The BlueGene/L Computing Environment

Mark Seager

Lawrence Livermore National Laboratory

October 15, 2003

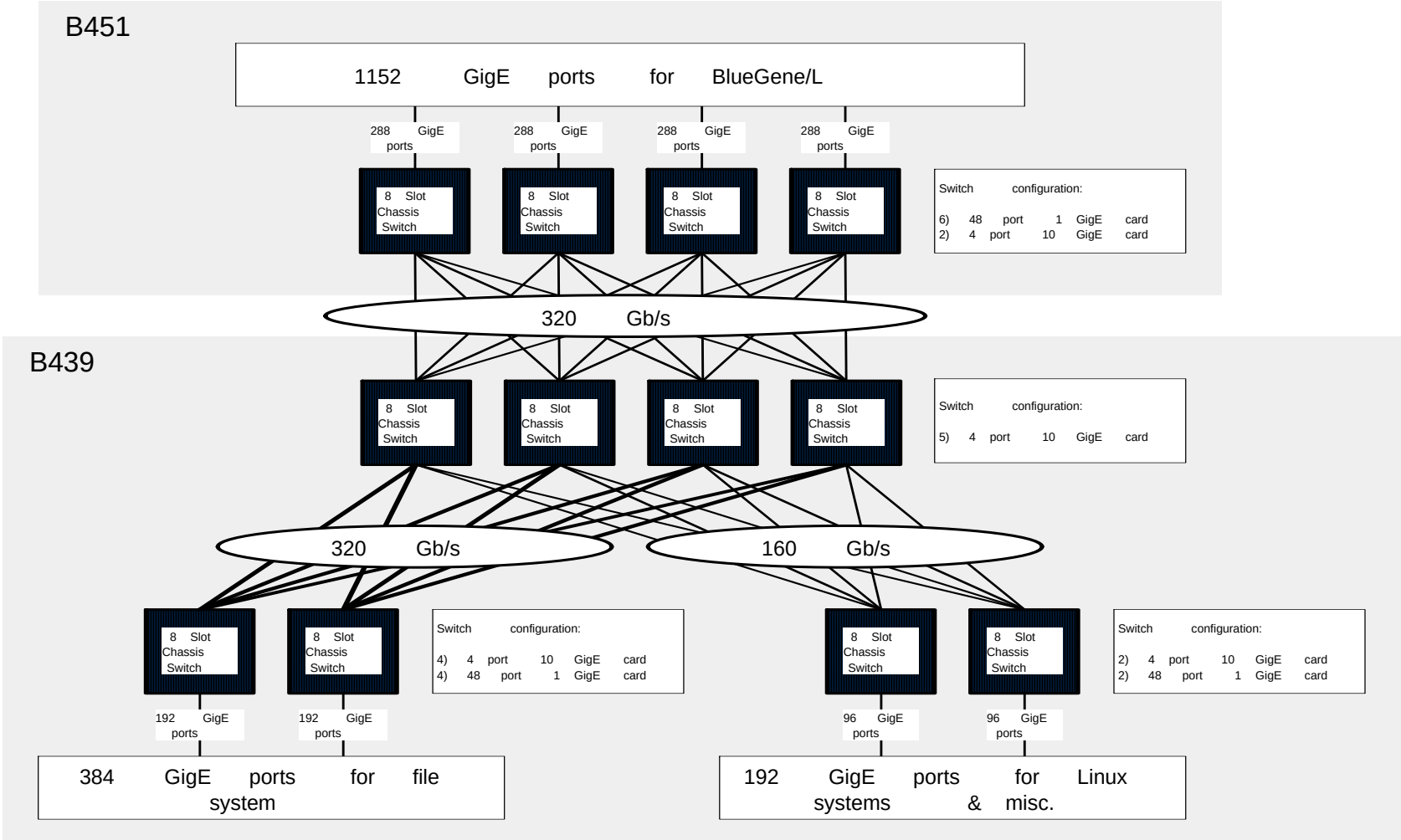
This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

UCRL-PRES-203309

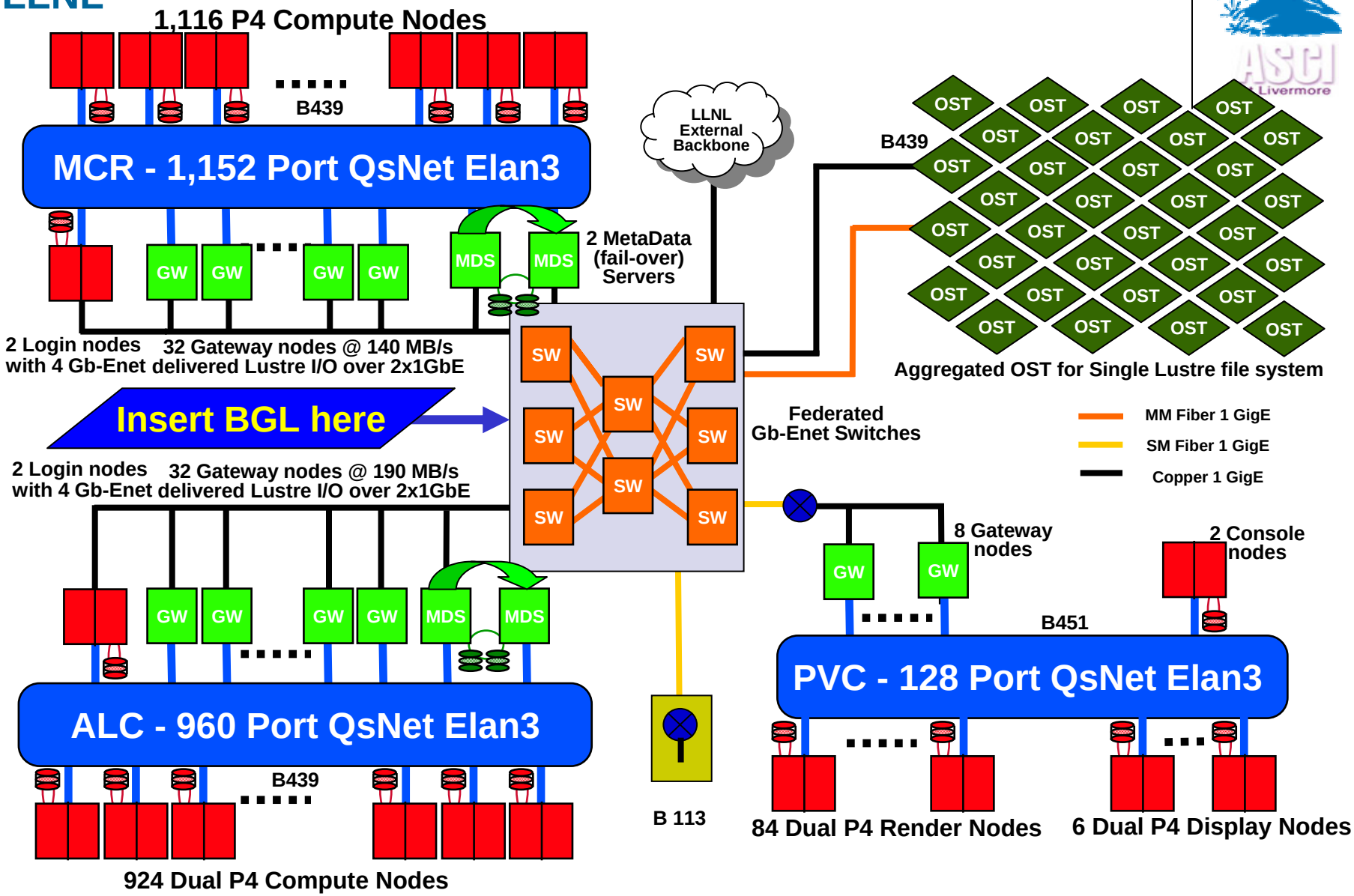
Overview

- BlueGene/L represents a fundamentally new computational science research platform for ASCI
  - Computation environment is critical for overall usefulness of machine
- LLNL is responsible for the BlueGene/L Computation Environment
  - Federated 1 Gb/s Ethernet switches
  - File system
  - HPSS archive
  - Parallel visualization
- BlueGene/L partners will all contribute technology
  - Partnership evaluating two possible file systems for BlueGene/L
- BlueGene/L computation environment will require a highly scalable solution

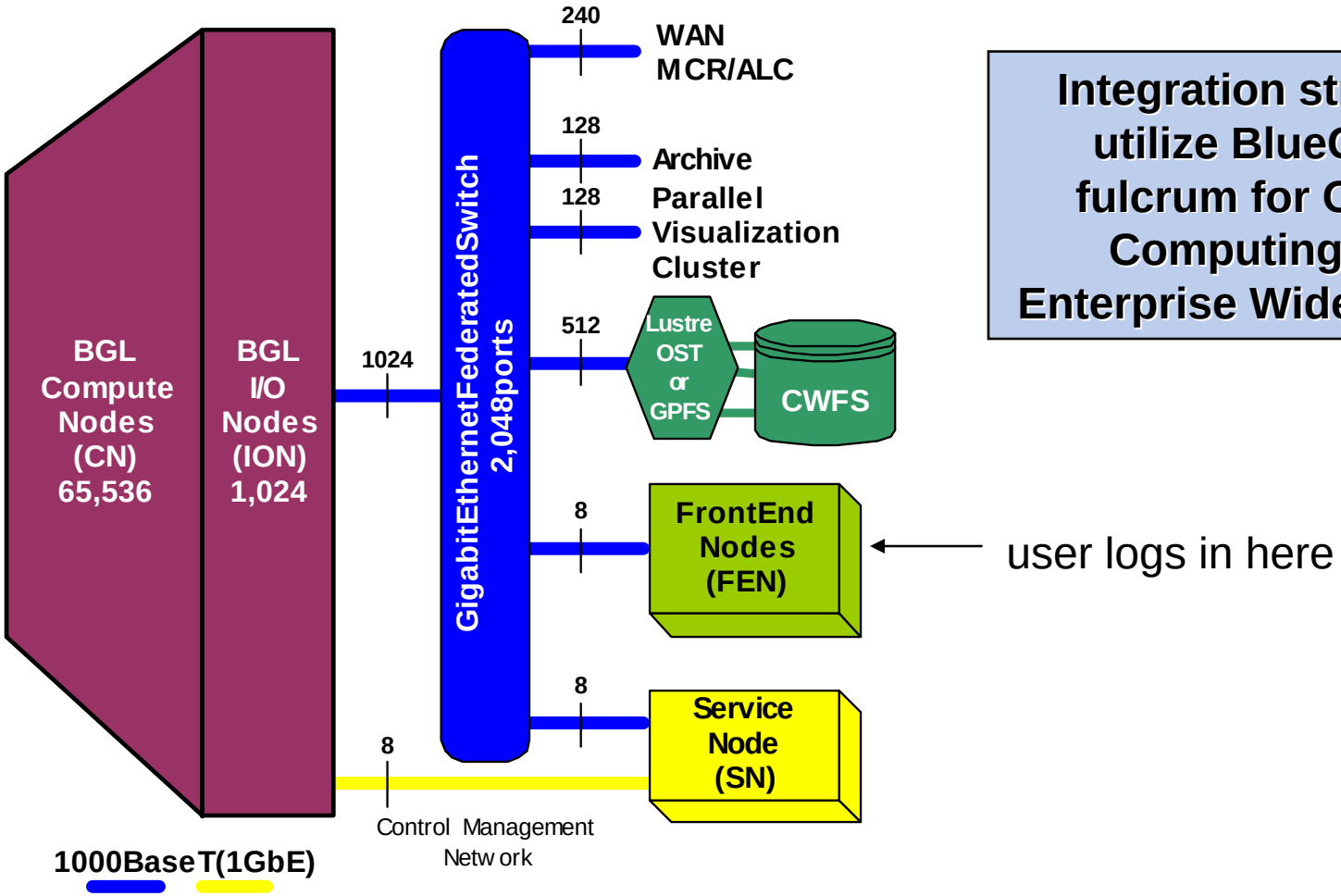
Federated 1 Gb Ethernet switching infrastructure requires distributed network between multiple building



Multi-cluster Simulation Environment based on a single Lustre File System is already impacting every program at LLNL



BlueGene/L IO architecture presents opportunity for highly integrated simulation environment



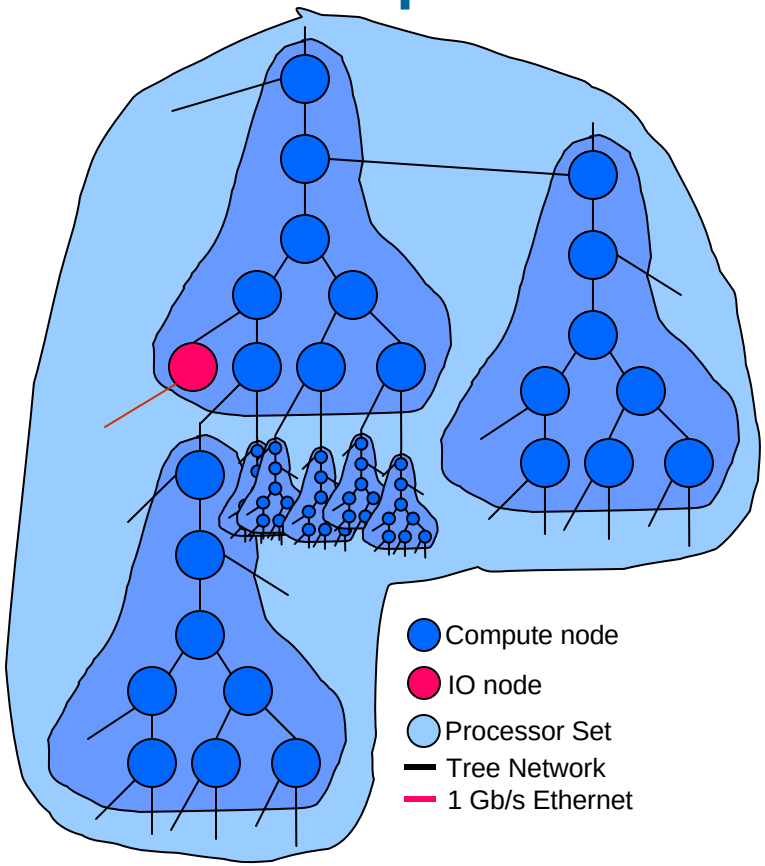
Integration strategy is to utilize BlueGene/L as fulcrum for OCF (Open Computing Facility) Enterprise Wide File System

user logs in here

BlueGene/L file system strategy

- Build machine with same file system as machine will be deployed with
- ASCI Linux Cluster (ALC) nodes and disk will be used for file system resources
- Choice between
  - GPFS – NFSv4 global export of GPFS file system on ALC. NFSv4 client on BGL IO nodes.
  - Lustre – Global file system with OST, MDS and client components. Lustre Client on BGL IO nodes.
- Evaluation 2H2003
- Following slides assume Lustre

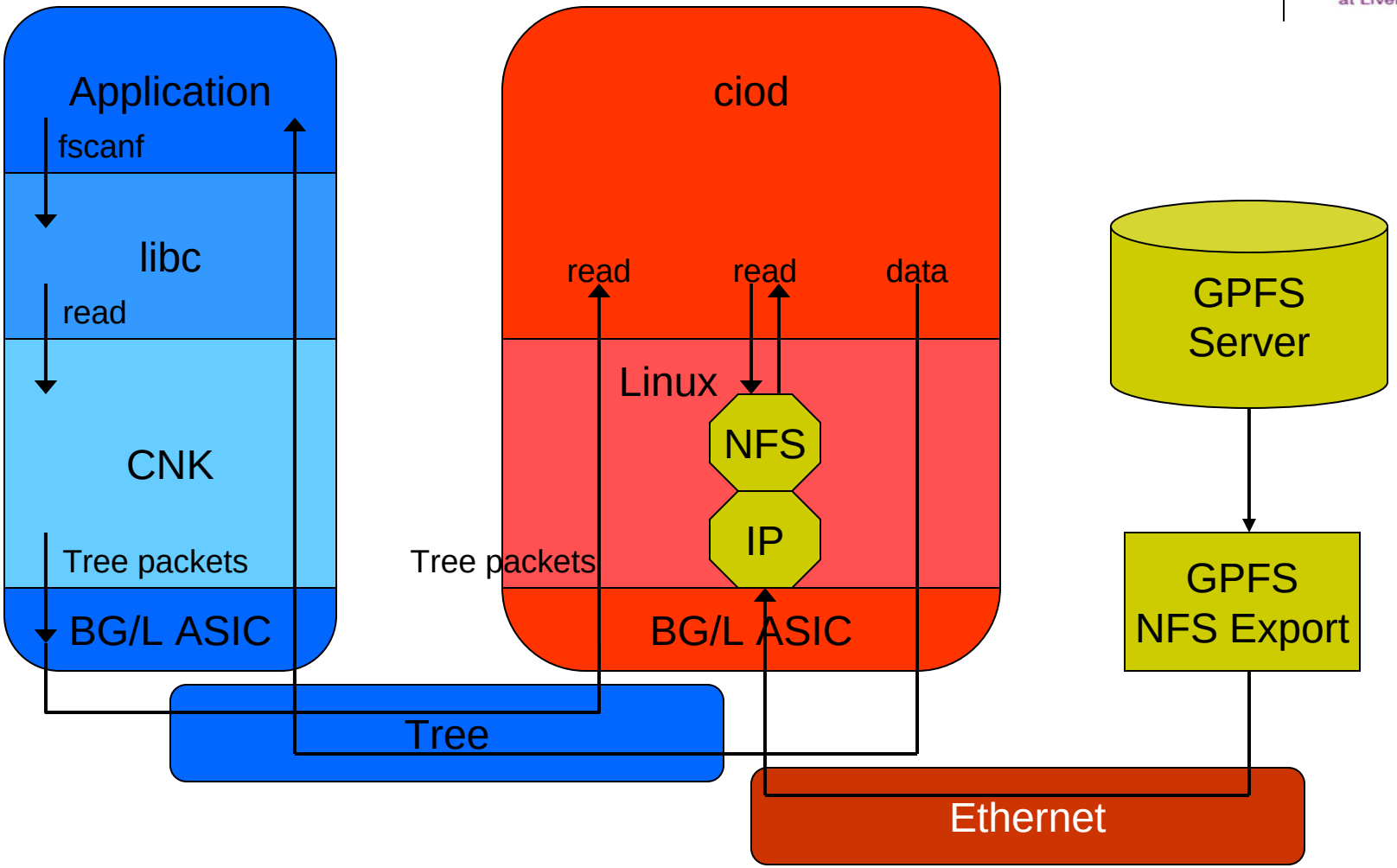
File I/O communication path between I/O node and its 64 compute nodes is the tree network



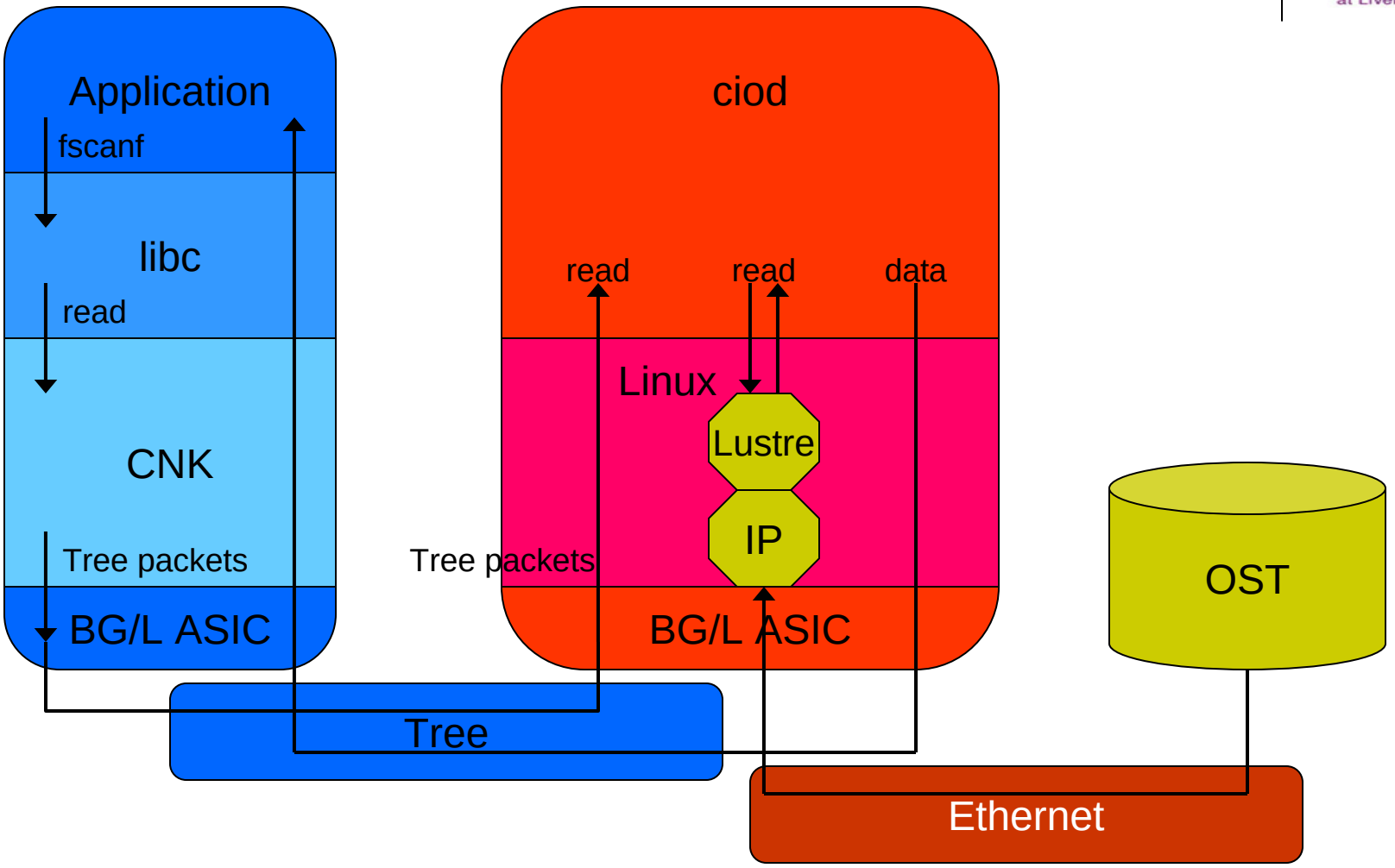
- **IO sub-tree consists of 8 compute nodes**
- **BGL has 8 IO sub-trees per I/O node**

- Tree link is 2b+2b @ 1.4GHz = 350+350MB/s
  - 350 MB/s broadcast from IO node to 64 compute nodes
  - 5.5+5.5 MB/s effective simultaneous point to point bandwidth/compute node
- I/O link to file system is 1 Gb/s Ethernet
  - Achievable bandwidth below about 90+90 MB/s
- Aggregate tree BW through a node is 2.1 GB/s
  - Achievable with multiple global combine operations
- Arithmetic operations implemented in tree
  - Integer/FP max/min
  - Integer add/subtract, bitwise logical ops
- Latency of tree less than 2.5 µs to top, additional 2.5 µs to broadcast to all
- Global sum over 64K in less than 2.5 µs (to top of tree)
- Partitioned with Torus boundaries
- Flexible local routing table

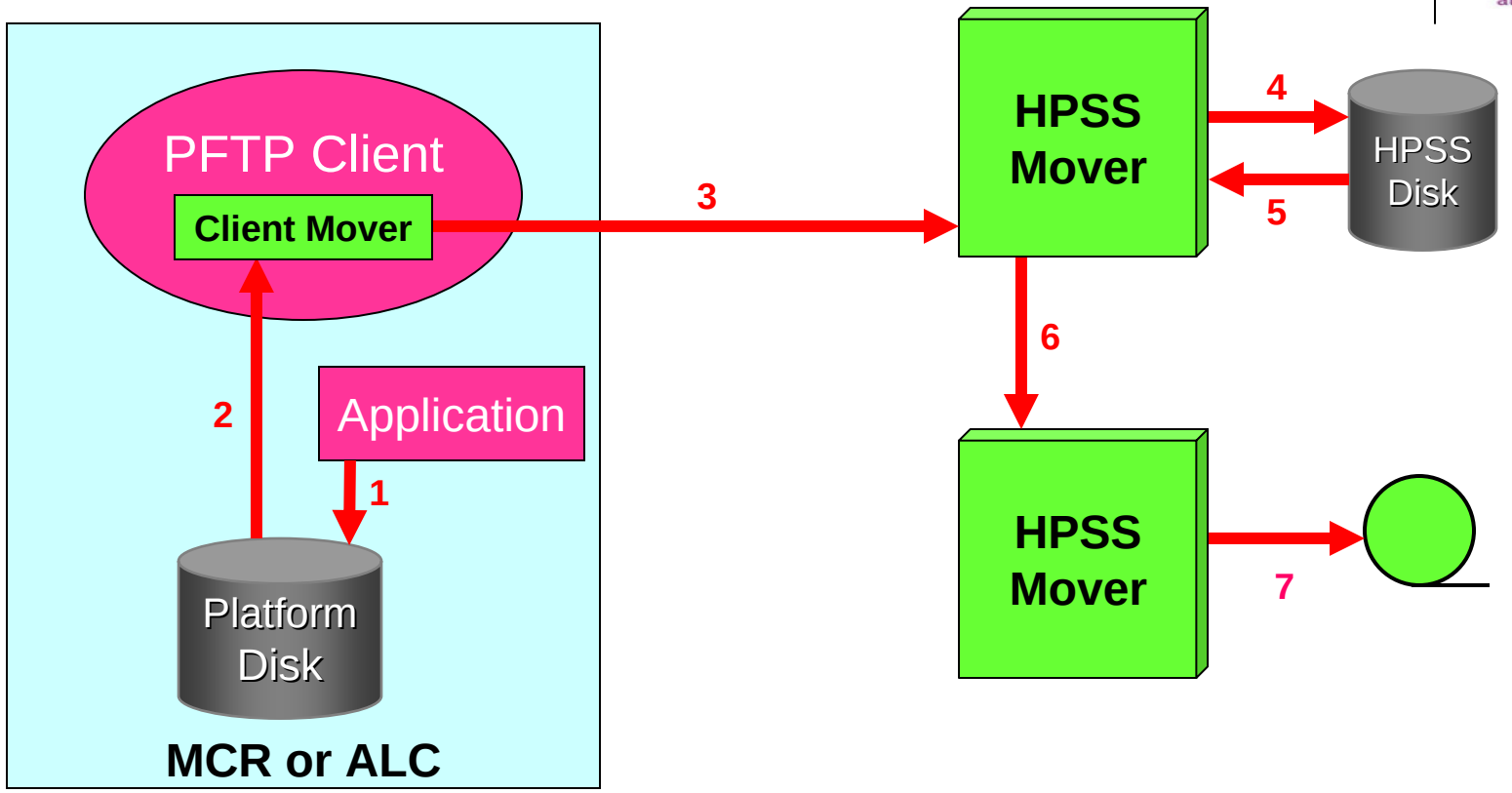
File I/O in BlueGene/L with NFS/GPFS



File I/O in BlueGene/L with Lustre

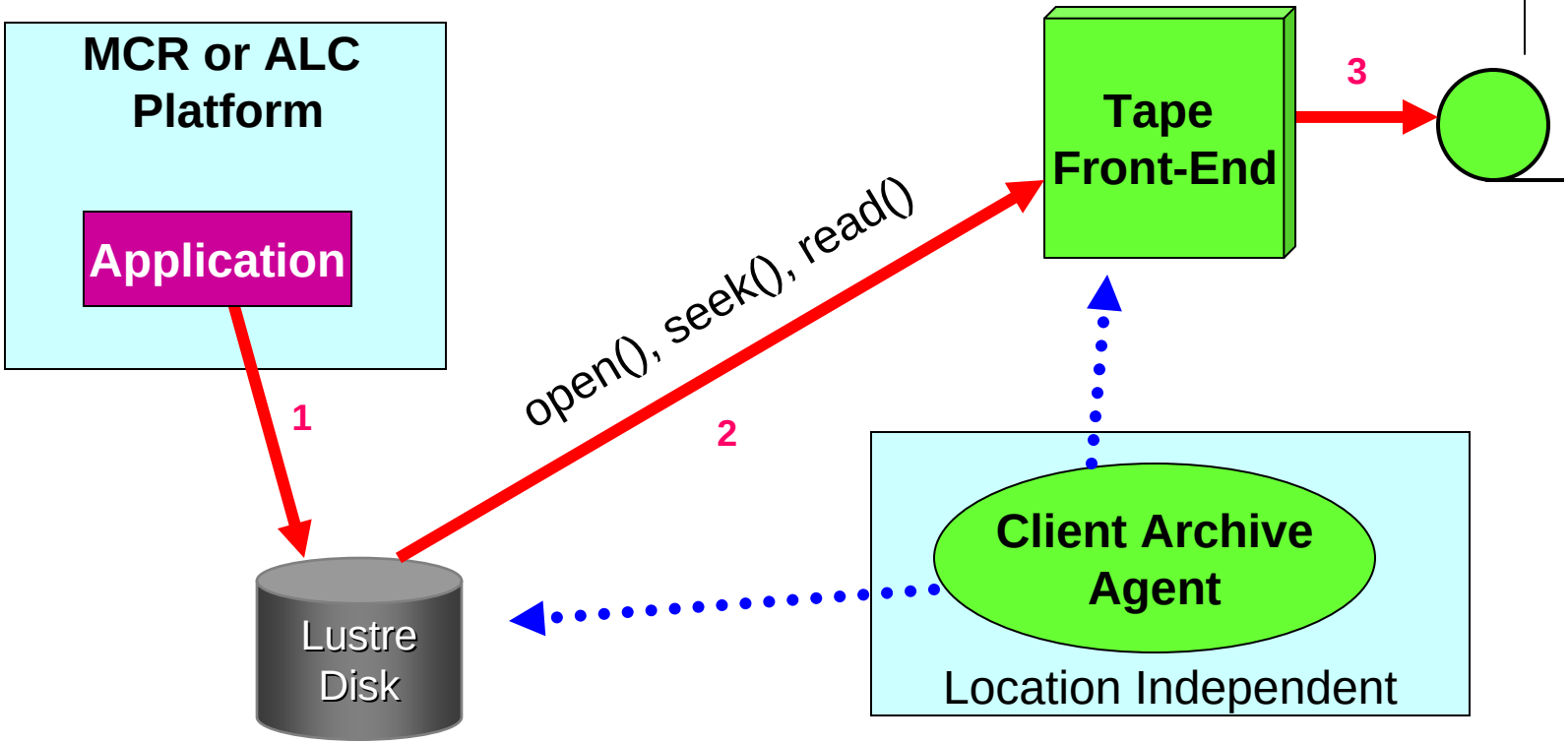


Current HPSS Data Movement



Key elements
1. User in the loop to force keep/delete decision
2. Large HPSS disk cache
3. Bandwidth limited by PFTP bandwidth off compute platform

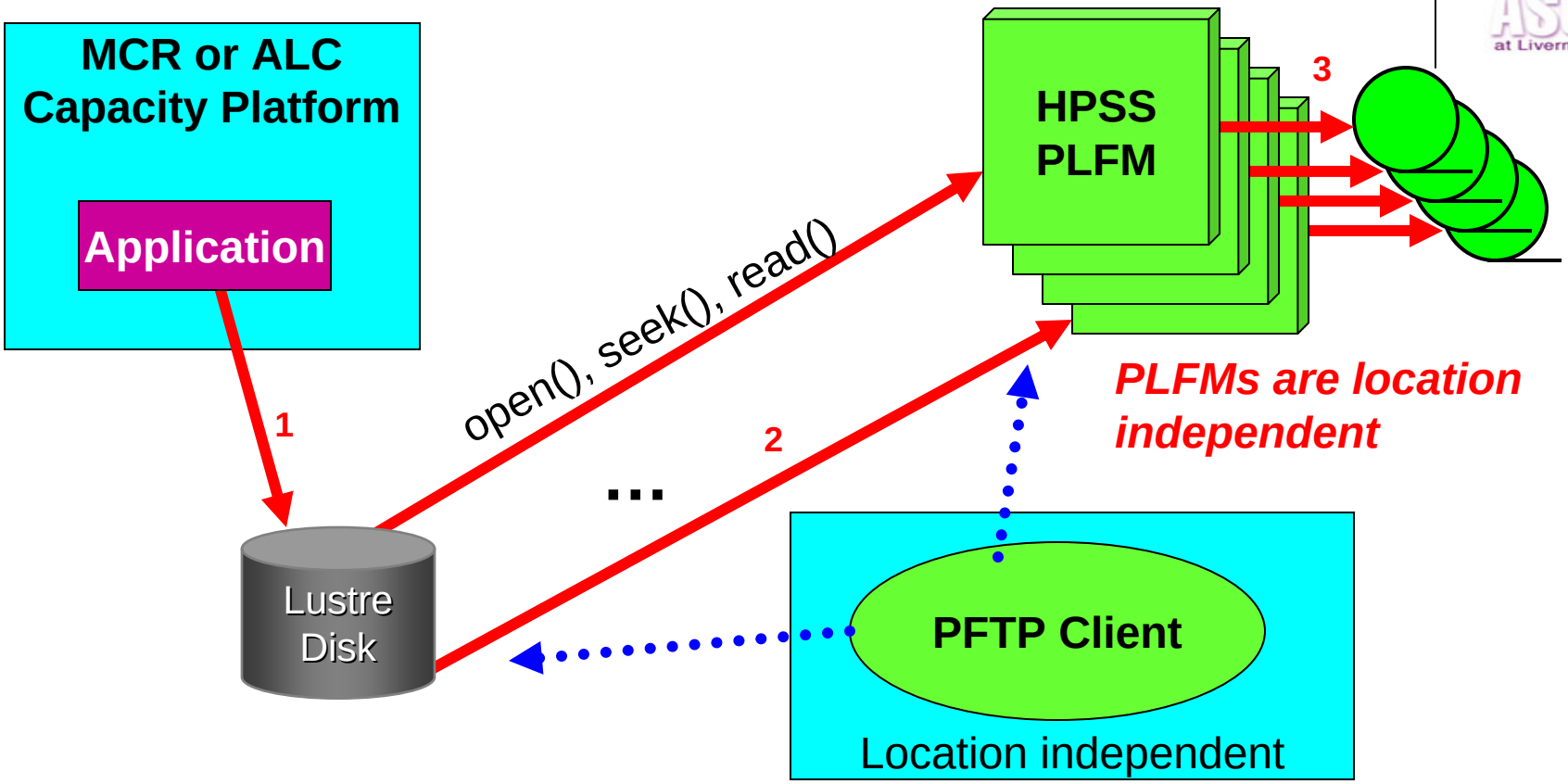
Lustre HPSS Data Movement Vision



Global Accessible Lustre OST

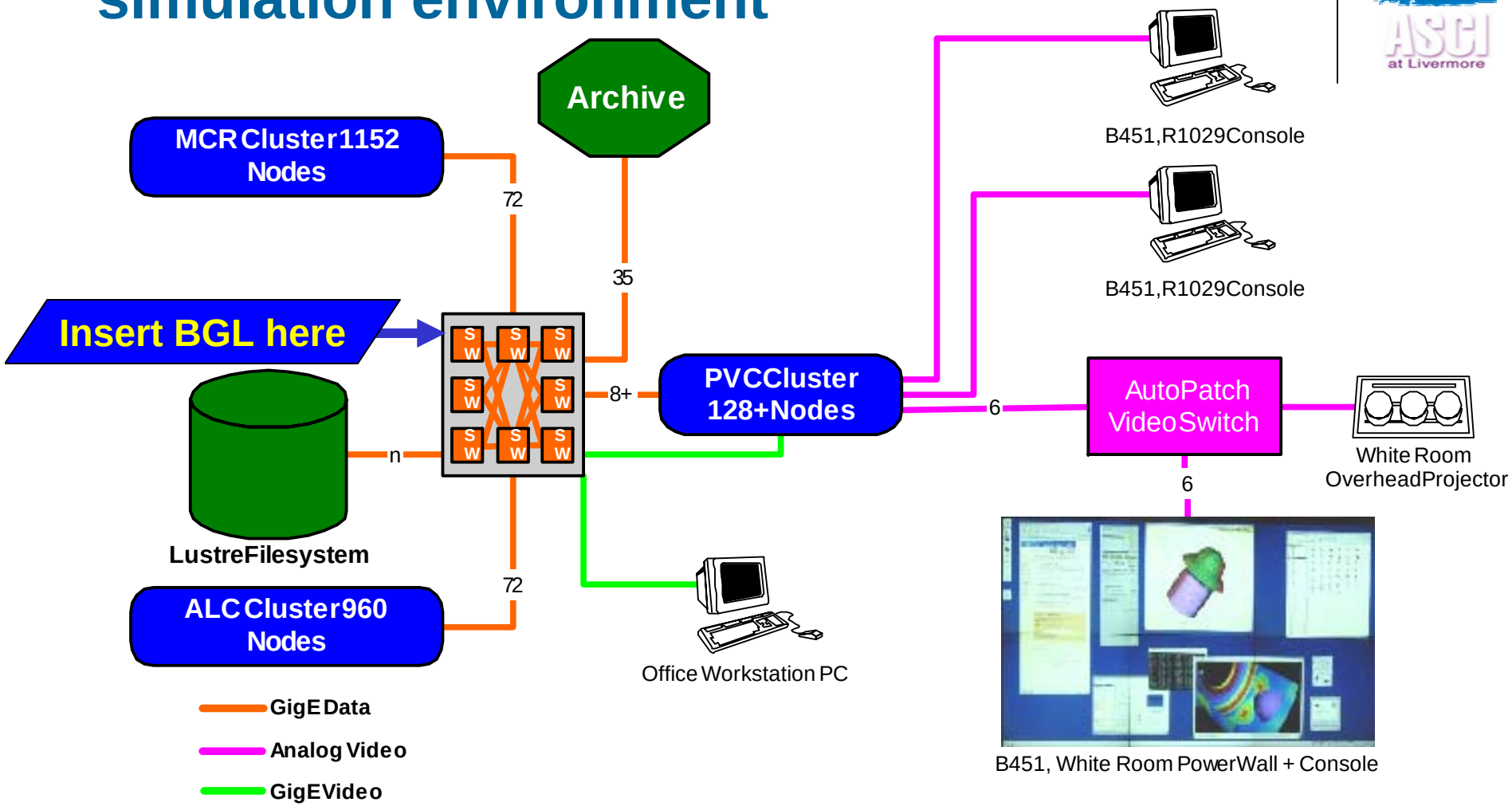
1. Key elements
2. User in the loop to force keep/delete decision
3. Direct to tape transfers eliminate HPSS disk cache
4. Bandwidth limited by file system bandwidth access from HPSS

Proposed HPSS Parallel Local File Movers



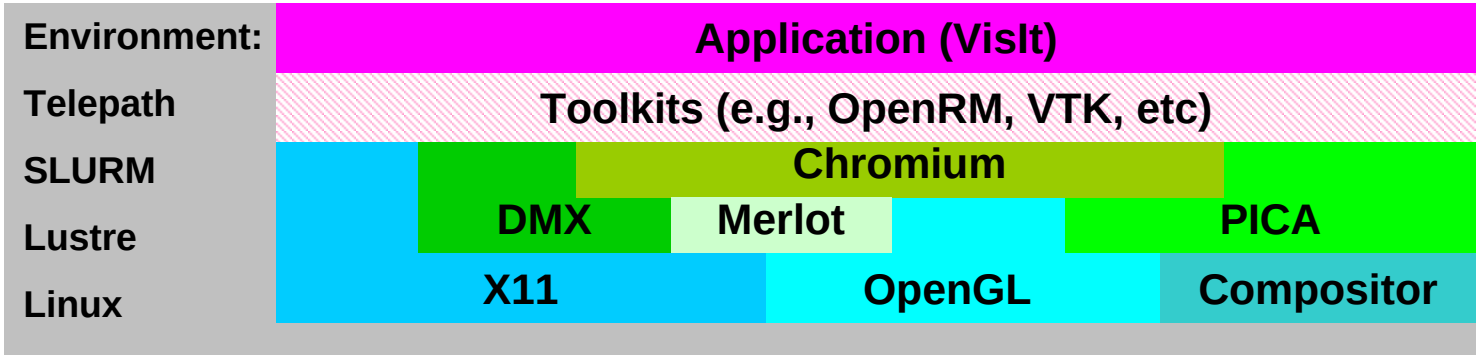
1. Key elements
2. User in the loop to force keep/delete decision
3. Direct to tape transfers eliminate HPSS disk cache
4. Bandwidth limited by parallel file system bandwidth access from HPSS

Next generation of parallel visualization technology will cleanly integrate into BGL simulation environment



All components of parallel visualization solution are built on commodity technology and are scalable

LLNL visualization based on ASCI tri-Laboratory distributed, parallel rendering software stack



- Multi-year Open Source development effort, delivering a distributed graphics API stack. The backbone of a portable, cluster based scalable rendering system:
  - Tiled X11 services
  - OpenGL
  - Visualization scheduling/display control
  - Applications
- Based on core local services
  - Unix/X11/OpenGL/Scheduling
  - Other HW: compositors, distance, etc
- Notable in the release are the Chromium distributed parallel OpenGL API and the DMX distributed multi-headed X11 server.
- Infrastructure
  - Telepath – Scheduling & display routing
  - VTK – Distributed visualization algorithms
  - Chromium – Parallel OpenGL rendering
  - DMX – 2D distributed display scaling
  - PICA – Parallel compositing model
  - Merlot – Image delivery infrastructure
- Applications
  - VisIt, ParaView – Full featured viz apps
  - TeraScale Browser – Specialized viz apps
  - MIDAS – Merlot based remote rendering
  - Blockbuster – Tiled animation display

Summary

- BlueGene/L will have a highly scalable computation environment based on four components:
  - Federated Gb/s Ethernet within and between multiple buildings at LLNL
  - Scalable parallel file system
  - HPSS archive with direct access to file system
  - Scalable parallel visualization
- Computation environment currently under development and will be ready for BlueGene/L
  - All components on the floor now