Efficiently Transferring Petabytes at ~70Gbps
ESCC meeting LBNL,

Chin Fang Zettar, Les Cottrell SLAC, May 5, 2017
Requirements for LCLS-II

- Beam Pulse rate 120HZ => 1 MHz
- Data rate increase by factor 1000 to Tbps by 2024

Plus
- LHC/ATLAS
- LSST
- The rest
Overview

- Transfer files over the 5000-mile OSCARS SLAC link
- **Shared** in the SLAC production 100Gbps border network
  - Need to keep ~ 20Gbps for other production traffic
- 0.1PB in 3.4 hours at ~70Gbps, **1PB in 34 hours**
- Using the following testbed:
  - Two 2 x1U storage tiers with 8 x Intel DC P3700 U.2 1.6TB NVMe SSDs (each 1U server has 4 x NVMe SSDs)
  - Connected by InfiniBand
  - Two 2-1U DTN clusters (one sending, one receiving), running Zettar zx.
  - Each DTN has 4x10G Ethernet ports, i.e. 2 x 4 x 10Gbps = 80Gbps
  - All ports are connected to 2 x Arista 7280SE-68 10/100G switches.
  - One of the Aristas connects to the SLAC Cisco 100GBps border router & thence to ESnet
- Note that due to the testbed hardware configuration, the max speed the testbed can attain is ~ 80Gbps.
The test bed collocated at SLAC

The Test-bed

- **DTNs**
  - 2 x Yahoo! C73E/64/960 1U servers/cluster
  - 4x10G/server

- **Storage tiers**
  - 2 x HPC all-NVMe storage tiers (2 x 1U AIC SB122A-PH 10Bay servers/tier)
    - 20GB/s read/tier
    - 12GB/s write/tier

- **OSCARS**
- **Other cluster or High speed Internet**
Impact on all ESnet traffic

When running data transfer contributes ~ 1/3 of total ESnet traffic
100TiBytes in 3.4 hours Testing

- 10 runs of 10 TiB each = 100TiB took 3.4 hours
- ~ 1PiB in 34 hours
- LCLS-II need to transfer 20PB SLAC => NERSC takes 680 hours with our testbed
Weathermap of ESnet during PB transfer

The 5000-mile SLAC loop

100*10TiByte transfers

4/12/2017 12:12:33
1 PebiByte in < 1.5 days
34hrs 16 mins 51 seconds

4/13/2017 22:29:24
Not just the network

How to Attain High Data Transfer Rates? Two Critical 1st Steps

A. Fully understand what are involved in the data transfer path! *It's not just network!*

B. Learn about your storage performance well using *fs* and realistic test data sets!

Bottle neck today is the IOPS needed for write
Conclusions

- Demonstrated **sustained 70Gbps over long distances**.
- **Today’s challenge is writing** the data to the files (IOPS)
  - Network not a problem **using standard TCP**
- We have been beating the 16 Intel DC P3700 NVMe SSDs since 2015 much harder and longer than most people in the world. But **Intel DC P3700 NVMe SSD performance has been consistent**
- The four AIC SB122A-PH 10Bay NVMe 1U **storage servers have proven to be highly cost-effective choices** as well. Do not need to spend big $$$ on the proprietary all-flash storage systems from NetApp, Dell/EMC, Hitachi etc.
- **InfiniBand just works.** The use of a Mellanox EDR (100Gbps) in each of the AIC SB122A-PH 10Bay NVMe 1U storage server, and a Mellanox FDR (56Gbps) HCA in each of the Yahoo! 1U C73E/64/960 DTN, together with the two Mellanox SB7700 IB switches has proven to be a quite cost-effective and reliable combination.
Future

Thinking about using test data sets with different file size distribution patterns, also even bigger test data sets (> 10TiB each, e.g. 50TiB each would be good)

Upgrade 200Gbps border at SLAC to two *100Gbps

Discuss with Intel about testing different CPU models.

• On the LCLS side, the modern (Broadwell) but low-end Intel E5-2620v4 8 core @ 2.1 Ghz CPU is used on all five LCLS DTNs
• NERSC DTNs use the older (Ivy Bridge) E5-2680 v2 20 cores @ 2.80GHz
• How do the CPU choices on DTNs affect the transfer performance,

Look at impact of LCLS Lustre file system on performance

Then onto NERSC
Summary

What is special:

- **Scalable.** Add more NICs, more DTNs, more storage servers, links as needed/available…
- **Power & space efficient; low cost**
- **HA tolerant to loss of components**
- **Storage tiering friendly**
- **Reference designs**
- **Easy to use software available commercially**

Proposed Future PetaByte Club

A member of the Petabyte Club **MUST** be an organization that is capable of using a **shared production** point-to-point WAN link to attain a **production** data transfer rate >= **150PiB-mile/hour**
Other information, questions

Who needs it

- **LCLS Exascale requirements**, Thayer and Perazzo, Tbit/s 2014
  - [https://confluence.slac.stanford.edu/download/attachments/178521813/ExascaleRequirementsLCLSCaseStudy.docx](https://confluence.slac.stanford.edu/download/attachments/178521813/ExascaleRequirementsLCLSCaseStudy.docx)

- **Focus more on data migration when moving to the cloud**, 

- **Amazon**, ship a PByte **in a week** (168 hours). They manually ship appliances around to get the data from A to B.

Progress

- **186 Gbps Data Transfer Sets New Record, 2011**
  - SC11 Seattle <-> U Victoria, 97Gbps/direction, 2 racks at SC11

- **LCLS SLAC->NERSC 2013, 116TB in 5 days**

- **The Petascale project**, Eli Dart, ESCC Winter 2016
  - Goal Pbyte/week using Cosmology data

- **Moving a Petabyte of data** June 13, 2015, identifies why it is difficult.