Lustre File System Evaluation at FNAL

Stephen Wolbers

for

Alex Kulyavtsev, Matt Crawford, Stu Fuess, Don Holmgren, Dmitry Litvintsev, Alexander Moibenko, Stan Naymola, Gene Oleynik,Timur Perelmutov, Don Petravick, Vladimir Podstavkov, Ron Rechenmacher, Nirmal Seenu, Jim Simone

Fermilab

CHEP'09, Prague

March 23, 2009

Outline

- Goals of Storage Evaluation Project and Introduction
- General Criteria
  - HPC Specific Criteria
  - HSM Related Criteria
  - HEP Specific Criteria
- Test Suite and Results
- Lustre at FNAL HPC Facilities: Cosmology and LQCD
- Conclusions

Storage Evaluation

- Fermilab's Computing Division regularly investigates global/high-performance file systems which meet criteria in:
  - capacity, scalability and I/O performance
  - data integrity, security and accessibility
  - usability, maintainability, ability to troubleshoot & isolate problems
  - tape integration
  - namespace and its performance
- Produced a list of weighted criteria a system would need to meet (HEP and HPC) – FNAL CD DocDB 2576

  http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=2576
- Set up test stands to get experience with file systems and to perform measurements where possible

Storage Evaluation

- Additional input for evaluation:
  - File system documentation
  - Design and performance of existing installations
  - Communications with vendor/organization staff
  - Training
- The focus of this talk is our evaluation of Lustre:
  - Most effort so far concentrated on general functionality and HPC (Cosmological Computing Cluster and Lattice QCD)
  - We are in process of evaluating Lustre for HEP
- Lustre Installations:
  - Preproduction and production systems on Cosmology Computation Cluster
  - LQCD Cluster

Storage Evaluation Criteria

- ✓ Capacity of 5 PB scalable by adding storage units
- ✓ Aggregate I/O > 5 GB/s today scalable by adding I/O units
  - LLNL BlueGene/L has 200,000 clients processes access 2.5PB Lustre fs through 1024 IO nodes over 1Gbit Ethernet network at 35 GB/sec
- ✓ Disk subsystem should impose no limit on sizes of files. The typical file used in HEP today are 1GB to 50GB

Legend for the criteria:
- ✓ Means satisfies criteria
- ➢ Means either doesn't satisfy or partially satisfies criteria
- ? Not tested

green - example exists
purple – coming soon
red – needs attention

Criteria: Functionality

- ✓ Storage capacity and aggregate data IO bandwidth scale linearly by adding scalable storage units
- ✓ Storage runs on general purpose platform. Ethernet is primary access medium for capacity computing
- ➢ Easy to add, remove, replace scalable storage unit. Can work on mix of storage hardware. System scales up when units are replaced by ones with advanced technology.
  - ✓ addition and replacement are fine
  - ➢ removal still has issues

Criteria: Namespace

- ✓ Provides hierarchical namespace mountable with POSIX access on apx. 2000 nodes (apx. 25,000 nodes on RedStorm at SNL)
- ✓ Supports millions of online (and tape resident) files
  - ✓ 74 million @ LLNL.
- ? Client processes can open at least 100 files, and tens of thousands files can be open for read in the system. *We have not tested this requirement, but Lustre's metadata server does not limit the number of open files*
- ✓ Supports hundreds metadata ops/sec without affecting I/O
  - ✓ The measured meta data rate is by factor 10-100 better
- ➢ It must be possible to make a backup or dump of the namespace & metadata without taking the system down
  - ➢ We perform hourly LVM snapshots, but we really need the equivalent of transaction logging

Criteria : Scaling Data Transfers and Recovery

- ➢ Must support at least 600 WAN and 6000 LAN transfers simultaneously, with a mixture of writes and reads (perhaps 1 to 4 ratio). One set must not starve the other
- ➢ Must be able to control number of WAN and LAN transfers independently, and/or set limits for each transfer protocol
- ✓ Must be able to limit striping across storage units to contain the impact of a total disk failure to a small percentage of files
- Serving “hot files” to multiple clients may conflict with “less striping”
- We are interested in some future features:
  - **file Migration for Space Rebalancing**
  - **set of tools for Information Lifecycle Management**

Criteria: Data Integrity & Security

- ➢ Support for Hashing or checksum algorithms.
  - Adler32, CRC32 and more will be provided on Lustre v2.0. End-to-end data integrity will be provided by ZFS DMU
- ➢ System must scan itself or allow or allow scanning for the silent file corruption without undue impact on performance
  - under investigation
- ✓ Security over the WAN is provided by WAN protocols such as GridFTP, SRM, etc.

  We require communication integrity rather than confidentiality
  - Kerberos support will be available in Lustre v2.0
- ✓ ACLs, user/group quotas

  Space management (v2.x ?)

HPC Specific Criteria and Lustre

Lustre on Computational Cosmology and LQCD clusters:

- Large, transparently (without downtime) extensible, hierarchal file system accessible through standard Unix system calls
- Parallel file access:
  - File system visible to all executables, with possibility of parallel I/O
  - Deadlock free for MPI jobs
- POSIX IO
- More stable than NFS (Computational Cosmology only)
- Ability to run on commodity hardware and Linux OS to reuse existing hardware

HSM-Related Criteria:

➔ Integration of the file system site's HSM (e.g., Enstore, CASTOR, HPSS) is required for use at large HEP installations

- The HSM shall provide transparent access to 10 to 100 PB of data on tape (growing in time)
- Must be able to create file stubs in Lustre for millions of files already existing in HSM in a reasonable time
- Automatically migrates designated files to tape
- Transparent file restore on open()
- Pre-stage large file sets from HSM to disk. Enqueue many read requests (current CMS FNAL T1 peak: 30,000) with O(100) active transfers
- Evict files already archived when disk space is needed

Lustre HSM Feature

- Lustre does not yet have HSM feature. Some sites implement simple tape backup schemes
- HSM integration feature is under development by CEA and Sun

HSM version v1.0

– "Basic HSM" in a future release of Lustre — beta in fall 2009 ?

- Integration with HPSS (v1), others will follow
- Metadata scans to select files to store in HSM v1
  - File store on close() on-write in HSM v2

Integration work

- Work specific to the HSM is required for integration

HEP and general Criteria and Lustre

- ✓ GridFTP server from Globus Toolkit v4.0.7 worked out of the box
- ✓ BeStMan SRM gateway server is installed on LQCD cluster
- ✓ Storm SRM performance on top of Lustre is reported on this conference
- ✓ Open source, training and commercial support available
- ✓ Issues reported to Lustre-discuss list are quickly answered

Lustre Tests at Fermilab

Developed test harness and test suite to evaluate systems against criteria

- "Torture" tests to emulate large loads for large data sets
- metadata stressor – create millions of files

Used standard tests against a Lustre filesystem:

data I/O : IOZone, b_eff_io
metadata I/O : fileop/IOZone, metarates, mdtest

Pilot test system was used to get experience and validate system stability

Initial throughputs were limited by the disks used for Lustre's data storage. Subsequent installations used high-performance disk arrays and achieved higher speeds.

Lustre Test System

Three to five client nodes

Two data servers

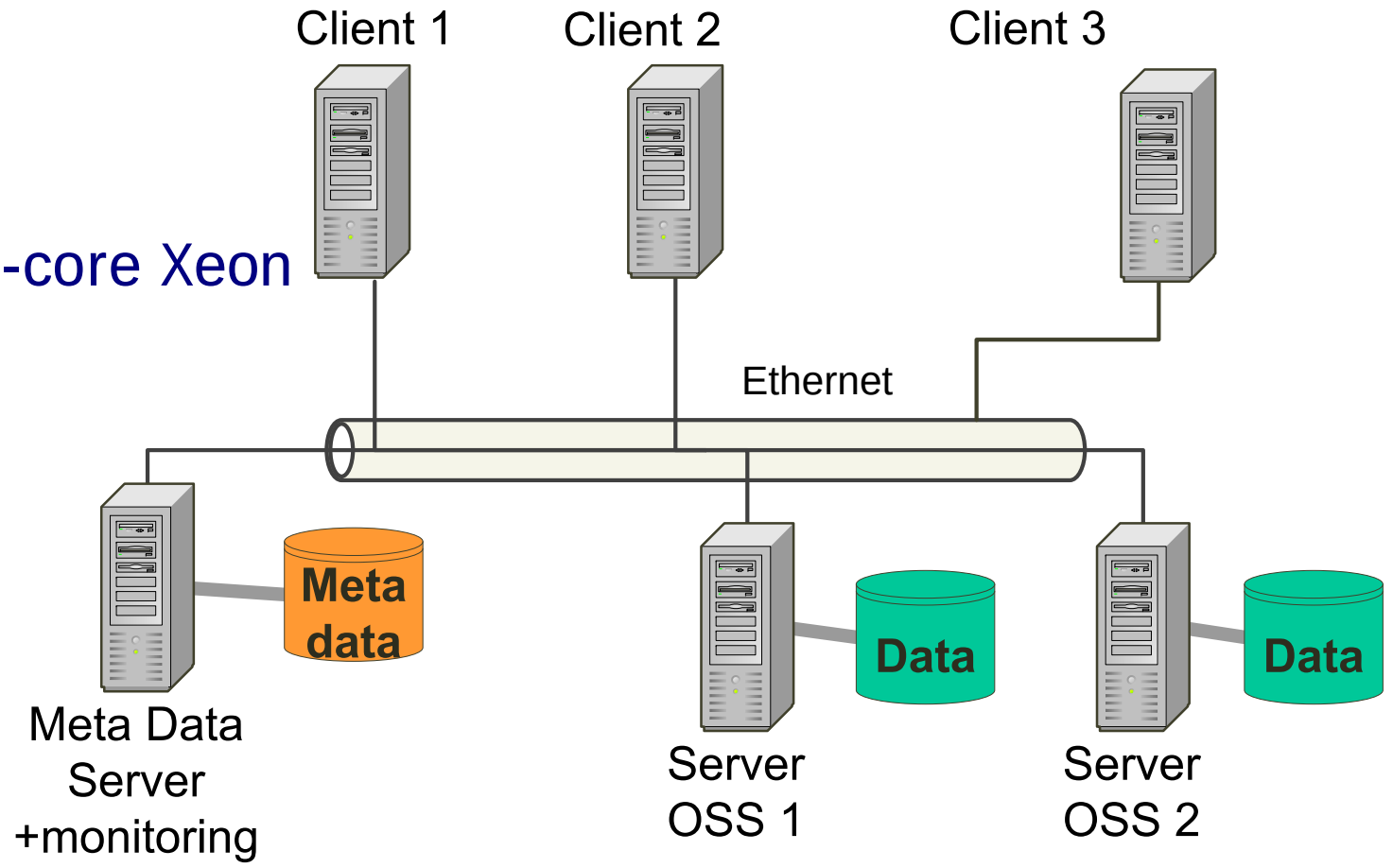
Each node :

Dual CPU quad-core Xeon

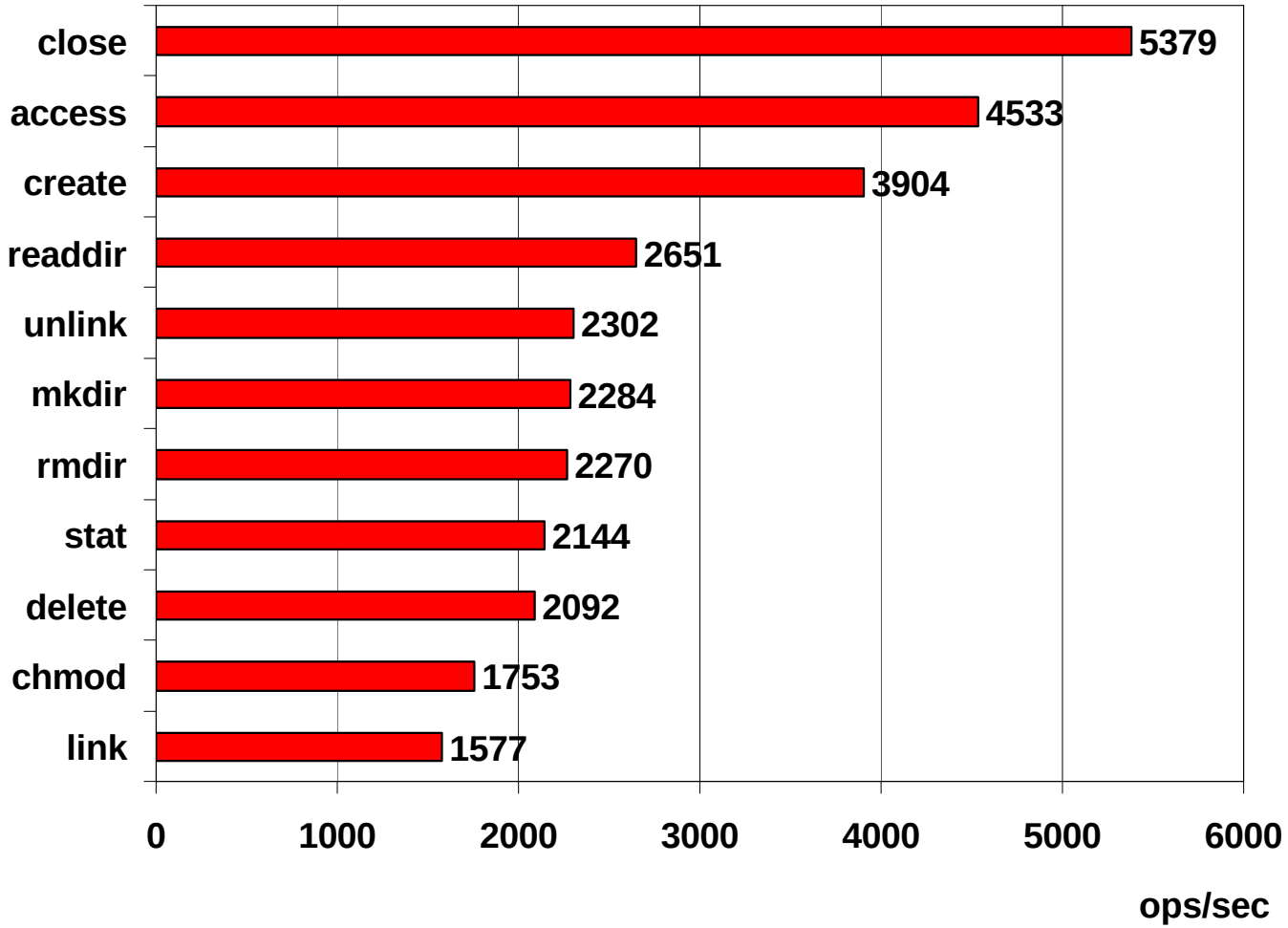
16 GB mem

Local SATA

disk 500 GB

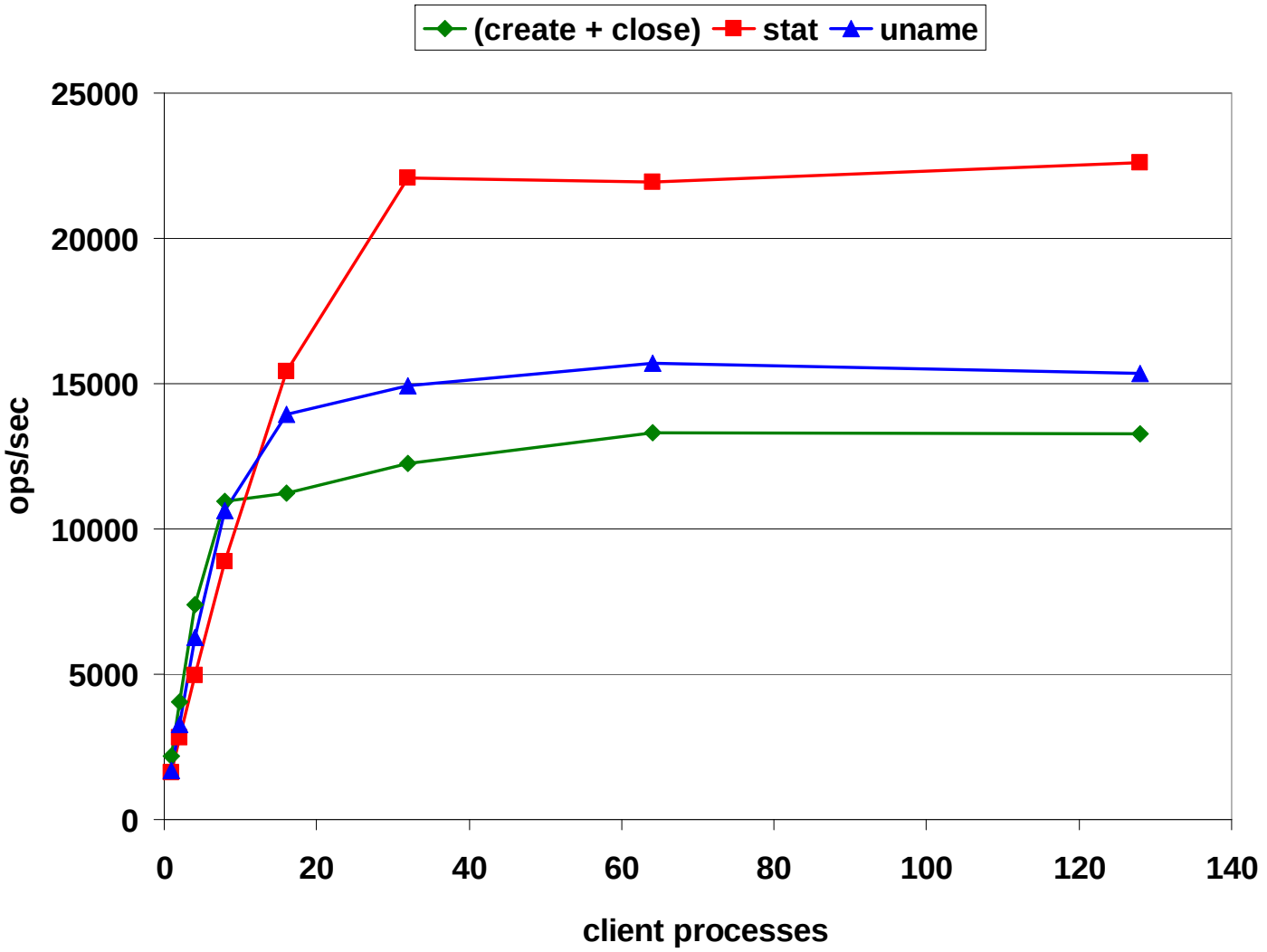


Lustre Metadata Rates



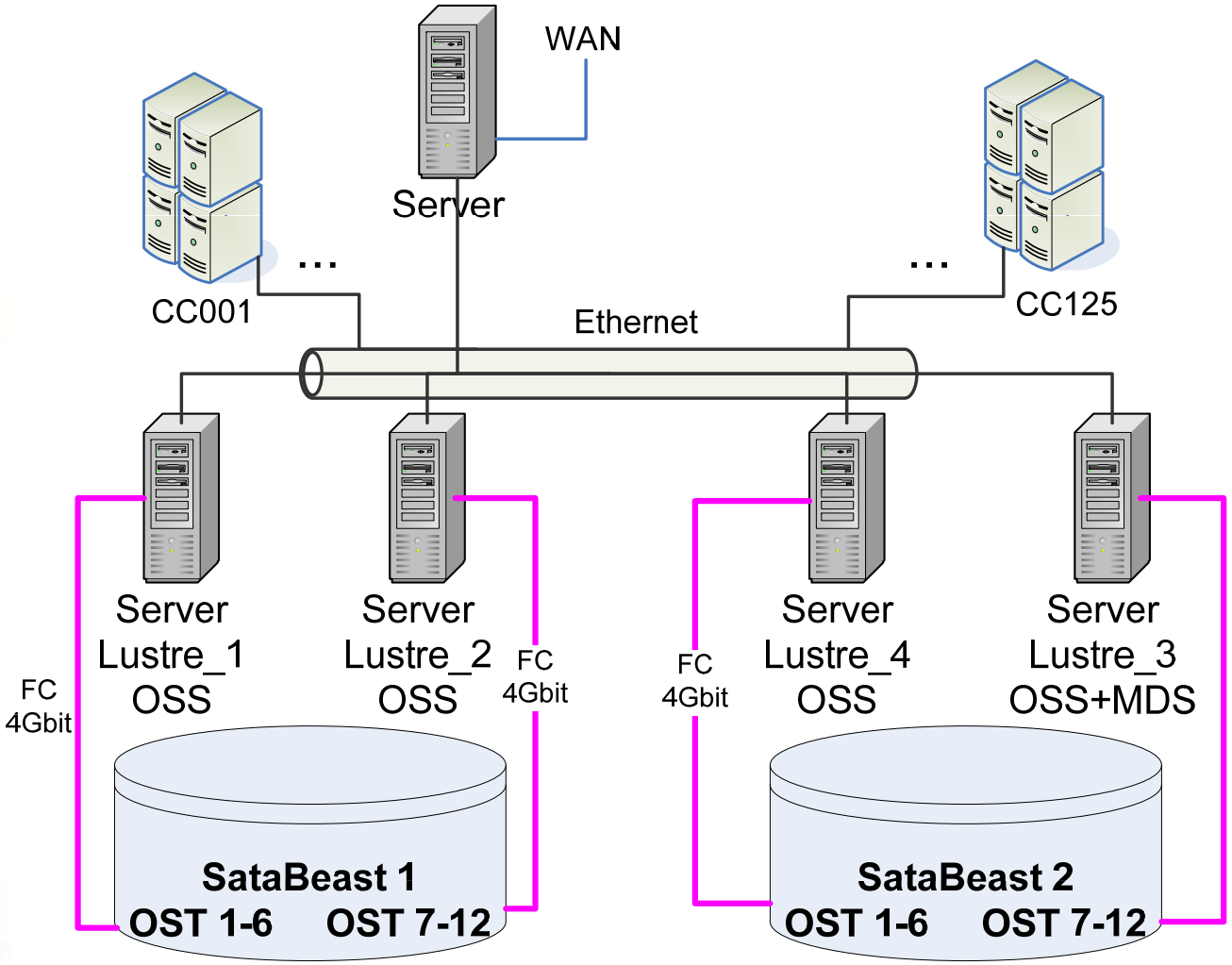
Single client metadata rates measured with *fileop/IOZone* benchmark

Lustre Metadata Rates



Aggregate metadata rate measured with *metarates* benchmark for one to 128 multiple clients running on 5 nodes * 8 cores

Lustre on Computational Cosmology Cluster



125 Compute nodes

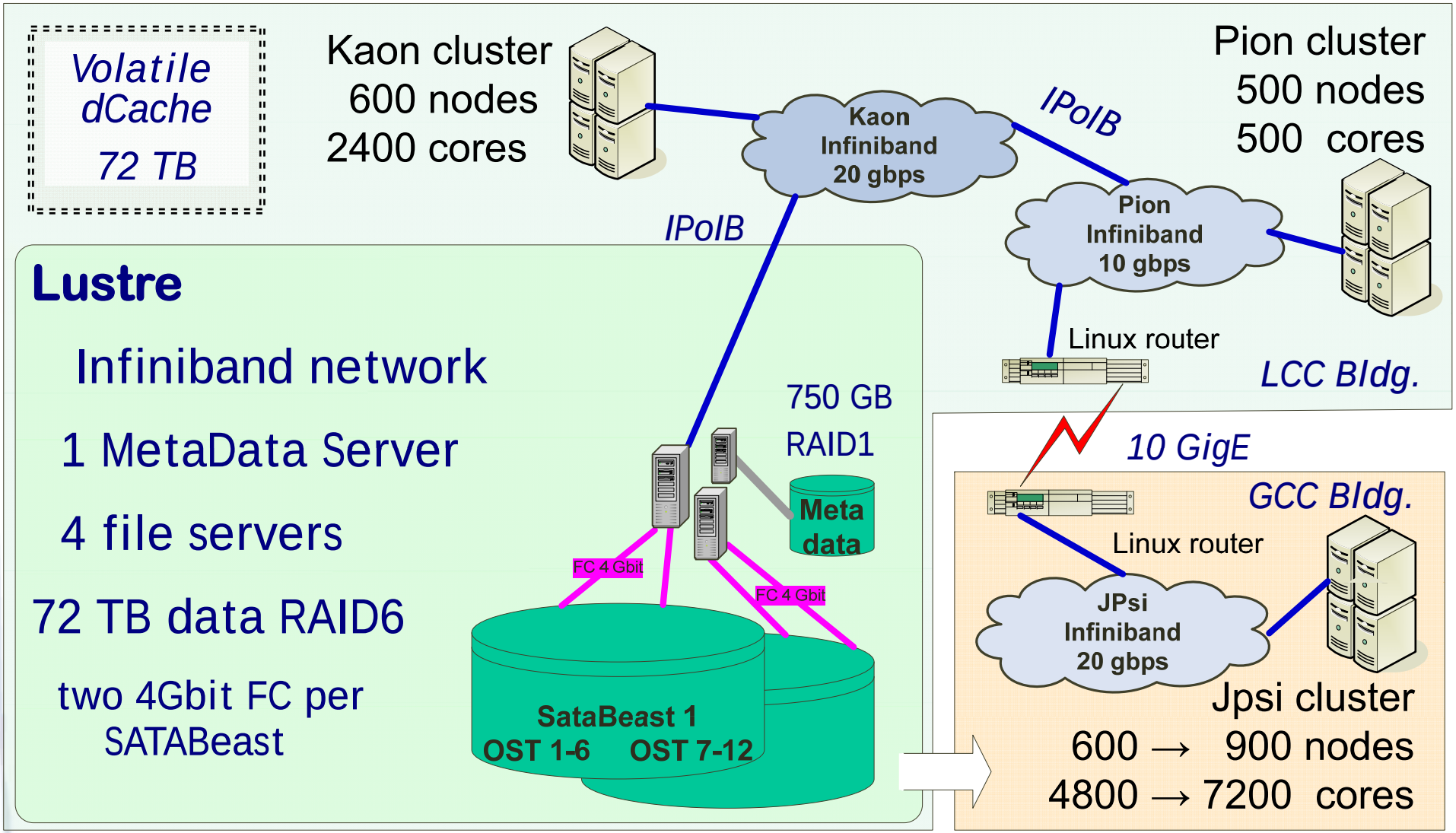
1Gb Ethernet
Stackable Switch
SMS 8848M

4 Lustre data Servers
one shared with
Metadata Server
250 GB on LVM

Lustre DATA
2 SATABeasts
= 72 TB RAID6

2* 12 LUNS = 2* 3 vol. *4 partitions

Lustre on FNAL LQCD clusters



Lustre Experience - HPC

- From our experience in production on Computational Cosmology Cluster (starting summer 2008) and limited pre-production on LQCD JPsi cluster (December 2008) the Lustre File system:
  - Lustre doesn't suffer the MPI deadlocks of dCache
  - direct access eliminates the staging of files to/from worker nodes that was needed with dCache (Posix IO)
  - improved IO rates compared to NFS and eliminated periodic NFS server "freezes"
  - reduced administration effort

Conclusions - HEP

- Lustre file system meets and exceeds our storage evaluation criteria in most areas, such as system capacity, scalability, IO performance, functionality, stability and high availability, accessibility, maintenance, and WAN access.
- Lustre has *much* faster metadata performance than our current storage system.
- At present Lustre can only be used for HEP applications not requiring large scale tape IO, such as LHC T2/T3 centers or scratch or volatile disk space at T1 centers.
- Lustre near term roadmap (about one year) for HSM in principle satisfies our HSM criteria. Some work will still be needed to integrate any existing tape system.

Backup Slides

Lustre Jargon

Client – client node where user application runs. It talks to MetaData Server and data server (OSS)

MDS – MetaData Server - the *node*, one active per system

MDT – MetaData Target – *disk storage* for metadata, connected to MDS

OSS – Object Store Server – the *node* serving data files or file stripes

OST – Object Store Target – *disk storage* for data files or file stripes, connected to OSS

What is Lustre?