UCSD CMS T2 Center Storage System

2010 OSG Storage Forum

Presenter: Terrence Martin

UCSD Storage Breakdown

- Hadoop Based Distributed Storage System
- Bestman SRM
- HDFS Gridftp
- 958 TB Raw Storage
- 147 Storage/Compute Nodes

Storage Hardware

- 147 Compute/Storage Nodes
  - Capacities from 3.4TB to 21.55TB actual capacity
- Cisco 6500 series router/switch
  - 1Gbps copper to storage nodes
  - 10Gbps fiber to Internet2 via Layer 3
  - 10Gbps fiber to FNAL, Layer 2 to Chicago

Latest Node Deployed

- Intel SR2612UR 2U
- 12 3.5 SAS/SATA Disk Bays
- 1 2.5” Internal Disk bay (SSD)
- Intel 5500/5600 CPU Support
- Rear panel access
- 21.55TB Capacity deployed

Networking 2010+

- Cisco 16 Port RJ45 Copper 10Gbps Line Card

- Initial deployment of 8 10Gbps NIC in 8 Nodes for storage performance testing
- Future support for 10Gbps Copper networking within each rack with 4x10Gbps back to the central switch

Hadoop

- Apache Hadoop Project http://hadoop.apache.org/
- UCSD Deployed HDFS Summer 2009
- Why? UCSD required a more reliable and stable storage solution than what we had at the time (dcache)
- Hadoop offers reliable flexible storage access, ease of administration, full data replication (hadoop block level)

Hadoop Day to Day

- Conceptually simple design to manage consisting of a Name Node, Secondary Name Node and Data Nodes.
- Extremely tolerant of disk or node level failure, even more so in very recent versions
- Replication begins quickly and has proven very reliable
- Excellent performance
- Individual nodes can be removed from the cluster from the Name Node
- Consistency checks are fast even with 1 million+ blocks

Hadoop Storage Access

- Local job level access to Hadoop is via HDFS Fuse mount.
- Fuse storage access is read only on worker nodes
- Remote write and read available via Bestman SRM (gateway mode)
- Actual data is read and written with HDFS gridftp
- UCSD also runs a test xrootd hdfs install

UCSD Bestman SRM

- UCSD runs an OSG VDT install version of Bestman in gateway mode
- Storage GUMS Authentication Server
- Bestman SRM server is a dedicated Quad Core Xeon 5345 8GB RAM. Bestman heap size 4GB.
- UCSD developed a custom gridftpd selector module
- A custom selector was required to support 80+ gridftp servers

UCSD Gridftp Selector

- Replaces the default gridftp selector entirely
- Java component reads a list of gridftp servers from a text file, default is once a minute
- Gridftp server selection is random
- Text file is updated via separate external gridftp server tester
- Tester can be as simple as a tcpping, a more complex transfer test, or any combination

Most recent UCSD Gridftp Node Tester

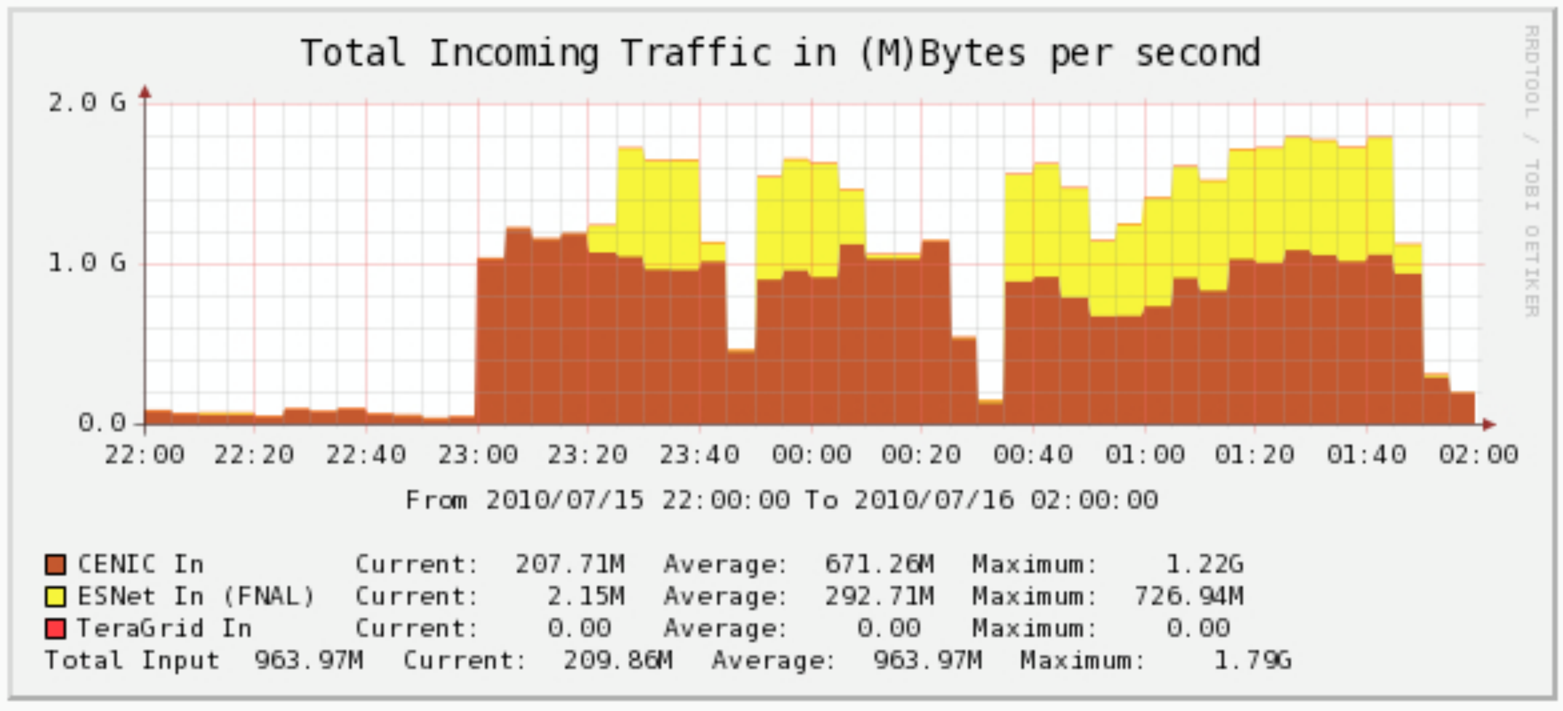
- UCSD has developed a new transfer based tester
- Tester can detect problems like authentication or problems with an active but non-functional server
- Capable of testing the many gridftp servers in parallel extremely quickly
- Can be used to feed a list of nodes to a light weight tcpping based tester.

UCSD SRM Users

- UCSD has a wide selection of users accessing SRM storage
- Users include CMS members, CMS data, Dzero, SBGrid, Scripps Oceanographic HARP group

Throughput Capacity

- In July 2010 UCSD Performed throughput test. Using both of our links and grabbing data from FNAL, Caltech and UNL we hit 14.32Gbps.



From 2010/07/15 22:00:00 To 2010/07/16 02:00:00

```
🟫 CENIC In          Current:  207.71M   Average:  671.26M   Maximum:    1.22G
🟨 ESNet In (FNAL)   Current:    2.15M   Average:  292.71M   Maximum:  726.94M
🟥 TeraGrid In       Current:    0.00    Average:    0.00    Maximum:    0.00
Total Input  963.97M  Current:  209.86M   Average:  963.97M   Maximum:    1.79G
```

Bestman Scalability Study

- UCSD is involved in a scalability study of Bestman and Bestman2
- Haifeng Pi is the lead investigator in this study
- The study makes use of production and test resources at the UCSD T2 center including storage and job submission resources (GlideinWMS)
- Initial Deployment of 10Gbps copper is meant to facilitate this test at high wire speeds

Production Storage Experience

- Bestman and Hadoop experience at UCSD has be very positive
- All main components of the system are reliable and require relatively little support once configured
- Bestman like most Java applications likes to have a lot of heap
- Developing a custom gridftp selector was required