



Hadoop experience in USCMS

Nov. 12, 2009

Michael Thomas



What is Hadoop

Map-Reduce plus the HDFS filesystem implemented in java

Map-Reduce is a highly parallelized distributed computing system

HDFS is the distributed cluster filesystem

*** This is the feature that we are most interested in**

Open source project hosted by Apache

Commercial support available from Cloudera

Used throughout Yahoo. Cloudera and Yahoo are major contributors to the Apache Hadoop project.



HDFS



Distributed Cluster filesystem

Extremely scalable – Yahoo uses it for multi-PB storage

Easy to manage – few services and little hardware overhead

Files split into blocks and spread across multiple cluster datanodes

- ★ **64MB blocks default, configurable**
- ★ **Block-level decomposition avoids 'hot-file' access bottlenecks**
- ★ **Block-level decomposition means the loss of multiple data nodes will result in the loss of more files than file-level decomposition**

Not 100% posix compliant

- ★ **non-sequential writes not supported**
- ★ **Not a replacement for NFS**



HDFS Services



Namenode – Manages the filesystem namespace operations

- ★ File/directory creation/deletion
- ★ Block allocation/removal
- ★ Block locations

Datanode – Stores file blocks on one or more disk partitions

Secondary Namenode – Helper service for merging namespace changes

Services communicate through java RPC, with some functionality exposed through http interfaces



Namenode (NN)

Purpose is similar to dCache PNFS

Keeps track of entire fs image

- * The entire filesystem directory structure
- * The file block -> datanode mapping
- * Block replication level
- * ~1GB per 1e6 blocks recommended

Entire namespace is stored in memory, but persisted to disk

- * Block locations not persisted to disk
- * All namespace requests served from memory
- * Fsync across entire namespace is really fast



Namenode Journals

NN fs image is read from disk only once at startup.

Any changes to the namespace (mkdir, rm) are written to one or more journal files (local disk, NFS, ...)

Journal is periodically merged with the fs image

Merging can temporarily require extra memory to store two copies of fs image at once.



Secondary NN

The name is misleading... this is NOT a backup namenode or hot spare namenode. It does NOT respond to namespace requests.

Optional checkpoint server for offloading the NN journal -> fsimage merges

- **Download fs image from namenode (once)**
- **Periodically download journal from namenode**
- **Merge journal and fs image**
- **Uploaded merged fs image back to namenode**

Contents of merged fsimage can be manually copied to NN in case of namenode corruption or failure.



Datanode (DN)

Purpose is similar to dCache pool

Stores file block metadata and file block contents in one or more local disk partitions. Datanode scales well with # local partitions

- * Caltech is using one per local disk (2-4 per datanode)**
- * Nebraska has 48 individual partitions on Sun Thumpers**

Sends heartbeat to namenode every 3 seconds

Sends full block report to namenode every hour

Namenode uses report + heartbeats to keep track of which block replicas are still accessible



Client access

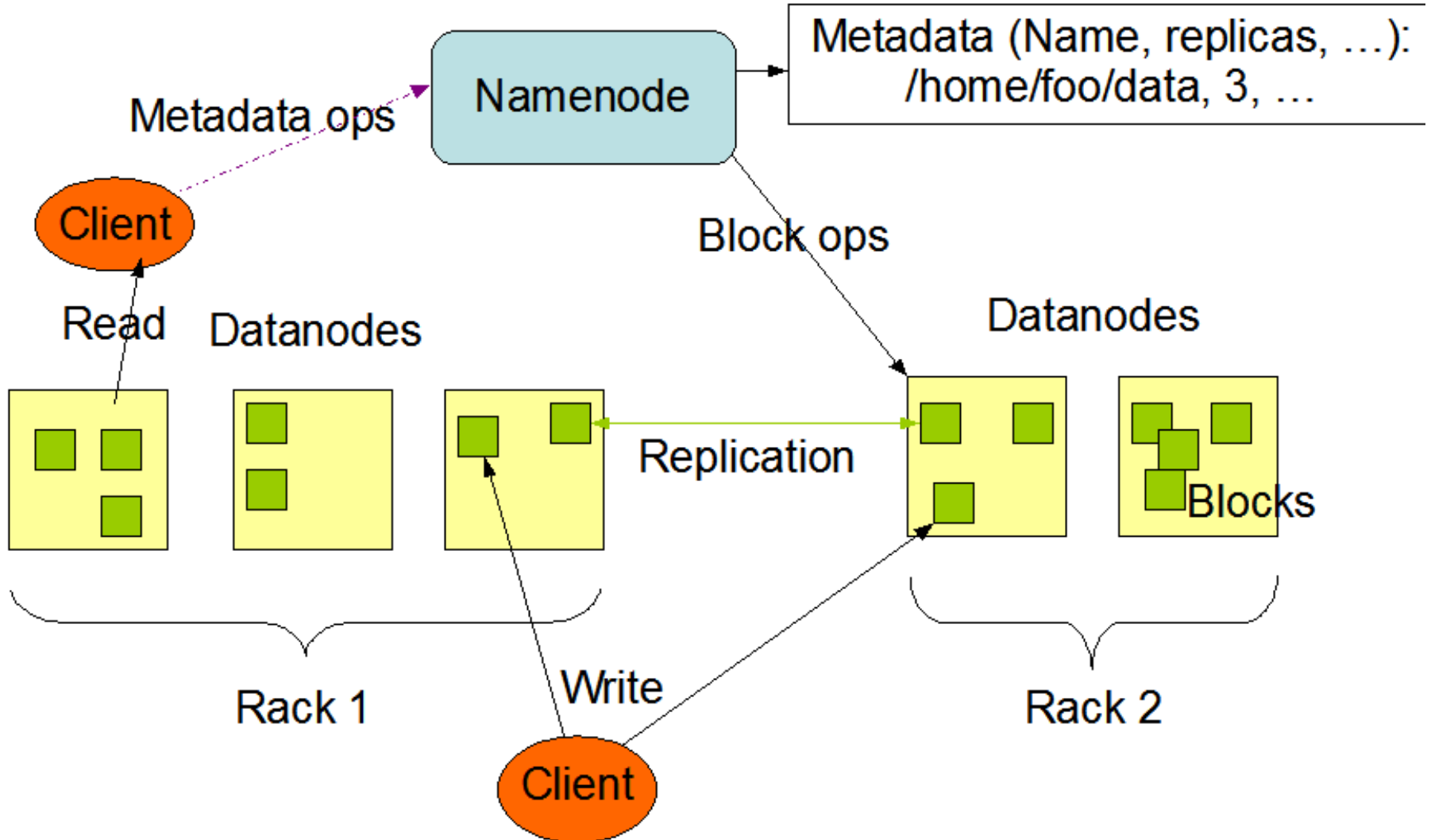
When a client requests a file, it first contacts the namenode for namespace information.

The namenode looks up the block locations for the requested files, and returns the datanodes that contain the requested blocks

The client contacts the datanodes directly to retrieve the file contents from the blocks on the datanodes



Hadoop Architecture





Native client



A native java client can be used to perform all file and management operations

All operations use native Hadoop java APIs



FUSE client

FUSE == Filesystem in Userspace

Presents a posix-like interface to arbitrary backend storage systems (ntfs, lustre, ssh)

HDFS fuse module provides posix interface to HDFS using the HDFS APIs. Allows the use of rm, mkdir, cat, and other standard filesystem commands on HDFS.

HDFS does not support non-sequential (random) writes

- * root TFile can't write directly to HDFS fuse, but not really necessary for CMS**

Random reads are ok



Gridftp/SRM clients



Gridftp could write to HDFS+FUSE with a single stream

Multiple streams will fail due to non-sequential writes

UNL developed a GridFTP dsi module to buffer multiple streams so that data can be written to HDFS sequentially

Bestman SRM can perform namespace operations by using FUSE

- ★ Running in gateway mode
- ★ srmrm, srmls, srmkdir
- ★ Treats hdfs as local posix filesystem



Hadoop monitoring



Nagios

- * `check_hadoop_health` – parses output of 'hadoop fsck'
- * `check_jmx` – blockverify failures, datanode space
- * `check_hadoop_checkpoint` – parses secondary nn logs to make sure checkpoints are occurring

Ganglia

- * Native integration with Hadoop
- * Many internal parameters

MonALISA

- * Collects Ganglia parameters

gridftpspy

Hadoop Chronicle

jconsole

hadoop native web pages




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/PhEDEx_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





Hadoop NameNode compute-13-1.local:9000 - Mozilla Firefox 3.5 Beta 4

File Edit View History Bookmarks Tools Help

http://t2-headnode.ultralight.org:50070/dfshealth.jsp

Most Visited EVO docs Caltech T2 Fedora

Hadoop NameNode com...



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

Started: Tue May 26 12:12:00 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

66825 files and directories, 359972 blocks = 426797 total. Heap Size is 269.38 MB / 888.94 MB (30%)

Configured Capacity : 174.45 TB
DFS Used : 81.13 TB
Non DFS Used : 98.23 GB
DFS Remaining : 93.23 TB
DFS Used% : 46.51 %
DFS Remaining% : 53.44 %
Live Nodes : 65
Dead Nodes : 6

Live Datanodes : 65

Node	Last Contact	Admin State	Configured Capacity (TB)	Used (TB)	Non DFS Used (TB)	Remaining (TB)	Used (%)	Used (%)	Remaining (%)	Blocks
compute-11-11	2	In Service	1.61	0.75	0	0.86	46.87		53.13	7579
compute-11-12	1	In Service	1.61	0.82	0	0.79	50.93		49.07	8252
compute-11-9	1	In Service	1.61	0.8	0	0.81	49.54		50.46	7998
compute-14-10	0	In Service	1.61	0.82	0	0.79	50.87		49.13	8092
compute-14-11	2	In Service	1.61	0.82	0	0.79	51		49	8432
compute-14-12	1	In Service	1.61	0.81	0	0.8	50.17		49.83	8325
compute-14-13	0	In Service	1.61	0.83	0	0.78	51.36		48.64	8465
compute-14-14	2	In Service	1.61	0.81	0	0.8	50.26		49.74	8156
compute-14-15	1	In Service	1.61	0.83	0	0.78	51.27		48.73	8342
compute-14-16	2	In Service	1.61	0.79	0	0.82	49.2		50.8	8057
compute-14-17	1	In Service	1.38	0.71	0	0.67	51.51		48.49	6999
compute-14-18	0	In Service	1.61	0.82	0	0.79	51.21		48.79	8379
compute-14-19	2	In Service	1.61	0.8	0	0.81	49.97		50.03	8182



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.ultraflight.org/fgc-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	162646	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_FINAL/Caltech93/LoadTest07_FINAL_53_Ja2B1u9YpQEIOH_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	172521	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_CE_00CgZ3q43nD7UVX_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	172760	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_AD_mknvV3UrZ9VW61R0_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	130957	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_4D_S5dLzI9mDmcpv4_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	479726	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_59_XgDRxKvG3Kp9Q_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	476821	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_57_9yYosK14ivulci3_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	726762	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_FD_BwUlaU3GmElNny4_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	71745	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech634/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	502594	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech633/LoadTest07_UCSD_F2_W64wVwHpoE5j6ZA_633
23654	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	274477	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_C3_KjWhaNoKhZOp2oXTe_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	391431	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_BD_Sjllj6itsyTyOc_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	314437	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_C7_xKqgI6uqjZArcc_635
23953	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	342616	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech633/LoadTest07_UCSD_4D_yuk8Uoc6DBHbtA_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	1580	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/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	657857	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_57_KN145xkklRAlkKW5_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	359697	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech634/LoadTest07_UCSD_CD_NYXVYkqOGjW08U_634
24637	cmsprod	Andrea Sciaba	cihep200.ultraflight.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	cihep252.ultraflight.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/1246400880-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	118804	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech633/LoadTest07_UCSD_6D_kwPFKcaJJEI004_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	601685	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech633/LoadTest07_UCSD_40_hRWRTRRZKOVsjjSx_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.ultraflight.org/fgc-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.ultraflight.org/fgc-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	663663	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_FINAL/Caltech93/LoadTest07_FINAL_D1_KkLlaQ0u0qIHfG_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.ultraflight.org/fgc-util/testfile-is-notoken-20090701-00018.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	543901	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_59_76gMyxWUHRHn6l_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	461727	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_A8_35oXp8GkGsW0R8_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	590681	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_0D_zimru4Mh5PawcP_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	145876	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech634/LoadTest07_UCSD_4D_aAlpKlPaOXJDrfaD_634
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	812948	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech633/LoadTest07_UCSD_0C_VCav5Z7ralgYFmYJ_633
10179	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:11:37 PM PDT 2009	Tue Jun 30 03:11:27 PM PDT 2009	2684354560	51.2MB/s	550829	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech634/LoadTest07_UCSD_4D_gWkHdIoVa3sW8D_635
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	685896	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech634/LoadTest07_UCSD_5B_80lwm9enuplanXIE_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	129601	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_1D_jnXmHmlYdLveL3_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	653853	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_DA_xbKPA07vrsFIdjr_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	320858	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_FA_pydHbbyVzcdPcZ_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	493927	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_34_06zOrvFDHhJM3E_635
10888	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	51.2MB/s	224488	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_07_q255UtaRQrBPF87_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	308855	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_0C_FrJhyDK4d3WY1Cu_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	369916	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech635/LoadTest07_UCSD_8D_FrJhyDK4d3WY1Cu_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	677733	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech633/LoadTest07_UCSD_40_7iGmV12TSNbv4S_633
11640	uscms0377	Michael Thomas	cihep252.ultraflight.org	Tue Jun 30 03:28:03 PM PDT 2009	Tue Jun 30 03:28:03 PM PDT 2009	128		1/1	in	/rs/v/1246400880-storage-probe-test-file-remote.31742
11761	cmsprod	Andrea Sciaba	t2-headnode.ultraflight.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	336952	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech634/LoadTest07_UCSD_FC_dBuYsJhL7KdGfDra_634

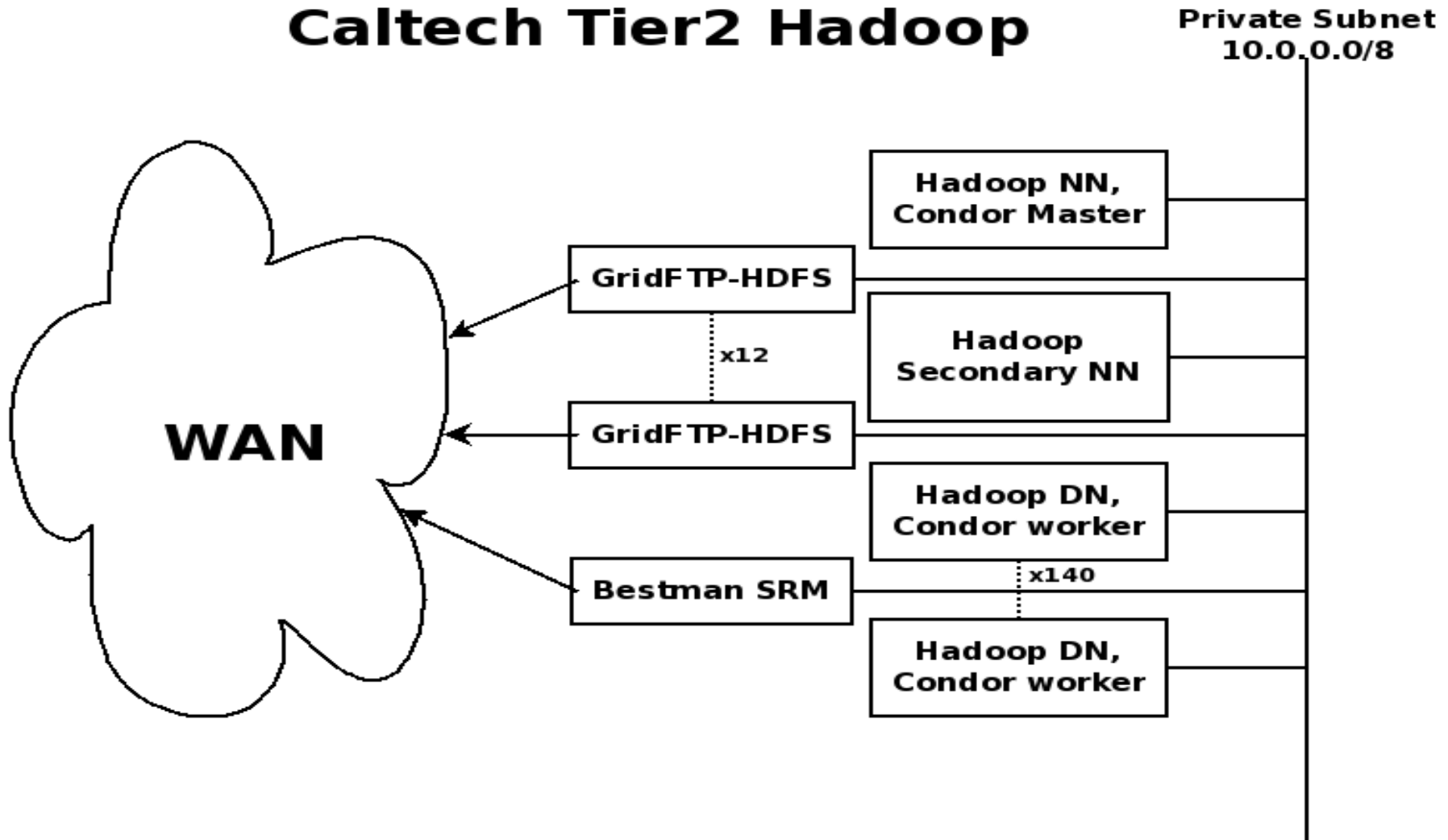
[11882] Tue Jun 30 15:29:44 2009 :: Closed connection from cmsfts1.fnal.gov:49535



Caltech Setup



Caltech Tier2 Hadoop





Caltech Setup



- **Namenode runs on same system as Condor negotiator/collector**
 - ★ 8 cores, 16GB RAM
 - ★ System is very over-provisioned. Load never exceeds 1.0, JVM uses ~2GB out of 4GB
 - ★ Plenty of room for scaling to more blocks
- **Secondary NN runs on a mostly dedicated server**
 - ★ Used to OOM when run on a worker node
- **140 data nodes, 560TB available space**
 - ★ Includes 2 Sun Thumpers running Solaris
 - ★ Currently 470TB used
 - ★ Most datanodes are also condor batch workers
- **Single Bestman(-gateway) SRM server using FUSE for file ops**
- **12 gridftp-hdfs servers**
 - ★ 4 with 2 x 10GbE, 8 with 2 x 1GbE



Deployment history

T2_US_Nebraska first started investigating Hadoop in late 2008. They performed a lot of R&D to get Hadoop to work in the CMS context

- **Two SEs in SAM**
- **Gridftp-hdfs DSI module**
- **Use of Bestman SRM**
- **Many internal Hadoop bug fixes and improvements**

Presented this work to the USCMS T2 community in February 2009



Caltech Deployment



Started using Hadoop in Feb. 2009 on a 4-node testbed

Created RPMS to greatly simplify the deployment across an entire cluster

Deployed Hadoop on new RHEL5 cluster of 64 nodes

Basic functionality worked out of the box, but performance was poor.

Attended a USCMS Tier2 hadoop workshop at UCSD in early March



Tier2 Hadoop Workshop



- **Held at UCSD in early March 2009**
- **Intended to help get interested USCMS Tier2 sites jump-start their hadoop installations**
- **Results:**
 - ★ **Caltech, UCSD expanded their hadoop installations**
 - ★ **Wisconsin delayed deployment due to facility problems**
 - ★ **Bestman, GridFTP servers deployed**
 - ★ **Initial SRM stress tests performed**
 - ★ **UCSD <-> Caltech load tests started**
 - ★ **Hadoop SEs added to SAM**
 - ★ **Improved RPM packaging**
 - ★ **Better online documentation for CMS**

<https://twiki.grid.iu.edu/bin/view/Storage/HdfsWorkshop>



Caltech Deployment



Migrated OSG RSV tests to Hadoop in mid-march

Migrated data from dCache to Hadoop over the course of 6 months (Apr. - Oct.). Operating with two SEs during this time.

Added read-only http interface in mid-May

CMS review of Hadoop on Sep. 16. Formal approval given on Oct. 21.

Decommissioned dCache on Oct. 22.



Current Successes

- **SAM tests passing**
- **All PhEDEx load tests passing**
- **RPMs provide easy installs, reinstalls**
 - ★ **hadoop, gridftp, bestman, xrootd (under development)**
- **Bestman + GridFTP-HDFS have been stable**
- **Good Nagios coverage**
- **Great inter-node transfer rates (4GB/s aggregate)**
- **Adequate WAN transfer rates (7Gbps peaks)**
- **Extensive Install/config documentation**
 - ★ **<https://twiki.grid.iu.edu/bin/view/Storage/Hadoop>**
- **Primary storage system at 3 USCMS T2 sites**



Why Hadoop?

Caltech: “Lower operational overhead due to fewer moving parts. The simple architecture is relatively easy to understand”

UCSD: “Scalable SRM and replication that just works, and the FUSE interface is simple for admins and users to work with”

UNL: “Manageability and reliability”



Not without problems...

- **OSG RSV tests required patch to remove ":" from filenames. This is not a valid character in hadoop filenames. (resolved in OSG 1.2)**
- **Bestman dropped VOMS FQAN for non-delegated proxies, caused improper user mappings and filesystem permission failures for SAM, PhEDEx (resolved)**
- **Bestman error messages incompatible with lcg-utils (resolved)**
- **TFC not so "t" with multiple SEs**
- **Datanode/Namenode version mismatches (improved)**
- **Initial performance was poor (400MB/s aggregate) due to cluster switch configuration (resolved)**



Not without more problems...



FUSE was not so stable

- * Boundary condition error for files with a specific size crashed fuse (resolved)
- * df sometimes not showing fuse mount space (resolved)
- * Lazy java garbage collection resulted in hitting ulimit for open files (resolved with larger ulimit)
- * scp, tar, rsync didn't work (resolved)

GridFTP servers crashing

- * Excessive memory usage for large files (resolved)
- * temp file not configurable (resolved)
- * Unstable NIC driver (resolved)

Running two CEs and SEs requires extra care so that both CEs can access both SEs

- * Some private network configuration issues (resolved)
- * Lots of TFC wrangling (obsolete)



Daily Operations

Wait for Nagios to alert for node-specific problems

Balance datanode disk usage daily with cron:

* `hadoop balance -threshold 5`

Set replicas based on file path daily with cron:

* `Default replication == 2`

* `hadoop fs -setrep -R 3 /store/user`

gridftpspy running on desktop to watch for errors

Decommission node for maintenance

* `vi /etc/hadoop/hosts_exclude && hadoop dfsadmin -refreshNodes`

Just for fun: `hadoop fsck /`



Next Steps



xrootd interface

Improve application IO profiles

Collaborations with Cloudera

Transition packaging and testing to VDT/OSG

Hadoop 0.20.1, 0.21.0

Evangelize



Overall Impressions



Management of HDFS is simple relative to other SE options

Performance has been more than adequate

Scaled from 4 to 64 to 140 nodes with minimal problems

~50% of our initial problems were related to Hadoop, the other 50% were Bestman, TFC, PhEDEx agent, or caused by running multiple SEs

Hadoop now serves 100% of the CMS data at Caltech, UNL, and UCSD