

USCMS Hadoop Review

Brian Bockelman

Sept 15, 2009

Acknowledgements

- This review greatly benefitted from the input of (in no particular order):
 - Michael Thomas
 - Frank Wuerthwein
 - Terrence Martin
 - Haifeng Pi
 - Garhan Attebury
 - Carl Lundstedt
- And a big thanks to Ken for helping put everything together!

Today's Format

- We will be covering “special topics” and details about each site.
 - We assume familiarity with the basic architecture!
- 15 min: The HDFS SE: Support, Risks, and Rewards.
- 10 min: UCSD
- 10 min: Caltech
- 10 min: Nebraska
- 45 min: Q & A

The HDFS SE

- This short presentation concentrates on three issues: support, risks, and rewards.
 - We break HDFS support into software support, packaging and testing, and operational support
 - Risks are broken into technical and non-technical issues.
 - Rewards are quickly highlighted at the end, but I'd rather site admins speak for themselves.

HDFS Software Support

- The software packaging is currently done at Caltech using the RedHat build system, mock.
 - Packaging-related add-ons are developed by Caltech (configuration generation scripts, gmond.conf integration, logrotate, etc)
 - Nebraska selects the patches we want to run and develops new ones as-needed.
 - As much as possible, we want Hadoop to look and feel like a native Linux application.

Software Support FY2010

- In FY2010, our focus will be:
 - Combine efforts with Cloudera on HDFS patches.
 - Always work on upstreaming patches to reduce necessary involvement of our experts.
 - Utilize the R&D capital we have in place at UCSD, Caltech, and Nebraska, but make sure we don't depend on it.

Packaging, FY2010

- RPM packaging will migrate to the VDT team.
 - This will be done when VDT has the infrastructure to support RPM builds (not in place!)
- We will at least attempt integration with the Cloudera distribution. If nothing else, share techniques (i.e., Cloudera has written man pages. We want to incorporate those).

Packaging and Testing, FY2010

- There will be effort at UCSD to maintain small testbeds for all platforms and a scalability testbed.
- We will attempt to minimize changes to decrease support costs; feature releases every 6 months, otherwise only bugfixes.
 - All updates will be tested on small testbeds
 - Feature releases will be scalability-tested as necessary.
- There will still be a “bleeding edge” for sites that want it, but we won’t support it or recommend it.

Operations

- For support questions and ticketing, we will be joining forces with the OSG-Storage team.
 - OSG-Storage currently maintains a HDFS install, but won't be the testbed – they only should be using the released versions.
 - We will use the ticketing system and weekly meetings already in place.
 - However, support will be community-based and depend heavily on the individuals on osg-hadoop@opensciencegrid.org; more success means more available support.

Technical Risks

- FUSE-DFS glue layer has been buggy – in terms of memory leaks and serving incorrect data.
 - Lots of effort here; currently reliable.
- BeStMan has at least one known failure mode due to the Globus container it uses.
 - BeStMan2 should be much better because it uses a Tomcat container instead.
- Memory Usage of gridftp-hdfs. Whereas we previously used many, many streams without care, care must be taken to not exhaust memory with multiple streams or too many processes per node.
 - D0 is a big abuser of resources.

Technical Risks

- An aside:
 - Assuming a binomial probability distribution, the expected value of number of blocks lost during a simultaneous 2-disk loss is:
$$E = (\# \text{ of blocks}) * 4 * (\text{size disk 1}) * (\text{size disk 2}) / (\text{size HDFS})^2$$
 - For a file-based system, take # of blocks = # of files.
 - Hence, the expected ratio of loss in a file-based system compared to block based system is simply the average # of blocks per file.
 - For our HDFS deploys, this varies between 2:1 (lots of user files) and 10:1 (mostly CMS files).

Non-technical Risks

- “Expert dependency”: We are growing our own pool of local experts. During this growth phase, it would be a big setback to lose an expert.
 - Relatively short “expert training” time; ~6 months.
 - Documentation always helps.
- Funding streams: To provide a quality product, we need a way to do packaging and testing.
 - We are mitigating this by increased collaboration with Cloudera so we might be able to use their distribution.
- Upstream risk: Facebook, Yahoo, and Cloudera contribute a huge % of the code.
 - This is mitigated by the fact it is mostly “feature complete” for things we are interested in.

Rewards

- Reliability: stop loss of user files at T2. Stop file access errors.
- Simplified setup and install
 - Possible eventual inclusion in upstream distro.
- Reduce administrative costs.
 - Software has been more reliable; no real scaling issues.
 - At 2x replicas, we can pick and chose when we feel like recovering lost nodes.
- Outside improvement. Like the Linux kernel, when another company invests in it, we gain the improvements for free. Total investment per year is in the millions of dollars.
- Scalability.

Summary

- Transition from volunteer, bottom-up support structures to more stable, organized ones.
 - Reduces risk from personnel churn.
- Short-term risks include a few technical issues (we will heavily benefit from an open-source BeStMan2); Long-term risks tend toward personnel and funding issues.
- Rewards include decreased cost of operation, increased reliability, and constant improvements through heavy investment by commercial companies.

Questions?

- Up next, UCSD

UCSD Reasons

- **BestMan is the most scalable srm implementation available.** For T2 operations this is important as the required scale depends on the total number of CPUs users of a given T2 can obtain worldwide.
- **Hdfs is the only storage implementation with replication that works.** For T2 operations this is important because T2s have no archival backup, nor does CMS support any backup scheme for its normal users. At UCSD, every disk that is lost leads to file losses unless we deploy hdfs.

Responsibilities of UCSD T2

- **Support** 3 physics groups with 50TB each
 - 2008/09: EWK, top, e/gamma
 - 2009/10: tracking, top, e/gamma
- **Support** USCMS user community
 - Assigned to: UCSB, UCSD, UCR => ~50 people.
 - Personal disk space, hosting data, ...
 - Currently: 58 TB across 600k files of private data
- **Support** the entire CMS community for data analysis and MC production.

UCSD hdfs system

390 TB (Total)

106 Data Nodes

106 Gridftp Servers

1 Bestman Server

1 Namenode

1 Secondary Namenode

All data nodes are also worker nodes.

Typically 2x4 core CPU plus 4xSATA disk.

Disks range from 1-2TB each.

Disks are software RAID0 for performance.

Everything on public IP.

Presently purchasing upgrade to allow for 2x350TB,
i.e. full replication of the entire space of 350TB usable.

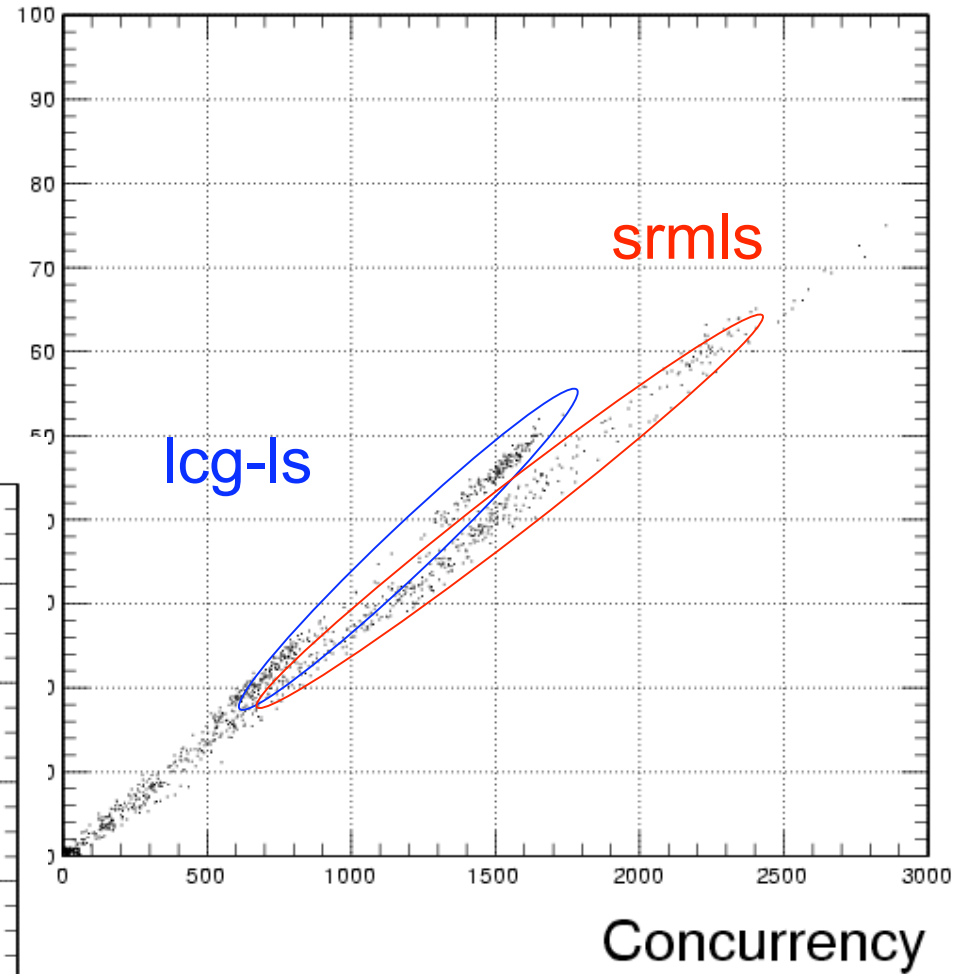
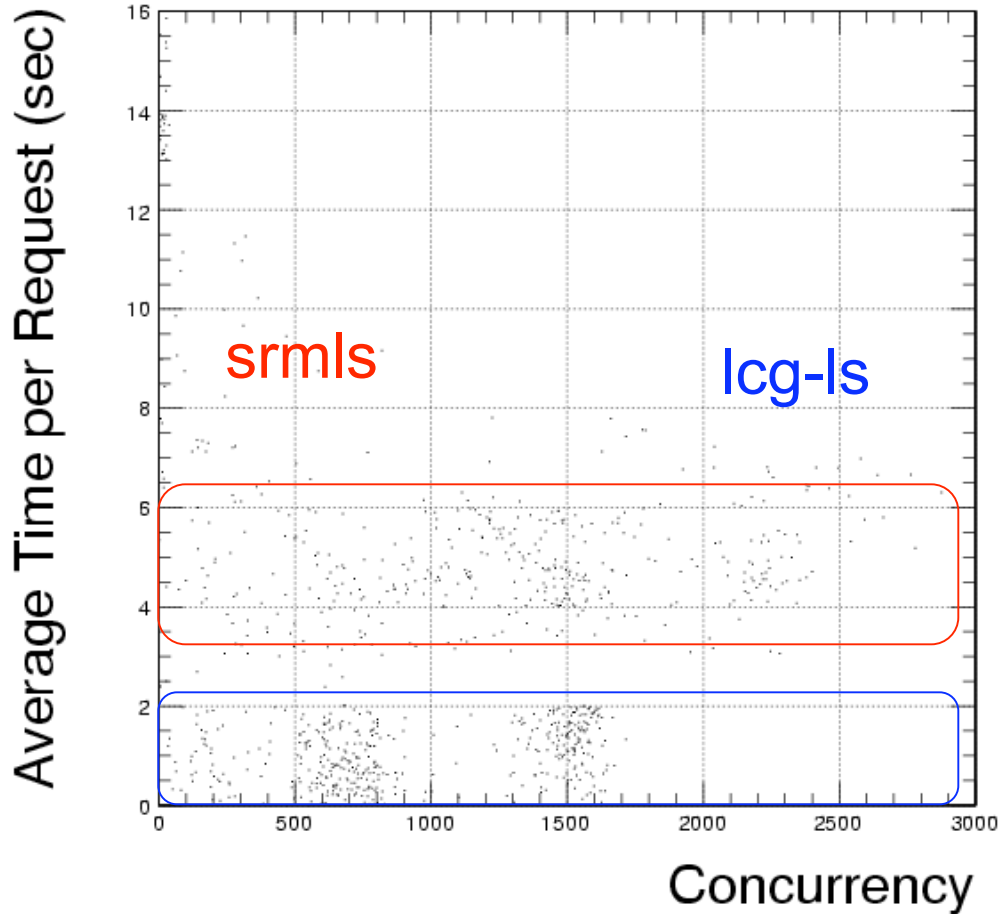
Note: CMS data analysis is presently IO limited at all sites worldwide. The full replication may thus have unintended additional positive impact on overall performance.

Scalability

Concurrency = # of simultaneous Is.

Measured via WAN access from jobs submitted to grid.

Processing Rate



Avg time per request
roughly independent of
of simultaneous Is.

Gridftp Plugin for BestMan

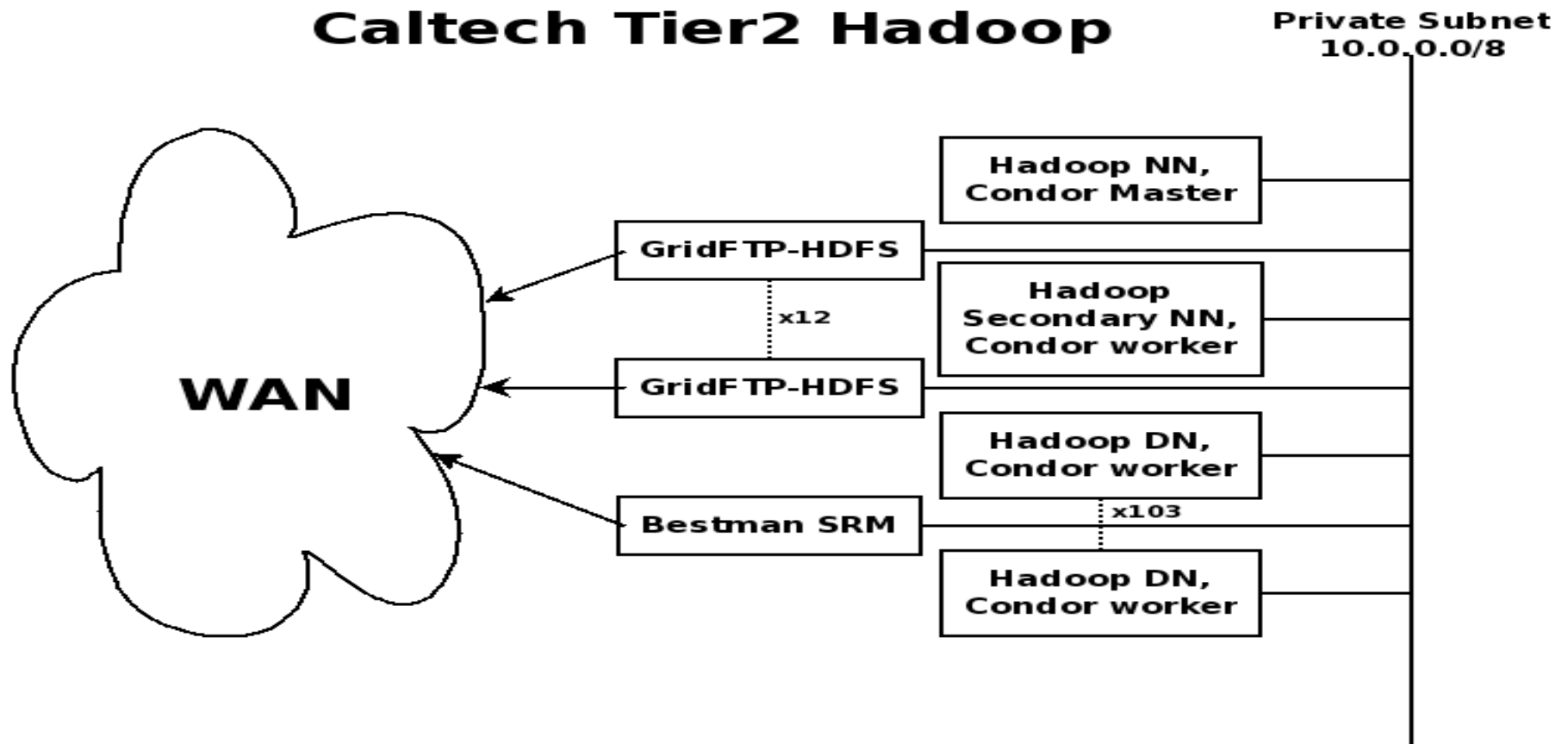
- Implements the Bestman Gridftp selection class
- Gridftp servers selected randomly from list of available
 - List of available read from a file (once every 10 requests or once every minute whichever is sooner)
- File updated every minute in an atomic way with an external program which tests if the gridftp servers are listening
 - Decision about which gftp servers to include is independent of the selection among the included gftp servers!
- Allows a flexible way to manage 100+ Gridftp servers
- More advanced tests possible using user proxy authentication

With 100+ gftp servers, it is important to automatically exclude those that are not presently working.

Questions?

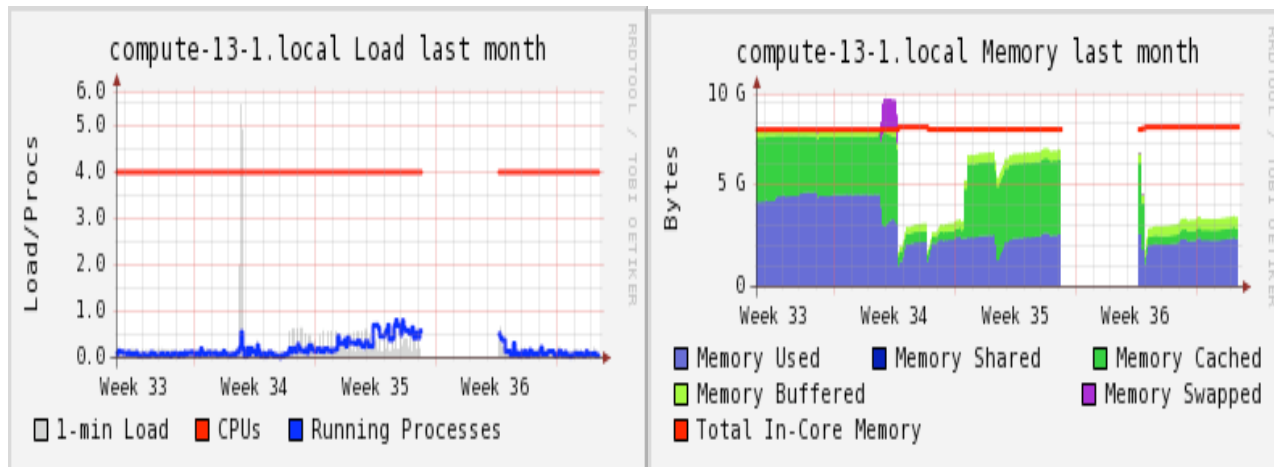
- Up next, Caltech

Caltech Tier2 Hadoop



Namenode

- 8 cores, 8GB RAM
- Runs on same host as condor collector/negotiator
- NN JVM currently using 1.5GB/4GB
- private network access only; no public IP



Namenode UI

NameNode 'compute-13-1.local:9000'

Started: Thu Sep 03 22:46:17 PDT 2009
Version: 0.19.2-dev, r748415
Compiled: Mon Mar 23 15:21:37 PDT 2009 by wart
Upgrades: There are no upgrades in progress.

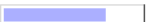
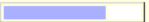








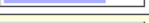


[Browse the filesystem](#)
[Namenode Logs](#)

Cluster Summary

159010 files and directories, 1539622 blocks = 1698632 total. Heap Size is 1.67 GB / 3.56 GB (46%)

Configured Capacity : 342.64 TB
DFS Used : 299.95 TB
Non DFS Used : 269.64 GB
DFS Remaining : 42.43 TB
DFS Used% : 87.54 %
DFS Remaining% : 12.38 %
Live Nodes : 103
Dead Nodes : 1

Live Datanodes : 103

Node	Last Contact	Admin State	Configured Capacity (TB)	Used (TB)	Non DFS Used (TB)	Remaining (TB)	Used (%)	Used (%)	Remaining (%)	Blocks
compute-10-19	0	In Service	3.41	2.99	0	0.42	87.6		12.4	26929
compute-10-20	0	In Service	2.5	2.19	0	0.31	87.7		12.3	19527
compute-10-21	2	In Service	5.23	4.4	0	0.83	84.18		15.82	38801
compute-10-22	2	In Service	3.41	2.99	0	0.42	87.63		12.37	26773
compute-10-23	0	In Service	5.23	4.49	0	0.74	85.85		14.15	39665
compute-10-24	0	In Service	5.23	4.7	0	0.53	89.91		10.09	42132
compute-11-11	2	In Service	1.61	1.42	0.01	0.18	88.1		11.38	13198
compute-11-12	1	In Service	1.61	1.43	0	0.17	88.91		10.86	12919
compute-11-9	2	In Service	1.61	1.47	0	0.14	91.04		8.96	13197
compute-12-35	1	In Service	2.5	2.13	0	0.37	85.21		14.79	18635
compute-12-36	0	In Service	2.5	2.24	0	0.26	89.75		10.25	19802
compute-12-37	0	In Service	2.5	2.2	0	0.3	88.15		11.85	19515
compute-12-38	2	In Service	2.5	2.21	0	0.3	88.2		11.8	19864

Datanodes

- 103 active datanodes
 - ★ Most are worker nodes with 2,3,4 data disks, 8 cores, 2GB per job slots
 - ★ 1 16-disk JBOD, 21TB
 - ★ 4 dedicated 2U disk servers, 6.5TB each
 - ★ 4 dedicated 4U disk servers, 13TB each
 - ★ 2 Sun x4500 Thumpers with Solaris, 36TB each
 - ★ private network access only; no public ips
 - ★ fully automated installation with Rocks
 - ★ Rack-aware: blocks get replicated on separate racks

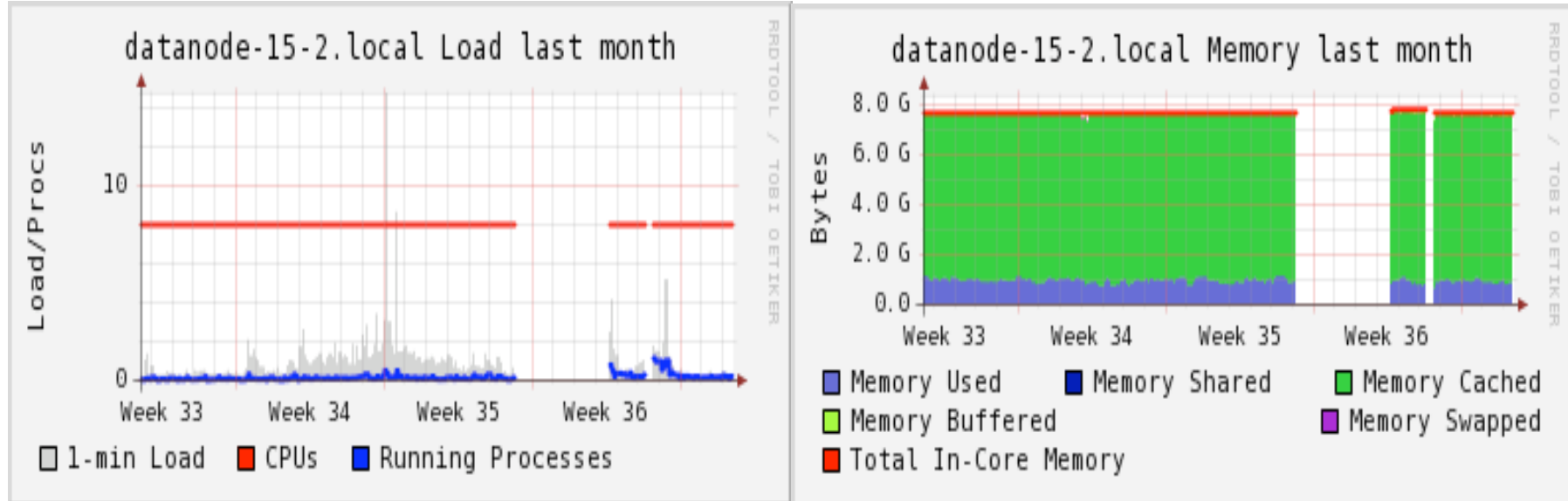
Dedicated datanodes use raid5 for large partitions

- ★ JBOD uses 16 independent disks

Worker nodes use independent disks

Datanodes

- Relatively little resource consumption, even on large datanodes



Other nodes

- Secondary NN
 - ✳ Shared with worker node
 - ✳ OOM twice (fixed)
 - ✳ processes checkpoints once per hour
 - ✳ private network access only; no public IP
 - ✳ Fully automated installation with Rocks

Bestman SRM

- ✳ Dedicated 8 core, 8GB
- ✳ Fully automated installation with Rocks
- ✳ Clocked at 200Hz for lcg-ls

Gridftp

- ✳ 4 dedicated 8-core, 16GB, 2 x 10GbE
- ✳ 8 dedicated 4-core, 8GB, 2 x 1GbE
- ✳ Fully automated installation with Rocks
- ✳ Also act as https doors to hdfs using apache + mod_ssl + fuse

Local Effort

- Responsible for RPM packaging
- Repond to support requests on osg-hadoop list
- Original authors of hadoop_chronicle (now part of OSG storage toolkit)
- Authors of gridftpspy
- FDT + HDFS integration


The Hadoop Chronicle - Mozilla Firefox 3.5 Beta 4

File Edit View History Bookmarks Tools Help

caltech.edu https://cms.hep.caltech.edu/hadoop/ Google

Most Visited EVO docs Caltech T2 Fedora

The Hadoop Chronicle



Selected or last chronicle

2009_06_26_08:30

=====

The Hadoop Chronicle | 46 % | Fri Jun.26.2009 08:30

=====

Global storage

Configured Capacity: 191813069336576 (174.45 TB)
 Present Capacity: 191707600577536 (174.36 TB)
 DFS Remaining: 102718547960182 (93.42 TB)
 DFS Used: 88989052617354 (80.94 TB)
 DFS Used%: 46.42%

/store/ area

Path	Size(GB)	#Files	#Dirs
/store/PhEEx_LoadTest07	667	262	668
/store/data	24337	33474	462
/store/mc	2353	2146	15
/store/unmerged	438	3416	172
/store/user	8461	25234	433

User area

Path	Size(GB)	#Files	#Dirs
/store/user/burt	0	2	1
/store/user/chiorbo	902	5341	127
/store/user/dkcira	0	17	13
/store/user/dorian	41	286	1
/store/user/hpi	2	6	21
/store/user/ligioi	0	2	1
/store/user/litvin	0	3	8
/store/user/oatramen	3178	1036	77
/store/user/ssekmen	0	4	4
/store/user/test	0	2	5
/store/user/tucker	0	17	6
/store/user/uscms0377	10	7	1
/store/user/uscms0755	614	2754	3
/store/user/vlitvin	3709	15754	153
/store/user/wart	1	3	1

System health

Total size: 38929932017766 B (Total open files size: 3087007744 B)
 Total dirs: 1765
 Total files: 64551 (Files currently being written: 2)
 Total blocks (validated): 358503 (avg. block size 108590254 B) (Total open file blocks (not validated): 23)
 Minimally replicated blocks: 358503 (100.0 %)
 Over-replicated blocks: 63 (0.017573075 %)
 Under-replicated blocks: 0 (0.0 %)
 Mis-replicated blocks: 0 (0.0 %)
 Default replication factor: 2
 Average block replication: 2.349721
 Corrupt blocks: 0
 Missing replicas: 0 (0.0 %)

All Chronicles

- 2009_06_26_08:30
- 2009_06_25_19:42
- 2009_06_25_08:30
- 2009_06_24_19:42
- 2009_06_24_08:30
- 2009_06_23_19:42
- 2009_06_23_08:30
- 2009_06_22_19:42
- 2009_06_22_08:30
- 2009_06_21_19:42
- 2009_06_21_08:30
- 2009_06_20_19:42
- 2009_06_20_08:30
- 2009_06_19_19:42
- 2009_06_19_08:30
- 2009_06_18_19:42
- 2009_06_18_08:30
- 2009_06_17_19:42
- 2009_06_17_08:30
- 2009_06_16_19:42
- 2009_06_16_08:30
- 2009_06_15_19:42
- 2009_06_15_08:30
- 2009_06_14_19:42
- 2009_06_14_08:30
- 2009_06_13_19:42
- 2009_06_13_08:30
- 2009_06_12_19:42
- 2009_06_12_08:30
- 2009_06_11_19:42
- 2009_06_11_08:30
- 2009_06_10_19:42
- 2009_06_10_08:30
- 2009_06_09_19:42
- 2009_06_09_08:30
- 2009_06_08_19:42
- 2009_06_08_08:30
- 2009_06_07_19:42
- 2009_06_07_08:30
- 2009_06_06_19:42
- 2009_06_06_08:30
- 2009_06_05_19:42
- 2009_06_05_08:30
- 2009_06_04_19:42
- 2009_06_04_08:30
- 2009_06_03_19:42
- 2009_06_03_08:30
- 2009_06_02_19:42
- 2009_06_02_08:30
- 2009_06_01_19:42
- 2009_06_01_08:30
- 2009_05_31_19:42
- 2009_05_31_18:55
- 2009_05_31_18:33
- 2009_05_31_17:52

cms.hep.caltech.edu FoxyProxy: Patterns

gridftpspy

gridftpspy

File

Main

Log file:

Job PID	userid	User Name	Remote host	Start Date	End Date	Size	Rate	# buffers	direction	file
21996	cmsprod	Andrea Sciaba	vocms36.cern.ch	Tue Jun 30 03:02:46 PM PDT 2009	Tue Jun 30 03:02:47 PM PDT 2009	41.472	40.5KB/s	1/1	in	/store/unmerged/SAM/testSRM/SAM-cit-se2.ultraight.org/cg-util/testfile-gt-mm-gt-notoken-20090701-000239.txt
22129	uscms0994	Paul Rossman	cmsrvr45.fnal.gov	Tue Jun 30 03:03:29 PM PDT 2009		0		1/1	in	/mnt/hadoop/store/PhEDEx_LoadTest07/LoadTest07_Prod_Caltech/LoadTest07_Caltech_A6
22254	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:04:27 PM PDT 2009	Tue Jun 30 03:05:03 PM PDT 2009	2740771472	72.61MB/s	162/645	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_FINAL/Caltech/33/LoadTest07_FINAL_53_Ja2B17u9pYQEIOH_93
22408	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:51 PM PDT 2009	Tue Jun 30 03:07:39 PM PDT 2009	2684354560	53.33MB/s	172/521	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_CE_oOMCgZ3g43nD7UVX_635
22499	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:51 PM PDT 2009	Tue Jun 30 03:07:44 PM PDT 2009	2684354560	48.3MB/s	172/560	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_AD_mknV3Ufz9vV61fO_635
22590	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:51 PM PDT 2009	Tue Jun 30 03:07:35 PM PDT 2009	2684354560	58.18MB/s	130/957	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_AD_S5dLzjI9Dmcpvc4_635
22771	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:54 PM PDT 2009	Tue Jun 30 03:07:46 PM PDT 2009	2684354560	49.23MB/s	479/726	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_59_XgDRxKvG9kP9G_635
22894	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:56 PM PDT 2009	Tue Jun 30 03:07:59 PM PDT 2009	2684354560	40.63MB/s	476/821	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_F3_9yYosK1I4vulci3_635
23220	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:09:25 PM PDT 2009	Tue Jun 30 03:09:59 PM PDT 2009	2684354560	75.29MB/s	726/762	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_ED_8wUlaU3GmmElNny4_633
23363	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:11:37 PM PDT 2009	Tue Jun 30 03:12:07 PM PDT 2009	2684354560	85.33MB/s	71/745	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/634/LoadTest07_UCSD_05_ImlB4NoAuGr0V6A_634
23513	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:20:11 PM PDT 2009	Tue Jun 30 03:20:43 PM PDT 2009	2684354560	80.0MB/s	502/594	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_F2_W64wVwHpoESj6ZA_633
23645	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:37 PM PDT 2009	Tue Jun 30 03:23:33 PM PDT 2009	2684354560	45.71MB/s	274/477	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_C3_kWhaNoKhZOp2oXTe_635
23745	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:35 PM PDT 2009	2684354560	45.71MB/s	391/431	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_BD_SlIjIw6itsyTyOo_635
23836	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:38 PM PDT 2009	Tue Jun 30 03:23:45 PM PDT 2009	2684354560	38.21MB/s	314/437	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_C7_xKqgI6uqjIzArcC_635
23953	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:37 PM PDT 2009	Tue Jun 30 03:23:37 PM PDT 2009	2684354560	44.14MB/s	342/616	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_4D_yuk8Uc0d6B4HbTA_633
24136	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:37 PM PDT 2009	2684354560	44.14MB/s	1/580	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_A5_zimTZkhaDhFXMmx_635
24036	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:45 PM PDT 2009	2684354560	38.79MB/s	657/875	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_57_KNt4sxxkLrAltkW5_635
24380	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:42 PM PDT 2009	Tue Jun 30 03:23:47 PM PDT 2009	2684354560	39.38MB/s	359/697	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/634/LoadTest07_UCSD_CD_NYXVYkG0GjV008U_634
24637	cmsprod	Andrea Sciaba	cithep200.ultraight.org	Tue Jun 30 03:26:46 PM PDT 2009	Tue Jun 30 03:26:46 PM PDT 2009	20000		1/1	in	/store/unmerged/SAM/StageOutTest-610-Tue-Jun-30-15-26-39-2009
24758	uscms0377	Michael Thomas	cithep252.ultraight.org	Tue Jun 30 03:28:06 PM PDT 2009	Tue Jun 30 03:28:06 PM PDT 2009	128		1/1	out	/mnt/hadoop/rsv/124640080-storage-probe-test-file-remote.31742
24877	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:28:55 PM PDT 2009	Tue Jun 30 03:29:28 PM PDT 2009	2684354560	77.58MB/s	118/804	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_6D_kwPFKcxalJEI0D4_633
25017	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:29:46 PM PDT 2009	Tue Jun 30 03:30:17 PM PDT 2009	2684354560	82.58MB/s	601/685	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_40_hRWRTRfzK0VsjSx_633

[25017] Tue Jun 30 15:30:19 2009 :: Closed connection from cmsfts1.fnal.gov:49876

gridftpspy #2

File

Main

Log file:

Job PID	userid	User Name	Remote host	Start Date	End Date	Size	Rate	# buffers	direction	file
8634	cmsprod	Andrea Sciaba	vocms36.cern.ch	Tue Jun 30 03:02:34 PM PDT 2009	Tue Jun 30 03:02:34 PM PDT 2009	41.472		1/1	out	/mnt/hadoop/store/unmerged/SAM/testSRM/SAM-cit-se2.ultraight.org/cg-util/testfile-cp-notoken-20090701-000144.txt
8963	cmsprod	Andrea Sciaba	vocms36.cern.ch	Tue Jun 30 03:03:29 PM PDT 2009	Tue Jun 30 03:03:30 PM PDT 2009	41.472	40.5KB/s	1/1	in	/store/unmerged/SAM/testSRM/SAM-cit-se2.ultraight.org/cg-util/testfile-gt-notoken-20090701-00023.txt
9085	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:04:25 PM PDT 2009	Tue Jun 30 03:07:58 PM PDT 2009	2563227279	1.73MB/s	663/663	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_FINAL/Caltech/33/LoadTest07_FINAL_D1_KkIaOaQoJmFq_93
9176	cmsprod	Andrea Sciaba	vocms36.cern.ch	Tue Jun 30 03:04:27 PM PDT 2009	Tue Jun 30 03:04:28 PM PDT 2009	41.472	40.5KB/s	1/1	in	/store/unmerged/SAM/testSRM/SAM-cit-se2.ultraight.org/cg-util/testfile-is-notoken-20090701-000418.txt
9361	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:51 PM PDT 2009	Tue Jun 30 03:07:56 PM PDT 2009	2684354560	39.38MB/s	543/901	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_59_76fgMyxWURRHn6l_635
9452	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:51 PM PDT 2009	Tue Jun 30 03:07:40 PM PDT 2009	2684354560	52.24MB/s	461/727	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_A8_3soXpGgkGsX0sR_635
9552	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:51 PM PDT 2009	Tue Jun 30 03:07:40 PM PDT 2009	2684354560	52.24MB/s	590/681	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_0D_zimxU0h5PoawP_635
9818	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:06:56 PM PDT 2009	Tue Jun 30 03:07:45 PM PDT 2009	2684354560	52.24MB/s	1456/76	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_97_q255UTaRQrBPF87_635
10024	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:09:10 PM PDT 2009	Tue Jun 30 03:09:52 PM PDT 2009	2684354560	60.95MB/s	812/948	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_0C_VCAvs7ZnagVfmyJ_633
10179	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:11:37 PM PDT 2009	Tue Jun 30 03:12:27 PM PDT 2009	2684354560	51.2MB/s	550/829	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/634/LoadTest07_UCSD_4D_aAlpkiCaOXIDraD_634
10325	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:13:47 PM PDT 2009	Tue Jun 30 03:14:25 PM PDT 2009	2684354560	67.37MB/s	685/896	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/634/LoadTest07_UCSD_5B_80lwm9enupanXIE_634
10500	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:38 PM PDT 2009	Tue Jun 30 03:23:42 PM PDT 2009	2684354560	40.0MB/s	129/601	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_1D_gWkHmlIyeCLVeL3_635
10591	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:38 PM PDT 2009	Tue Jun 30 03:24:06 PM PDT 2009	2684354560	29.09MB/s	653/853	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_DA_xbKPA07vsFfjdr_635
10682	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:51 PM PDT 2009	2684354560	35.56MB/s	320/858	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_FA_pydlHbqVzcfDpZ_635
10789	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:57 PM PDT 2009	2684354560	32.82MB/s	493/927	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_94_06zOrvFDHhJM3E_635
10888	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:29 PM PDT 2009	2684354560	51.2MB/s	224/488	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_A1_gWkHmlIyeCLVeL3_635
11008	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:56 PM PDT 2009	2684354560	33.25MB/s	308/855	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_8C_FPhyDK4dKwy1C1u_635
11116	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:22:39 PM PDT 2009	Tue Jun 30 03:23:48 PM PDT 2009	2684354560	37.1MB/s	369/916	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_07_FNhyYMLVAF80_635
11488	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:24:37 PM PDT 2009	Tue Jun 30 03:25:09 PM PDT 2009	2684354560	80.0MB/s	677/733	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_40_7iGmV125TnBnVs_633
11640	uscms0377	Michael Thomas	cithep252.ultraight.org	Tue Jun 30 03:28:03 PM PDT 2009	Tue Jun 30 03:28:03 PM PDT 2009	128		1/1	in	/rsv/124640080-storage-probe-test-file-remote.31742
11761	cmsprod	Andrea Sciaba	t2-headnode.ultraight.org	Tue Jun 30 03:28:35 PM PDT 2009	Tue Jun 30 03:28:35 PM PDT 2009	20000		1/1	in	/store/unmerged/SAM/StageOutTest-9462-Tue-Jun-30-15-28-27-2009
11882	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:29:00 PM PDT 2009	Tue Jun 30 03:29:43 PM PDT 2009	2684354560	59.53MB/s	336/952	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/634/LoadTest07_UCSD_FC_dBUyJhL7Kd9FDRa_634

[11882] Tue Jun 30 15:29:44 2009 :: Closed connection from cmsfts1.fnal.gov:49535

Concerns

- gridftp-hdfs is memory hungry
 - ★ Increase per-server memory for reordering data streams
 - ★ Deploy multiple servers
 - ★ Max rate of 300MB/s per server, but only in a controlled environment
 - ★ However, have seen peaks up to 7.5Gbps across all gridftp servers, so performance is acceptable
- gridftp client error messages under load are wrong
 - ★ “end of file” and “out of disk space” when xinetd refuses connections
- replicas = 1 can be risky
 - ★ No different than what we have always done, except that # files affected in case of disk loss is greater

Questions?

- Up next, Nebraska

Hadoop SE Review @ Nebraska



- History

- Started exploring HDFS side by side with dCache
 - dCache on vaults, HDFS on workers
- Fixed up the 'glue' to make HDFS work with SRM / GridFTP
- Maintained two SEs via TFC magic (still doing this today)
- Lots of testing, trial and error, admin experience
- Slowly migrated bulk of data to HDFS on worker nodes
- HDFS as primary SE, vaults slowly being migrated to HDFS

Hadoop SE Review @ Nebraska



- Current HDFS Deployment for CMS
 - All worker nodes + handful of vaults as datanodes
 - ~110 datanodes in total
 - Non-optimal datanode sizes (vault vs worker sizes)
 - ~370TB raw for past few months
 - ~275k files
 - RPM based deployments of HDFS, Bestman, and GridFTP
 - Still a few 'manual' holdovers from our initial attempts in place
- Future plans
 - HA deployment for namenode
 - Rack awareness (to help with non-optimal datanode sizes / availability)
 - More datanodes as needed to meet storage requirements

Hadoop SE Review @ Nebraska



- Other local work with Hadoop
 - Two non-CMS Hadoop instances (Prairiefire / Bugeater)
 - Used for both mass storage and mapreduce functionality
 - Students working with mapreduce in their research
 - Students working on HDFS improvements / utilities
 - Collective knowledge and experience among all our personnel.

Questions?

- We now have a 45 minute timeslot for reviewer questions.