Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.
Hyperion Overview

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

<table>
<thead>
<tr>
<th>Intel, Dell, Supermicro</th>
<th>Qlogic, Cisco, Mellanox</th>
<th>DDN, Sun, LSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Processors</td>
<td>• IB &amp; Ethernet switches</td>
<td>• Storage Hardware</td>
</tr>
<tr>
<td>• Nodes</td>
<td>• IB &amp; Ethernet HCAs</td>
<td></td>
</tr>
<tr>
<td>• Racks</td>
<td>• IB- Ethernet Routers</td>
<td></td>
</tr>
<tr>
<td>• Integration</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Redhat
- Linux Testing
- System Admin

### Tri-Lab
- 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 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 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

288 Port IB Switch 1
288 Port IB Switch 2
288 Port IB Switch 3
288 Port IB Switch 4
288 Port IB Switch 5
288 Port IB Switch 6

SU1  SU2  SU3  SU4

144 Nodes  144 Nodes  144 Nodes  144 Nodes

Two 4x IB DDR

Coming with Phase Two

SU 5-8 Coming with Phase Two (Integrating right now)
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 segregated 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

12TB RAW
MDS
OSS1
OSS2
OSS3
OSS4
OSS5

Lustre IB SAN
172.16.10.0/24
LNET: o2ib1

IB LNET Router

Lustre eth SAN
172.16.11.0/24
LNET: tcp0

Eth LNET Router

10Gb Eth

4x IB DDR

12TB RAW

12TB RAW

12TB RAW

12TB RAW

12TB RAW

Hyperion IB core
MPI network
192.168.120.0/21
LNET: o2ib0

IB LNET Router

Eth LNET Router

4x IB DDR

4x IB DDR

4x IB DDR

4x IB DDR

Login

RPS

Compute
LSI Lustre File System – 284TB

LSI

OSS1

OSS2

4x IB DDR

80TB Raw

80TB Raw

80TB Raw

OSS3

OSS4

4x IB DDR

OSS5

OSS6

4x IB DDR

OSS7

OSS8

MDS

4x IB DDR

Lustre IB SAN 172.16.10.0/24
LNET: o2ib1

LNET: o2ib0

IB LNET Router

IB LNET Router

4x IB DDR

4x IB DDR

4x IB DDR

LNET: o2ib1

4x IB DDR

4x IB DDR

4x IB DDR

x4

Hyperion IB core MPI network
192.168.120.0/21
LNET: o2ib0

Login

RPS

Compute
Terascala Lustre File System – 14TB
Discussion