Florida Tier2 Site Report

Presented by Yu Fu

For the Florida CMS Tier2 Team:

Paul Avery, Dimitri Bourilkov, Yu Fu, Bockjoo Kim, Yujun Wu

USCMS Tier2 Workshop

Fermilab, Batavia, IL

March 8, 2010

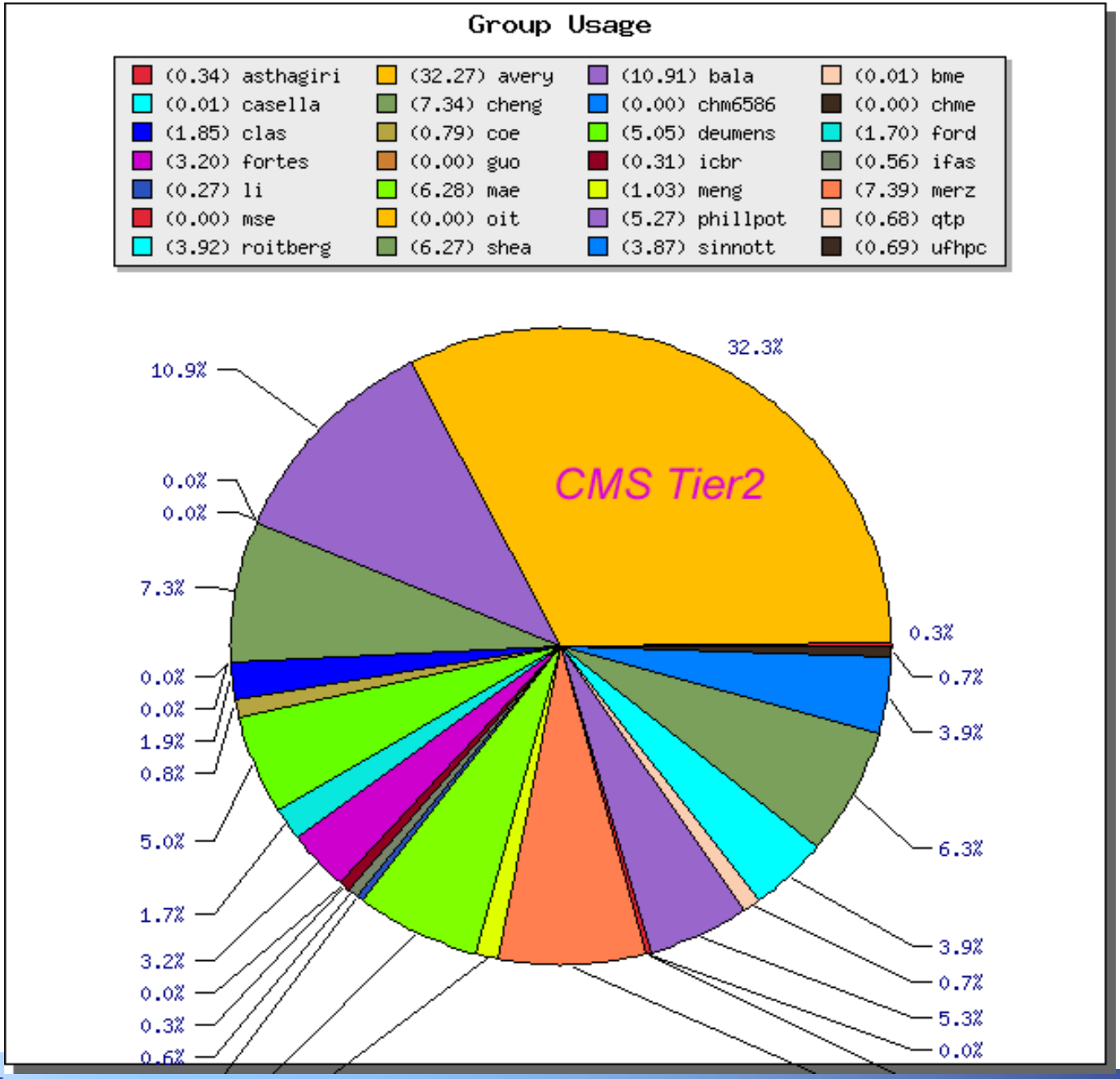
Hardware Status

- Computational Hardware:
  - UFlorida-PG (dedicated)
    - 126 worker nodes, 504 cores (slots)
    - 84 * dual dual-core Opteron 275 2.2GHz + 42 * dual dual-core Opteron 280 2.4 GHz, 6GB RAM, 2x250 (500) GB HDD
    - 1GbE, all public IP, direct outbound traffic
    - 3760 HS06, 1.5 GB/slot RAM.

Hardware Status

- UFlorida-HPC (shared)
  - 642 worker nodes, 3184 cores (slots)
  - 112 * dual quad-core Xeon E5462 2.8GHz + 16 * dual quad-core Xeon E5420 2.5GHz + 32 * dual hex-core Opteron 2427 2.2GHz + 406 * dual dual-core Opteron 275 2.2GHz + 76 * single-core Opteron 250 2.4GHz
  - Infiniband or 1GbE, private IP, outbound traffic via 2Gbps NAT.
  - 25980 HS06, 4GB/slot, 2GB/slot or 1GB/slot RAM.
  - Managed by UF HPC Center, Tier2 invested partially in three phases.
  - Tier2's official quota is 845 slots, actual usage: typically ~1000 slots (~30% total slots).
  - Tier2's official guaranteed HS06: 6895 (normalized 845 slots).

Typical UFlorida-HPC cluster usage (30-day average)



Hardware Status

- Interactive analysis cluster for CMS
  - 5 nodes + 1 NIS server + 1 NFS server
  - 1 * dual quad-core Xeon E5430 + 4 * dual dual-core Opteron 275 2.2GHz, 2GB RAM/core, 18 TB total disk.

- Roadmap and plans for CE's:
  - Total Florida CMS Tier2 official computing power (Grid only, not including the interactive analysis cluster):

    10655 HEP-SPEC06, 1349 batch slots.
  - Have exceeded the 2010 milestone of 7760 HS06.
  - Considering new computing nodes in FY10 to enhance the CE as the worker nodes are aging and dying (many are already 5 years old).
  - AMD 12-core processors in testing at UF HPC, may be a candidate for the new purchase.

Hardware Status

- Storage Hardware:
  - User space data RAID: gatoraid1, gatoraid2, storing CMS software, $DATA, $APP, etc. 3ware controller based RAID5, mounted via NFS.
  - Resilient dCache: 2 x 250 (500) GB SATA drives on each worker node.
  - Non-resilient RAID dCache: FibreChannel RAID5 (pool03, pool04, pool05, pool06) + 3ware-based SATA RAID5 (pool01, pool02), with 10GbE or bonded multiple 1GbE network.
  - Lustre storage, our main storage system now: Areca-controller-based RAID5 (dedicated) + RAID Inc Falcon III FibreChannel RAID5 (shared), with InfiniBand.
  - 2 dedicated production GridFTP doors: 2x10Gbps, dual quad-core processors, 32 GB memory.
  - 20 test GridFTP doors on worker nodes: 20x1Gbps
  - Dedicated 2*PhEDEx servers, 2*SRM servers, 2*PNFS servers, dCacheAdmin server, dCap server, dCacheAdminDoor server.

Hardware Status

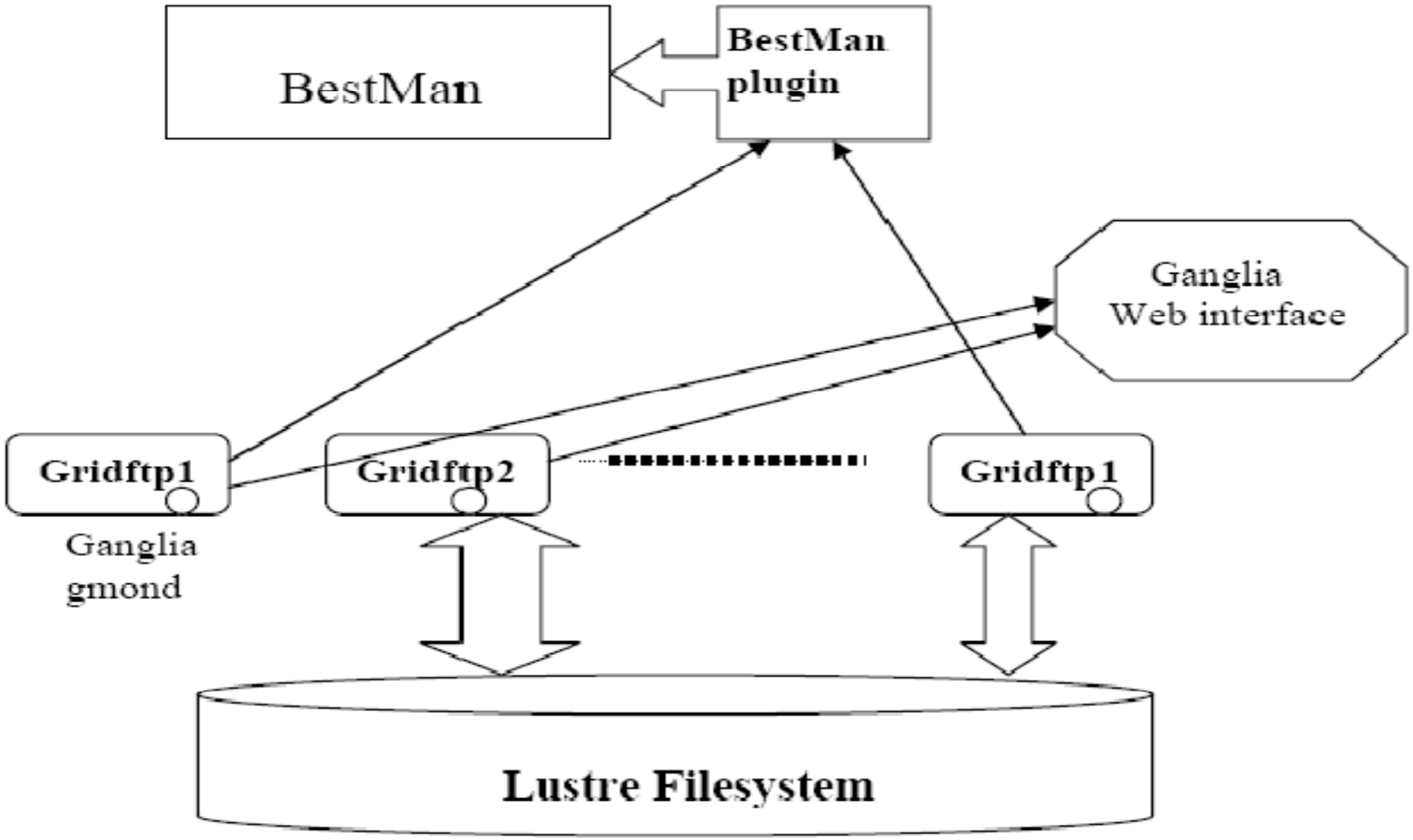
- SE Structure:
  - Co-existing of dCache and Lustre + BestMan + GridFTP currently, migrating to Lustre.
- Motivation of choosing dedicated RAID's:
  - Our past experience shows resilient dCache pools on worker nodes relatively often went down or lost data due to CPU load and memory usage as well as 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.

Hardware Status

- Motivation of choosing Lustre:
  - 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 in a chassis.

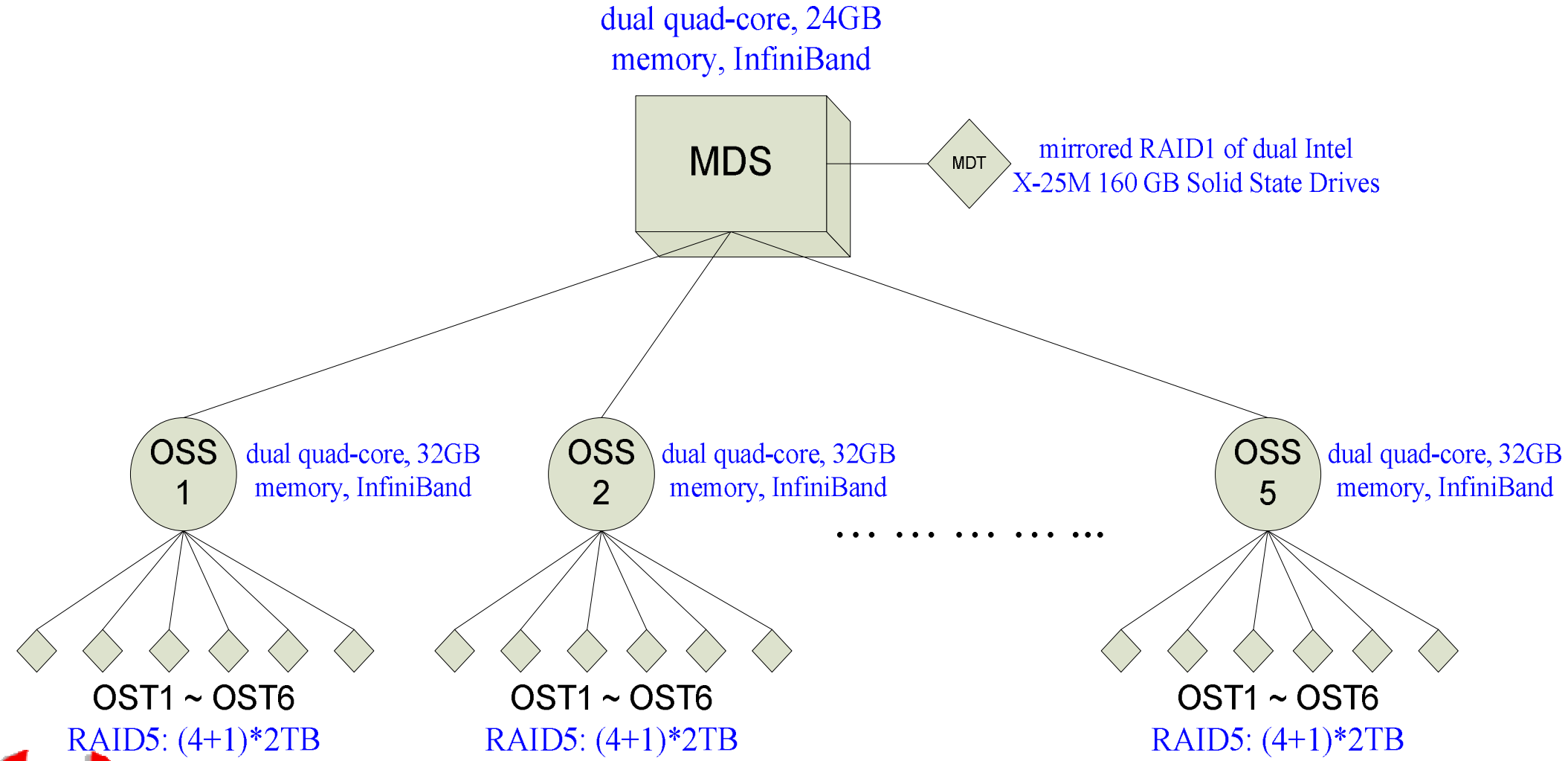
Hardware Status

Lustre + BestMan + GridFTP



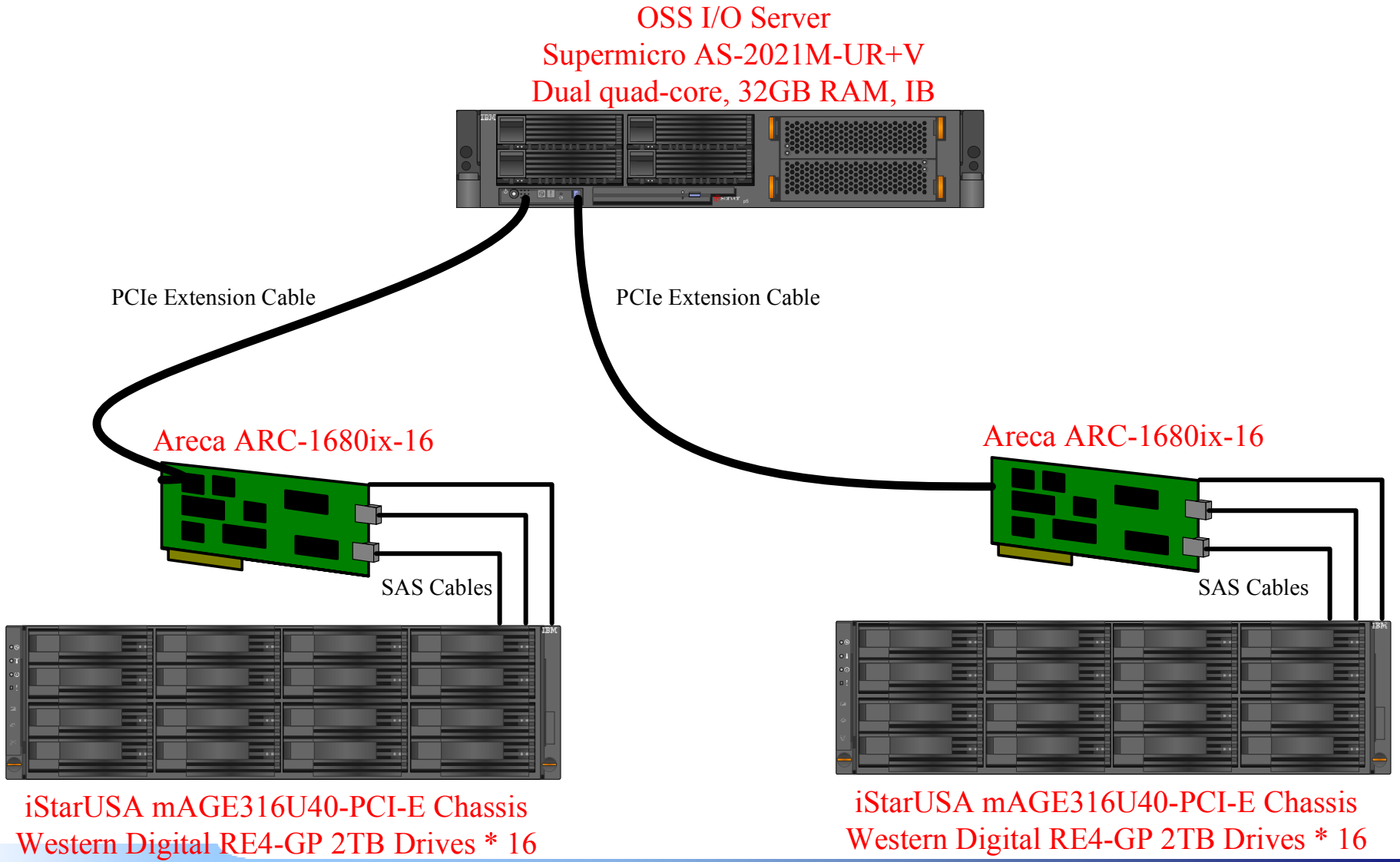
Hardware Status

UF Tier2 Lustre Structure



Hardware Status

UF Tier2 Luster OSS/OST



Hardware Status

- Total SE Capacity:

Resource	Raw	Usable (after redundancy, overhead and/or resilient)
dCache non-resilient pools (RAID's)	181.2TB	144TB
dCache resilient pools (worker nodes)	93.0TB	35.6TB
Dedicated Tier2 production Lustre	320TB	215TB
Shared HPC Lustre (Tier2's share)	N/A	30TB
Total	**594.2TB**	**424.6TB**

(Disks on the UFlorida-HPC worker nodes are not counted since they are not deployed in dCache system or the Lustre system.)

Hardware Status

- Roadmap and plans for SE's:
  - **Total raw ~600 TB, net usable 425 TB.**
  - There is still a little gap to meet the 570 TB 2010 milestone.
  - Planning to deploy more new RAID's (similar to current OSS/OST configuration) in the production Lustre system in FY10 to meet the milestone.
  - 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.

Software Status

- Most systems running 64-bit SLC5/CentOS5.
- OSG 1.2.3 on UFlorida-PG and interactive analysis clusters, OSG 1.2.6 on UFlorida-HPC cluster.
- Batching system: Condor on UFlorida-PG and Torque/PBS on UFlorida-HPC.
- dCache 1.9.0
- PhEDEx 3.3.0
- BestMan SRM 2.2
- Squid 4.0
- GUMS 1.3.16
- ……
- All Tier2 resources managed with a 64-bit customized ROCKS 5, all rpm's and kernels are upgraded to current SLC5 versions.

Network Status

- Tier2 Cisco 6509 switch
  - All 9 slots populated
  - 2 blades of 4 x 10 GigE ports each
  - 6 blades of 48 x 1 GigE ports each
- 20 Gbps uplink to the campus research network.
- 20 Gbps to UF HPC.
- 10 Gbps to UltraLight via FLR and NLR.
- Running perfSONAR.
- GbE for worker nodes and most light servers, optical/copper 10GbE for servers with heavy traffics.
- Florida Tier2's own domain and DNS.
- All Tier2 nodes including worker nodes on public IP, directly connected to outside world without NAT.
- UFlorida-HPC worker nodes and Lustre on InfiniBand.
- UFlorida-HPC worker nodes on private IP with a 2Gbps NAT for communication with outside world.

Analysis Operations

- Provide local CMS analysis users with interactive analysis cluster and direct access to the batch system of both UFlorida-HPC (PBS torque) and UFlorida-PG (condor flocking) clusters.
- Associated with Muon, SUSY, Particle Flow, and JetMet analysis groups.
- Able to afford Higgs group now, where one of the conveners was asking us to include Higgs datasets, as we have obtained new storage space.

Group	Used Space (TB)
AnalysisOps	21.11
DataOps	1.44
Higgs	2.70
Local	0.72
Muon	11.91
SUSY	42.99
Tau-pflow	5.11
Undefined	32.76
Total	**118.76**

Analysis Operations

- Experience with associated analysis groups
  - Muon

    Mainly interacted with Alessandra Fanfani to arrange write access to the dCache and the fair share configuration during the OctX Dataset request. Local users also requested large portion of muon group datasets for charge ratio analysis.
  - SUSY

    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. Recently Jim Smith from Colorado is managing dataset requests for the SUSY group and we approved one request.
  - Particle Flow

    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.
  - JetMet

    Not much interaction, only with the validation datasets were hosted for this group. Recently one of our postdocs (Dayong Wang) is interacting with us to coordinate the necessary dataset hosting.
  - Higgs

    Andrey Korytov was asking if we could afford Higgs dataset before our recent Lustre storage was added. We could not afford Higgs group datasets at the time. Now we can do it easily.

Summary

- Already exceeded the 7760-HS06 2010 CE milestone, considering to enhance the CE and replace old dead nodes.
- Near the 570TB 2010 SE milestone, planning more SE to fulfill this year's requirement.
- Established a production Lustre+BestMan+GridFTP storage system.
- Satisfied with Lustre, will put dCache non-resilient pools and future storage in Lustre.
- Associated with 4 analysis groups and can take more groups as storage space increases.