XCache - XRooTd
Cache for CMS in SoCal

DOMA Caching Meeting

Sep 18 2018

Edgar Fajardo (Presenting),
Terrence Martin, Frank Würthwein, Alja Tadel, Matevž Tadel, Justas Balcas
Motivation

• Have all the MiniAOD of Run 2 easily available for CMS users in SoCal
• First step on potentially merging UCSD and Caltech’s namespace.
• Profit from the PRP 100 Gbit connection and 3ms latency between sites.
The idea of an Xrootd cache cluster was presented on Xrootd Tokyo meeting

XRootd Proxy File Cache V2

XRootd @ ICEPP Tokyo, 11/08/2016
Alja & Matevž Tadel, UCSD

Full talk available: here
Xrootd Local Cache

[Diagram of local cache system with labels for local cluster, caching proxy, RAM, disk, and federation redirector.]
On ACAT 2017 the Xrootd cache cluster was scale tested. Full paper [here](#).
Two caches become one

- 120 miles
- 100 Gbit/sec
- 3ms
Jobs at UCSD and Caltech transparently use the cache

SoCal Xrootd Cache (2018)
## SoCal Xrootd Cache specs

<table>
<thead>
<tr>
<th></th>
<th>UCSD</th>
<th>Caltech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Disk Capacity node</td>
<td>12x2TB = 24TB.</td>
<td>30 x 6TB disks (HGST Ultrastar 7K6000)</td>
</tr>
<tr>
<td>Network Card</td>
<td>10 Gbps</td>
<td>40 Gbps</td>
</tr>
<tr>
<td>Total Capacity</td>
<td>264 TB</td>
<td>360 TB</td>
</tr>
</tbody>
</table>
## Space Needs

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Size (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>/Run2016</em>-03Feb2017*/MINIAOD</td>
<td>182.8</td>
</tr>
<tr>
<td><em>/RunIISummer16MiniAODv2-PUMoriond17_80X_</em>/MINIAODSIM</td>
<td>502.5</td>
</tr>
<tr>
<td>/*/<em>RunIIFall17MiniAODv2</em>/MINIAODSIM</td>
<td>211</td>
</tr>
<tr>
<td>/<em>/</em>-31Mar2018*/MINIAOD</td>
<td>137.9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1041</strong></td>
</tr>
</tbody>
</table>
Scaling tests

- We had a Monash University student (Ben Steer) that performed some scale testing on the UCSD cache.
- Thorough results of the scale tests can be found [here](#).
- The most important aspects are in the following slides.
- Tests were revisited for HEPIX on spring 2018.
- UCSD student (Caitlin Hung) performed acceptance testing on the Caltech side.
Scaling Tests Description

• Baseline test used jobs reading at twice the CMS job usual spec 2MB/sec.
  • We ran at different scales 1000/2000/3000/4000 jobs.
• Tests run jobs that simulate actual use of the system by clients using tool xrdfragcp (like xrdcp but controlling read rate)
How to test it?
Use a sleeper pool

Not like this one:

Like This one:

A worker node
Real Cores
Sleeper cores
Aggregate view of all 9 NICs

Peak of 45 Gbps received

Bits received means IO from WAN to cache

Peak of 57.8 Gbps sent

Bits send means IO from cache to jobs

Bits read from disk cache = Bits send – Bits received

3/20/17  CENIC17  38
write to disk peak at ~3Gbps

Reads from disk pretty steady at ~2Gbps

Note: both reads and writes are limited by complicated interplay of cache behavior, hardware performance, and requests from jobs.
Scaling tests (Caltech Only)

- 8k jobs reading at 2MB/sec
- 2Gbps
- Out of the Cache
- 25 Gbps
- Into the Cache
Scaling tests (Both sites)

Serving at 50 Gbps

1.5k jobs at 4MB/sec
Production performance

Serving at 30 Gbps

Writing to disk only 10 Gbps
How to bring the cache in production

Storage: Rule in storage.xml

```xml
<lfn-to-pfn protocol="direct" destination-match=".*"
   path-match="/+store/(data/Run2016[A-Z]/[^/]+/MINIAOD/03Feb2017.*)"
   result="root://xrootd.t2.ucsd.edu:2050//store/$1"/>
<lfn-to-pfn protocol="direct" destination-match=".*"
   path-match="/+store/(mc/RunIISummer16MiniAODv2/[^/]+/MINIAODSIM/
PUMoriond17_80X_.*)"
   result="root://xrootd.t2.ucsd.edu:2050//store/$1"/>
```

Computing: New group in frontend

```xml
<group name="overflow-xcache-socal" enabled="True"><start_expr='(regexp("/\*/Run2016.*-03Feb2017.*/MINIAOD", DESIRED_CMSDataset) || regexp("/\*/RunIISummer16MiniAODv2-PUMoriond17_80X_.*/MINIAODSIM", DESIRED_CMSDataset))'>
```

DOMA Caching Meeting
Future Work

- Having other Tier3’s in SoCal use the cache: i.e T3_US_UCR (OnGoing)
- Going out of California: Colorado
- What about NorCal? T3_US_UCD
- Learning how to operate a collection of Caches at all US Tier 2’s, which polices? each site own policy?
- Grow cache size on the fly with kubernetes.
Questions?

Contact us at:

1-900-Xrootd-Cache-Masters
Just Kidding

Contact us:

emfajard@ucsd.edu

Thank You