Any Data, Anytime, Anywhere

Dan Bradley <dan@hep.wisc.edu>
representing the AAA Team
At OSG All Hands Meeting
March 2013, Indianapolis
AAA Project

Goal

- Use resources more effectively through remote data access in CMS

Sub-goals

- Low-ceremony/latency access to any single event
- Reduce data access error rate
- Overflow jobs from busy sites to less busy ones
- Use opportunistic resources
- Make life at T3s easier
xrootd: Federating Storage Systems

- Step 1: deploy seamless global storage interface
- But preserve site autonomy:
  - xrootd plugin maps from global logical filename to physical filename at site
    - Mapping is typically trivial in CMS: /store/* → /store/*
  - xrootd plugin reads from site storage system
    - Example: HDFS
  - User authentication also pluggable
    - But we use standard GSI + Icmaps + GUMS
Status of CMS Federation

US

- T1 (disk) + 7/7 T2s federated
- Covers 100% of the data for analysis
- Does not cover files only on tape

World

- 2 T1s + 1/3 T2s accessible
- Monitored but not a “turns your site red” service (yet)
WAN xrootd traffic
Opening Files

![Aggregated Xrootd connection counts and rates graph](image)
# Microscopic View

<table>
<thead>
<tr>
<th>File Path</th>
<th>User Hash</th>
<th>Server Domain</th>
<th>Client Domain</th>
<th>Open Age</th>
<th>Update Age</th>
<th>Read [MB]</th>
<th>Read [%]</th>
<th>Rate [MB/s]</th>
<th>Avg Read [MB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>/store/user</td>
<td>9C73E43</td>
<td>pi.infn.it</td>
<td>cern.ch</td>
<td>00:00:45</td>
<td>00:00:05</td>
<td>1040.000</td>
<td>50.137</td>
<td>2.719</td>
<td>0.199</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>spacre.org.br</td>
<td>00:09:31</td>
<td>00:01:16</td>
<td>1595.292</td>
<td>54.211</td>
<td>2.791</td>
<td>1.538</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>spacre.org.br</td>
<td>00:58:38</td>
<td>00:00:00</td>
<td>7337.478</td>
<td>113.713</td>
<td>2.085</td>
<td>0.161</td>
</tr>
<tr>
<td>/store/reval</td>
<td>17EE54D2</td>
<td>unlnl.edu</td>
<td>hep.wisc.edu</td>
<td>00:07:40</td>
<td>00:04:30</td>
<td>943.557</td>
<td>100.000</td>
<td>2.048</td>
<td>117.945</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>ihep.ac.cn</td>
<td>01:01:28</td>
<td>00:00:06</td>
<td>7111.948</td>
<td>124.163</td>
<td>1.928</td>
<td>0.148</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>unlnl.edu</td>
<td>00:58:29</td>
<td>00:00:29</td>
<td>6104.501</td>
<td>97.811</td>
<td>1.739</td>
<td>0.144</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>kfki.hu</td>
<td>00:11:48</td>
<td>00:00:26</td>
<td>6224.495</td>
<td>104.975</td>
<td>1.678</td>
<td>0.151</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>oeaw.ac.at</td>
<td>00:01:20</td>
<td>00:00:26</td>
<td>5659.936</td>
<td>94.998</td>
<td>1.563</td>
<td>0.144</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>grid.hep.ph.ic.ac.uk</td>
<td>01:01:00</td>
<td>00:00:10</td>
<td>5439.597</td>
<td>86.131</td>
<td>1.486</td>
<td>0.153</td>
</tr>
<tr>
<td>/store/reval</td>
<td>17EE54D2</td>
<td>unlnl.edu</td>
<td>hep.wisc.edu</td>
<td>00:03:34</td>
<td>00:01:34</td>
<td>309.099</td>
<td>32.568</td>
<td>1.440</td>
<td>103.033</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>t2.ucsd.edu</td>
<td>grid.helsinki.fi</td>
<td>00:01:21</td>
<td>00:00:21</td>
<td>4739.671</td>
<td>82.013</td>
<td>1.287</td>
<td>0.134</td>
</tr>
<tr>
<td>/store/data/Run2012D</td>
<td>96DC057E</td>
<td>fnal.gov</td>
<td>colorado.edu</td>
<td>00:10:19</td>
<td>00:00:09</td>
<td>7755.694</td>
<td>24.519</td>
<td>1.252</td>
<td>0.910</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>unlnl.edu</td>
<td>00:07:23</td>
<td>00:00:05</td>
<td>529.325</td>
<td>17.730</td>
<td>0.393</td>
<td>0.117</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>unlnl.edu</td>
<td>00:56:16</td>
<td>00:02:46</td>
<td>3898.405</td>
<td>96.028</td>
<td>0.880</td>
<td>0.080</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>unlnl.edu</td>
<td>00:04:51</td>
<td>00:00:51</td>
<td>334.437</td>
<td>8.207</td>
<td>0.147</td>
<td>0.064</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>unlnl.edu</td>
<td>00:23:42</td>
<td>00:00:21</td>
<td>1617.640</td>
<td>39.931</td>
<td>1.137</td>
<td>0.078</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>unlnl.edu</td>
<td>00:58:53</td>
<td>00:04:53</td>
<td>3861.664</td>
<td>95.972</td>
<td>1.093</td>
<td>0.079</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>hpcc.ttu.edu</td>
<td>00:07:01</td>
<td>00:00:01</td>
<td>3049.141</td>
<td>47.980</td>
<td>1.081</td>
<td>0.099</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>cmsaf.mit.edu</td>
<td>00:58:30</td>
<td>00:00:50</td>
<td>3772.923</td>
<td>63.406</td>
<td>1.075</td>
<td>0.166</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>grid.hep.ph.ic.ac.uk</td>
<td>01:00:30</td>
<td>00:00:20</td>
<td>3795.141</td>
<td>58.114</td>
<td>1.045</td>
<td>0.146</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>iihe.ac.be</td>
<td>00:01:50</td>
<td>00:00:20</td>
<td>3515.779</td>
<td>56.985</td>
<td>0.963</td>
<td>0.159</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>jinru</td>
<td>00:58:23</td>
<td>00:00:47</td>
<td>3312.677</td>
<td>53.456</td>
<td>0.946</td>
<td>0.467</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>unlnl.edu</td>
<td>00:02:47</td>
<td>00:00:12</td>
<td>156.925</td>
<td>4.505</td>
<td>0.936</td>
<td>0.056</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>lnl.infn.it</td>
<td>00:01:20</td>
<td>00:00:50</td>
<td>3374.807</td>
<td>52.881</td>
<td>0.932</td>
<td>0.131</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>grid.hep.ph.ic.ac.uk</td>
<td>00:00:58</td>
<td>00:00:06</td>
<td>3391.159</td>
<td>58.117</td>
<td>0.927</td>
<td>0.154</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>lal.in2p3.fr</td>
<td>00:01:29</td>
<td>00:00:49</td>
<td>3355.611</td>
<td>56.928</td>
<td>0.909</td>
<td>0.129</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>kfki.hu</td>
<td>00:58:40</td>
<td>00:02:40</td>
<td>3156.378</td>
<td>49.767</td>
<td>0.897</td>
<td>0.157</td>
</tr>
<tr>
<td>/store/group/snt</td>
<td>51B44D6</td>
<td>fnal.gov</td>
<td>sinp.msu.ru</td>
<td>00:58:00</td>
<td>00:00:10</td>
<td>3120.052</td>
<td>50.683</td>
<td>0.896</td>
<td>0.135</td>
</tr>
</tbody>
</table>
Problem

- Access via xrootd overloads site storage system
  - Florida to Federation, “We are seceding!”
- Terms of the Feb 2013 treaty:
  - Addition of local xrootd I/O load monitoring
  - Site can configure automatic throttles
    - When load too high, rejects new transfer requests
    - End-user only sees error if file unavailable elsewhere in federation
- But these policies are intended for the exception, not the norm, because ...
Regulation of Requests

- To 1\textsuperscript{st} order, jobs still run at sites with the data
  - \(~0.25\) GB/s average remote read rate
  - \(O(10)\) GB/s average local read rate
  - \(~1.5\) GB/s PhEDEx transfer rate
- Cases where data is read remotely:
  - Interactive - limited by \# humans
  - Fallback - limited by error rate opening files
  - Overflow - limited by scheduling policy
  - Opportunistic - limited by scheduling policy
  - T3 - watching this
At the Campus Scale

Some sites are using xrootd for access to data from across a campus grid

• Examples: Nebraska, Purdue, Wisconsin
More on Fallback

- On file open error, CMS software can retry via alternate location/protocol
  - Configured by site admin
  - We fall back to regional xrootd federation
    - US, EU
    - Could also have inter-region fallback
      - Have not configured this ... yet
- Can recover from missing file error, but not missing block within file error (more on this later)
- Has more uses than just error recovery ...
More about Overflow

- GlideinWMS scheduling policy
  - Candidates for overflow:
    - Idle jobs with wait time above threshold (6h)
    - Desired data available in a region supporting overflow
  - Regulation of overflow:
    - Limited number of overflow glideins submitted per source site
- Data access
  - **No reconfiguration of job required**
    - Uses fallback mechanism
    - Try local access, fall back to remote access on failure
Overflow

- Small but steady overflow in US region
Running Opportunistically

- To run CMS jobs at non-CMS sites, we need
  - Outbound network access
  - Access to CMS datafiles
    - Xrootd remote access
  - Access to conditions data
    - http proxy
  - Access to CMS software
    - CVMFS (also needs http proxy)
CVMFS Anywhere

But non-CMS sites might not happen to mount the CMS CVMFS repository

→ **Run the job under Parrot** (from cctools)
  - Can now access CVMFS without FUSE mount
  - Also gives us identity boxing
    - Privilege separation between glidein and user job
  - Has worked well for guinea pig analysis users
    - Working on extending it to more users

What about in the cloud?

- If you control the VM image, just mount CVMFS
Fallback++

- Today we can recover when file is missing from local storage system
- But missing blocks within files cause jobs to fail
  - And job may come back and fail again ...
  - Admin may need to intervene to recover the data
  - User may need to resubmit the job
- Can we do better?
Yes, We Hope

- Concept
  - Fall back on read error
  - Cache remotely read data
  - Insert downloaded data back into storage system
File Healing Status

- Currently have it working via whole-file caching
  - Still only triggered by file open error
- Plans to support partial-file healing
  - Will need to fall back to local xrootd proxy on all read failures
  - Current implementation is HDFS-specific
    - Modifies HDFS client to do the fallback to xrootd
    - But it's not CMS-specific
Cross-site Replication

● Once we have partial-file healing …
  ● Could reduce HDFS replication level from 2 to 1 and use cross-site redundancy instead
    - Would need to enforce the replication policy at higher level
    - May not be good idea for hot data
    - Need to consider impact on performance
Performance

Mostly CMS application-specific stuff

- Improved remote read performance by combining multiple reads into vector reads
  - Eliminates many round-trips
- Working on bit-torrent-like capabilities in CMS application
  - Read from multiple xrootd sources
  - Balance load away from slower source
  - React in $O(1)$ minute time frame
HTCondor Integration

- Improved vanilla universe file transfer scheduling and monitoring in 7.9
  - Used to have one file transfer queue
    - One misconfigured workflow could starve everything else
    - Difficult to diagnose
  - Now one per user
    - Or per arbitrary attribute (e.g. target site)
  - Equal sharing between transfer queues in case of contention
  - Reporting transfer status, bandwidth usage, disk load, and network load
  - And now you can condor_rm those malformed jobs that are transferring GBs of files :)

Any data, Anytime, Anywhere
Summary

- xrootd storage federation rapidly expanding and proving useful within CMS
- We hope to do more
  - Automatic error recovery
  - Opportunistic usage
  - Improving performance