Nebraska Storage

Holland Computing Center
University of Nebraska-Lincoln
Garhan Attebury

Site Overview

T2_US_Nebraska

- ~ 1600 job slots
- ~ 1.5PB raw storage, ~700TB really usable
- CentOS 5.5 x86_64
- Condor 7.4.3
- Hadoop 0.19.2-dev (from Caltech)
- Bestman + GridFTP / xrootd / fdt
- Pair of CEs + Interactive Tier3 + other HCC CEs

SE Status

- ~1.5PB raw and usually working, or is it 700TB ?
- Variety pack of worker and datanodes
  - 60x workers with small scratch, no HDFS, and <=1GB/core
  - 114x workers with 2x 1.5 or 2TB drives (X2200 / SC 1435s)
  - 28x workers with 6x 2TB drives no RAID (R710)
  - 40x workers with 12x 2TB with RAID + 2x 147GB (R510)
  - 12x workers with 4x 2TB drives no RAID (R410)
  - 5x workers with 12x 2TB drives without RAID (X4275)
  - 6x dedicated datanodes (10-20TB SCSI RAID vaults)
- Single gigabit connectivity (BI-RX16)
- Namenode/Secondary on commodity hardware (X2200)
- No rack awareness, need to upgrade to 0.20.x

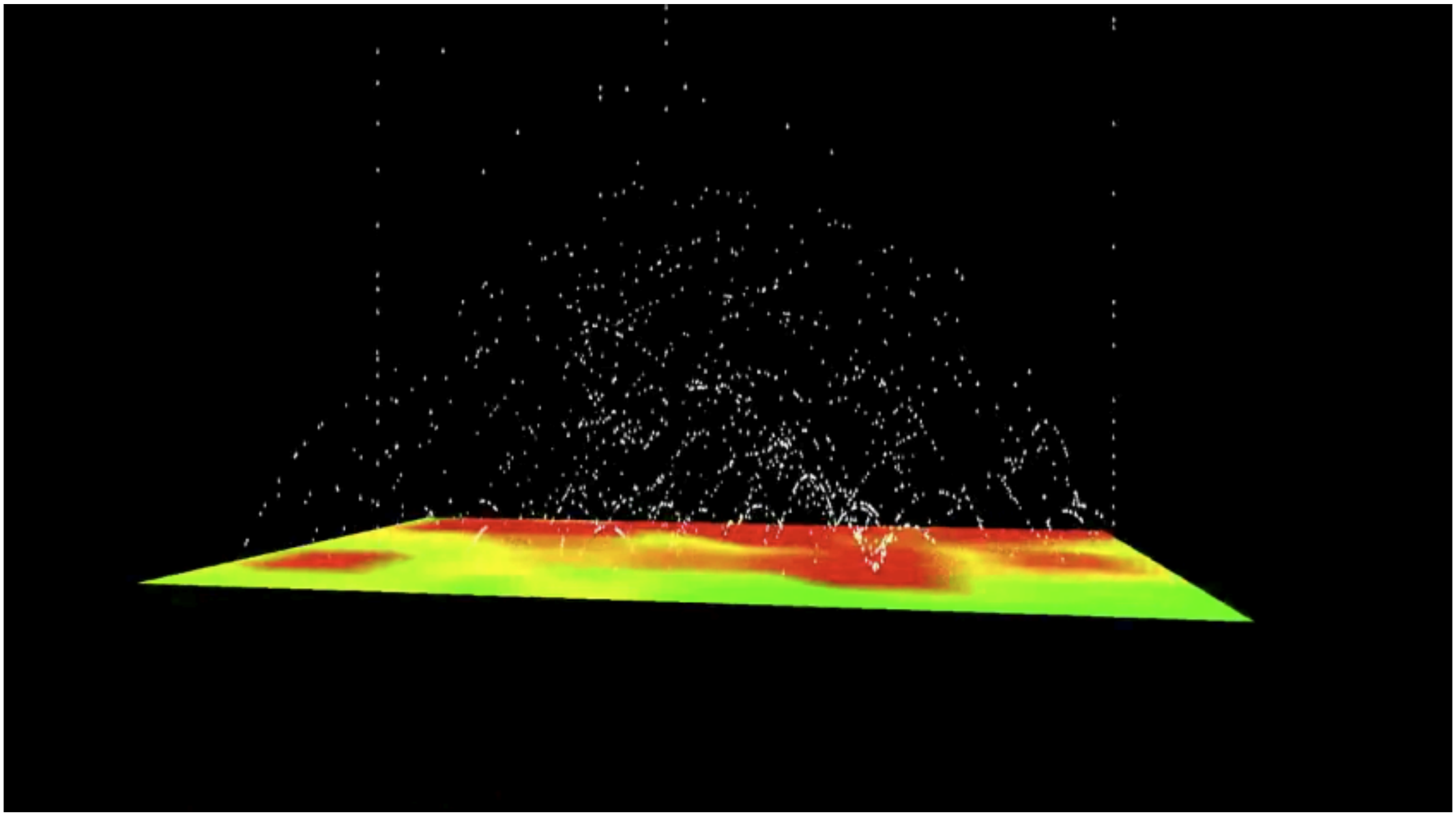
HDFS positives

What works well for us

- Automatic replication (no seriously, it works well!)
- View of current state and activity
  - HadoopVIZ
  - Hadoop Chronicles
- Easy to detect errors and corruptions
  - hfscker [-c, -m, -u] > bad_files.txt
  - hadoop fsck [-files -locations -blocks]
- Manual recovery techniques (not fun for most, but awesome when you want them)
- Masques the commodity hardware (::cough::1.5TB Seagates::cough)

HadoopViz

http://www.youtube.com/watch?v=qoBoEzOkeDQ



```
============================================================
  The Hadoop Chronicle | 42 % | 2010-09-20
============================================================

--------------------
| 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 |

--------------------
| Pool Information |
------------------------------------------------------------
|     Statistic      | Today | 1 Day Change | 7 Day Change |
------------------------------------------------------------
| 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-DF11-8D9F-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-DF11-8D9F-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 filesystem under path '/' is CORRUPT
```

```
[root@hadoop-name ~]# hadoop fsck /user/uscms01/pnfs/unl.edu/data4/cms/store/mc/Summer09/
PhotonJets_Pt40to100-madgraph/GEN-SIM-RECO/MC_31X_V3_7TeV-v2/0000/36C76CCF-4216-
DF11-8D9F-00D0680BF8C2.root -files -locations -blocks
/user/uscms01/pnfs/unl.edu/data4/cms/store/mc/Summer09/PhotonJets_Pt40to100-madgraph/GEN-SIM-
RECO/MC_31X_V3_7TeV-v2/0000/36C76CCF-4216-DF11-8D9F-00D0680BF8C2.root 1986657375 bytes, 15 block
(s):
/user/uscms01/pnfs/unl.edu/data4/cms/store/mc/Summer09/PhotonJets_Pt40to100-madgraph/GEN-SIM-
RECO/MC_31X_V3_7TeV-v2/0000/36C76CCF-4216-DF11-8D9F-00D0680BF8C2.root: CORRUPT block
blk_5149773138535264325
 MISSING 1 blocks of total size 134217728 B
...
4. blk_3599431394020861l7_4966136 len=134217728 repl=2 [172.16.3.37:50010, 172.16.3.13:50010]
5. blk_5149773138535264325_4966139 len=134217728 MISSING!
6. blk_4036939530067920146_4966141 len=134217728 repl=2 [172.16.1.235:50010, 172.16.3.14:50010]
...
```

1) Grep 5149773138535264325 from namenode logs
2) Poke around ext3 filesystem on a datanode that should have that block (recursive find command)
3) Copy the block and it's .meta to a good datanode
4) Restart the good datanode process to trigger a block report to the namenode
5) Namenode replicates as needed and we're good to go!

HDFS negatives

- Balancing act (not really a fault)
- Datanodes on workers restricts freedom
  - Time, time, and more time!
- Reliance on FUSE and its ... quirks
- HA, tweaking
- Real cost vs availability vs performance ?
  - Everything works fine! ... +/- 10%

HDFS Improvements

- Rack awareness (we keep saying it)
- **Real time performance monitoring**
- Data location tweakability? More knobs...

(Almost) Dedicated GridFTP

- 12x PowerEdge SC 1435 w/16GB RAM
  - gridftp-hdfs
  - xrootd-hdfs
  - fdt-hdfs
- Reordering data in both RAM and on disk
- Easy to admin and treat as its own entity
- ~9-9.5Gb/sec sustained (good enough)

Nebraska Storage

- 600+TB via GridFTP, xrootd, fdt - serving our ~1600 slots + Firefly well so far
- Very heterogenous environment
- HDFS (and loving it)