CMS Florida T2 Storage Status

Bockjoo Kim
for the CMS Florida T2

Storage Hardware Status

- ✓ **User space data RAID**: gatoraid1, gatoraid2, storing CMS software, $DATA, $APP, etc. 3ware controller based RAID5, mounted via NFS. 7.3TB
- ✓ **Resilient dCache**: 2 x 250 (500) GB SATA drives on 126 worker nodes. 63TB
- ✓ **Non-resilient RAID dCache**: FibreChannel RAID5 (pool03, pool04, pool05, pool06) + 3ware-based SATA RAID5 (pool02), with 10GbE or bonded multiple 1GbE network. 163TB
- ✓ **Lustre storage**: Our main storage system now. Serves Bestman. Areca-controller-based RAID5 (dedicated) + RAID Inc Falcon III FibreChannel RAID5 (shared), with InfiniBand+3ware SATA RAID5 (pool01) 215TB+20TB+6.3TB
- ✓ **dCache GridFTP servers**: Running on 20 worker nodes: 20x1Gbps
- ✓ **Dedicated GridFTP servers**: Serving the Lustre storage. 3 dedicated GridFTP servers. 3x10Gbps, dual quad-core processors, 32 GB memory.
- ✓ **Server Nodes**: 2*SRM servers (for dCache and Bestman), 2*PNFS servers (one backup), dCacheAdmin server, dCap server, dCacheAdminDoor server.

Storage Software Status

A. dCache: 1.9.6-4

- **SRM node**: 64-bit Centos 5.5 2GB memory. Serves **srm** and **gPlazma**
- **Admin node:** 32-bit Centos 4.2 2GB memory. Serves **lm, dCache(PoolManager), dir, httpd, utility, infoProvider, statistics, and replica**.
- **Chimera node:** 64-bit Centos 5.3 3.3GB memory. Serves **Chimera**
- **Dcap node:** 32-bit Centos 4.2 2GB memory. Serves **dcap and gsidcap**
- **Admin door node:** 32-bit Centos 4.2 2GB memory. Serves **admin interface**
- **Pool/gridftp nodes**: 64-bit Centos 5 Resilient and RAID pools. Serves **pool**

B. Bestman 2.2.2.1.3.10 + Lustre: 1.8.3

- **SRM node**: 64-bit Centos 5.5 4GB memory. Serves **Bestman srm + plugin for the gridftp selection mechanism.**
- **GridFTP node:** 64-bit Centos 5 16 - 32GB memory. Serves **gridftp (VDT)**
- **MDS node:** 64-bit Centos 5.3 24GB memory. Serves **MDS and MDT, infiniband**
- **OSS node:** 64-bit Centos 5.3 32GB memory. Serves **OSS and OST, infiniband**.

Storage Hardwares

RAID dCache

Resilient dCache

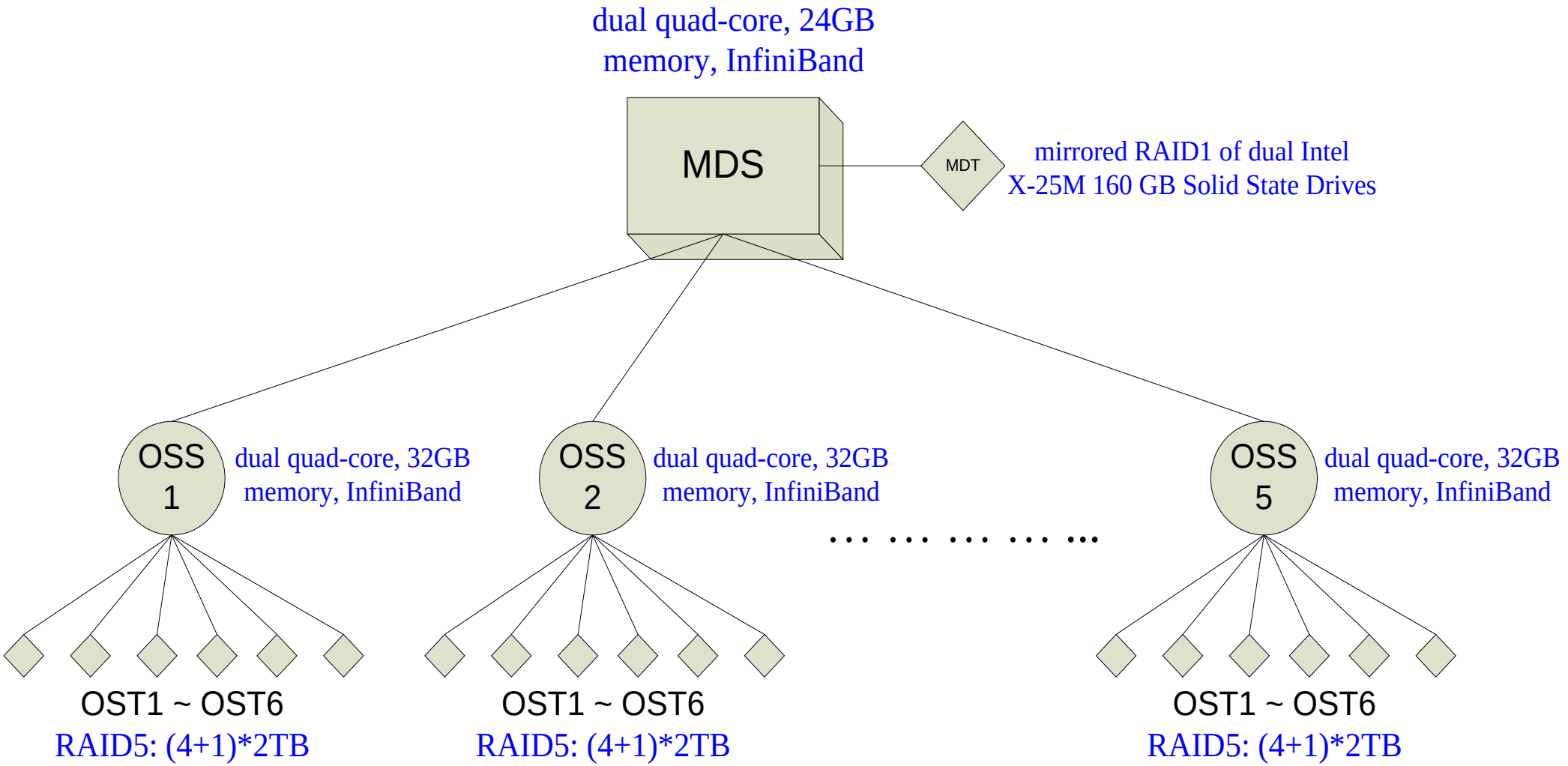
Areca+WD
Lustre

Adaptec+Hitachi
Lustre

Motivation for Lustre

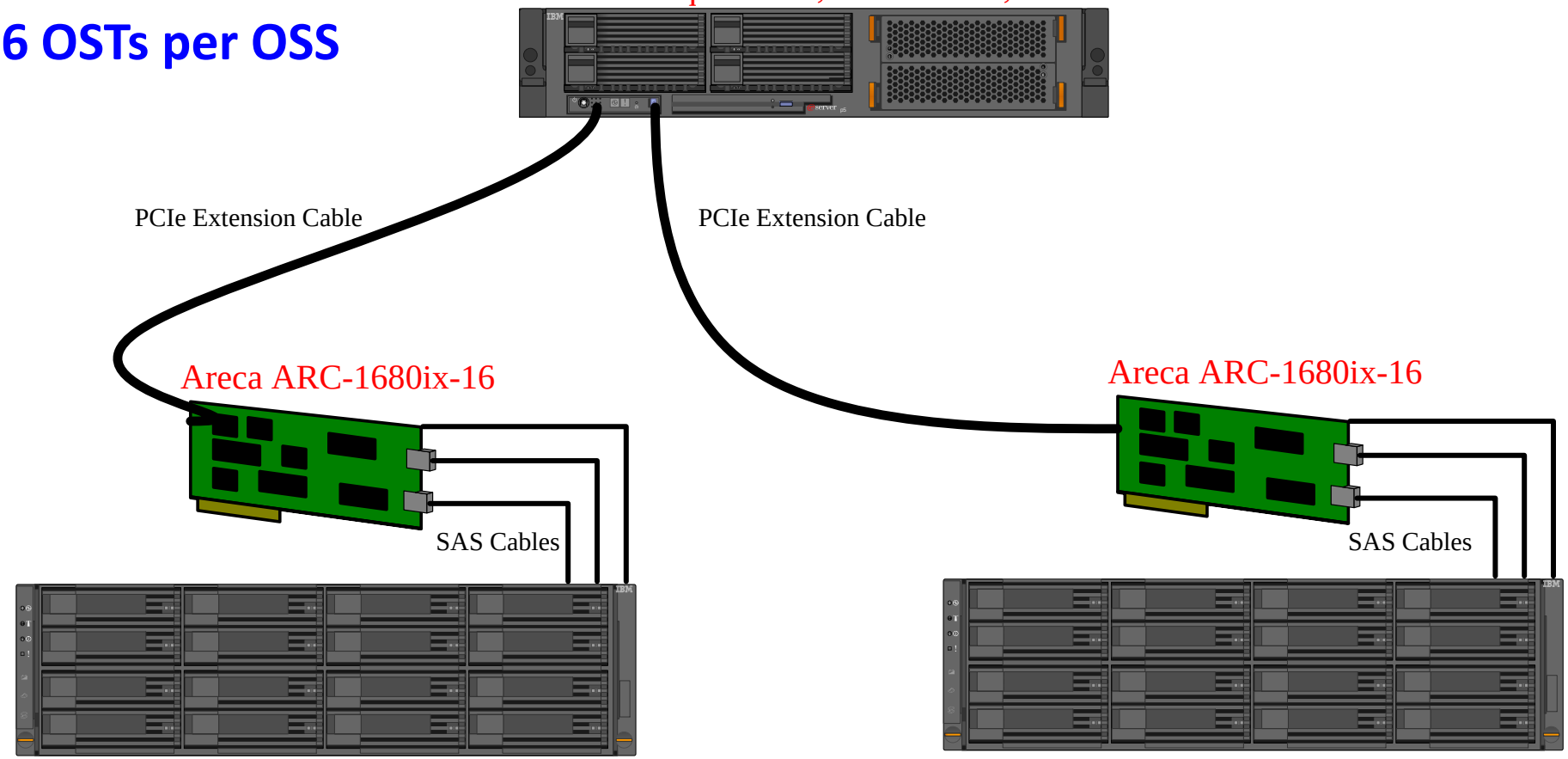
- Motivation
  - Best in comparison with various available parallel filesystems according to UF HPC's tests.
  - Widely used, proven performance, reliability and scalability.
  - Relatively long history. Mature and stable.
  - Existing local experience and expertise at UF Tier2 and UF HPC.
  - Excellent support from the Lustre community, SUN/Oracle and UF HPC experts, frequent updates, prompt patch for new kernels.
- Hardware selection:
  - With proper carefully chosen hardware, RAID's can be inexpensive yet with excellent performance and reasonably high reliability.
  - Areca ARC-1680ix-16 PCIe SAS RAID controller with 4GB memory, each controller supports 16 drives.
  - 2TB enterprise-class SATA drives, most cost effective at the time.
  - iStarUSA drive chassis connected to I/O server through extension PCIe cable: one I/O server can drive multiple chassis, more cost effective.
  - Cost: <$300/TB raw space, ~$400/TB net usable space with configuration of 4+1 RAID5's and a global hot spare drive for every 3 RAID5's.

Lustre at Florida T2



Florida T2 OSS/OST

We have a total of
5 OSSes and
6 OSTs per OSS



Florida SE Capacity Summary

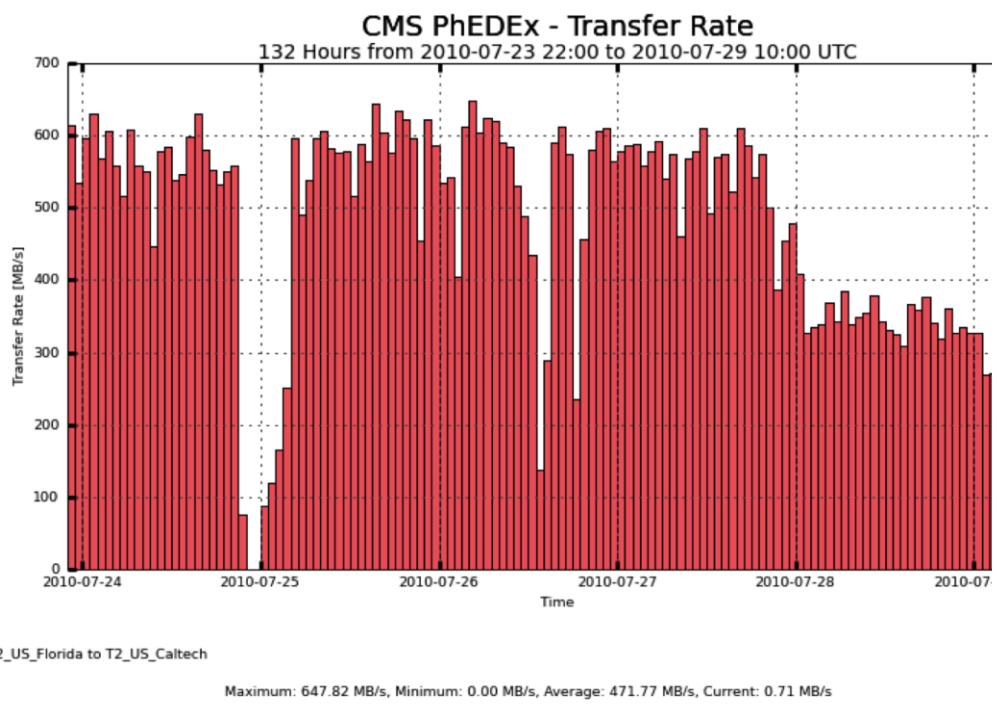
Resource	Raw	Usable (after redundancy, overhead and/or resilient)
dCache non-resilient pools (RAID's)	173.2TB	137.1TB
dCache resilient pools (worker nodes)	93.0TB	35TB
Dedicated Tier2 production Lustre	320TB	215TB
Test Lustre	8TB	6.3TB
Shared HPC Lustre (Tier2's share)	N/A	30TB
Total	**594.2TB**	**423.4TB**

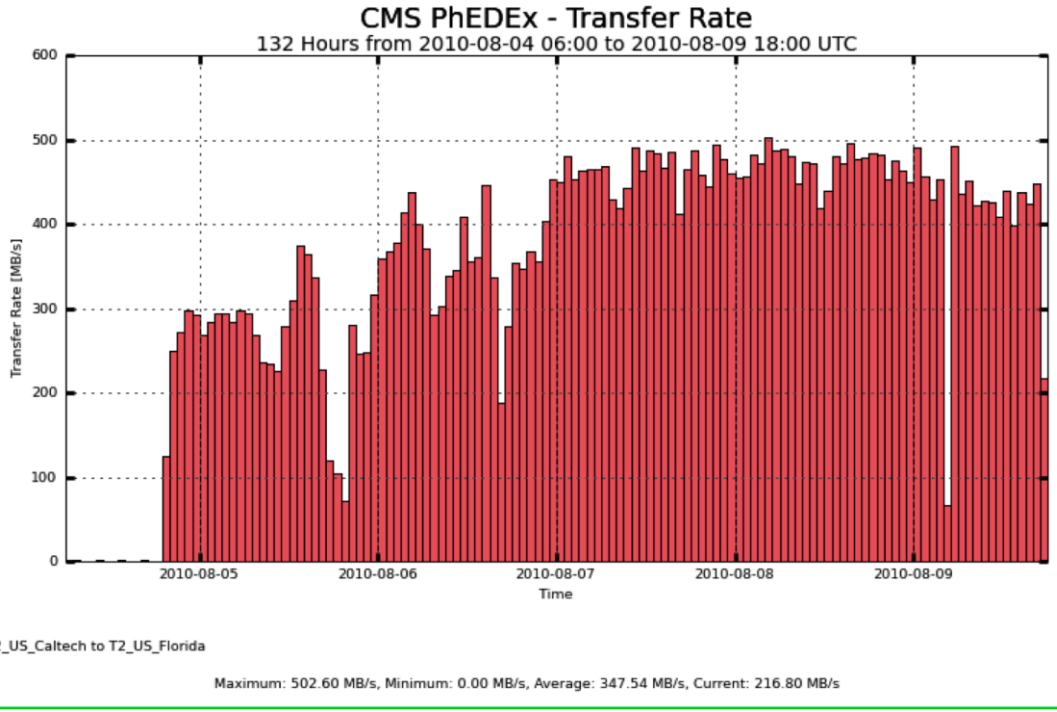
Experience with Lustre

- ✓ Remote Access Test with nearby institutions, FIU and FIT.
- ✓ Used Lustre as dCache pools (need localflock to make it work)
- ✓ It is not good for small large number of files (bad for $APP)
- ✓ It is not without problems: LBUG, expensive, metadata recovery time maybe long, etc

Performance: Data Moving

Four Day Data Transfer Test



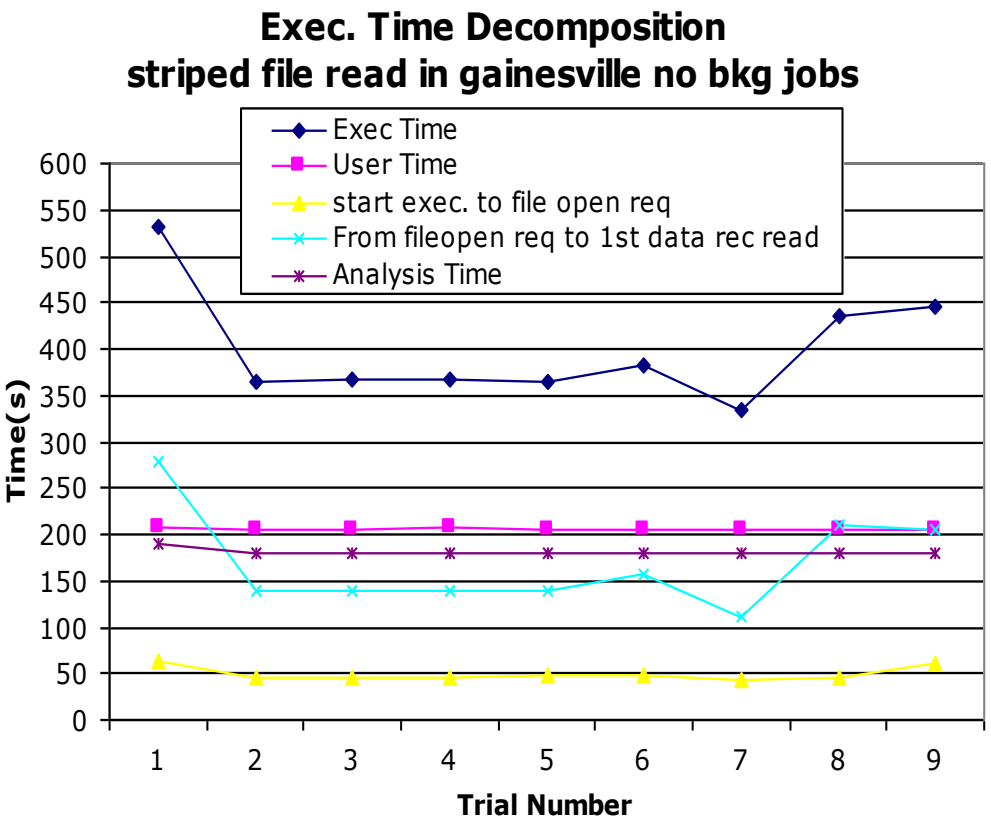


Florida -> Caltech 480MB/s average

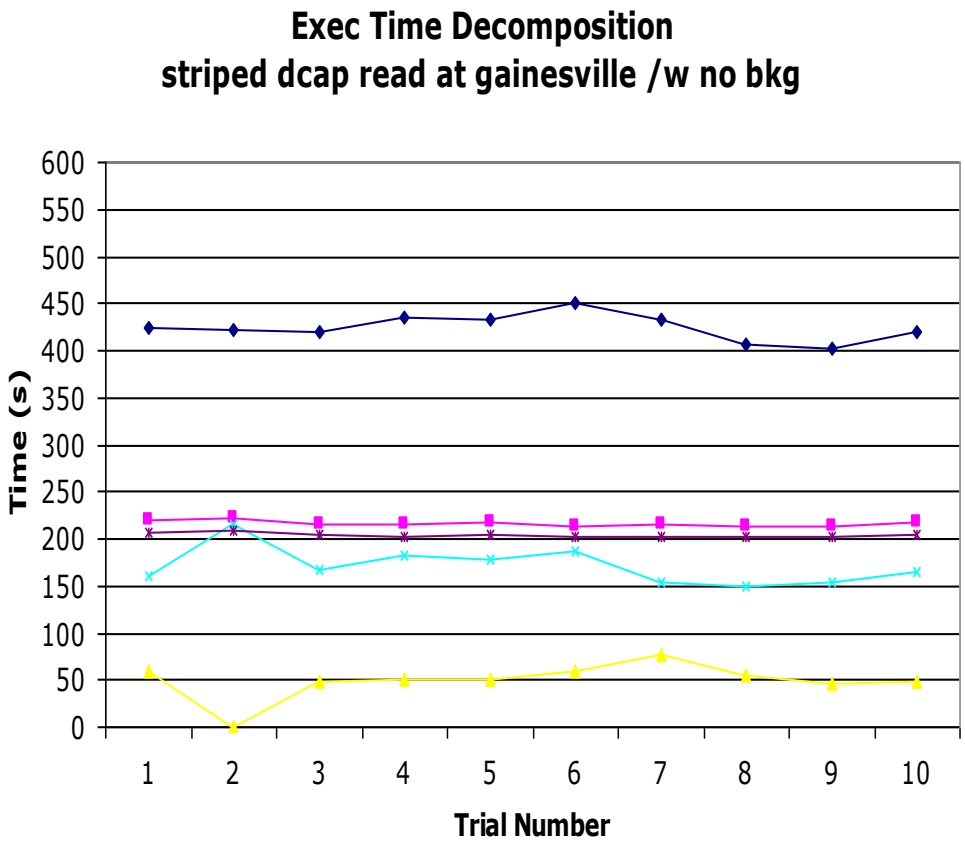
Caltech -> Florida 350MB/s average

Performance: dCache vs Lustre

File Read Performance



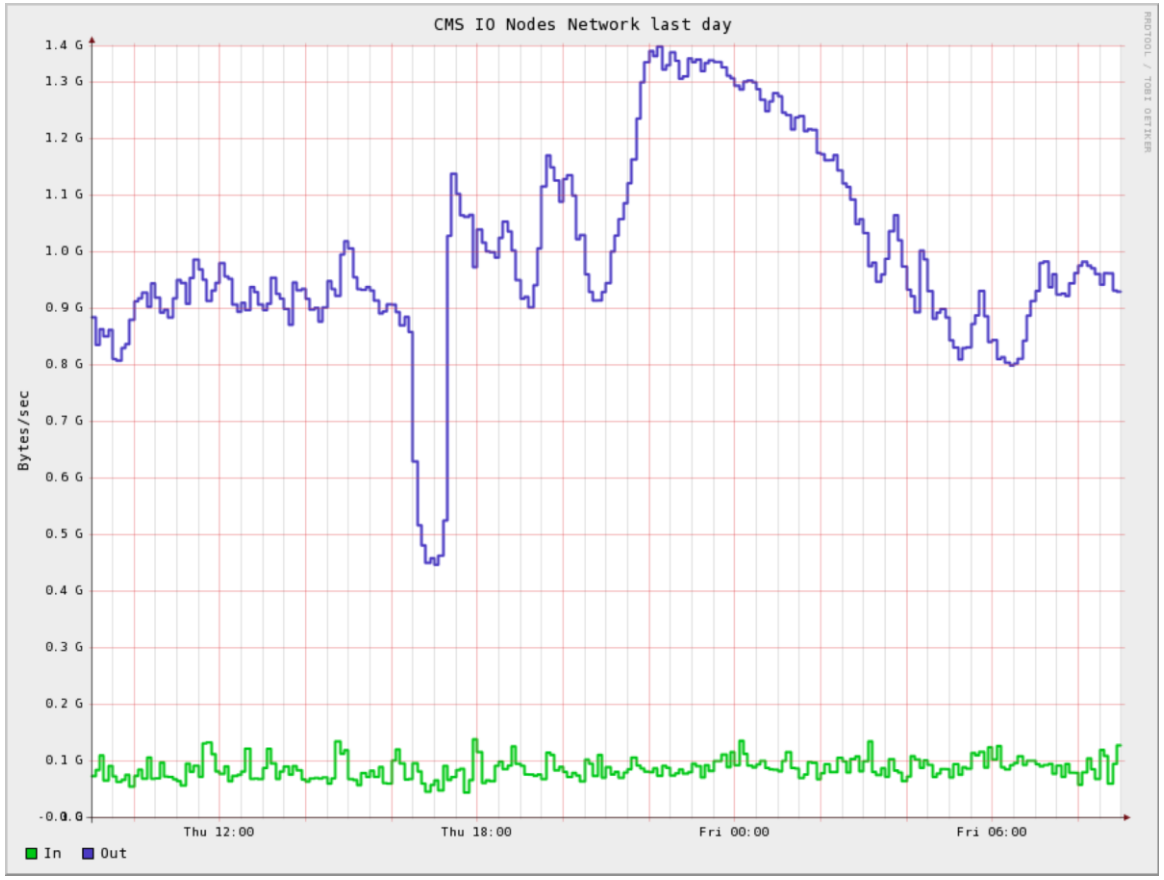
Lustre



dCache

Performance: Lustre I/O

Lustre I/O seen from Network Monitoring



1.4GB/s max

Out

In

Plans For Upgrade

- ✓ **Total raw ~600 TB, net usable 425 TB.**
- ✓ Trying to meet the 2010 USCMS T2 milestone: 570 TB
- ✓ Planning to deploy 256 TB (raw) new RAID's (similar to current OSS/OST configuration) in the production Lustre system in FY10. Budget planned: ~$80k.
- ✓ We tested a system based on Hitachi + Adaptec and Positive result. Upgrade will be based on this system.
- ✓ Satisfied with Lustre+Bestman+GridFTP and will stay with it.
- ✓ Will put future SE's into Lustre.
- ✓ Present non-resilient dCache pools (RAID's) will gradually fade out and migrate to Lustre. Resilient dCache pools on worker nodes will be kept.

Researches Supported

- Within CMS, we support 4 groups and the local CMS physicists
  - Muon Detector Group

    Arranged write access to the dCache and the fair share. Local users also requested large portion of muon group datasets.
  - SUSY Physicis Group

    In 2009, we provided the space for the private SUSY production. Large number of datasets that are produced by the local SUSY group are still in the analysis 2 DBS.
  - Particle Flow / JetMet Group

    We were not able to help much with this group because we did not have enough space at the time because most space was occupied by the Muon group.
    Not much interaction, only with the validation datasets were hosted for this group. Recently one of our postdocs is interacting with us to coordinate the necessary dataset hosting.
  - Higgs

    Thanks to addition of the Lustre storage.
- OSG Support
  - We only support SCEC via a development lustre system at the moment due to the space limitation and manpower.

XROOTD Servers

- We have recently deployed two xrootd servers
- They are customized to CMS by Brian
- One for dCache and the other for Lustre
- dCache : 50MB/s locally (dccp rate is 75MB/s)

Extenci

- ExTENCI: Extending Science Through Enhanced National CyberInfrastructure;
  - Funded by NSF/DOE for two years: $2M total
  - Work together between Open Science Grid and TeraGrid
  - The project just started
- Lustre will be one of the major projects for ExTENCI
- Project goal is to evaluate a Global wide-area file system for performance and robustness
- Florida Focus will be mainly on the secure Lustre WAN access and the project result will potentially allow secure access to remote (Tier-2 and other) resources in US-CMS with less administrative overhead

Summary

- 423TB of dCache+Lustre in production. Planning to increase the capacity to 570TB soon based on the Adaptec+Hitachi
- Established a production Lustre+BestMan+GridFTP storage system. A review document within USCMS being prepared.
- Satisfied with Lustre, will put most of current dCache non-resilient pools into the Lustre system.
- Supports 4 associated analysis groups within CMS, an OSG group, and the local CMS users.
- XROOTD deployed
- Extenci participation

Additional Slides

Motivation of RAID dCache

- Our past experience shows resilient pools on worker nodes relatively often went down or lost data due to CPU load and memory usage as well as hard drive etc hardware issues.
- Therefore we want to separate CE and SE in hardware so job load on CE worker nodes will not affect storage.
- High-performance-oriented RAID hardware is supposed to perform better.
- Dedicated, specially designed RAID hardware is expected to be more robust and reliable.
- Hardware RAID will not need full replicated copies of data like in resilient dCache and Hadoop, more efficient in disk usage.
- Fewer equipments are involved in dedicated large RAID's, easier to manage and maintain.

Diagram of a Lustre SE