All other file systems running 1.6.x
  - Evaluating storage hardware
- SAN
  - Fat-Tree like & TorMesh topologies
  - IBA DDR/QDR & 10/40GbE
- System software
  - Working on “Intel Cluster Ready” certification
  - RedHat pre-release testing

Hyperion petascale I/O testbed will test Sequoia I/O technologies at scale!



Petascale I/O Challenges

- Challenges
  - Stability, stability, stability
  - Scalability
  - Performance
  - Shared storage/Site wide file systems
- Disruptive technologies
  - Solid state storage
  - Virtualization
- New tier of solid state storage will expose new HW/SW bottlenecks

Hyperion Future Directions

- OpenFabrics OFED pre-release testing
- Intel Nehalem integration
- System Software
  - pNFS and other open source file systems
  - Virtualization
- Hyperion Data Intensive Computing
  - Upgrade to QDR
  - Expand filesystem to >512 GB/s & >100M IOPs
  - Expand SAN environments from 10's to 100's of ports

Questions?