ATLAS EXPERIMENT http://atlas.ch

SuperComputing 2011

ATLAS Great Lakes Tier-2 Experience

UM Personnel: Shawn McKee, Roy Hockett, Bob Ball, Ben Meekhof

Acknowledgements: Salvador Hernandez - Merit, Jim Feuerstein - ITS Comm, Mike Bear - ITS

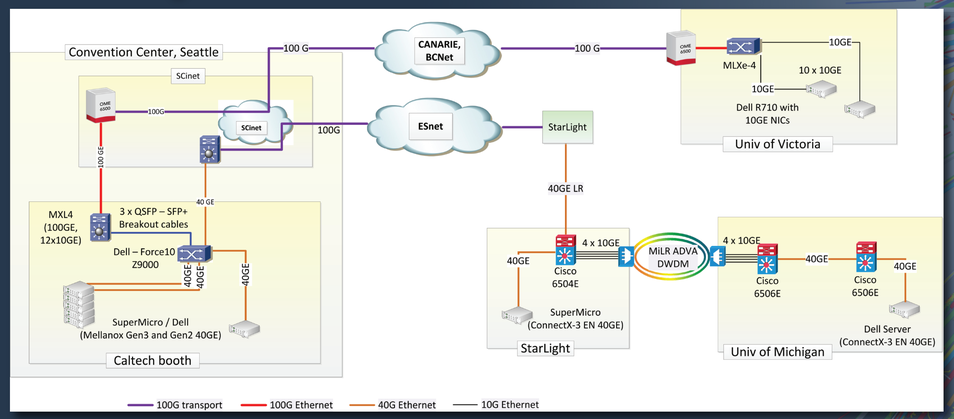
AGLT2

SC11 LHC Distributed Data Challenge

Caltech, the University of Michigan, and the University of Victoria joined forces to demonstrate LHC data movement at a rate of 100Gbps during the SuperComputing 2011 Conference. A peak of 151 Gbps was already demonstrated by Caltech, FNAL, Michigan and SLAC during the SC05 Conference in Seattle using multiple 10GE connections.

For these transfers a single 100GE connection was terminated in the booth using the Brocade 100GE router. A similar router was used in Univ of Victoria. The University of Michigan used its MiLR infrastructure, augmented with FSP-3000 DWDM equipment from ADVA and switches from Cisco to connect to a single Dell Server using the latest Mellanox 40GE NIC. ESnet carried the 40GE connection between Starlight, in Chicago, and SC11 in Seattle where it was merged into the 100GE wave.

Caltech's FDT Data transfer application was used to exchange data from storage to storage and demonstrate extremely high end-to-end bandwidth for sharing large LHC datasets in near realtime. This was a preview of a possible future LHC site configuration relying on a few powerful servers to enable low latency, large dataset sharing and access.



Caltech

MICHIGAN

University of Victoria

Setup in Seattle

The Showfloor for exhibitions is large

There are hundreds of booths from vendors, labs and universities. The LHC demonstrations are typically done from the Caltech booth but often involve others (Internet2, ESnet, FNAL , BNL and others).

Setup at SuperComputing is intense work

We go from boxes on the floor to a powerful, compact Tier-2 equivalent in 2 days. The two racks below contain switches from Brocade, Force10 and Cisco with a total of 2x100G ports, 48 40G ports, 72 10G ports. There are 12 servers with 40G (5 PCIe Gen3, 7 Gen2), 19 servers with 10G. There are a total of 472 processor-cores including 72 Intel "Sandy Bridge". Storage included 48 120G SSDs and the local SATA disk space totals 324 TB. All of this must be configured to connect to remote resources to demonstrate "future" capabilities for LHC data transfer

Azher Mughal
Caltech

Ian Gable
Univ. of Victoria

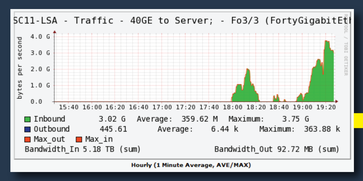
Shawn McKee
Univ. of Michigan

Team-effort is required

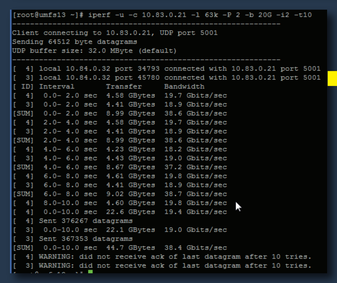
Having a successful SuperComputing (SC) demonstration takes a team effort. Lots of people with different skills in computing, networking, storage, operating systems and protocols need to work closely together to enable "cutting-edge" hardware to inter-operate. Many of the challenges we face are related to using the newest hardware and untested bios and firmware releases. We have a small time-window to come up to speed on what is new and how to best make it work in the context of our demonstrations.

MiLR Use for SC11

Michigan's connection to SC11 was enabled by contributions from ADVA, Cisco, Mellanox and ColorChip who all provided equipment and expertise.

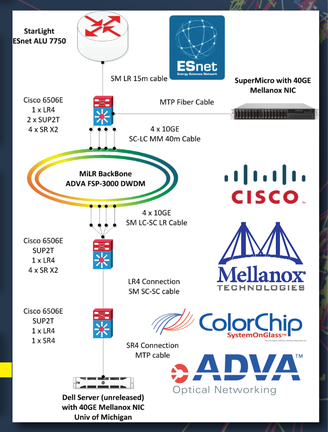


The Cacti graph on the left captures another test between AGLT2 and SC11



On the left is an example of using the 40G links (4x10G internally) to send test data from a single host at 38.4 Gbits/sec.

On the right is the diagram of the network from AGLT2 to Chicago.



Challenges and Lessons Learned

SSDs (we used OCZ Vertex3) have a write speed which assumes you have compressible data.
- We observed 510 MB/sec with highly compressible data.
- With incompressible data we only saw 165 MB/sec.

The Mellanox 40G NICs in a PCIe Gen3 slot using Sandy Bridge processors at 2.2 Ghz are capable of sending 38.4 Gbps by using two Iperf streams split over 2 CPU cores.
- Single process/stream rates were limited to about 31 Gbps (CPU core/interrupt limit)
- Fastest "sending" was by creating very large datagrams (63K) but this puts a large load on the receiver which then "drops" packets due to having to reassemble at high rate.
- Sending datagrams within the MTU size of the end-to-end path allows the receive side to process the most packets without loss. We demonstrated loss-less UDP transmission sending from 4 Iperf sessions, single stream at 8970 bytes achieving 4.7 Gbps each UM - SC11

Networking speeds beyond 10G don't yet allow full speed for **single flows.**
- Individual flows through the infrastructure are limited to 10G each (4x10G)
- The 100G setup we used also didn't allow single flows of more than 10G
- Multiple hashing algorithms on the path and the low # of flows made per-10GE-channel routing the most efficient choice.
- It is important to understand the details of the networking architecture if you want to take fullest advantage of it. We recommend waiting on future technology.

Results: 100G Mem-to-Mem

The results below are from 10 Servers at UVic exchanging data with 7 servers at SC11. Note that traffic flowed **into SC11 at ~97 Gbps** while simultaneously flowing **out around 87 Gbps.**

Servers were Dell R710s at UVic and a mix of Dell R510s, Supermicro Servers and 2 pre-release Dell Servers on the SC11 sides