CMS Data Transfer Challenges and Experiences with 40G End Hosts

NEXT Technology Exchange
Advanced Networking / Joint Techs
Indianapolis, October 2014

Azher Mughal, Dorian Kcira
California Institute of Technology
Caltech Tier2 Center

- Servers are spread across 3 Data Centers within campus
- Total Nodes including storage servers: 300
- CPU Cores: 5200
- Storage Space:
  - Raw: 3.1 PetaBytes
  - Usable: 1.5-2 PetaBytes
- Hadoop Storage Replication Factor: 50%
- GridFTPs are gateway to the rest of the CMS grid traffic
- Greater use of FDT with CMS PhEDEx and SDN for LHC Run2

Data Center remodeled in late 2013
Data Center Connectivity

- 100G uplink to CENIC (NSF CC-NIE award) with 10GE as a backup.
- 40G Inter building connectivity.
- Vendor neutral, switching hardware from Brocade, Dell and Cisco.
- Active Ports:
  - 8 x 40GE
  - ~40 x 10GE ports
  - ~500 x 1GE ports
- Core switches support OpenFlow 1.0 (OF 1.3 by 4th Qtr 2014).
- DYNES connected storage node.
- NSI ready using OSCARS NSI Bridge.

azher@hep.caltech.edu, dkcira@caltech.edu
100G LHCONE Connectivity, IP Peerings

- CC-NIE award helped purchase of 100G equipment. Strong support from CENIC.
- Tier2 connected with Internet2 LHCONE VRF since April 2014.
- In case of link failure, Primary 100G path fails over to the 10G link.
- Direct IP peering with UFL over AL2S and FLR.
- Ongoing point-to-point performance analysis experiments with Nebraska through Internet2 Advanced Layer2 Services (AL2S).
perfSONAR LHCONE Tests

- Participating in US CMS Tier2 and LHCONE perfSONAR mesh tests.
- Separate perfSONAR instances for OWAMP (1G) and BWCTL (10G)
- Uses TCP applications where RTT plays a major role in bandwidth throughput (single stream).
  - [Not a factor in other tests with FDT at high throughput shown in this talk.]

[root@perfsonar ~]# ping perfsonar-de-kit.gridka.de
64 bytes from perfsonar-de-kit.gridka.de (192.108.47.6): icmp_seq=1 ttl=53 time=172 ms

[root@perfsonar ~]# ping ps-bandwidth.lhcmon.triumf.ca
64 bytes from ps-bandwidth.lhcmon.triumf.ca (206.12.9.1): icmp_seq=2 ttl=58 time=29.8 ms
CMS Software Components Primer

- **PhEDEx**
  - Bookkeeping for CMS Data Sets. Knows the End points and manages high level aspects of the transfers (e.g. file router).

- **FTS**
  - Negotiates the transfers among end sites/points and initiates transfers through the GridFTP servers.

- **SRM**
  - Selects the appropriate GridFTP Server (mostly round-robin).

- **GridFTP**
  - Current workhorse or grid middleware for the transfers between end sites. Or, an interface between the storage element and the wide area network.
US CMS mandated a 20Gbps disk-to-disk test rate using PhEDEx load tests. (https://twiki.cern.ch/twiki/bin/view/CMSPublic/USCMSTier2Upgrades)

- The software stack is not ready *out of the box* to achieve higher throughputs among sites. Requires lot of tunings.
- Benchmarking the individual GridFTP servers to understand the limits.
- GridFTP uses 3 different file checksum algorithms for each file transfer. Consumes lot of CPU cycles.
- During the first round of tests, FTS (the older version) used 50 transfers in parallel. This limit was removed in the August release.
10 Gbps gridFTPs behave very well
- up to 88% capacity steady
- Peaking at 96% capacity
- Optimal CPU/Memory consumption

Dual 10Gbps
20G Total = 10G
Local in + 10G WAN out
- Increases transfers observed during the LHCONE ANA integration.

- Manual ramp up was a test because remote PhEDEx sites were not subscribing enough transfers to FTS links (e.g. Caltech - CNAF)
**Capacity vs Application Optimization**

- Tier2 traffic flows, Peaks of 43G over AL2S with 20G to CNAF (Bologna) alone.

![Graph showing network traffic data](image-url)
Testing High Speed Transfers

Logical Layout from CERN (USLHCNet) to Caltech (Pasadena)

- Caltech
  - Storage Server
  - 80Gbps

- Internet2
  - AL2S

- CERN / Amsterdam
  - Maximum traffic allowed (80Gbps)
  - 40Gbps

- CERN
  - Sandy1 (client)
  - 40Gbps

- Caltech
  - Sandy3 (client)
  - 40Gbps

azher@hep.caltech.edu, dkcira@caltech.edu
Testing High Speed Transfers

Servers, OS, Transfer Tools

Storage Server:
- RHEL 7.0, Mellanox OFED (2.2-1.0.1)
- SuperMicro (X9DRX+-F)
- Dual Intel E5-2690-v2 (IVY Bridge)
- 64GB DDR3 RAM
- Intel NVMe SSD drives (Gen3 PCIe) [1.7 Gbytes/sec each in the lab].
- Two Mellanox 40GE VPI NICs

Client Systems:
- SL 6.5, Mellanox OFED
- SuperMicro (X9DR3-F)
- Dual Intel E5-2670 (Sandy Bridge)
- 64GB DDR3 RAM
- One Mellanox 40GE VPI NIC

Transfer software:
- RDMA FTP
  http://ftp100.cewit.stonybrook.edu/rftp/
- Operates using RDMA or TCP
- RDMA Mode: Breaks source files in parallel data streams
- TCP Mode: Single stream for each source file
- Written in C

azher@hep.caltech.edu, dkcira@caltech.edu
Testing High Speed Transfers

Data writing over the SSD Drives, Destination server in Caltech

- After tunings, With RFTP and large RTT, single TCP stream was still limited to 360Mbps. So ~200 streams were used; still achieved 70 Gbps stable throughput.
- AL2S is a shared infrastructure. During these transfers, LA-Phoenix segment showed 97.03Gbps.
- CPU (E5 2670 V2) is a bottleneck when using TCP at this rate and number of flows (network & I/O processes compete).

azher@hep.caltech.edu, dkcira@caltech.edu
Based on the work of Samir Cury (Caltech)
Using the Open Exchange Software Suite APIs for tests on the production network
Faster transfers between LHC sites will require better choice of paths, and sometimes multiple paths across Internet2
  – Avoid overloading paths
  – Make sure transfers do not interfere with each other
Load balancing: Each site has 2-3 paths and can alternate between them, choosing the less utilized one in each instance

Illustration of the sdn agent choosing between paths

```bash
# samir@desktop sdn-agent]$ ./sdn-agent
<tests suppressed>

For path nebraska-southern, max occupancy is 25.4086778631806 Gbps
Threshold of 10 Gbps is exceeded on the default path
For path nebraska-central, max occupancy is 2.98024951174855 Gbps
Path nebraska-central has not exceeded the threshold and is a good option to nebraska-southern
```
Based on the work of Julian Bunn (Caltech)
Goal is to support application level intelligent path selection for high bandwidth international transfers of LHC data
Based on initial work at Caltech with the Floodlight SDN/OpenFlow controller
Working with OpenDaylight “Hydrogen” release
Using a variety of hardware and virtual switchgear, including Brocade, Pica, HP, Padtec, Mininet

Virtual Internet2 testbed used for Multipath/ODL tests
Multipath Plugins for OpenDaylight SDN Controller

CERN ODL Testbed featuring Pica8 and HP switches, Sandy Bridge hosts

<table>
<thead>
<tr>
<th>Node Name</th>
<th>Node ID</th>
<th>Ports</th>
</tr>
</thead>
<tbody>
<tr>
<td>ofs01-gva</td>
<td>0F:5e:3e:e8:9a:8f:d2:80:57</td>
<td>21</td>
</tr>
<tr>
<td>ofs01-gva</td>
<td>0F:00:64:00:21:14:12:c0</td>
<td>27</td>
</tr>
<tr>
<td>ofs02-gva</td>
<td>0F:5e:3e:60:eb:69:fe:49:14</td>
<td>21</td>
</tr>
</tbody>
</table>

azher@hep.caltech.edu, dkcira@caltech.edu
We have developed two plugins for ODL:

- Multipath: implements path calculators (e.g. modified Dijkstra) and a variety of path selectors, for example:
  - Shortest path
  - Random path
  - Round Robin path
  - Maximum Available Bandwidth path
  - Path with fewest flows
  - Path with highest capacity

- Multipath.Northbound: provides REST endpoints for proactive control of routing, as well as monitoring. For example:
  - Select path calculator
  - Set up flow entries in switches between two hosts (according to current path calculator)
  - Query flow traffic, topology of network, link activation
These plugins, as well as multilayer dynamic circuits under software control, will be tested at ~1Tbps (with servers using many 40GE interfaces) during SuperComputing 2014.
Global Software-Defined Dynamic Circuits for Data Intensive Science (PhEDEx – ANSE – PANDA – OpenDayLight)
Conclusions / Moving Forward

- The LHC experiments (and other major science programs) will need to be aware of the impact of their large flows across the R&E networks in the US, Europe and across the Atlantic. As shown, with modern day off the shelf equipment, 100GE paths can be fully loaded relatively easily.

- Caltech achieved 20 Gbps+ in production-mode, using its Tier2 in a first set of trials (CC-NIE 100G, Internet2 ALS, ANA-100); this led to 43G peak on AL2S.

- We are integrating an OpenDaylight multi-path controller (based on a Floodlight controller in the OLiMPS project) with CMS’ mainstream data transfer management system PhEDEx using OSCARS circuits:
  - to create either parallel paths to the same destination (to help avoid backbone congestion) or
  - Choosing paths to destinations by looking at the load on the WAN.

- Benchmarking next generation CPUs and memory, keeping the software stack tuned and knowing its limitations under different set of application requirements.

- SDN Multipath demonstrations over the WAN and on the show floor will be showcased during the Supercomputing Conference 2014 in New Orleans.
Acknowledgments

- **DYNES**: NSF 0958998
- **ANSE**: NSF 1246133
- **OliMPS**: DOE DE-SC0007346
- **Cisco Research**: 2014-128271
- **US LHCNet** – Fermilab subaward # 608666 under DOE # DE-AC02-07CH11359
- **Caltech Tier2**: Princeton subaward # 00002009 under NSF award # 1121038

azher@hep.caltech.edu, dkcira@caltech.edu