Lawrence Livermore National Laboratory

Lustre on Hyperion

Marc Stearman

marc@llnl.gov

April, 2009

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-PRES-370578

Disclaimer

Hyperion Overview

- Goals and Purpose
- Design and Description
- Networks and Lustre File Systems
- Ease of Management and Flexibility

Contributions

Intel, Dell, Supermicro	Qlogic, Cisco, Mellanox	DDN, Sun, LSI
• Processors • Nodes • Racks • Integration	• IB & Ethernet switches • IB & Ethernet HCAs • IB⇔Ethernet Routers	• Storage Hardware

Redhat	Tri-Lab
• Linux Testing • System Admin	• Knowledge • Expertise • Facilities

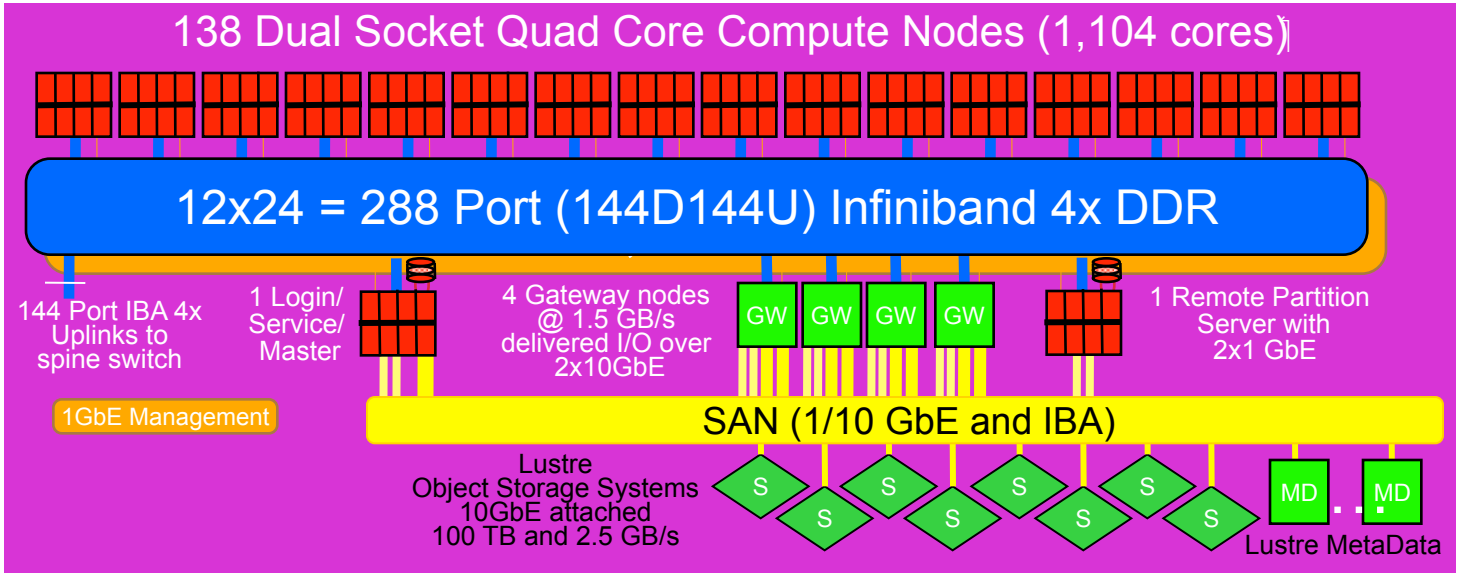
Hyperion: A Large Scalable Testbed

- Development and Testing Environment
  - Infiniband Open Source Software - OFED
  - Lustre Production Scaling
  - TOS software stack development & testbed
- Evaluation Testbed For New Hardware & Software
  - Petascale I/O scaling for Sequoia and beyond
  - Processor, memory, networking, storage, etc.
  - Designed for future technology refresh
  - Virtualization
- Vehicle for long term vendor and customer partnerships
  - OpenFabrics Alliance
  - Lustre Center of Excellence
  - Sun

Hyperion Scalable Unit

- RPS (Management node)
  - Contains diskless images for nodes
- Login
  - Interactive use nodes
  - Good for compiles, launching jobs, examining files, etc.
- Lustre LNET Routers
  - 2 Ethernet
  - 2 Infiniband
- Batch compute nodes: 138 nodes available

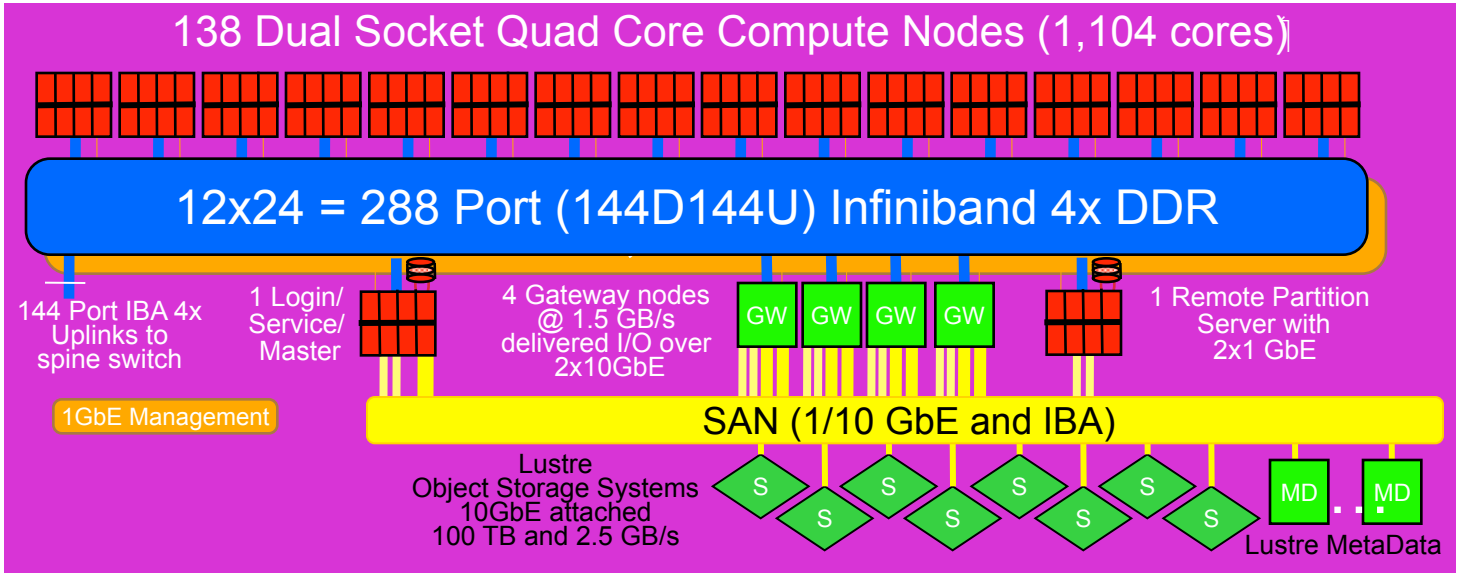
Hyperion Phase 1
12 TF/s Scalable Unit



Hyperion Phase 1 Deployment is an 4 SU 48 TF/s Cluster with full IBA

- 576 Total nodes and 4,608 cores, 12.1 TB/s memory bandwidth, 9.2 TB capacity
- IBA is expandable to 1,728 IB ports single plane and can double to dual plane
- Dual socket 2.5 GHz quad-core Intel 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
- 250 kW of power, 70 tons of cooling.
- Storage Scalable Units (SSU) from DDN, LSI and Sun yielding >36GB/s and 1.6 PB disk

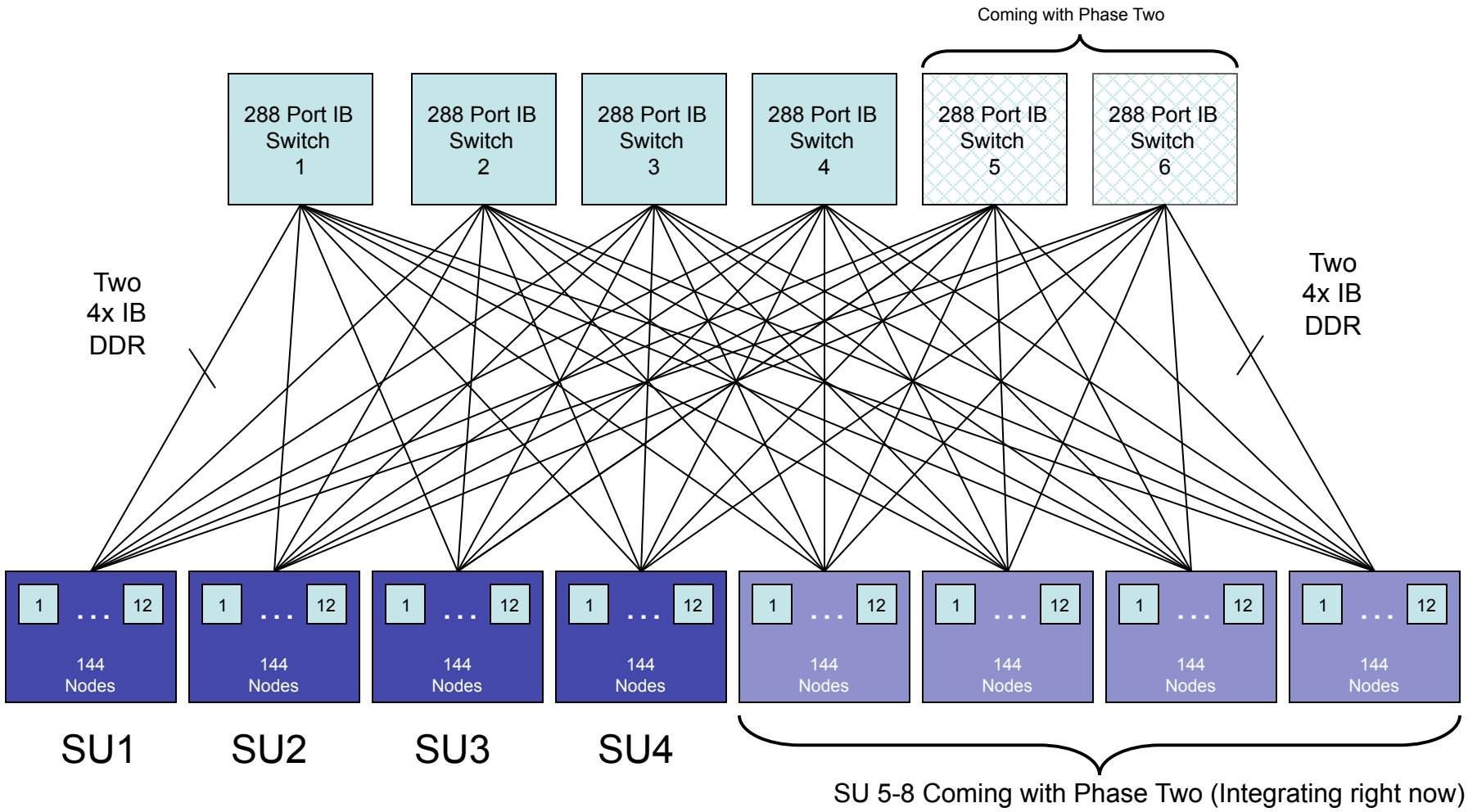
Hyperion Phase 2
18.4 TF/s Scalable Unit



Hyperion Phase 2 Deployment is an 4 SU 74 TF/s Cluster with full IBA

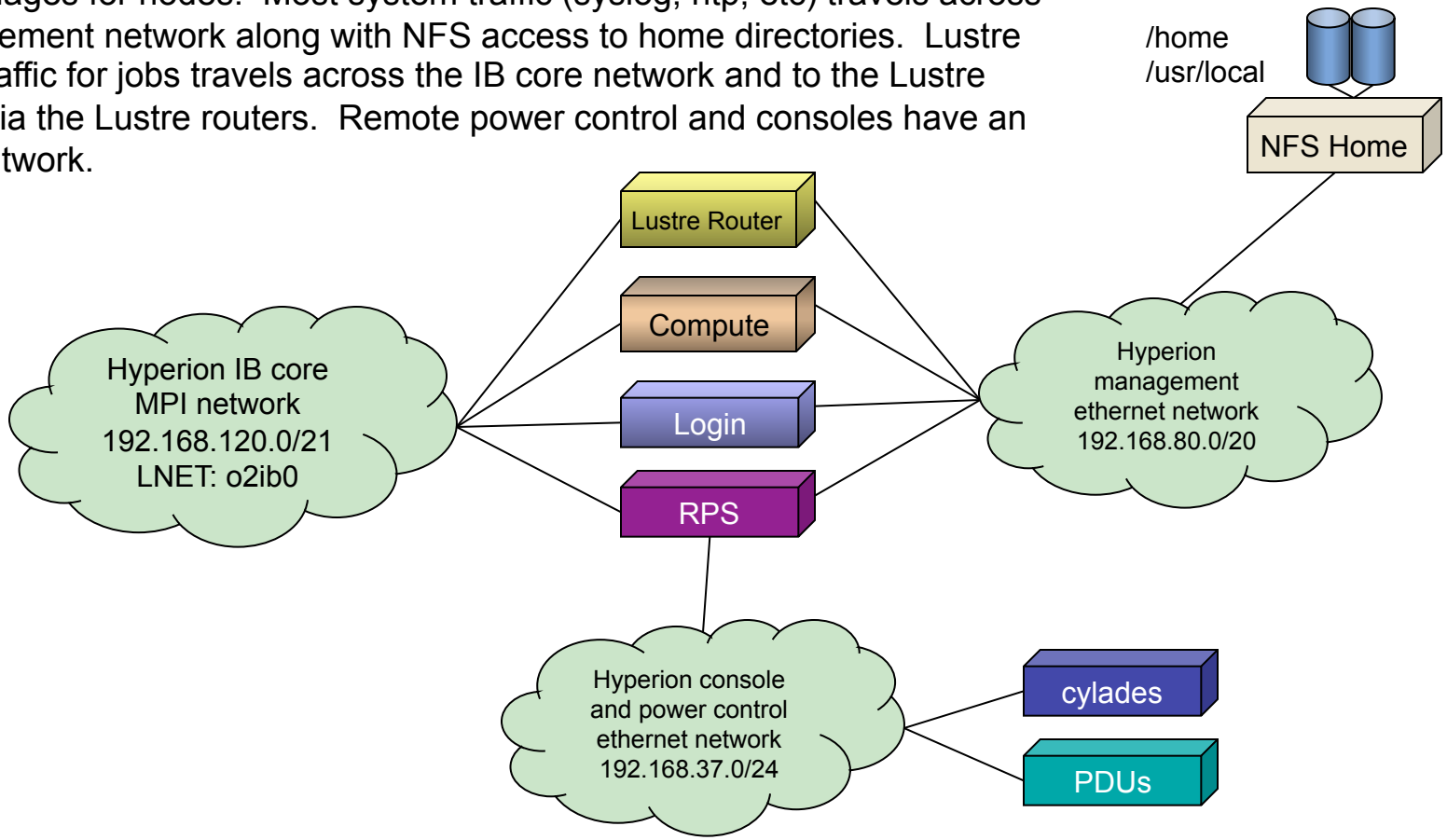
- 576 Total nodes and 4,608 cores, 37 TB/s memory bandwidth, 9.2 TB capacity
- IBA is expandable to 1,728 IB ports single plane and can double to dual plane
- Dual socket 2.4 GHz quad-core Intel Nehalem nodes
  - 12 GB from 6 channels of 1333 DDR3 SDRAM @ 64 GB/s bandwidth
- Nodes utilize PCI-Express generation 2 I/O which provides an upgrade path to IBA 4x QDR
- 250 kW of power, 70 tons of cooling.
- Storage Scalable Units (SSU) from DDN, LSI and Sun yielding >36GB/s and 1.6 PB disk

Hyperion Infiniband Core Diagram

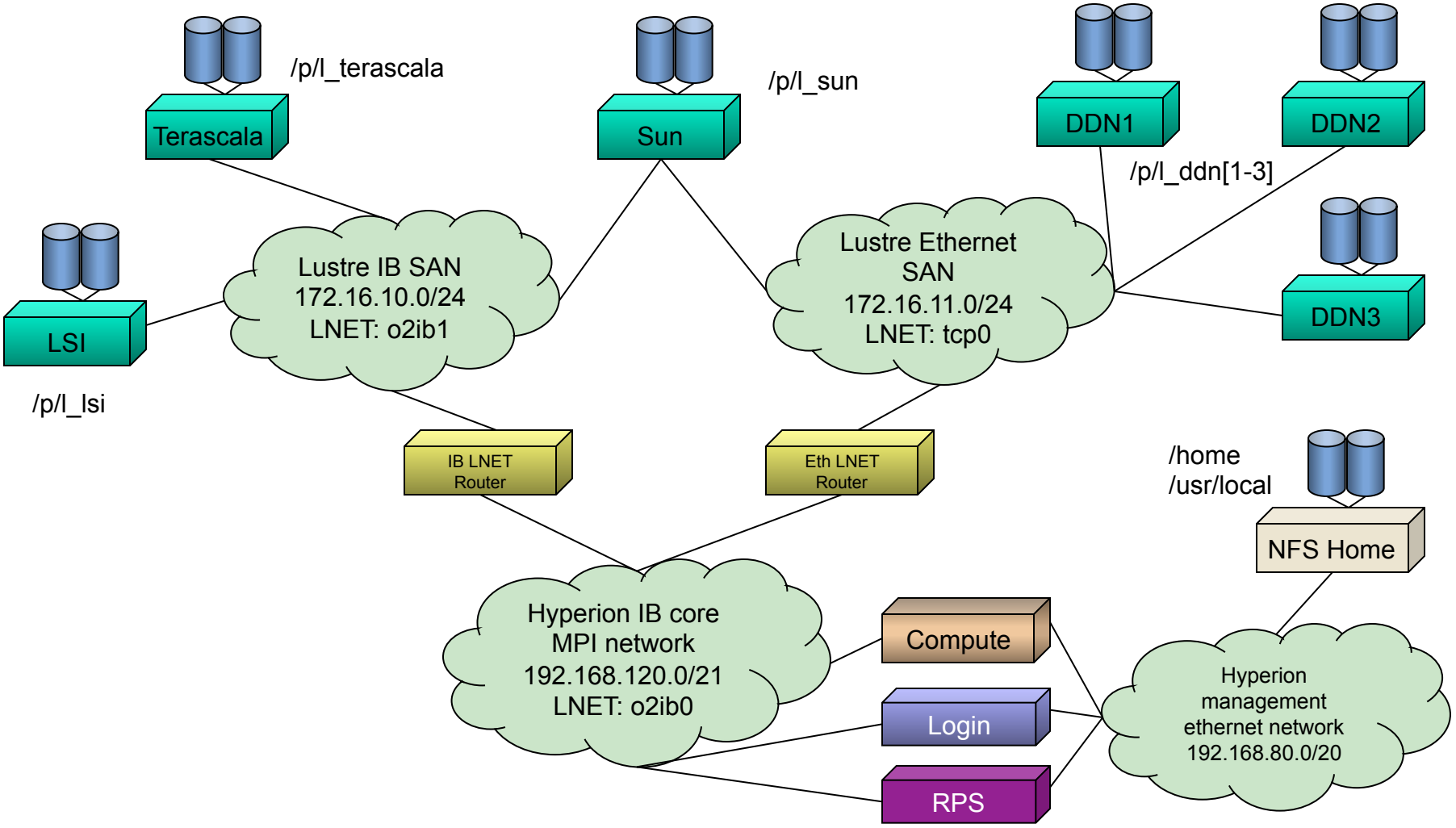


Hyperion Compute Core

Nodes dhcp boot across the management network. The RPS nodes contain the diskless images for nodes. Most system traffic (syslog, ntp, etc) travels across the management network along with NFS access to home directories. Lustre and MPI traffic for jobs travels across the IB core network and to the Lustre networks via the Lustre routers. Remote power control and consoles have an isolated network.



Hyperion File Systems and Networks



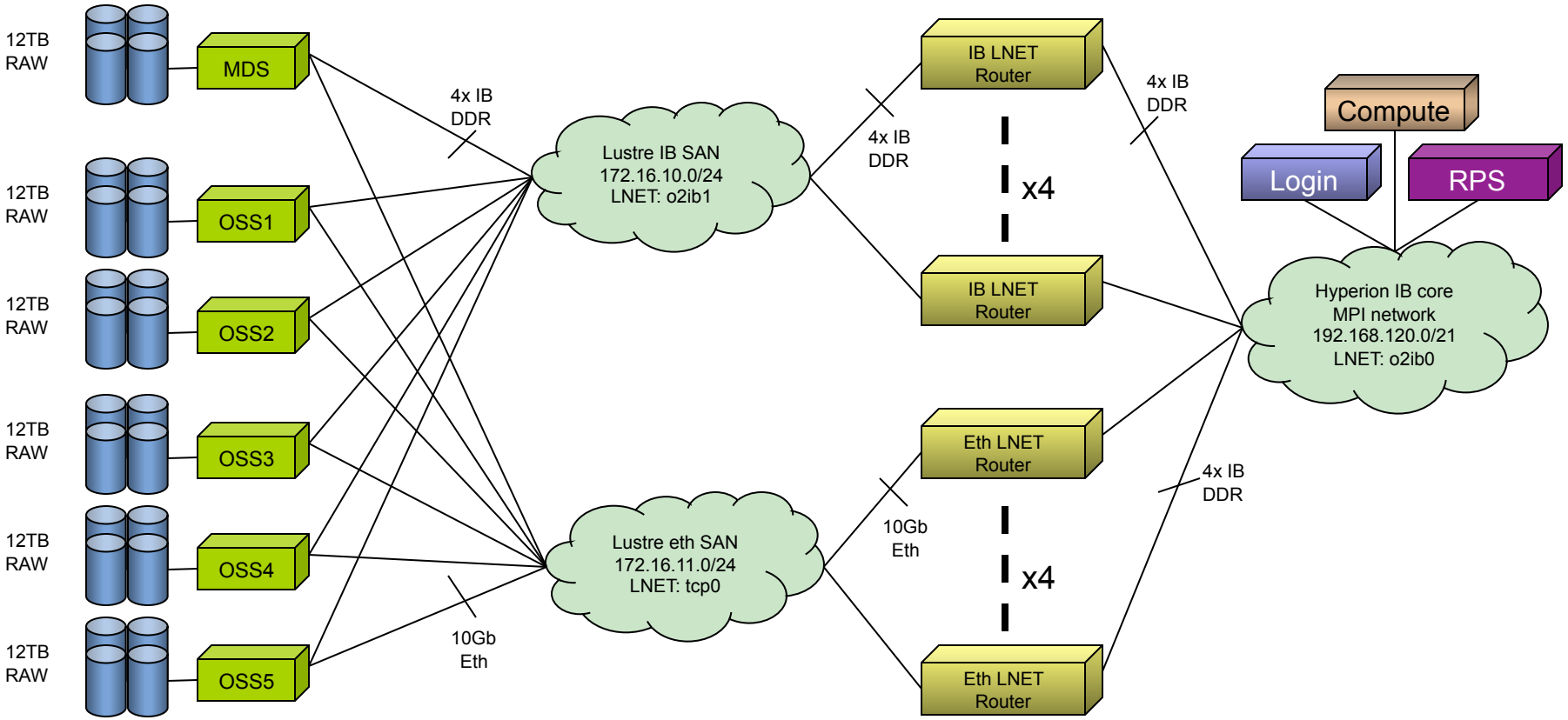
Multiple Lustre Versions

- Sun File system testing lustre 1.8
- All other file systems running 1.6.x
- Clients and routers have been segrated to run 1.6 or 1.8. Sun and LLNL started Interoperability testing a couple weeks ago.
- A slurm partition called *lustre_18* has been created for compute nodes that mount lustre 1.8. The 1.6 clients are still in the *pbatch* partition

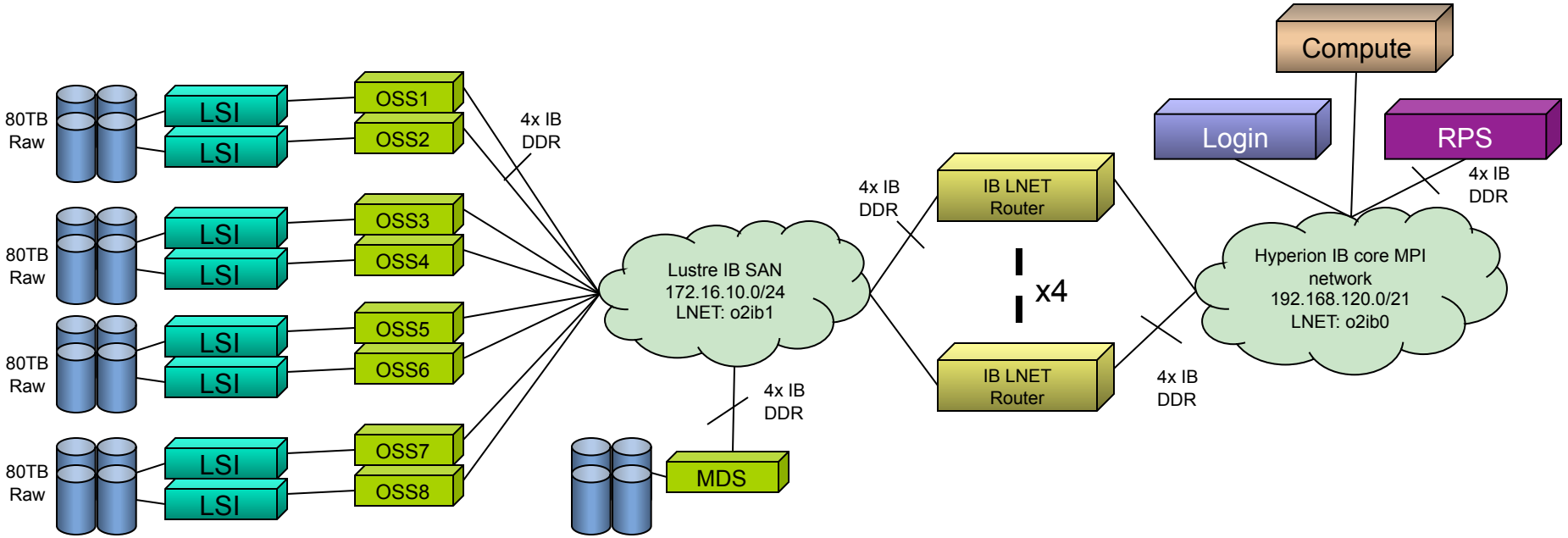
Flexible Configuration

- Servers, Clients and Routers run diskless nfsroot images
- Can be quickly switched to new software stack with a reboot
- Slurm partitions can be created to group nodes with similar features/software stacks
- Having six lustre file systems allows us to test many different versions
- Size of Hyperion allows us to test and reproduce bugs we see in production

Sun Lustre File system – 35TB



LSI Lustre File System – 284TB

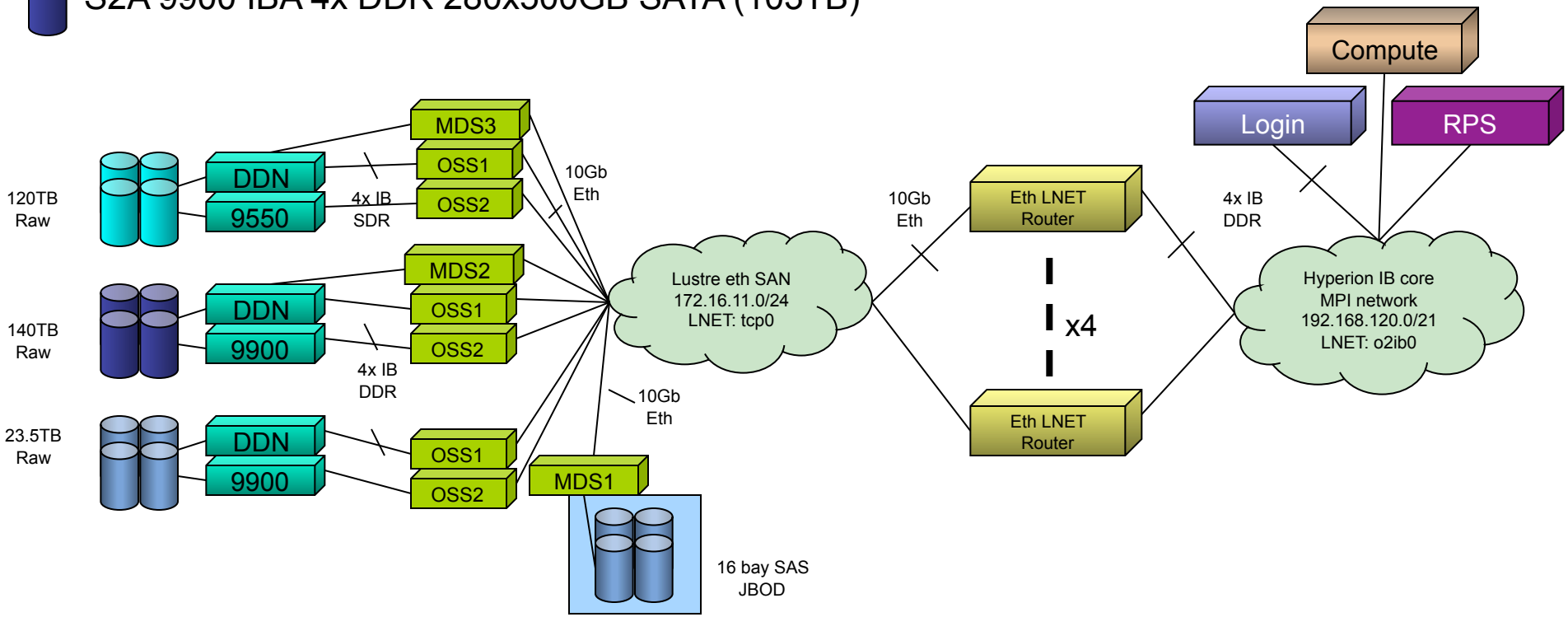


DDN Lustre Network Diagram

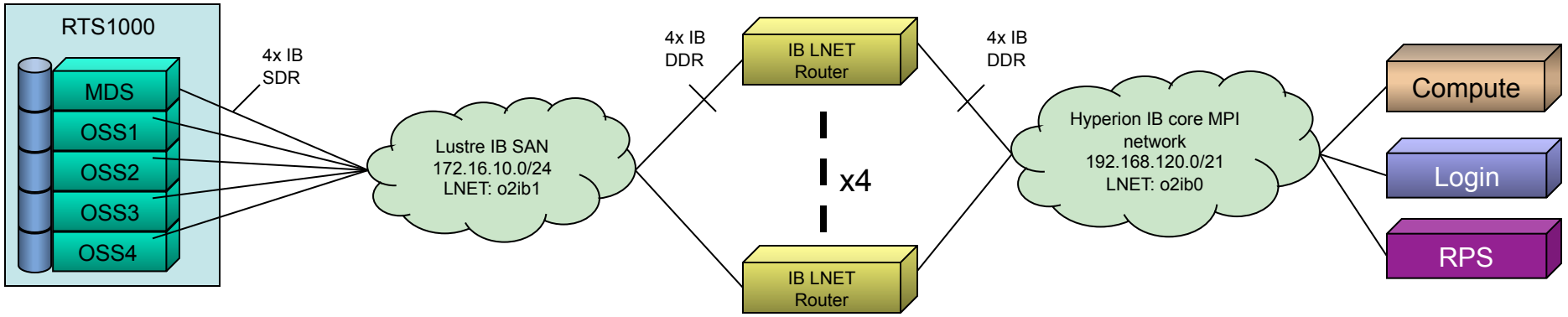
S2A 9900 IBA 4x DDR 160x147GB SAS (19TB)

S2A 9550 IBA 4x SDR 512x250GB SATA (91TB)

S2A 9900 IBA 4x DDR 280x500GB SATA (103TB)



Terascala Lustre File System – 14TB



Discussion