OSG Hadoop 2010

Brian Bockelman
OSG Storage Forum

Hadoop in General

- Many of you have probably heard my HDFS talk before.
  - If not, check out the web material.
- HDFS = Hadoop Distributed File System; developed primarily by Facebook and Yahoo! to store petabytes of data.
- Adopted in 2009 as a SE on the OSG.

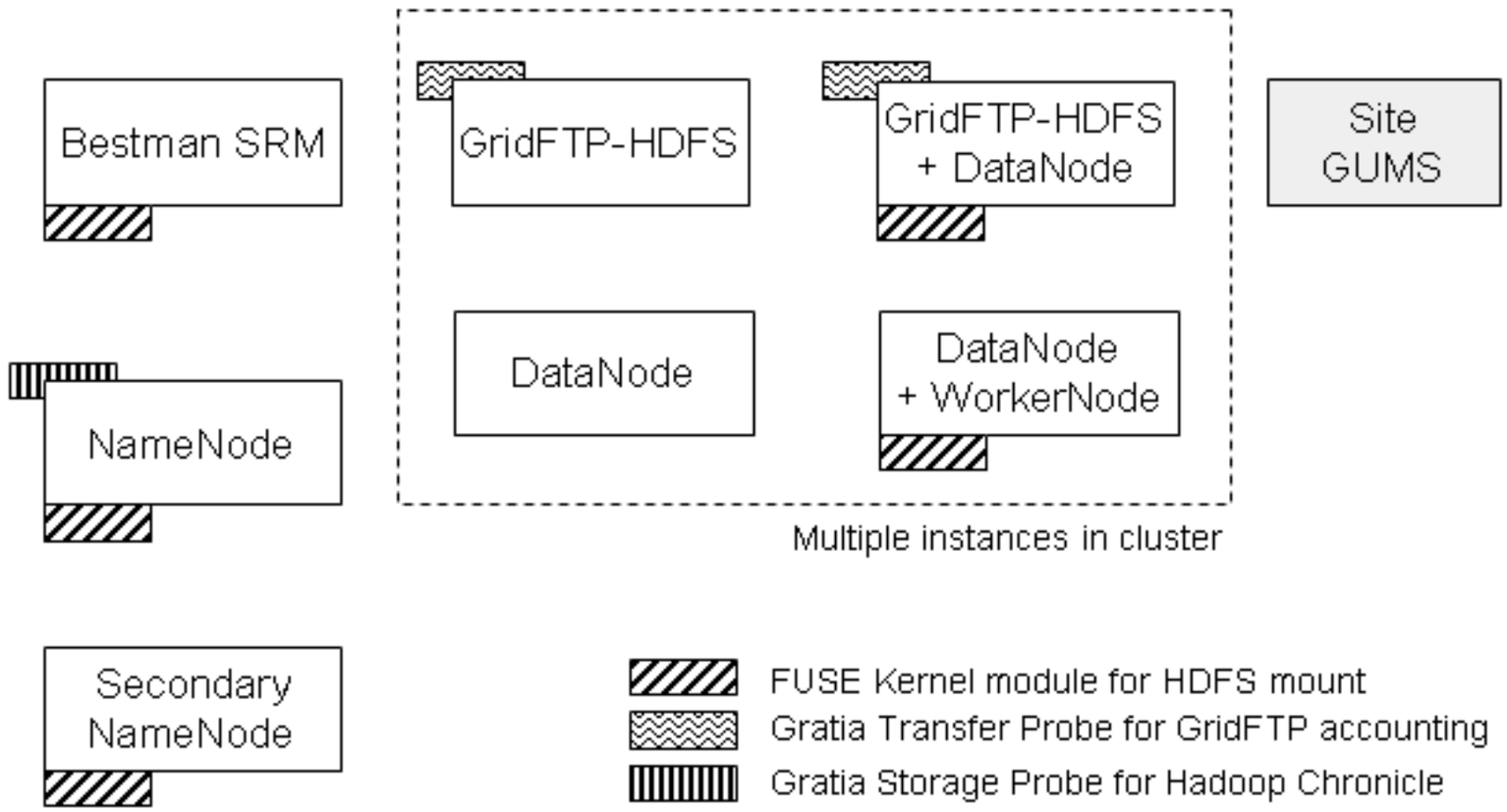
Users

- Big external: Yahoo (25,000 nodes; largest cluster is 4,000 nodes @ 16PB), Facebook (largest cluster, 13PB)
- LHC T2: UCSD (1PB), Nebraska (1.7PB), Caltech (1PB), Estonia (?)
- T3: UCD, UColorado, T3_ES_Oviedo (backup only)

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.

Architecture in a Slide



SRM Hadoop storage system: Example topology at an OSG Site

Management Highlights

- 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!
  - **No more data hotspots**.
- Setting **quotas** (*per directory*).
- **Backups of namespace**.
- **Balancer** is included.

Demo: HadoopViz

- The data from this demo is additionally used to feed the Gratia probes for this site.

Featured Ramblings...

- In this talk, I want to focus on the "Hadoop news" in 2010.

Hadoop 0.20

- Hadoop 0.20 support is coming!
  - This is mostly a bugfix/stability release - but bugfix and stability is important!
- There's already a release candidate available.
- You will be able to use the RHEL 'alternatives' command to manage different configurations.
  - For example, you can roll your own site's configuration RPM and set it as the highest-priority alternative.

Hadoop 0.20 Timeline

- Expect your site to be "encouraged" to upgrade to 0.20 by the end of this year.
- Upgrade should be a "less than 1 day" event.
- Nebraska and Caltech first - will update documentation if needed - then others.

Hadoop 0.22

- Probably next year sometime.
- Between 0.20 and 0.21, 1,000 bugs and new features. 0.22 will probably have 0.22 tickets closed.
- The big OSG-related news will be the underlying support for kerberos security.
  - Kerberos is just a stone's throw away from GSI.

New Packaging

- Good packaging has been near and dear to our heart.
- OSG Hadoop has only ever been officially distributed using source-based RPMs, installing into locations according to the official Fedora recommendations.
- This isn't changing.

New Packaging

- As the scope of HDFS grows to include things like Xrootd and Gratia probes, it is harder to "hand-maintain" build information.
- We have recently switched to Koji, the Fedora release-engineering tool:
  - http://koji.hep.caltech.edu/koji

Koji Info

- Koji allows us to go from updated source code to a development repo in a single click.
- Helps us manage patch sets, build artifacts.
  - Given a version number, I can definitively tell you what patches were applied. Prevents accidental reverting of patches.
- Builds each RPM from a “clean” chroot environment. Means that dependencies are better handled.

More Koji

- Because we're using native packages, we can take advantage of the source packaging of Globus gridftp in Fedora.
  - In the Hadoop 0.20 time frame, we will update from VDT binary RPMs to Fedora source RPMs.
- We will also switch from our developed RPMs to Cloudera-developed RPMs.
- Goal is to get out of packaging as much of the base as possible.

Increased Firepower with Xrootd

- HDFS is a cluster-oriented filesystem.
  - It assumes that all your users are inside your cluster.
- We "open it up" partially by adding grid translation layers based on existing software:
  - BestMan SRM.
  - Globus GridFTP.

What about Xrootd?

Xrootd Integration

- We have aggressively pushed Xrootd integration with HDFS.
  - We believe this is a great way to provide X509-secured access to collaborating physicists outside your LAN.
- Scales and load-balances access among several servers as needed.

Demo

- Enables:
  - Recursive download of files.
  - Downloading in parallel streams.
  - Doing analysis from your laptop.

The Hadoop Chronicle

- Actually, something done in 2009 but not widely advertised.
  - Uses Gratia space accounting

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

Delete Junk Reply Reply All Forward Print To Do

```
--------------------
| 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 |
```

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
```

Hadoop, Thoughts

- HDFS usage continues to grow in total space and number of sites.
  - We've had the same version in production - very stable - for 18 months.
  - Other teams contribute to the core; we work on extending the admin tools and grid integration.
  - The focus is *production*.

HDFS has lead the way in terms of commodity storage elements for the LHC!