Hadoop Overview and Installation
August 7, 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

Used throughout Yahoo. Yahoo is a major contributor 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 multliple 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
- Fsck 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

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, srmmkdir
- Treats hdfs as local posix filesystem

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 ~1GB out of 2GB
  - Plenty of room for scaling to more blocks
- Secondary NN runs on same system as condor batch worker
  - OOM twice (fixed)
- 84 data nodes, 277TB available space
  - Includes 2 Sun Thumpers running Solaris
  - Currently 207TB used
  - Most datanodes are also condor batch workers
- Single Bestman(-gateway) SRM server using FUSE for file ops
- Four gridftp-hdfs servers with 2 x 10GbE

Hadoop SE Tutorial

Prerequisites

- 1 server needed to run the Hadoop namenode, bestman SRM, and gridftp
  - globus account exists
  - host/service certificate exists in /etc/grid-security/globuscert.pem, globuskey.pem

- 1 server needed to run the Hadoop datanode
- gums is being used for user mappings
- fuse + fuse kernel module is installed on both servers
- Sun java 1.6 is installed from rpm on both servers
- No firewall is blocking traffic between the two servers

- Root access on both servers
- Read https://twiki.grid.iu.edu/bin/view/Storage/Hadoop

Assumptions

- **You will run bestman as the 'globus' user**
- **You use gums for user mappings**
- **Only one hadoop service runs per server**
- **Certificates will be installed in `/etc/grid-security` and managed via rpm**

Filesystem layout

/etc/sysconfig/* – init.d/cron configuration files

/etc/hadoop/*
/etc/gridftp-hdfs/* – hadoop/gridftp configuration files

/var/log/hadoop/*
/var/log/bestman/*
/var/log/gridftp*.log – Log files

/usr/share/java/hadoop/* – Hadoop jar files

/usr/bin/* - user/system binaries
...but...
/opt/bestman/* – All bestman files

Set up the hadoop repository

On both the namenode and datanode servers:

RHEL5 (32 and 64 bit):

```
rpm -ivh http://newman.ultralight.org/repos/hadoop/5/x86_64/caltech-hadoop-5-1.noarch.rpm
```

RHEL4 (32 and 64 bit):

```
rpm -ivh http://newman.ultralight.org/repos/hadoop/4/x86_64/caltech-hadoop-4-1.noarch.rpm
```

Namenode Installation

```
# yum install hadoop
```

Edit `/etc/sysconfig/hadoop`

```
# service hadoop-firstboot start
# service hadoop start
```

Browse to http://namenode:50070

/etc/sysconfig/hadoop

```
HADOOP_CONF_DIR=/etc/hadoop
HADOOP_NAMENODE=cithep196
HADOOP_NAMEPORT=9000
HADOOP_PRIMARY_HTTP_ADDRESS=${HADOOP_NAMENODE}:50070
HADOOP_REPLICATION_DEFAULT=2
HADOOP_REPLICATION_MIN=1
HADOOP_REPLICATION_MAX=4
HADOOP_USER=hadoop
HADOOP_DATADIR=/wntmp/hadoop
HADOOP_DATA=${HADOOP_DATADIR}/data
HADOOP_LOG=/var/log/hadoop
HADOOP_SCRATCH=${HADOOP_DATADIR}/scratch
HADOOP_GANGLIA_ADDRESS=
HADOOP_GANGLIA_PORT=8649
HADOOP_GANGLIA_INTERVAL=10
HADOOP_SECONDARY_NAMENODE=
HADOOP_SECONDARY_HTTP_ADDRESS=${HADOOP_SECONDARY_NAMENODE}:50090
HADOOP_CHECKPOINT_DIRS=${HADOOP_SCRATCH}/dfs/namesecondary
HADOOP_CHECKPOINT_PERIOD=3600
HADOOP_DATANODE_BLOCKSIZE=134217728
HADOOP_NAMENODE_HEAP=8192m
HADOOP_MIN_DATANODE_SIZE=300
HADOOP_RACKAWARE_SCRIPT=
HADOOP_SYSLOG_HOST=
```

Datanode Installation

```
# yum install hadoop
```

Edit `/etc/sysconfig/hadoop`

```
# service hadoop-firstboot start
# service hadoop start
```

Browse to http://cithep196:50070/

```
# hadoop fs -copyFromLocal /etc/hosts hdfs://cithep196:9000/test.file
# hadoop fs -ls /
```

Fuse installation

Can be installed on both NN and DN; must be installed on bestman servers

```
# yum install hadoop-fuse
```

If running selinux on RHEL5:

```
# yum install hadoop-fuse-selinux
```

Add to `/etc/fstab`:

```
hdfs# /mnt/hadoop fuse server=namenode,port=9000,rdbuffer=131072,allow_other 0 0
```

```
# mkdir /mnt/hadoop
# mount /mnt/hadoop
# ls /mnt/hadoop
```

gridftp installation

```
# yum install gridftp-hdfs osg-ca-certs fetch-crl
```

Edit `/etc/grid-security/prima-authz.conf` with your gums server url

Edit `/etc/gridftp-hdfs/gridftp-hdfs-local.conf` with your temp directory

Set proxy in `/etc/sysconfig/fetch-crl`, if necessary:

```
    http_proxy=http://your.proxy.com:3128
    export http_proxy
```

If xinetd is not already running, start it:

```
# service xinetd start
```

Service listens on port 2811

Bestman installation

```
# yum install bestman
```

Edit `/opt/bestman/conf/bestman.rc` to set:

- `GUMS_HOST`
- `supportedProtocolList`
- `localPathListAllowed`

Append to `/etc/sudoers` for file operations:

```
Cmnd_Alias SRM_CMD = /bin/rm, /bin/mkdir, /bin/rmdir, /bin/mv, /bin/ls
Runas_Alias SRM_USR = ALL, !root
globus ALL=(SRM_USR) NOPASSWD:SRM_CMD
```

Set proxy in `/etc/sysconfig/fetch-crl`, if necessary:

```
http_proxy=http://your.proxy.com:3128
export http_proxy
```

```
# service bestman start
```

Service listens on port 8443

Misc. Tools installation

Hadoop space usage summary

```
# yum install hadoop-chronicle
```

gridftp-hdfs server log viewer (requires epel repository)

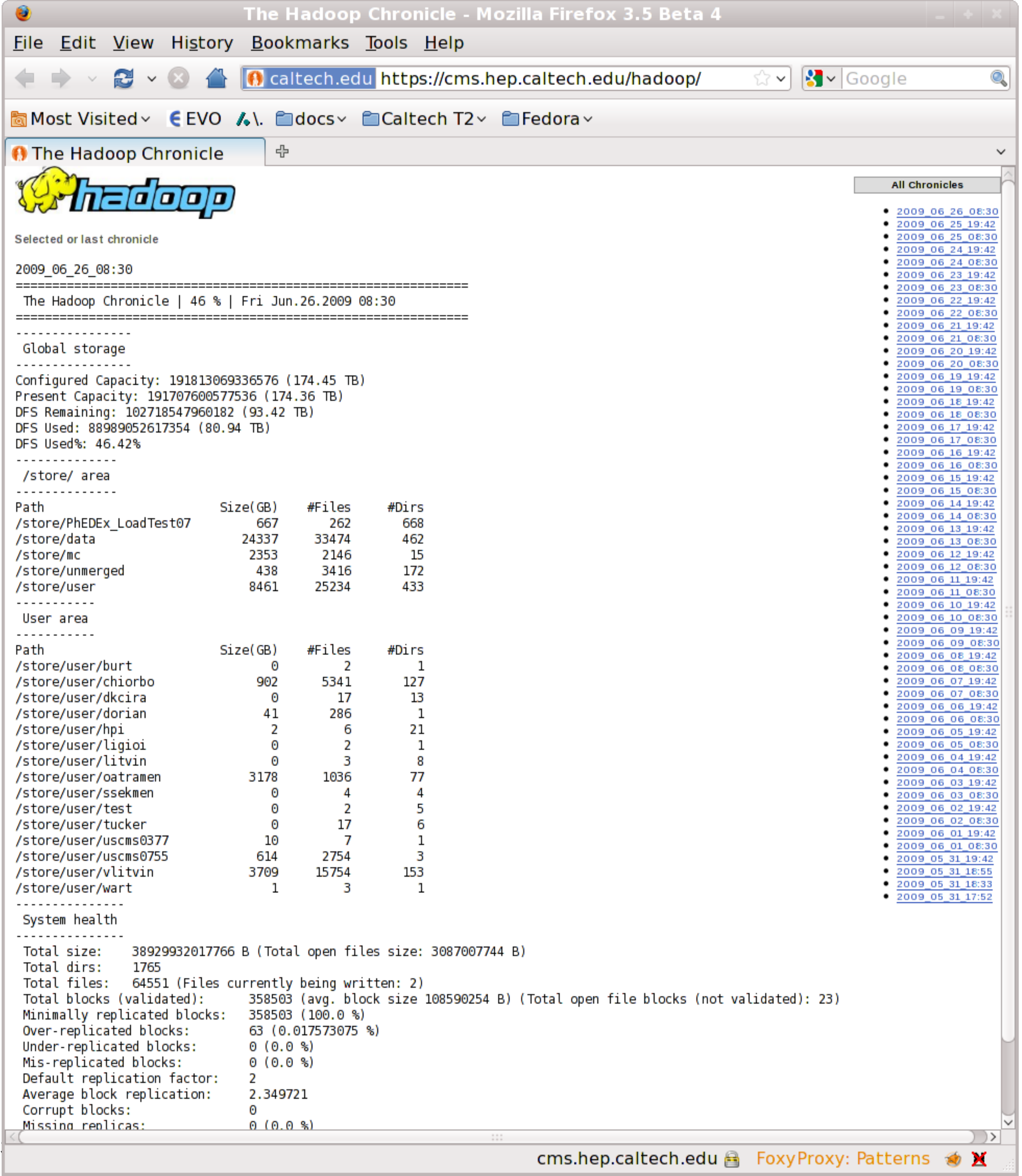
```
# yum install gridftpspy
```

JMX nagios plugin

```
# yum install nagios-plugins-jmx
```

Hadoop nagios plugins (not yet available)

```
# yum install nagios-plugins-hadoop
```

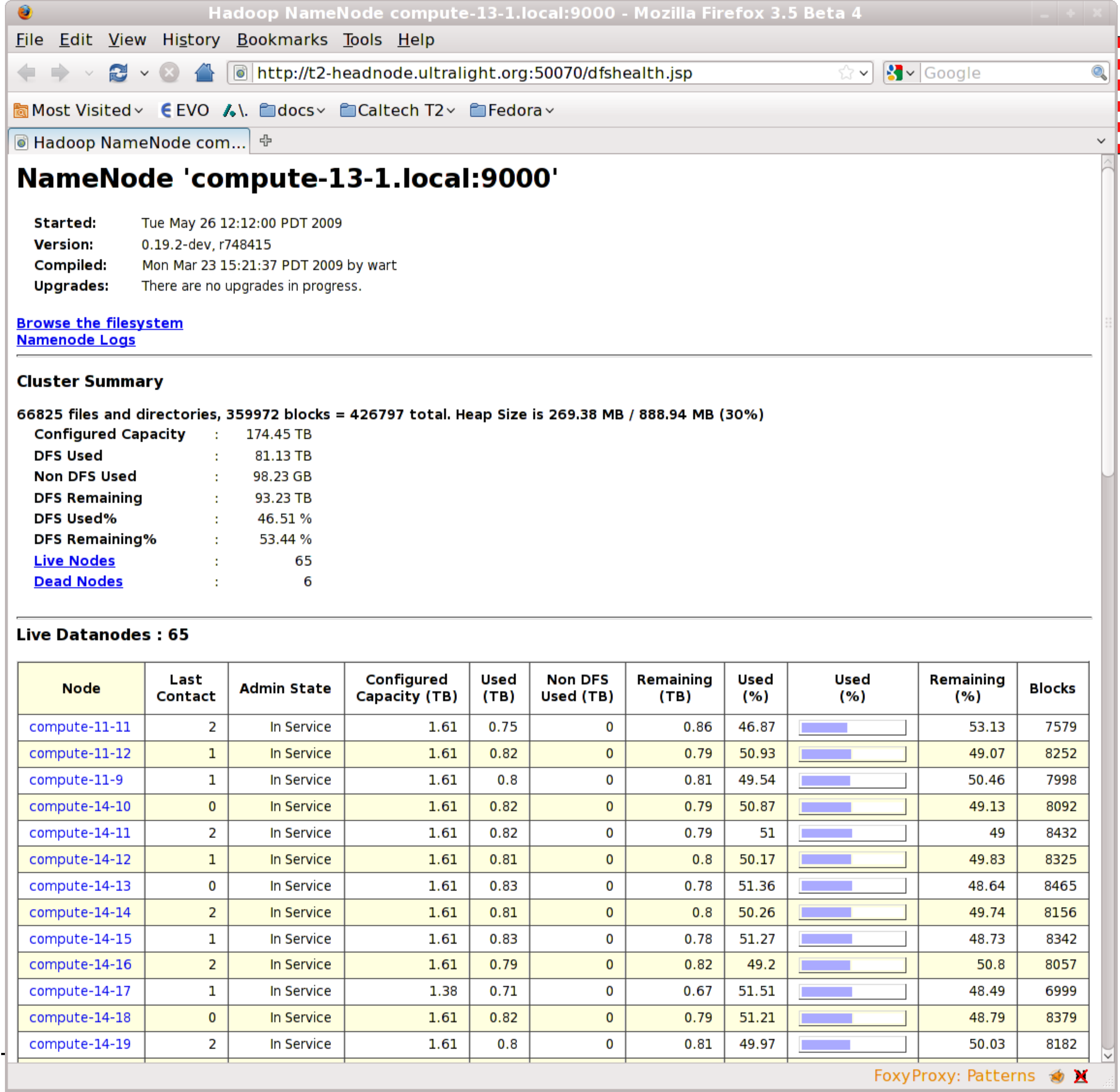


Selected or last chronicle

```
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
Google
Most Visited
EVO
docs
Caltech T2
Fedora
Hadoop NameNode com...
FoxyProxy: Patterns

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

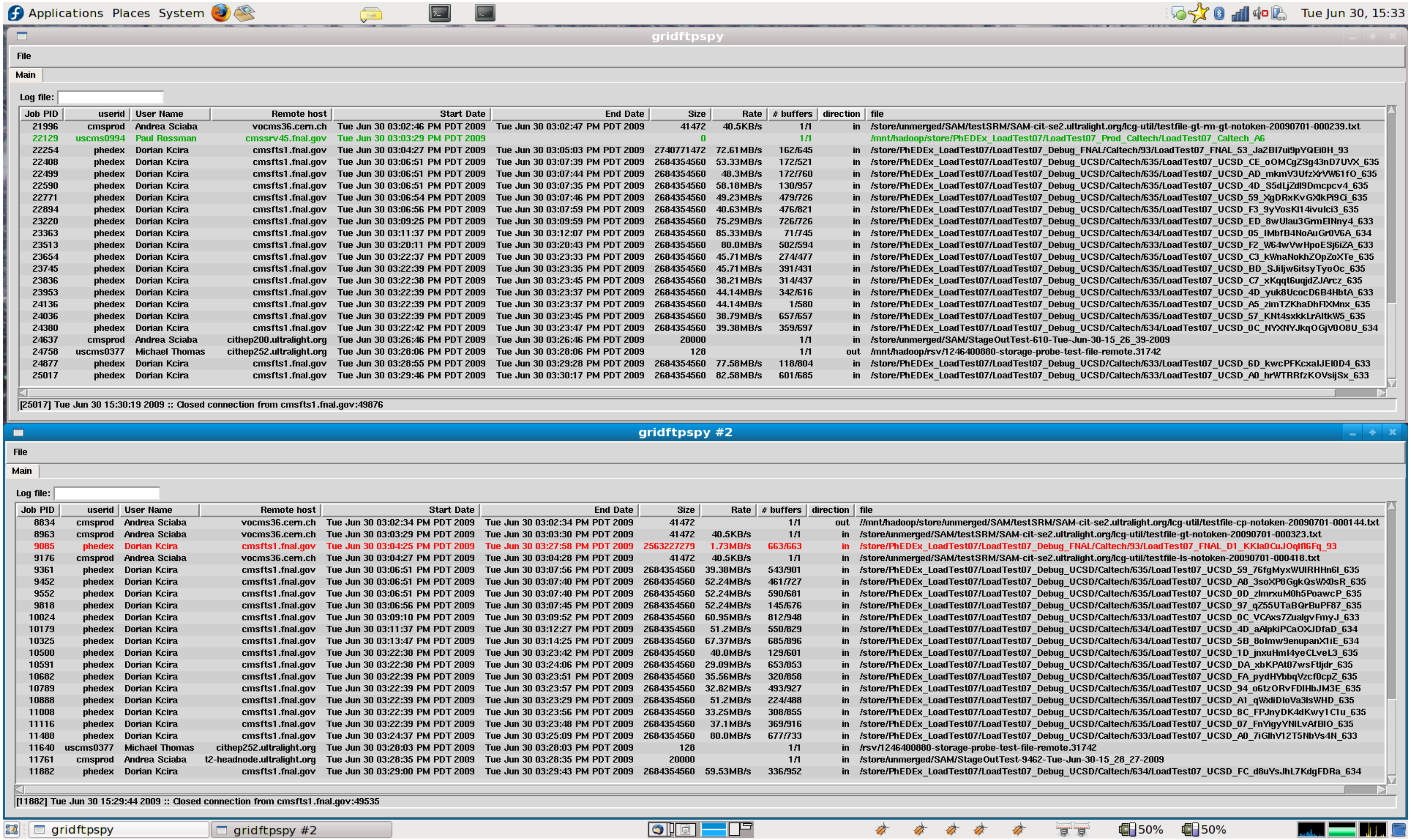
66825 files and directories, 359972 blocks = 426797 total. Heap Size is 269.38 MB / 888.94 MB (30%)

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

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	41472	40.5KB/s	1/1	in	/store/unmerged/SAM/testSRM/SAM-cit-se2.ultralight.org/lcg-util/testfile-gt-rm-gt-notoken-20090701-000239.txt
22129	uscms0994	Paul Rossman	cmssrv45.fnal.gov	Tue Jun 30 03:03:29 PM PDT 2009		0		1/1		/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_FNAL/Caltech/93/LoadTest07_FNAL_53_Ja2BI7ui9pYQEi0H_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_oOMCgZSg43nD7UVX_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	172/760	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_AD_mkmV3UfzXrVW61fO_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_4D_S5dLjZdl9Dmcpcv4_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_XgDRxKvGXlkPi9Ci_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_9yYosKl14ivulci3_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	726/726	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_ED_8wUlau3GrmEINny4_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_IMbfB4NoAuGr0V6A_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/633/LoadTest07_UCSD_F2_W64wVwHpoESj6iZA_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	274/477	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_C3_kWnaNokhZOpZoXTe_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_SJiljw6itsyTyoOc_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_xKqqt6uqjdZJArcz_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	342/616	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_4D_yuk8UcocD6B4HbtA_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_zimTZKhaDhFlXMnx_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	657/657	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_57_KNt4sxkkLrAltkW5_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_0C_NYXNYJkqOGjV0O8U_634
24637	cmsprod	Andrea Sciaba	cithep200.ultralight.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.ultralight.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	118/804	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/633/LoadTest07_UCSD_6D_kwcPFKcxalJEI0D4_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_A0_hrWTRRfzKOVsijSx_633

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

gridftpspy #2

Log file:

Job PID	userid	User Name	Remote host	Start Date	End Date	Size	Rate	# buffers	direction	file
8834	cmsprod	Andrea Sciaba	vocms36.cern.ch	Tue Jun 30 03:02:34 PM PDT 2009	Tue Jun 30 03:02:34 PM PDT 2009	41472		1/1	out	//mnt/hadoop/store/unmerged/SAM/testSRM/SAM-cit-se2.ultralight.org/lcg-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	41472	40.5KB/s	1/1	in	/store/unmerged/SAM/testSRM/SAM-cit-se2.ultralight.org/lcg-util/testfile-gt-notoken-20090701-000323.txt
9085	phedex	Dorian Kcira	cmsfts1.fnal.gov	Tue Jun 30 03:04:25 PM PDT 2009	Tue Jun 30 03:27:58 PM PDT 2009	2563227279	1.73MB/s	663/663	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_FNAL/Caltech/93/LoadTest07_FNAL_D1_KKla0CuJOqIfl6Fq_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	41472	40.5KB/s	1/1	in	/store/unmerged/SAM/testSRM/SAM-cit-se2.ultralight.org/lcg-util/testfile-ls-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_76fgMyxWUIRHHn6l_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_3soXP8GgkQsWX0sR_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_zlmrxuM0h5PoawcP_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	145/676	in	/store/PhEDEx_LoadTest07/LoadTest07_Debug_UCSD/Caltech/635/LoadTest07_UCSD_97_qZ55UTaBQrBuPF87_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_VCAxs7ZualgvFmyJ_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_aAlpkiPCaOXJDfaD_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_8olmw9enupanX1iE_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_jnxuHml4yeCLveL3_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_xbKPAt07wsFtljdr_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_pydHYbbqVzcf0cpZ_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_o6tzORvFDIHbJM3E_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_qWxliDIoVa3lsWHD_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_FPJnyDK4dKwy1C1u_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_FnYigyYNILvAfBIO_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_A0_7iGlhV12T5NbVs4N_633
11640	uscms0377	Michael Thomas	cithep252.ultralight.org	Tue Jun 30 03:28:03 PM PDT 2009	Tue Jun 30 03:28:03 PM PDT 2009	128		1/1	in	/rsv/1246400880-storage-probe-test-file-remote.31742
11761	cmsprod	Andrea Sciaba	t2-headnode.ultralight.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_d8uYsJhL7KdgFDRa_634

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

Tunings

- **Increase the HDFS block size**
  - ✷ Reduces memory footprint of namenode
- **Change replication**
  - ✷ Reduce space usage, or increase block availability
- **Increase ulimit for hadoop processes**
  - ✷ lazy garbage collection results in lots of open files
- **Put gridftp-hdfs tmp dir on fast RAID0**
  - ✷ Increases throughput for large files
- **Use proxy with fetch-crl**
  - ✷ /etc/sysconfig/fetch-crl

Tunings (cont.)

- **Use a Secondary Namenode (SNN)**
  - Offloads expensive journal merge from NN

- **Run the balancer in a cron job**
  - hadoop balancer -threshold 5

Rack Awareness

Hadoop can spread replicated blocks to different racks for added safety

Set in `/etc/hadoop/hadoop-site.xml:`

```
<property>
  <name>topology.script.file.name</name>
  <value>/usr/bin/rocks-hostname-to-rack.sh</value>
</property>
```

Points to a script that maps IP addresses to Rack ids

```
#!/bin/sh

# The default rule maps systems based on the rack id in the hostname.
# For example, compute-14-1 is in Rack 14.  Only exceptions to this rule
# need to be explicitly listed.

for ip in $@ ; do
    hostname=`nslookup $ip | grep "name =" | awk '{print $4}' | sed -e
's/\.local\.$//' `
    case $hostname in
        compute-0-*)  rack="/Rack10" ;;
        *)
            rack=`echo $hostname | sed -e 's/^[a-z]*-\([0-9]*\)-[0-
9]*.*/\/Rack\1/'`
            ;;
        esac
        echo $rack
done
```

syslog

- Hadoop uses log4j for logging. Set `SyslogHost` in `/etc/hadoop/log4j.properties:`

```
log4j.appender.SYSLOG.SyslogHost=10.3.1.1
```

- gridftp-hdfs can use syslog for logging. Set `GRIDFTP_SYSLOG` in `/etc/gridftp-hdfs/gridftp-hdfs-local.conf:`

```
export GRIDFTP_SYSLOG=10.3.1.1
```

Daily Operations

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

Reboot gridftp nodes when they crash

Decommission node for maintenance

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

Just for fun: `hadoop fsck /`