New Computing Models and LHCOOne

22 May 2012
Ian Fisk
Over the development the evolution of the WLCG Production grid has oscillated between structure and flexibility.

- Driven by capabilities of the infrastructure and the needs of the experiments.
Old Model

- To greater and lesser extents LHC Computing model are based on the MONARC model
  - Developed more than a decade ago
  - Foresaw Tiered Computing Facilities to meet the needs of the LHC Experiments
- Assumes poor networking
- Hierarchy of functionality and capability

Fig. 4-1. Computing for an LHC Experiment Based on a Hierarchy of Computing Centers. Capacities for CPU and disk are representative and are provided to give an approximate scale.
Old Model Impact

- Connections are seen as not sufficient or reliable
  - Data needs to be pre-placed
  - Nothing can happen utilizing remote resources on the time of a running job
  - Data comes from specific places
  - Sites have specific functions
New Model

- Strict hierarchy of connections becomes more of a mesh
- Divisions in functionality especially for chaotic activities like analysis become more blurry
- More access over the wide area

Model changes have been an evolution
- Not all experiments have emphasized the same things
- Each pushing farther in particular directions
During the evolution the low level services are largely the same.

Most of the changes come from the actions and expectations of the experiments.

Grid Services

- ** Experiment Services **
  - WMS
  - BDII
  - FTS
  - VOMS

- ** Site **
  - CE
  - SE

Connection to batch (Globus and CREAM based)

Connection to storage (SRM or xrootd)

Lower Level Services
Providing Consistent Interfaces to Facilities

Higher Level Services
Mesh Transfers

- Change from

Tier-1 → Tier-2

- To

Tier-1 → Tier-2

CMS PhEDEx - Transfer Rate
120 Days from Week 35 of 2010 to Week 00 of 2011

Transfers West

Transfers East
Completing the Mesh

Tier-2 to Tier-2 transfers are now similar to Tier-1 to Tier-2 in CMS.

T1->T2 (500MB/s Average)

T2->T2 (300MB/s Average)
Access

- Situation is improved
  - less artificial separation in where data comes from
  - But data is still placed at sites

In CMS 30 % of samples subscribed by physicists not used for 3 months during 2010
Services like the Data Popularity Service track all the file accesses and can show what data is accessed and for how long.

- Over a year, popular data stays that way for reasonable long periods of time.
Dynamic Data Placement

- ATLAS uses the central queue and popularity to understand how heavily used a dataset is
  - Additional copies of the data can be made at sites
  - Jobs re-brokered to use them
- Unused copies are cleaned
- Reduction in the amount of disk needed
Wide Area Access

- With optimized IO other methods of managing the data and the storage are available
  - Sending data directly to applications over the WAN
- Not immediately obvious that this increases the wide area network transfers
  - If a sample is only accessed once, then transferring it before hand or in real time are the same number of bytes sent
  - If we only read a portion of the file, then it might be fewer bytes
Current Xrootd demonstrator in CMS is intended to support the university computing

- Facility in Nebraska and Bari with data served from a variety of locations
- CERN xrootd server to export large EOS pools
- Tier-3 receiving data runs essentially diskless

Similar installation being prepared in ATLAS
Performance

- This Tier-3 has a 10Gb/s network
- CPU Efficiency competitive

Data Served
- Average 1.5TB/day, Max 8TB/day
- Won’t win records, but shows it’s not a joke.

Omaha Analysis
- CPU efficiency about 60% in Omaha
- Best USCMS Tier 2 efficiency about 80%
Failover to sites

- We’ve introduced the concept of that when we see a long queue at a site we can divert additional requests to a site that doesn’t host the data

- Triggered by operators and only triggered to sites with reasonable network connectivity to the hosting site

- Job failover to xrootd

- Efficiency loss is measured in 10s of percent and not factors
Networking

→ ALICE Distributes Data in this way

- Rate from the ALICE Xrootd servers is comparable in peaks to other LHC experiments
Content Delivery

Why is our problem harder than Netflix?

- Netflix delivers streaming video content to about 20M subscribers
- Routinely quoted as the single largest user of bandwidth in the US
By the numbers

<table>
<thead>
<tr>
<th></th>
<th>Netflix</th>
<th>HEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth per client</td>
<td>1.5Mbit</td>
<td>1MB</td>
</tr>
<tr>
<td>Clients</td>
<td>1M*</td>
<td>100k cores</td>
</tr>
<tr>
<td>Serving</td>
<td>1.5Tbits</td>
<td>0.8Tbits</td>
</tr>
<tr>
<td>Total Data Distributed</td>
<td>12TB</td>
<td>20PB</td>
</tr>
</tbody>
</table>

We have a smaller number of clients, less distribution, and higher bandwidth per client.

They have much less data.

Similar Problems Not all files are equally accessed

Forward Physics
Challenge of HEP

- High Energy Physics has a lot of data in a highly distributed environment
  - Hard to make many multiple static copies
  - Need to be able to make dynamic replicas and clean up
  - Need to access data over long distances

- Trying to make networking more predictable
  - Enter LHCOOne
LHCONE in a Nutshell

LHCONE was born (out the 2010 transatlantic workshop at CERN) to address two main issues:

- To ensure that the services to the science community maintain their quality and reliability
- To protect existing R&E infrastructures against overuse by our traffic

LHCONE is expected to

- Provide some guarantees of performance
  - Large data flows across managed bandwidth that would provide better determinism than shared IP networks
  - Segregation from competing traffic flows
  - Use all available resources, especially transatlantic
  - Provide Traffic Engineering and flow management capability
- Leverage investments being made in advanced networking
Start with virtual routing and forwarding (VRF)
- To bring up a system with more predictable networking
- Improve the monitoring

Migrate toward more advanced technologies
- Dynamic circuits
- Scheduled bandwidth
Changes

- New model has less structure
  - More options for where data comes from
  - More flexibility in where activities happen

- This lack of structure places more expectations on the support services like networking and data management
  - Developing more advances service for management and transport
  - Network has been very reliable. Programs like LHCOOne try to maintain this
Outlook

- More flexible and dynamic use of resources available will make more efficient use of the resources
- All the actions trying to ensure that we don’t make artificial separation
- Should put us in a better situation to make use of Computing Services we don’t control
  - Clouds and Opportunistic Computing