Data Transfers Between LHC Grid Sites

Dorian Kcira
dkcira@caltech.edu

Caltech High Energy Physics Group
hep.caltech.edu/cms
CERN Site: LHC and the Experiments

Large Hadron Collider
27 km circumference

CMS
ATLAS
LHCb
ALICE

Lake Geneva
Compact Muon Solenoid

Magnetic length 12.5 m
Free bore diameter 6 m
Central B Field 4 Tesla
Temperature 4.2° K
Nominal current 20 kA
Radial Pressure 64 Atm.

**Stored energy** 2.7 GJ

CMS: A Nimitz Class
117,000 Ton Carrier
at 20 mph
Compact Muon Solenoid

• CMS: several tens of petabytes of data
• >70’000 cores spread globally in the LHC Grid
  ▪ all LHC experiments: >250K cores
• Computing resources: 24/7/365 basis
• High efficiency (estimated 85%)
• Some fluctuations due to data taking variations of the experiment and analysis variations due to interesting results and/or conferences
• Two main types of main workflows:
  1. Monte Carlo simulation of physics events: high CPU, small I/O
  2. Reconstruction meaningful physical quantities + Analysis: high CPU + high I/O
The LHC Data Grid Hierarchy: Developed at Caltech (1999)

A Global Dynamic System. Petabytes to Exabytes per Year
A New Generation of Networks

11 Tier1 and 300 Tier2 and Tier3 Centers

Tier 0 +1

CERN Center
PBs of Disk; Tape Robots

FNAL Center

Tier 1
10 – 40 to 100 Gbps

IN2P3 Center
RAL Center
INFN Center

Tier 2

Tier2 Center

Tier 3
10 to N X 10 Gbps

Physics data cache
Workstations

Tier 4
1 to 10 Gbps

Synergy with US LHCNet TA Network
State of the Art Data Transfer Tools
Location Independent Access: Blurring the Boundaries Among Sites + Analysis vs Computing

- Once the archival functions are separated from the Tier-1 sites, the Tier2-Tier1 functional difference blurs. Similar for the analysis/computing-ops.
- Connections and functions of sites are defined by their capability, including the network.

AAA Scale tests ongoing: 20% of data across WAN: 200k jobs, 60k files, (100TB)/day
Any Data, Any Time, Any Where

• AAA aims at removing data locality and increase number of sites where jobs can run
  ➢ US CMS effort based at Nebraska/UCSD/Wisconsin

• The solution is a federated data storage
  ➢ collection of disparate storage resources transparently accessible across a wide area network via a common namespace

• Underlying technology is Xrootd
  ➢ Uniform interface in front of heterogeneous storage systems
  ➢ Applications query a redirector that points them to the sites
  ➢ Fallback mode very useful, access missing data at other sites
  ➢ Access is authenticated

• Possible because of reliable access of data over WAN and because of the globally consistent namespace of CMS
Any Data, Any Time, Any Where

Xrootd Local Region Redirection

Xrootd Fallback Access

Local Site

Site Storage

Q: Open file /store/foo?
A: File Not Found

CMSSW

Q: Open file /store/foo?
A: Contact Site A

Site A

Q: Open file /store/foo?
A: Success! Send Data...

Site Storage

Global Xrootd Service
Any Data, Any Time, Any Where

Aggregated Xrootd traffic over the last 3 months
Average: 1.757 GB/s => 148 TB / day over WAN
Any Data, Any Time, Any Where

- Xrootd at Caltech in September
- Peaks at 6.8 Gbps
- Non-negligible traffic
- Largest part from FNAL but contributions also from other T2 sites.
The Caltech Tier 2

- Very first LHC Tier 2 site
  - On Caltech Campus at 3 different locations
  - 3.1 PB Hadoop storage
  - 5200 job slots
  - 300 physical servers
  - Upgraded 100G link to Internet 2

- Lead key areas of the LHC computing and software aimed at enabling grid-based data analysis, as well as global-scale networking and collaborative systems
Caltech HEP Tier2

HEP Cluster in CACR

Dell E600

N.. 10GE

40GE

Dell s4810

HEP Cluster in IPAC

HEP Cluster in Lauritsen

CHOPIN 100G Backbone

100GE

20GE

40GE

1 x 10GE

Cisco 7606

100GE

20GE

20GE

40GE

WAN = 100GE Link from CHOPIN Project (CC-NIE)

WAN = 1 x 10GE Link (Production Network)
CMS Data Transfer Volume

CMS PhEDEx - Cumulative Transfer Volume
52 Weeks from Week 06 of 2012 to Week 05 of 2013

42 PetaBytes Transferred Over 12 Months
= 10.6 Gbps Avg. (>20 Gbps Peak)

2012 Versus 2011: +45%

Feb. 2012 to Jan 2013
Data to and from the Caltech Tier 2

Outgoing from Caltech
370 TB

- Transfers during 52 weeks
- Relatively calm period due to LHC shutdown

Incoming to Caltech
240 TB

- Outgoing data of the same order as incoming. This is due to
  - MC generation happening at Tier 2s
  - Data transfers not happening anymore centrally: “full mesh” model
- More transfers through, federated data system, AAA: any data anytime anywhere.
  - Fallback or overflow, especially at regional level
CMS Data Transfer Tools

- **PhEDEx**: CMS Data Transfer Management System
- **FTS**: Grid Data Transfer Service
- **SRM**: Storage Resource Manager. Grid storage services providing interfaces to storage resources
- **GridFTP**: high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks
- **FDT**: java application for efficient data transfers, capable of reading and writing at disk speed over wide area networks
Steering CMS Data Transfers

PhEDEx -> FTS -> CMS Central Services -> Site, e.g. T2_US_Caltech

PhEDEx client -> GridFTP

SRM -> GridFTP

Data transfers to other sites

Site Storage
20 Gbps and Beyond

• Links of USCMS sites were upgraded to 100 Gbps
• USCMS asked all sites to commission the links and demonstrate the ability to perform transfers of at least 20 Gbps in production
  ➢ hence the name “20 Gpbs Challenge”
  ➢ there are 8 US CMS Tier 2 sites and the Tier 1 site at Fermilab
• Caltech was the first site to pass the challenge and took a leading role in this effort.
• Results presented here are based on the work of Samir Cury of Caltech
20 Gbps and Beyond Transfer Commissioning

• As expected, 20 Gbps transfers did not work out of the box.
• Most problems related to configuration at different levels
  • PhEDEx
    ➢ At Caltech we had to change the number of transfers to the different sites in the client configuration
    ➢ These configuration changes were propagated to other sites
  • FTS
    ➢ FTS 3 migration on June 2014. Transition transparent until we started with the high rate transfers
    ➢ FTS optimizer changes the number of active transfers (GridFTP processes) depending on transfer rates (or failures)
    ➢ Work went into tuning the FTS Optimizer
## FTS Optimizer

<table>
<thead>
<tr>
<th>Time</th>
<th>Mode</th>
<th>Efficiency</th>
<th>Throughput</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-07-26T22:45:29</td>
<td>46</td>
<td>100.00%</td>
<td>10.772 MB/s</td>
<td>Good link efficiency, throughput steady, retrials steady or decreasing</td>
</tr>
<tr>
<td>2014-07-26T22:44:28</td>
<td>48</td>
<td>100.00%</td>
<td>9.813 MB/s</td>
<td>Good link efficiency, throughput deterioration as caused by congestion due to too many transfer</td>
</tr>
<tr>
<td>2014-07-26T22:43:28</td>
<td>49</td>
<td>100.00%</td>
<td>13.001 MB/s</td>
<td>Good link efficiency, throughput steady, retrials steady or decreasing</td>
</tr>
<tr>
<td>2014-07-26T22:42:28</td>
<td>49</td>
<td>100.00%</td>
<td>14.289 MB/s</td>
<td>Good link efficiency, current average throughput is larger than the preceding average throughput and retrials steady or decreasing</td>
</tr>
<tr>
<td>2014-07-26T22:41:28</td>
<td>48</td>
<td>100.00%</td>
<td>13.174 MB/s</td>
<td>Good link efficiency, current average throughput is larger than the preceding average throughput and retrials steady or decreasing</td>
</tr>
<tr>
<td>2014-07-26T22:40:28</td>
<td>47</td>
<td>100.00%</td>
<td>10.474 MB/s</td>
<td>Good link efficiency, throughput deterioration as caused by congestion due to too many transfer</td>
</tr>
<tr>
<td>2014-07-26T22:39:27</td>
<td>48</td>
<td>100.00%</td>
<td>11.978 MB/s</td>
<td>Good link efficiency, throughput steady, retrials steady or decreasing</td>
</tr>
</tbody>
</table>
## FTS Optimizer

<table>
<thead>
<tr>
<th>Source</th>
<th>Destination</th>
<th>N Streams</th>
<th>Throughput</th>
<th>Tested</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>srm://cmssrm.fnal.gov</td>
<td>srm://storage01.lcg.cscs.ch</td>
<td>7</td>
<td>20.76 MB/s</td>
<td>1</td>
<td>2014-10-03T12:32:10</td>
</tr>
<tr>
<td>srm://cmssrm.fnal.gov</td>
<td>srm://storage01.lcg.cscs.ch</td>
<td>16</td>
<td>17.88 MB/s</td>
<td>1</td>
<td>2014-10-02T05:54:07</td>
</tr>
<tr>
<td>srm://cmssrm.fnal.gov</td>
<td>srm://storage01.lcg.cscs.ch</td>
<td>5</td>
<td>17.77 MB/s</td>
<td>1</td>
<td>2014-10-02T18:10:37</td>
</tr>
</tbody>
</table>
• FTS developers agreed that we can trade-off better rates for imperfect transfer efficiencies (failure rates)
• Performed tests bypassing the optimizer
There are three different auto-tuning modes for the FTS Optimizer:

- **Conservative**
  - Try to maintain 99%-100% transfer efficiency
  - No TCP streams optimization, finite range per file size

- **Moderate**
  - Allow transfer efficiency to go down to 97%
  - Probe no less than 15 min for 1-16 TCP streams
  - After converging, stick to the best settings for ½ day

- **Aggressive**
  - Allow transfer efficiency to go down to 95%
  - Probe for 15 mins time window with 1-16 TCP streams
  - After converging, maintain setting for ½ day
  - Set TCP buffer size to 8MB. N.B. this has not been tested in production.
20 Gbps and Beyond
Transfer Commissioning (PhEDEx)

Transfers from Caltech to other sites
20 Gbps and Beyond
Design Considerations

- 8 x 10 Gpbs GridFTP servers => High cost
- 2 x 40 Gbps GridFTP servers => at least ½ the cost
- Challenges
  - 40 Gbps do not perform as well
  - Evaluating their performance for different setups
  - Max rate so far: 18Gbps / GridFTP server
- Still a lot of tuning options to be tested
20 Gbps and Beyond

Next Steps

• USCMS Sites that passed the challenge already: Caltech, Purdue
• USCMS Sites very close to passing: Florida, Nebraska
  ➢ Coordinated efforts with the other USCMS sites
• Development needed in order for Xrootd to be used transparently through FTS at any site
• Work on optimizing GridFTP setup
Summary & Conclusions

• The LHC computing and data models continue to evolve towards more dynamic, richly structured, on-demand data movement.

• **Large data transfers** (requiring high throughput) are complemented by **remote data access** (latency sensitive).

• Uplinks of US CMS LHC Grid sites upgraded to 100 Gbps.

• Caltech coordinating effort to commission faster transfers over the new links. Work ongoing into optimization of transfer tools and development of new tools.

• Caltech leads projects that are engaging the LHC experiments to implement increasing network-awareness and interaction in their data management and workflow software stacks.

➢ See also presentation in this track, Wednesday at 9:15
Questions?