SC18 Network Research Exhibition: Demonstration Preliminary Abstract

Wide Area Workflows at 400 Gbps

Stephen Simms, Indiana University ssimms@iu.edu

Abstract

The increasing rate of data production from digital instruments and simulations makes it harder and harder to replicate that data due to the limitations of local resources. To address situations like these, as part of the NSF funded Data Capacitor project, Indiana University (IU) worked with Oak Ridge National Laboratory (ORNL) in 2006 to examine the feasibility of using the Lustre file system across 10 Gbps networks to compute in place. The success of these efforts led to the deployment of a Lustre WAN file system in 2009 that allowed researchers to compute against their data across distance, and was made available to the NSF TeraGrid project. These efforts continue to have relevance today; at IU a physics laboratory located across the Bloomington campus from the data center mounting a Lustre file system on their cluster to compute against data that's local to the university's larger supercomputing resources. Researchers at the Indianapolis campus run Docker on their local workstations to compute against that same file system. The Pittsburgh Supercomputing Center (PSC) mounts the Lustre file system on their Bridges supercomputer to support the efforts of the National Center for Genome Analysis Support in order to shift computational workload from IU's resources to Bridges without moving data.

ORNL has done the most recent work in this area, presenting data at the Lustre User Group 2018 conference showing how it could be possible to increase Lustre performance across a WAN through the use of LNET routers. Using LNET routers to aggregate client traffic on both sides of a WAN connection could greatly simplify the tuning required to achieve file system performance.

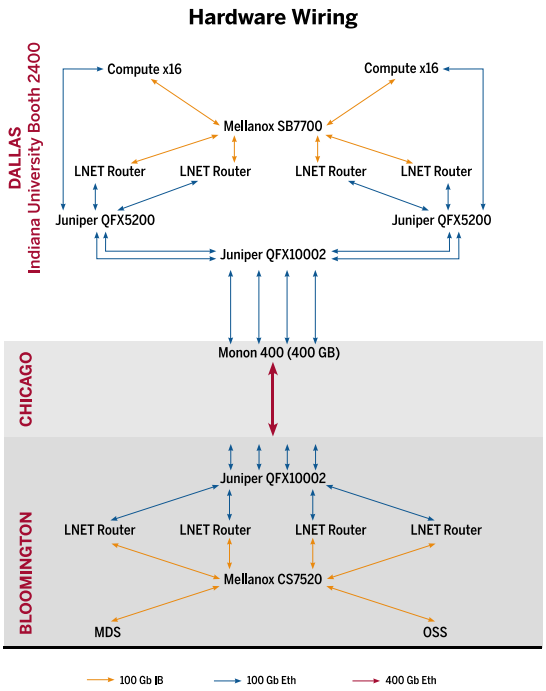
IU in partnership with ORNL will use a modest compute resource on the SC18 exhibit floor in Dallas, TX which will connect to a Lustre file system in Bloomington, IN via four 100 Gbps links to Chicago, IL and a single 400 Gbps channel connecting Chicago, IL to Bloomington, IN. In addition to benchmarking file system performance over distance using TCP and RoCE v2, we will be showcasing applications from IU physicist, Matthew Shepherd and IU computer scientist, David Crandall that could potentially benefit from computing in place across distance.

Goals

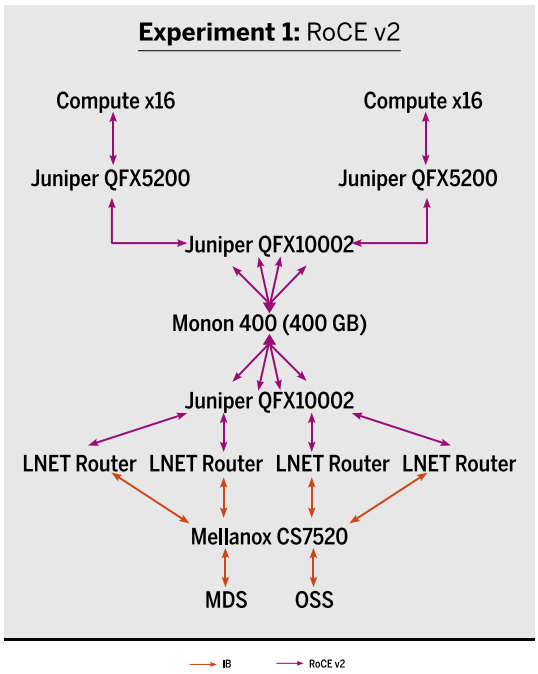
1. Using Ciena's Waveserver Ai, in conjunction with their Blue Planet orchestration platform, demonstrate and exercise a single 400 Gbps channel connecting Bloomington and Chicago from the exhibit floor in Dallas.
2. Demonstrate use of the Lustre file system across the wide area from 100 Gbps to 400 Gbps using an end to end link from Dallas to Bloomington.
3. Compare the performance of Lustre across the wide area using TCP and RoCE v2 using an end to end link from Dallas to Bloomington
4. Demonstrate applications that could potentially benefit from computation in place.

Resources

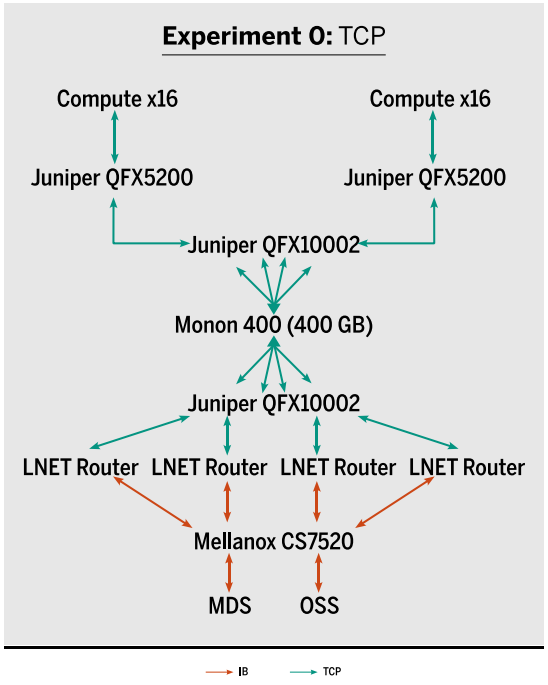
The following diagram shows how we will connect 32 HPE AMD EPYC compute nodes and 4 HPE AMD EPYC LNET routers to a Mellanox 36 port IB switch, on the exhibit floor in Dallas which will be wired into two separate Juniper QFX5200 switches, a Mellanox 36-port IB switch, and a Juniper QFX10002 switch. From the QFX10002, data will ride 4 x 100Gbps connections to Chicago. ESnet will provide one of the four connections which will be dedicated to this demonstration, one will be shared with the Naval Research Laboratory, and the other two will be scheduled through SCinet. Once traffic reaches Chicago, it will travel to Bloomington, IN via a single 400Gbps channel, the first non-commercial deployment of this technology. In the Bloomington data center, the single channel will be broken into 4 x 100Gbps connections which will connect to a Juniper QFX10002. Four LNET routers will convert the traffic to IB and push it to IU's new Lustre/ZFS file system via a Mellanox director switch.



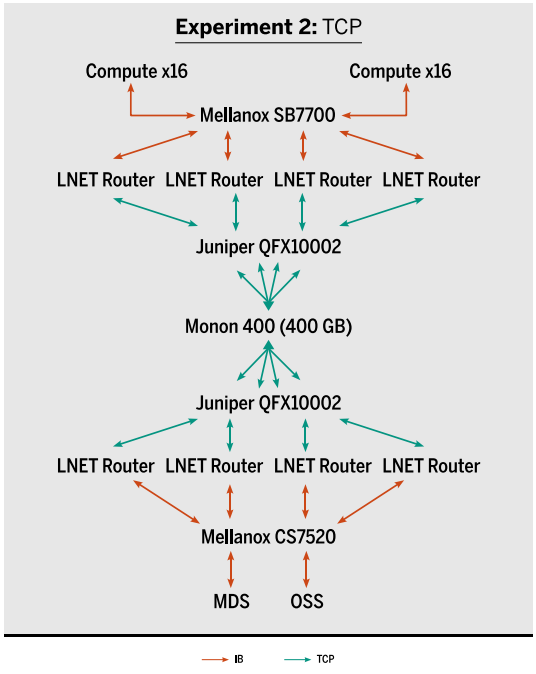
Experiment 0 will be an examination of the performance of Lustre clients across the WAN using TCP.



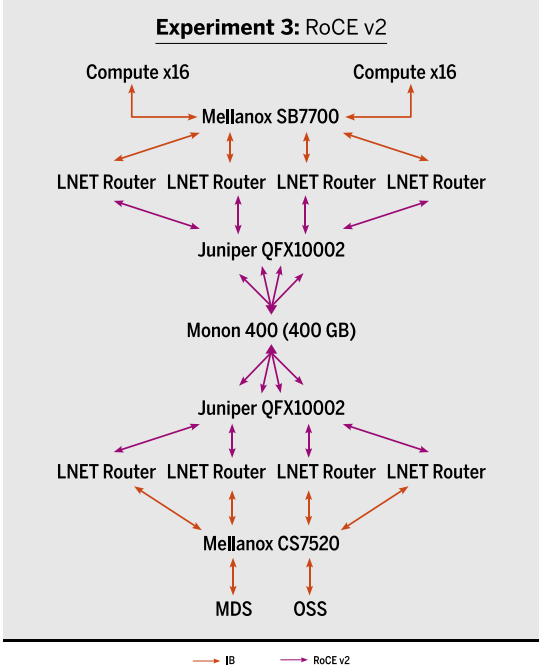
Experiment 2 will examine the performance of clients using LNET routers on each side of the WAN with TCP connections between them.



Experiment 1 will examine the performance of Lustre clients across the WAN using RoCE v2.



Experiment 3 will repeat experiment 2, changing the protocol from TCP to RoCE v2.



Involved Parties

- IU Research Technologies
  - Stephen Simms, ssimms@iu.edu
  - Matt Link, mrlink@iu.edu
  - David Hancock, dyhancoc@iu.edu
  - Tom Crowe, thcrowe@iu.edu
  - Chris Hanna, hannac@iu.edu
  - Nathan Heald, nheald@iu.edu
  - Nathan Lavender, nblavend@iu.edu
  - Ken Rawlings, kerawlin@iu.edu
  - Shawn Slavin, slavin@iu.edu
  - Peg Lindenlaub, plinden@iu.edu
  - Bret Hammond, bret@iu.edu
  - Robert Henschel, henschel@iu.edu
  - Laura Huber, lamhuber@iu.edu
  - Scott Michael, scamicha@iu.edu
  - Eric Wernert, ewernert@iu.edu
  - Jeff Rogers, jlrogers@iu.edu
  - Therese Miller, millertm@iu.edu
- IU Networking
  - Dave Jent, djent@iu.edu
  - Marianne Chitwood, chitwood@iu.edu
  - Caroline Weilhamer, cweilham@iu.edu
  - Tom Johnson, wtjohnso@iu.edu,
  - Shannon Dockter, sdockter@iu.edu
- IU Computer Science
  - David Crandall, djcran@iu.edu
  - Mingze Xu, mx6@iu.edu
- IU Physics
  - Matthew Shepherd, mashephe@indiana.edu
  - Jonathan Zarling, jzarling@iu.edu
- Oak Ridge National Laboratory
  - Neena Imam, imamn@ornl.gov
  - Sarp Oral, oralhs@ornl.gov
  - Nageswara S. Rao, raons@ornl.gov
- ESnet
  - Inder Monga, imonga@es.net
- Naval Research Laboratory
  - Linden Mercer, linden@cmf.nrl.navy.mil
- Starlight
- Ciena
  - Marc Lyonnais, mlyonnai@ciena.com
  - Rod Wilson, rwilson@ciena.com
- HPE
  - James Kovach, james.kovach@hpe.com
  - Andre Gardinalli, andre.gardinalli@hpe.com
- Juniper Networks
  - Shane Praay, spraay@juniper.net
- Mellanox
  - Gilad Shainer, shainer@mellanox.com
- NVIDIAmailto:mthomas@nvidia.com
  - Ronak Shah, ronaks@nvidia.com
  - Michael Thomas, mthomas@nvidia.com
- PIER Group
  - Chad Williams, cwilliams@piergroup.com
  - Dan Fennell, dfennell@piergroup.com
  - Michelle Keller, mkeller@piergroup.com
- Whamcloud
  - Peter Jones, pjones@whamcloud.com