HDFS and Remote Xrootd in a CMS T2

Brian Bockelman

SuperB Workshop, 6 July 2011

HDFS Basics*

*for us, anyway

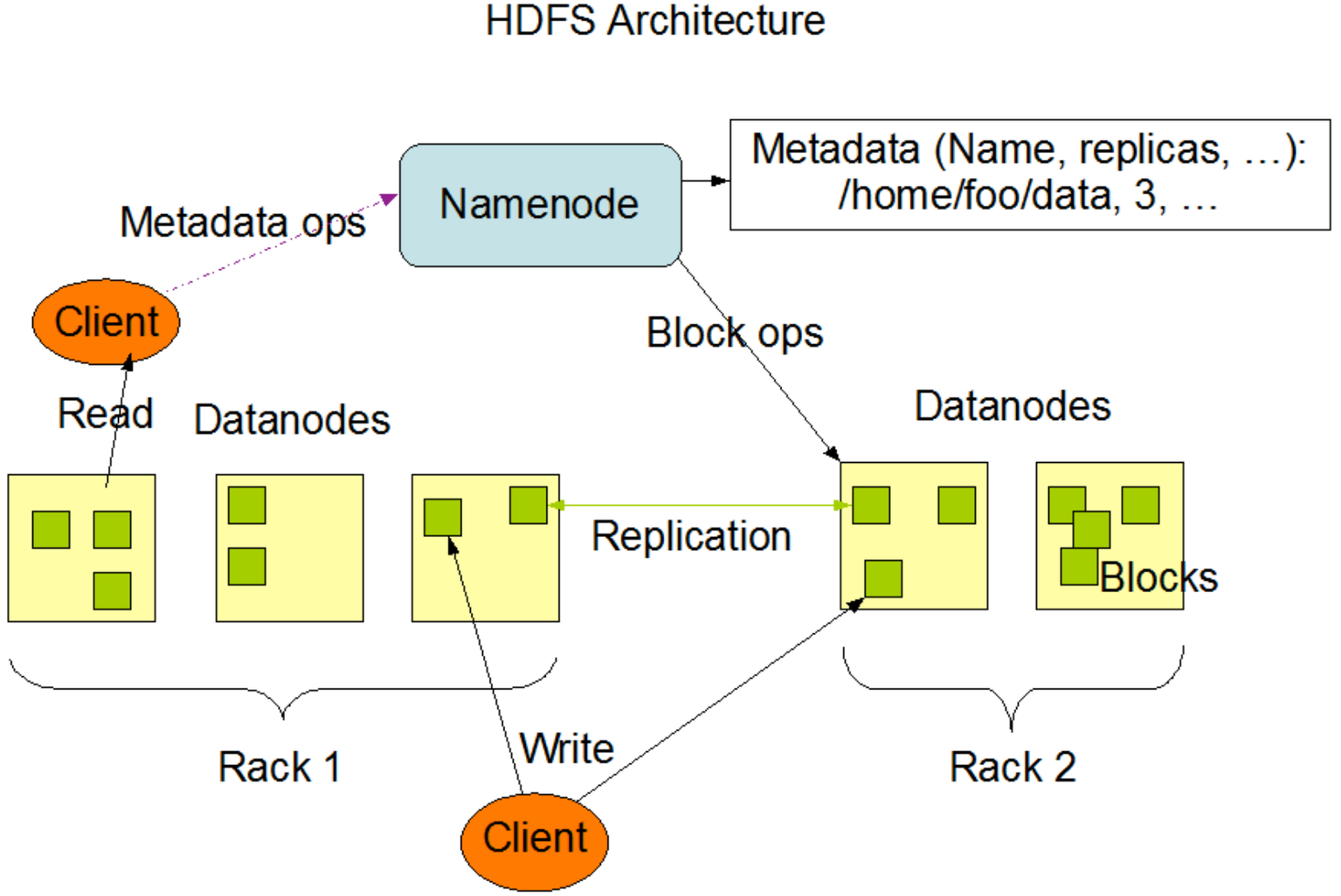
- Hadoop has many interesting part. In this talk, we're just interested in the Hadoop Distributed File System (HDFS).
- HDFS is a highly scalable distributed file system coming from the Apache Foundation (majority of developers come from Facebook and Yahoo) with an emphasis on reliability.

Users

- Big external: Yahoo (25,000 nodes; largest cluster is 4,000 nodes @ 80PB), Facebook (largest cluster, 13PB)
- LHC T2: UCSD, Nebraska, Caltech, Wisconsin, MIT (soon), Estonia. About 8PB in the US.
- T3: UCD, UColorado, UCR, others.

Architecture

- HDFS architecture is no surprise to anyone familiar with a grid SE.



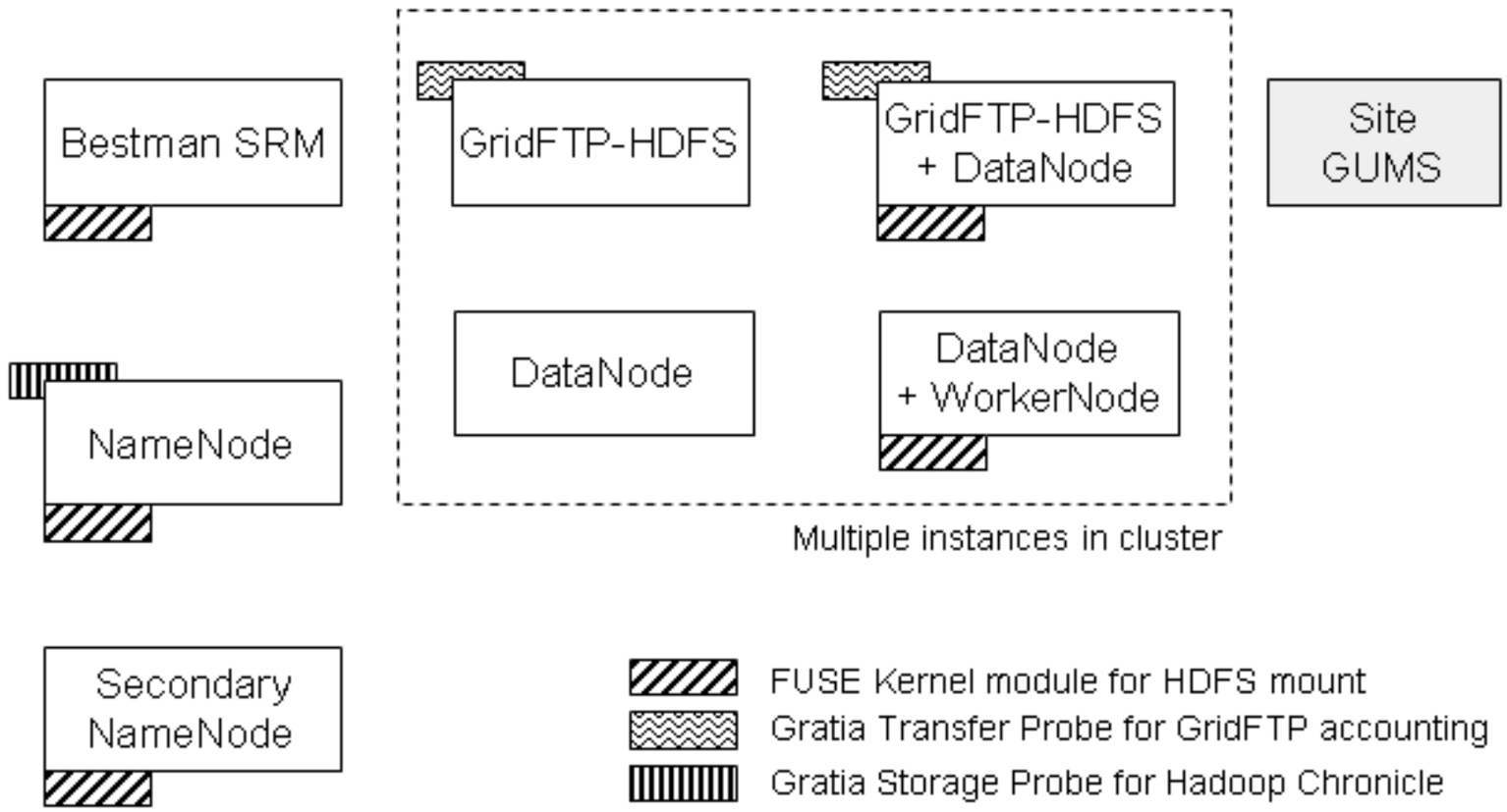
High points

- HDFS is designed to work with hard drives in worker nodes (we buy Dell r710s; 2U worker node with 6 x 2TB disks).
- Reliability is provided through replicating chunks on many datanodes.
- SRM/GridFTP provided by BestMan and Globus GridFTP, respectively.
- Completely YUM/RPM packaging is available; integrates in Linux like expected.

Focus

- The focus is on reliability:
  - Hard drive fails? No problem; data is re-replicated elsewhere.
  - Node disappears? No problem; the client re-routes its requests to the other node holding the data.
  - Rack loses power? No problem; HDFS can keep replicas on different racks.
- We are comforted by the fact that the private-industry investment in Hadoop is $5-10M / year.
- LHC will never be close to the largest users of this software.

Technology Details



SRM Hadoop storage system: Example topology at an OSG Site

Protocols

- We pride ourselves on our ability to *integrate* with other, modular protocol servers. Examples:
  - GridFTP
  - HTTP
  - SRM
- For each of these, support and development is done outside of Hadoop.
- Primary access method remains POSIX. *NOT* full POSIX semantics, but good enough for HEP.

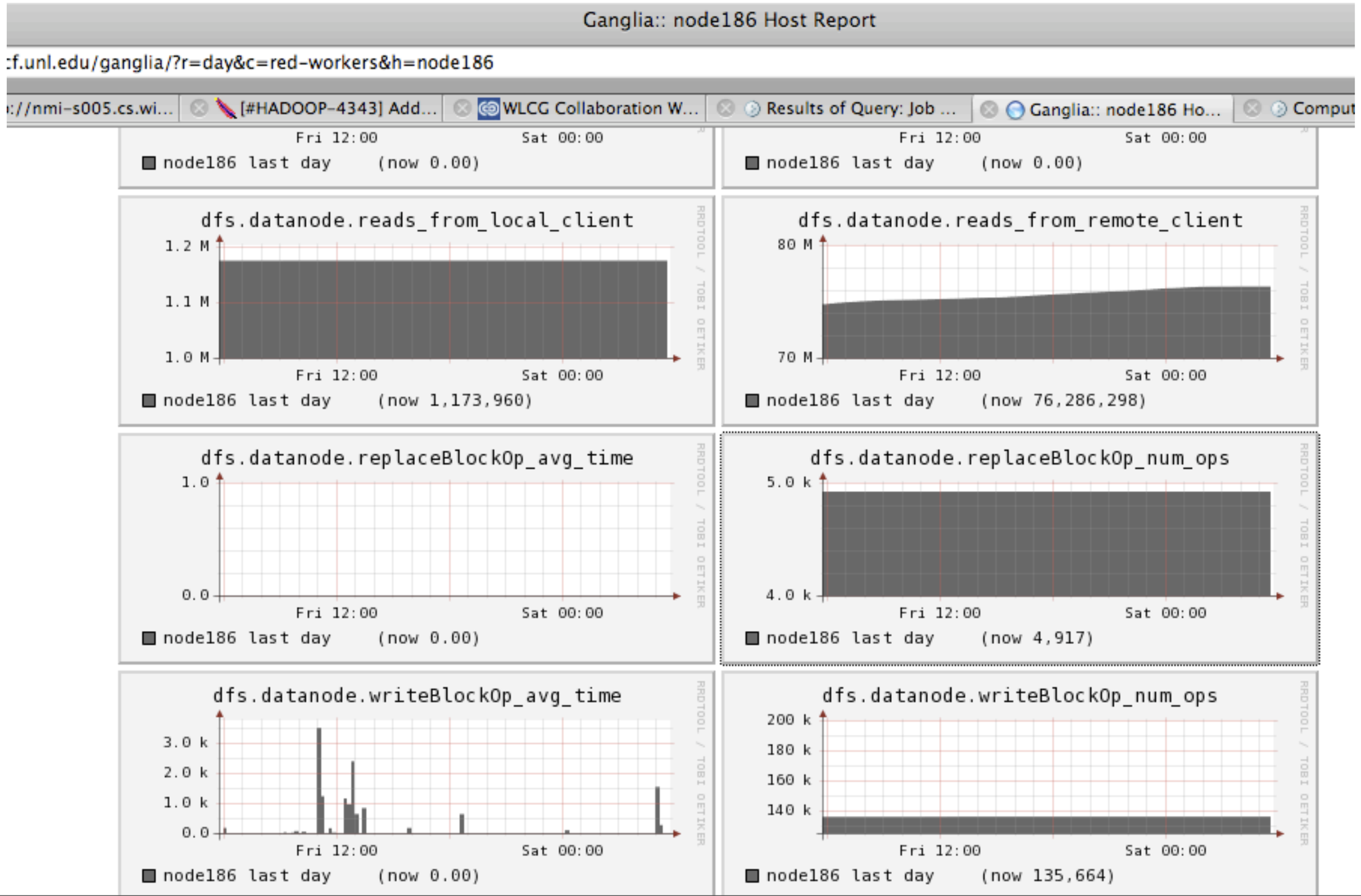
Why HDFS?

- In the next few slides, we'll discuss why we think HDFS means:
  - Less management.
  - More reliability.
  - Better scalability.
  - Usability

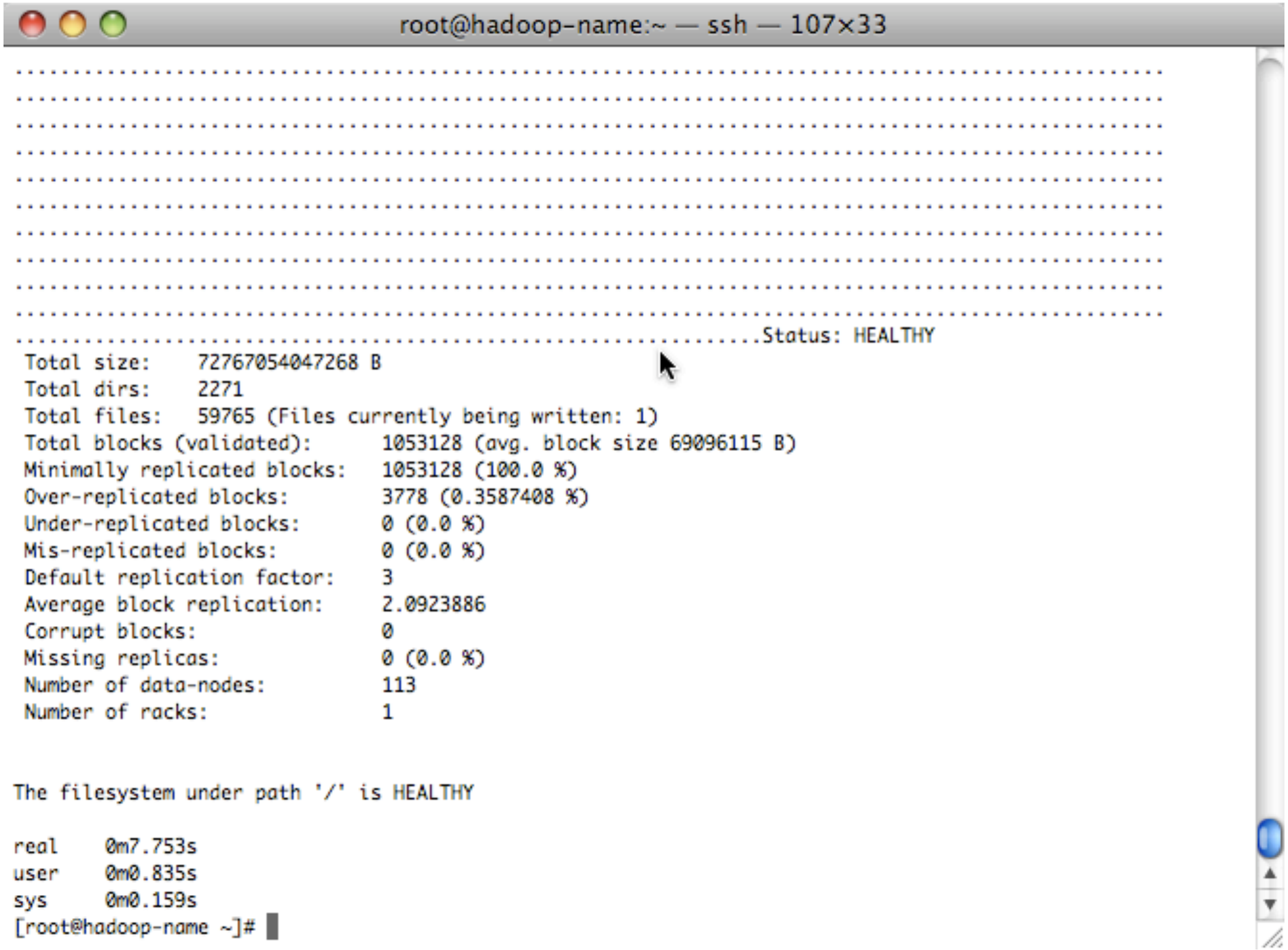
Management

- The following tasks are trivial:
  - Integration of statistics with **Ganglia**.
  - **Decommissioning** hardware.
  - **Recovery** from hardware failure.
  - **Fsck**!
    - Checks the current knowledge of the filesystem and counts how many block replicas there are per file, and highlights any which are under-replicated.
- **RPM** install (including Grid components).
- Many of our “well-known” problems are not possible.
  - **Don’t need a separate admin toolkit**! (one gremlin)
- Setting **quotas (***per directory***)**.
- **Backups of namespace**.
- **Balancer** is included.

Ganglia Graphs

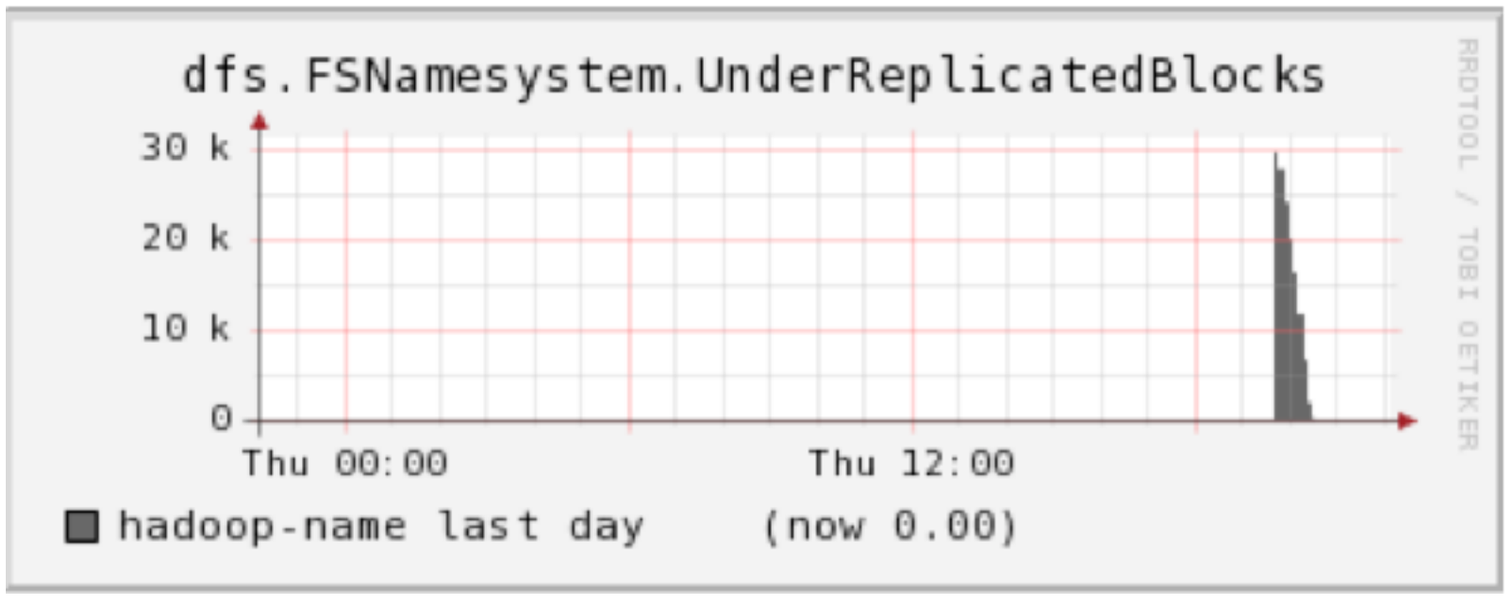


FSCK example



Reliability

- Clients will automatically connect to a different datanode if one fails during a read.
- Blocks will automatically re-replicate -- and quickly! Often, we will recover from a loss in an hour.
  - Namenode controls this. We have it set to re-replicate if a node hasn't checked in for 10 minutes.
- All data is checksummed on read.



Reliability

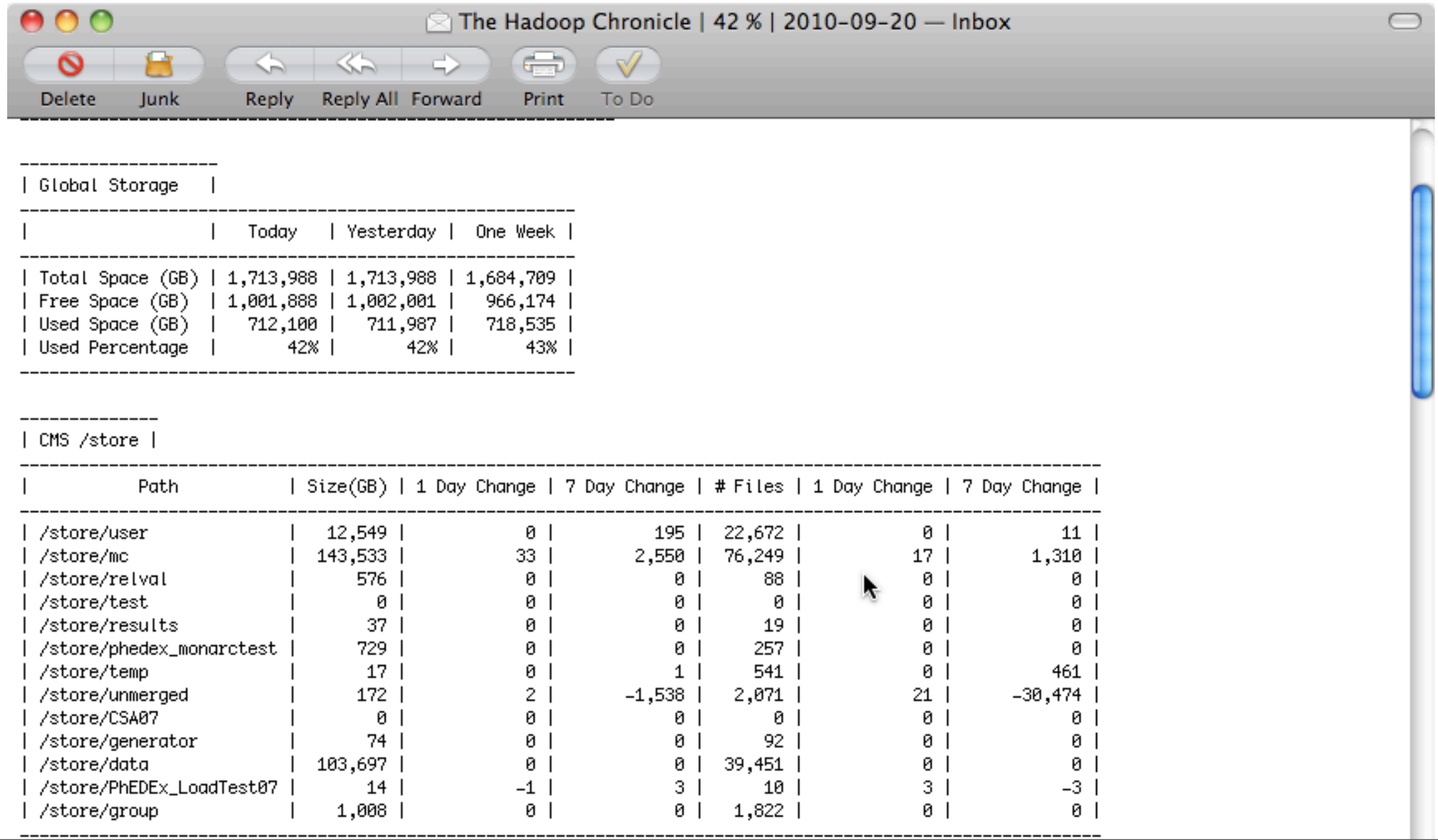
- Data node failures (on read or write) do not result in client failures!
- No hotspots: Due to block decomposition, access to a single file might be spread over 20 servers. Plenty of bandwidth and spindles!

Performance Stats

- We've clocked:
  - The filesystem at 80Gbps.
  - 23 Gbps for 300 CMSSW processes analyzing a *single file* @ 2 replicas (we picked a fake workflow to pump up the per-job rate).
  - SRM endpoints at ~200Hz (these SRMs are stateless; load-balancing is trivial). Done using GUMS auth.
  - fsck takes ~1 minute.
  - Decommissioning a pool <1hr.
  - Namenode restart in about 60s.
  - WAN transfers peak at 9Gbps, sustain 5Gbps.
  - 18,400 metadata ops / sec from the namenode.

The Hadoop Chronicle

Daily email summarizing Hadoop usage



Global Storage			
	Today	Yesterday	One Week
Total Space (GB)	1,713,988	1,713,988	1,684,709
Free Space (GB)	1,001,888	1,002,001	966,174
Used Space (GB)	712,100	711,987	718,535
Used Percentage	42%	42%	43%

CMS /store

Path	Size(GB)	1 Day Change	7 Day Change	# Files	1 Day Change	7 Day Change
/store/user	12,549	0	195	22,672	0	11
/store/mc	143,533	33	2,550	76,249	17	1,310
/store/relval	576	0	0	88	0	0
/store/test	0	0	0	0	0	0
/store/results	37	0	0	19	0	0
/store/phedex_monarctest	729	0	0	257	0	0
/store/temp	17	0	1	541	0	461
/store/unmerged	172	2	-1,538	2,071	21	-30,474
/store/CSA07	0	0	0	0	0	0
/store/generator	74	0	0	92	0	0
/store/data	103,697	0	0	39,451	0	0
/store/PhEDEx_LoadTest07	14	-1	3	10	3	-3
/store/group	1,008	0	0	1,822	0	0

The Hadoop Chronicle

The Hadoop Chronicle | 42 % | 2010-09-20 — Inbox

Delete Junk Reply Reply All Forward Print To Do

| /store/group | 1,888 | 0 | 0 | 1,022 | 0 | 0 |

| CMS /store/user |

Path	Size(GB)	1 Day Change	7 Day Change	Remaining	# Files	1 Day Change	7 Day Change	Remaining
/store/user/hpi	0	0	0	1,099	15	0	0	9,985
/store/user/clundst	0	0	0	NO QUOTA	809	0	0	NO QUOTA
/store/user/npanyam	188	0	0	2,922	84	0	0	9,916
/store/user/gattebury	0	0	0	1,100	1	0	0	9,999
/store/user/belforte	252	0	195	2,596	1,029	0	11	8,971
/store/user/bockjoo	2	0	0	3,295	2	0	0	9,998
/store/user/skhalil	317	0	0	2,665	218	0	0	9,782
/store/user/shruti	44	0	0	3,167	708	0	0	9,292
/store/user/mkirn	0	0	0	1,100	3	0	0	9,997
/store/user/spadhi	13	0	0	1,061	1,154	0	0	8,846
/store/user/creed	0	0	0	1,099	6	0	0	9,994
/store/user/jproulx	13	0	0	3,258	120	0	0	9,880
/store/user/zeise	100	0	0	2,998	393	0	0	9,607
/store/user/malik	0	0	0	1,099	3	0	0	9,997
/store/user/tkelly	0	0	0	3,299	0	0	0	10,000
/store/user/rossman	0	0	0	1,099	5	0	0	9,995
/store/user/bloom	1,933	0	0	NO QUOTA	2,907	0	0	NO QUOTA
/store/user/kaulmer	0	0	0	1,098	109	0	0	9,891
/store/user/ewv	7	0	0	1,081	284	0	0	9,716
/store/user/eluiggi	655	0	0	-211	231	0	0	9,769
/store/user/test	0	0	0	11	179	0	0	821
/store/user/iraklis	1,237	0	0	29,274	1,084	0	0	98,916
/store/user/bbockelm.nocern	1,259	0	0	634	4,796	0	0	5,204
/store/user/schiefer	752	0	0	1,043	3,265	0	0	6,735
/store/user/kellerjd	1,146	0	0	13,054	2,629	0	0	7,371
/store/user/zrwan	0	0	0	3,298	6	0	0	9,994
/store/user/malbouis	4,629	0	0	2,605	2,399	0	0	7,601
/store/user/dnoonan	0	0	0	3,298	3	0	0	9,997
/store/user/drell	1	0	0	1,097	221	0	0	9,779
/store/user/sarkar	0	0	0	NO QUOTA	9	0	0	NO QUOTA

The Hadoop Chronicle

The Hadoop Chronicle | 42 % | 2010-09-20 — Inbox

Delete Junk Reply Reply All Forward Print To Do

```
-------------------------------------------------------------
| Online Pool Count  |   185 |           0 |           5 |
| Offline Pool Count |    15 |           0 |         -14 |
| % Used Avg         |   43% |          -1% |         -4% |
| % Used Std Dev     |    5% |          -1% |         -4% |
-------------------------------------------------------------
No new pools today.
New pools this week: node114, red-d9n2, node079, node156, node120, node121, node181, node125
No new dead pools today.
New missing/dead pools this week: node074, node142, node148

-------------
| FSCK Data |
-------------
/user/uscms01/pnfs/unl.edu/data4/cms/store/mc/Summer09/PhotonJets_Pt40to100-madgraph/GEN-SIM-RECO/MC_31X_V3_7TeV-v2/0000/36C76CCF-4216-DF1
1-8D9F-00D0680BF8C2.root: CORRUPT block blk_5149773138535264325
/user/uscms01/pnfs/unl.edu/data4/cms/store/mc/Summer09/PhotonJets_Pt40to100-madgraph/GEN-SIM-RECO/MC_31X_V3_7TeV-v2/0000/36C76CCF-4216-DF1
1-8D9F-00D0680BF8C2.root: MISSING 1 blocks of total size 134217728 B.....................................................
 Total size:    284402226706795 B (Total open files size: 7970226176 B)
 Total dirs:    60154
 Total files:   781238 (Files currently being written: 9)
 Total blocks (validated):      2764454 (avg. block size 102878263 B) (Total open file blocks (not validated): 60)
  ********************************
  CORRUPT FILES:        1
  MISSING BLOCKS:       1
  MISSING SIZE:         134217728 B
  CORRUPT BLOCKS:       1
  ********************************
 Minimally replicated blocks:   2764453 (99.99997 %)
 Over-replicated blocks:        6750 (0.24417119 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    3
 Average block replication:     2.5729363
 Corrupt blocks:                1
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          185
 Number of racks:               1
```

The Economics of HDFS

- As far as my crystal ball reaches, any storage element can have sufficient performance (assuming sanity on part of the experiment). Selection criteria should be based on other things.
  - Probably the cheapest solution in terms of hardware *if* co-located with worker nodes. Somewhere between USD $100-$200 / TB.
  - Maintenance costs go down the more sites running it.
  - Disk-only: doesn't integrate with a tape system!
  - Twice as many replicas improves scalability *but* means twice as many disks are powered on. Inconsequential power cost in the US midwest, but might be significant in the EU.

Remote Access

- Although the current version of HDFS has KRB5 integration, we don't use KRB5 locally.
  - We wanted to offer secure access to on-site, but off-cluster physicists.
  - Physicists didn't seem to care for gsiftp, but xrootd was well-integrated into their applications (ROOT).

Remote Access

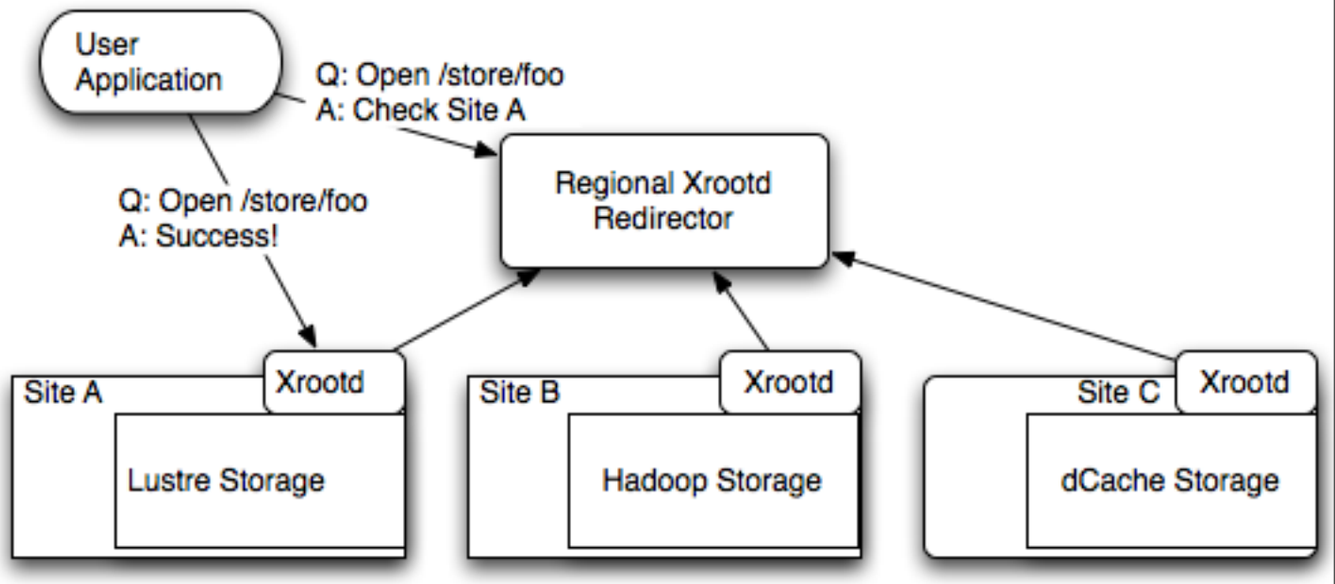
- Xrootd is a modular architecture; about a day's worth of work to integrate HDFS's libhdfs.so with Xrootd server.
- We were able to provide GSI access to physicists, and present a namespace that matched the CMS global namespace.
- We added multiple servers for load-balancing.

Xrootd Federation

- To Xrootd, there's little difference between aggregating several disk servers and servers that act as proxies for sites.
- We built a single redirector for 2 sites, then gradually expanded it to all USCMS T1/T2 sites.

Regional Architecture

- Same namespace, irrespective of data location:



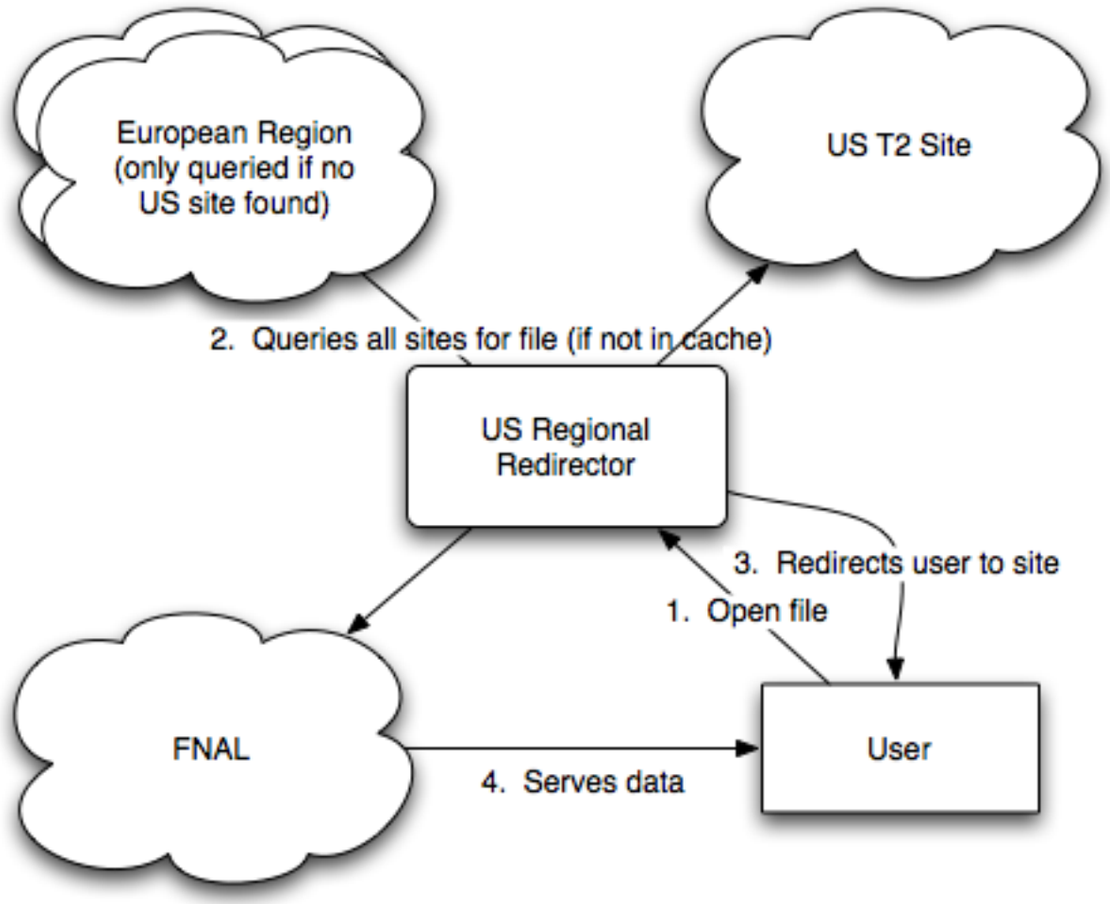
TFile::Open("root://xrootd.unl.edu//store/foo")

New: now integrates with native dCache xrootd door!

New: all CMS AOD data available via above URL!

Architecture

- Long term goal:



“Approved” Use Cases

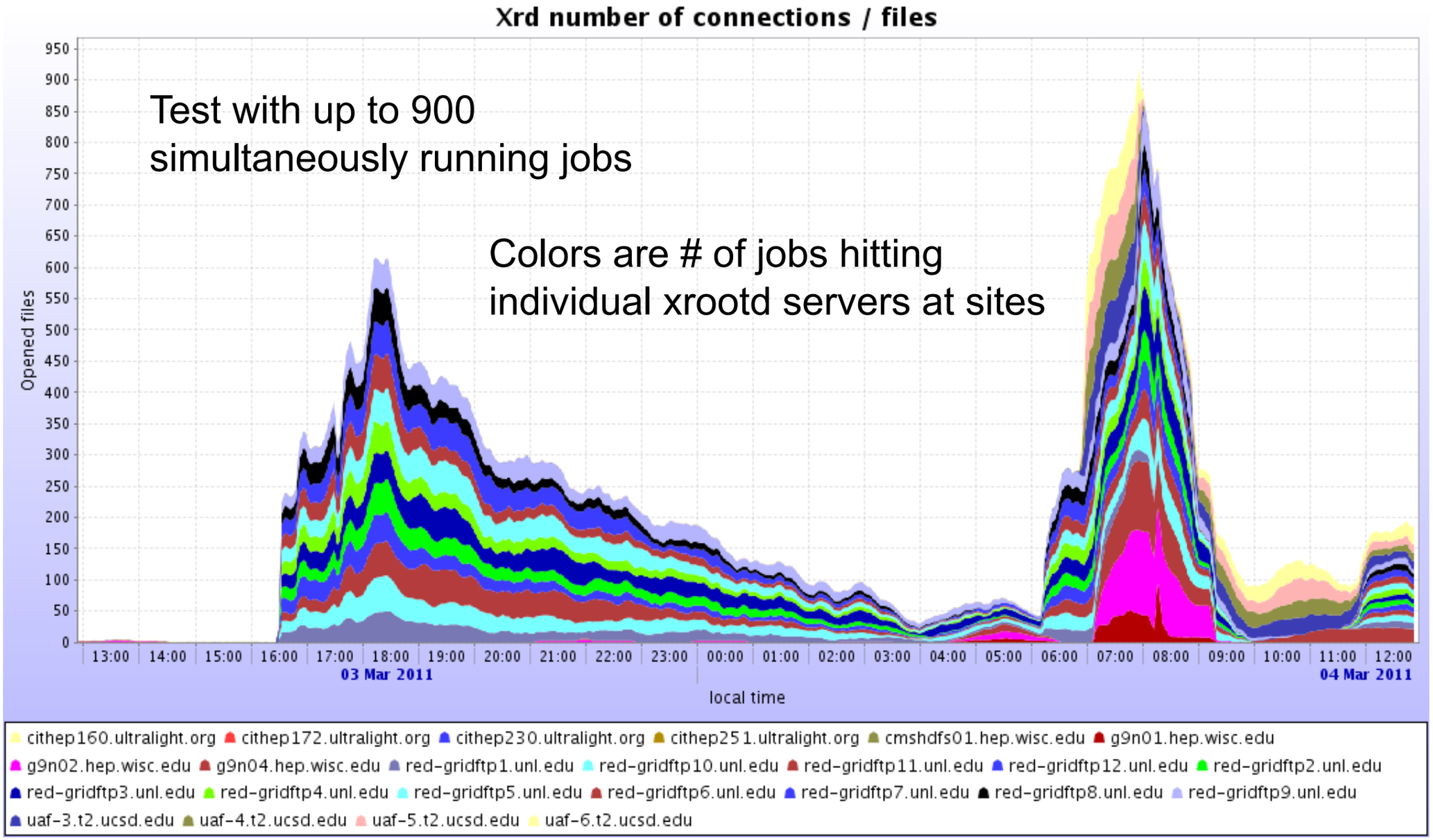
- **Fallback**: To avoid a crash, allow jobs to open file remotely if they fail in doing it locally.
- **Interactive use**: Debugging single files, event viewer.
- **Overflow**: Purposely allow jobs to go to sites without the data at a small scale (if they would otherwise be queued).
- I suspect one challenge will be able to protect against / support “surprise” use cases.

Speed

- High-latency CMSSW was impossible prior to 3_8.
- It was annoying until 4_1.
- As of 4_2, there's little difference for many workflows.
- Basically: the slowdown is acceptable unless you are doing something large-scale (which we discourage)

Message for SuperB: high-latency access is possible *after* investing the effort in the I/O layer.

MonALISA - Open Files



Currently, at Wisconsin, *most* access to HDFS is via Xrootd

Monitoring

- Status is monitored by Nagios heartbeat tests.

Xrootd Services from Nebraska (xrootd-services-nebraska)

Host	Status	Services	Actions
red-gridftp1	UP	1 OK	
red-gridftp10	UP	1 OK	
red-gridftp11	UP	1 OK	
red-gridftp12	UP	1 CRITICAL	
red-gridftp2	UP	1 OK	
red-gridftp3	UP	1 OK	
red-gridftp4	UP	1 OK	
red-gridftp5	UP	1 OK	
red-gridftp6	UP	1 OK	
red-gridftp7	UP	1 OK	
red-gridftp8	UP	1 OK	
red-gridftp9	UP	1 OK	
xrootd.unl.edu	UP	1 CRITICAL	

Xrootd Services from Purdue (xrootd-services-purdue)

Host	Status	Services	Actions
cmsdbs.rcac.purdue.edu	UP	1 OK	
crabserver.rcac.purdue.edu	UP	1 OK	
xrootd-itb.unl.edu	UP	1 OK	
xrootd.rcac.purdue.edu	UP	1 OK	

Xrootd Services from Xrootd (xrootd-services-ucsd)

Host	Status	Services	Actions
uaf-3.t2.ucsd.edu	UP	1 OK	
uaf-4.t2.ucsd.edu	UP	1 OK	
uaf-5.t2.ucsd.edu	UP	1 OK	
uaf-6.t2.ucsd.edu	UP	1 OK	
xrootd-itb.unl.edu	UP	1 OK	

Xrootd Services from Caltech (xrootd-services-caltech)

Host	Status	Services	Actions
cithep160.ultralight.org	UP	1 OK	
cithep172.ultralight.org	UP	1 OK	
cithep230.ultralight.org	DOWN	1 CRITICAL	
cithep251.ultralight.org	UP	1 OK	
gridftp-16-23.ultralight.org	UP	1 CRITICAL	
xrootd.unl.edu	UP	1 OK	

Xrootd Services from FNAL (xrootd-services-fnal)

Host	Status	Services	Actions
cmssrv32.fnal.gov	UP	1 OK	
xrootd.unl.edu	UP	1 OK	

Hadoop/Xrootd Thoughts

- We've been running HDFS for about 2.5 years. It's been a wonderful fit for our USCMS T2 site.
- Xrootd allows us to extend the cluster storage to external users.
- Further, xrootd use allows us to extend a uniform interface to users across all US sites. This will roll out to users in August: we are very excited!