

CMS Data Transfer Challenges

LHCOPN-LHCONE meeting
Michigan, *Sept 15/16th, 2014*

Azher Mughal
Caltech

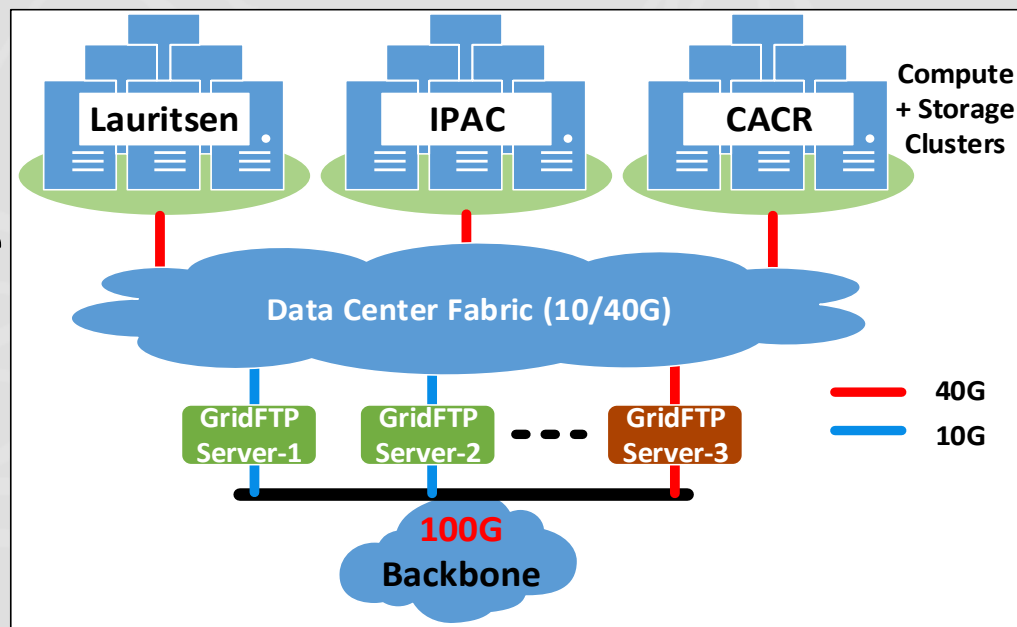
Data Center remodeled in late 2013

- Servers are spread across 3 Data Centers within campus
- Total Nodes including storage servers : 300
- CPU Cores : 4500
- Storage Space :
 - Raw : 3 PetaBytes
 - Useable : 1.5 PetaBytes
- Hadoop Storage Replication Factor : 50%
- GridFTP's are gateway to the rest of the CMS grid traffic



Data Center Connectivity

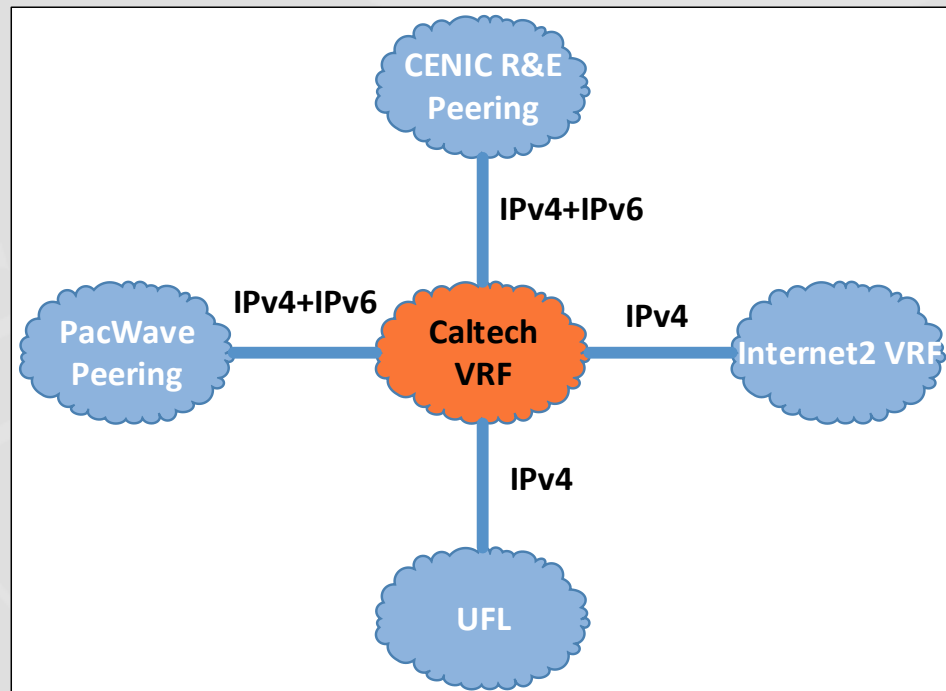
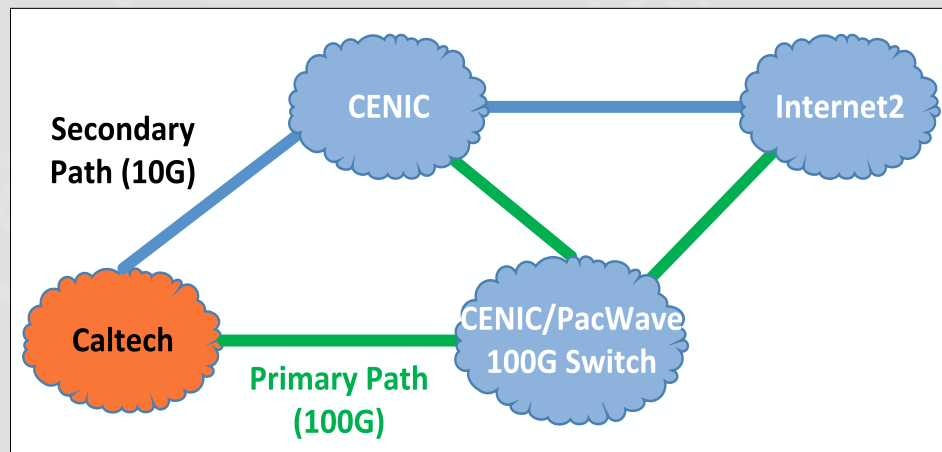
- 100G uplink to CENIC (NSF CC-NIE award) with 10GE as a backup.
- 40G Inter building connectivity.
- Vendor neutral, switching hardware from Brocade, Dell and cisco.
- Active Ports:
 - 8 x 40GE
 - ~40 x 10GE ports
 - ~500 x 1GE ports
- Core switches support OpenFlow 1.0 (OF 1.3 by 4th Qtr 2014).



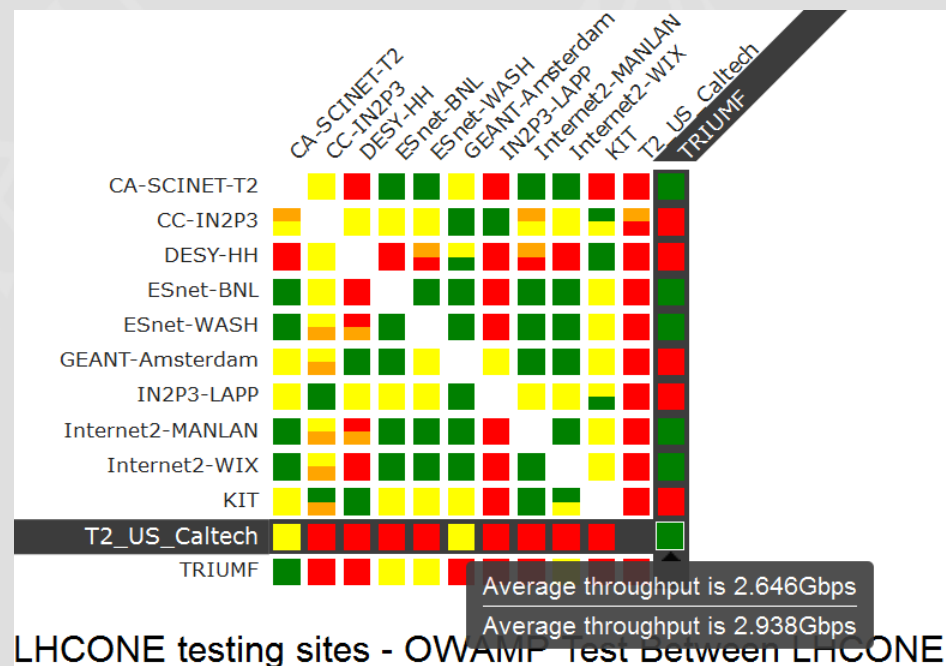
- OSCARS / OESS Controller in production since 2009. Peering with Internet2.
- DYNES connected storage node.
- NSI ready using OSCARS NSI Bridge.

100G LHCONE Connectivity, IP Peerings

- CC-NIE award helped purchase of 100G equipment
- 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 p-t-p performance analysis experiments with Nebraska through Internet2 Advanced Layer2 Services (AL2S)



- Participating in US CMS Tier2 and LHCONE perfSONAR mesh tests.
- Separate perfSONAR instances for OWAMP (1G) and BWCTL (10G)
- RTT plays major factor in bandwidth throughput (single stream).



```
[root@perfsonar ~]# ping perfsonar-de-kit.gridka.de
```

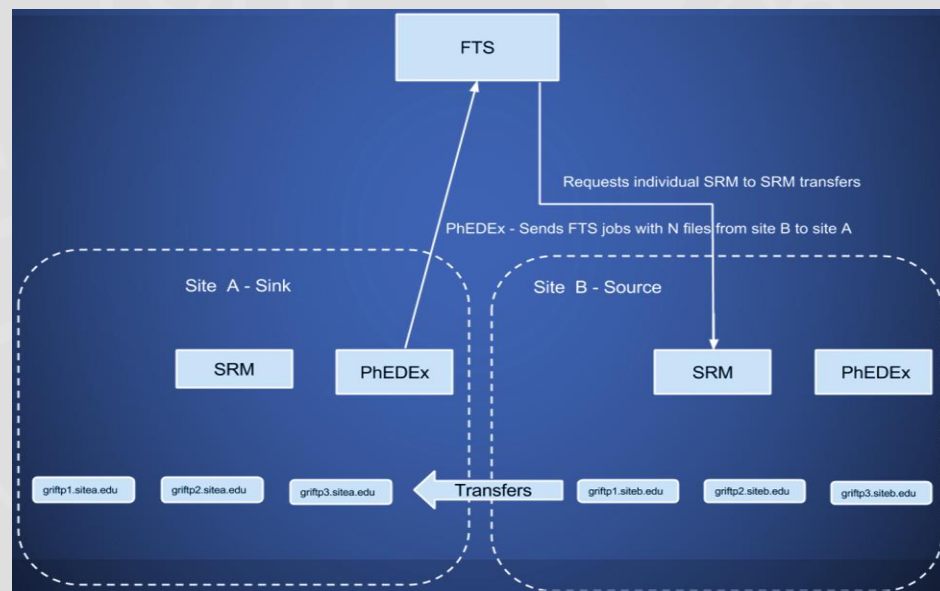
64 bytes from perfsonar-de-kit.gridka.de (192.108.47.6): icmp_seq=1 ttl=53 **time=172 ms**

```
[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
 - Book keeping 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
 - Actual workhorse or grid middleware for the transfers between end sites. Or, an interface between the storage element and the wide area network.



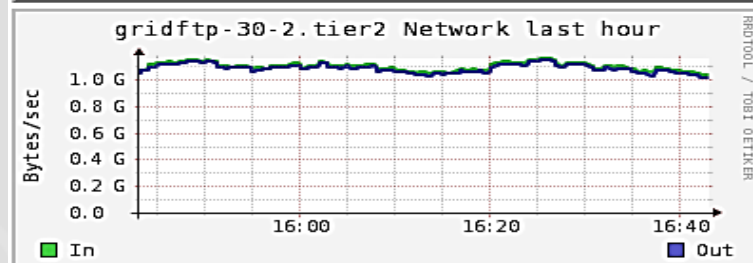
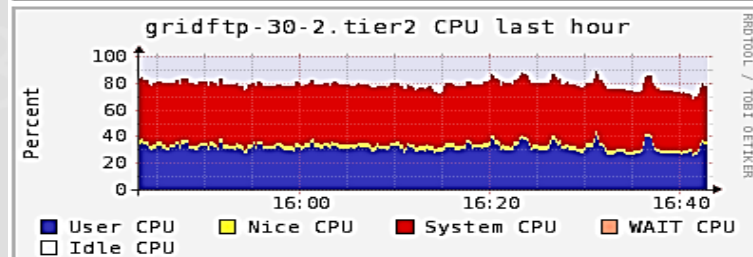
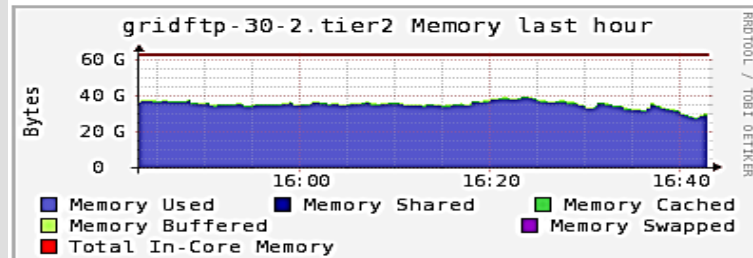
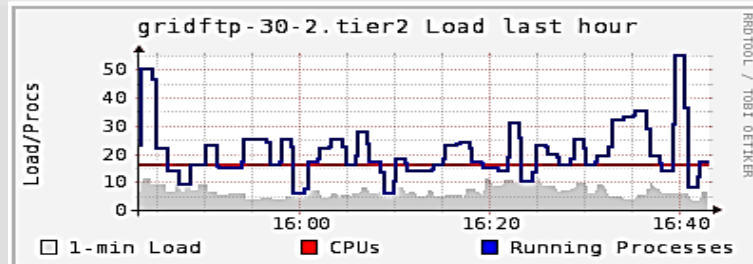
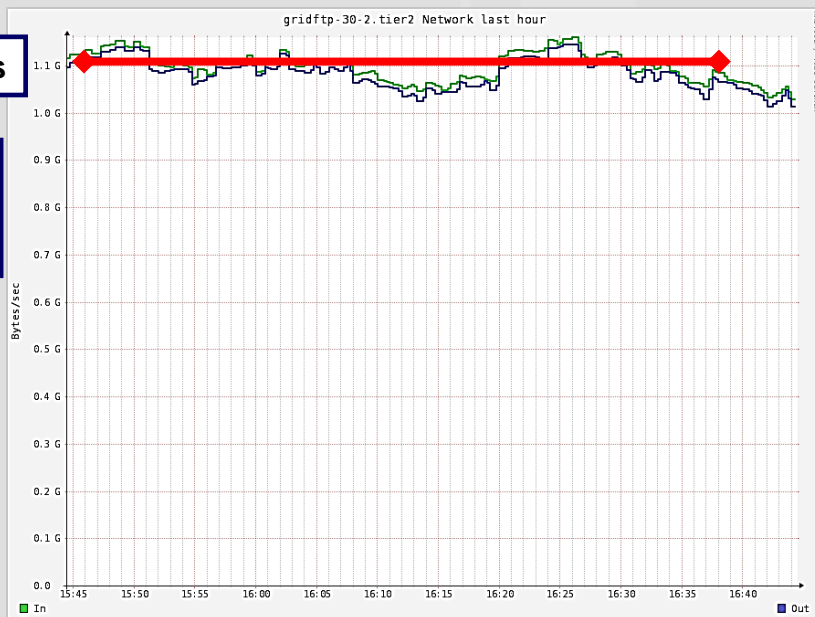
- US CMS mandated a 20Gbps disk-to-disk test rates using PhEDEx load tests.
<https://twiki.cern.ch/twiki/bin/view/CMSPublic/USCMSTier2Upgrades>
- The software stack is not ready as an *out of the box* to achieve higher throughputs among sites. Requires lot of tunings.
- Benchmarking the individual GridFTP 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.

Capacity vs Application Optimization

- 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

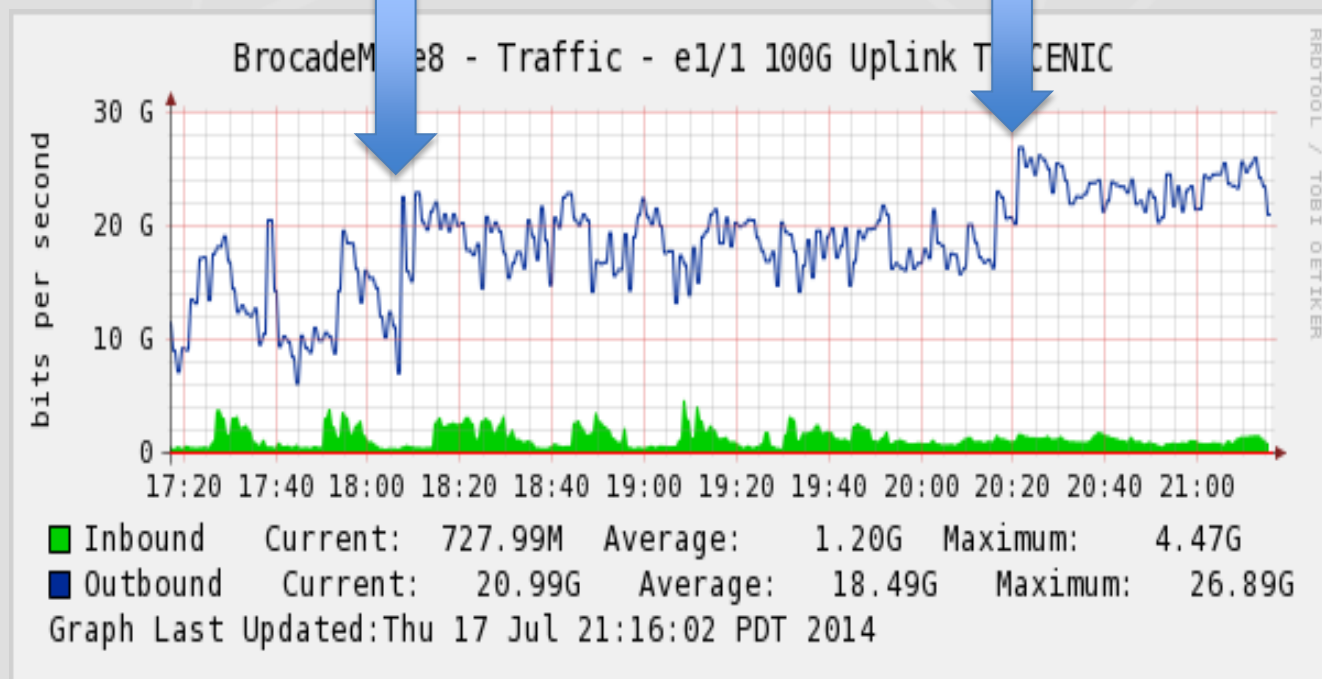


Capacity vs Application Optimization

- Increases transfers observed during the LHCONE ANA integration.
- Manual ramp up was just a test because remote PhEDEx sites were not subscribing enough transfers to FTS links (e.g. Caltech - CNAF)

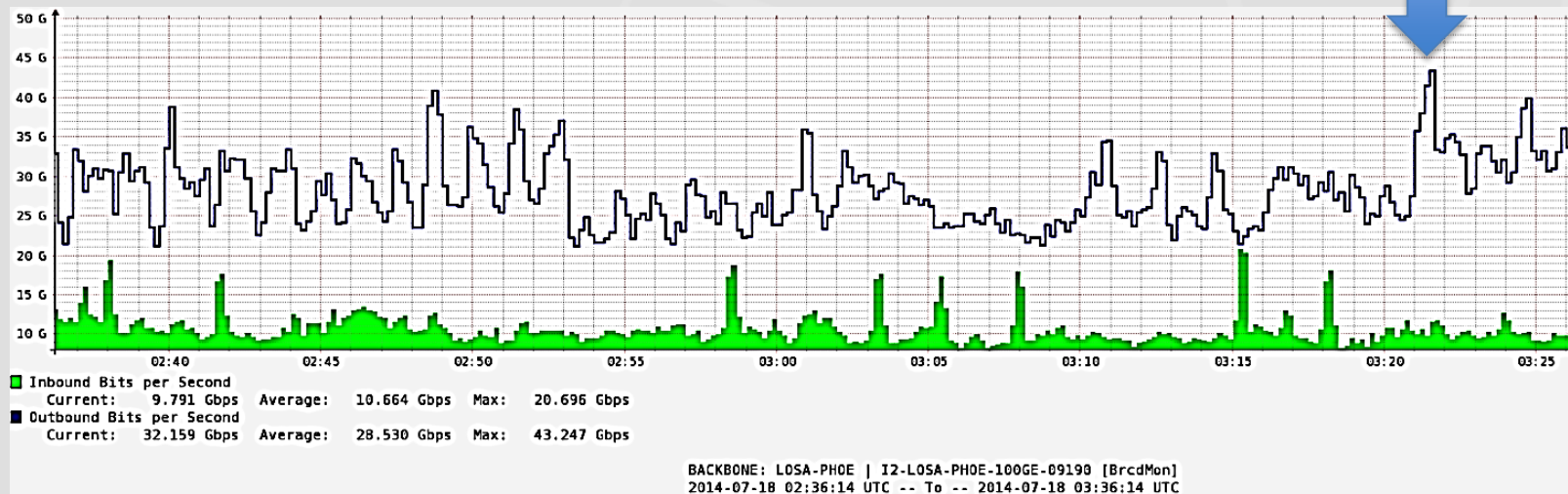
LHCONE -> ANA 100G

Manual Ramp up

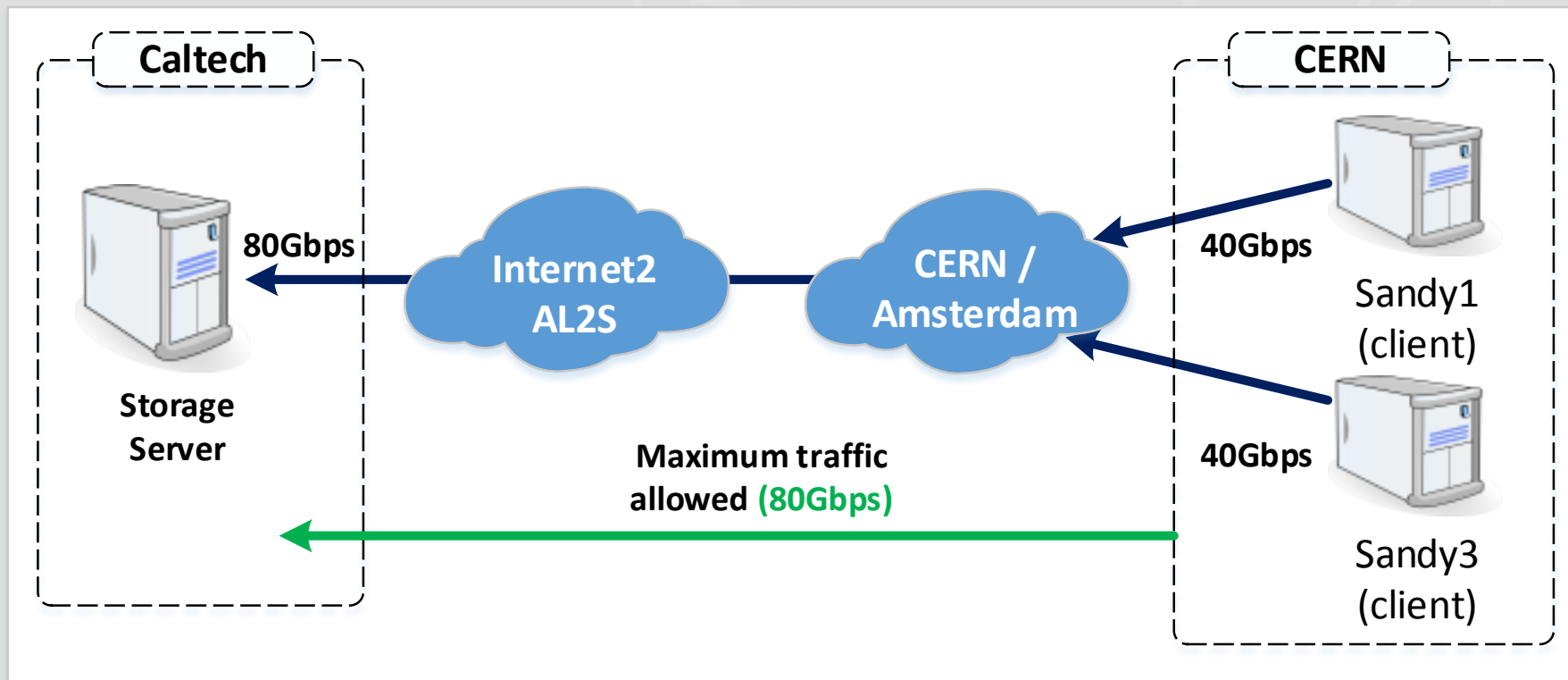


Capacity vs Application Optimization

- Tier2 traffic flows, Peaks of 43G over AL2S with 20G to just CNAF.



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



Server and Operating System Specifications

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)
- ☐ 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

- It is very important for the LHC experiments to be aware of the impact of their large data flows on the R&E networks, both in the US and Europe and across the Atlantic. With modern day off the shelf equipment, 100GE paths can be easily overloaded.
- Caltech is able to achieve the US CMS Tier2 milestones with peaks reaching to 43Gbps (CC-NIE 100G, LHCONE, ANA).
- We are looking forward on how to integrate OLiMPS multi-path controller with Internet2 flow space firewall (p-t-p services):
 - to create either parallel paths to same destination (avoid backbone congestion) or
 - individual paths to destinations by looking at 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 Louisiana.

Time for Questions!

