University of Southern California – NRE Submission for SuperComputing 2018

Azher Mughal, Celeste Anderson, David Galassi, Maureen C. Dougherty
University of Southern California
{amughal, celestea, dgalassi, mdougher }@usc.edu

Abstract

The decentralization of the Internet and increased collaboration in large-scale science projects has increased the demand for higher throughput to move their datasets. Traditional applications which rely only on the TCP protocol suffer from CPU saturation at high bit rates. With the help of SCinet, National and regional network providers, AutoGOLE path provisioning and Pacific Research Platform, we intend to demonstrate large-scale disk-to-disk data transfers using the data transfer nodes (DTN). We will be optimizing and demonstrating highly optimized, first ever-built data transfers nodes to take benefit of RDMA over an IP network at 100G and 200Gbps.

Overview

TCP/IP protocol was developed in the early 1970s, and still it is very dominant in the majority of the data transfer applications used on the DTNs nodes. TCP protocol relies on acknowledgements for each or selective packets based on how the protocol stack in the operating system has been tuned. Networks are getting an increased number of fat and clean pipes, routers and switches providing non-blocking packet switching. Thus, we feel there is an opportunity to start exploring alternate protocols to speed up the data transfer performance.

Non Volatile Memory Express over Fabric (NVMeoF) [1] was ratified in 2016 and was designed to map NVME devices including storage over the network. Since then NVMeoF has seen rapid innovation and adoption in the industry and currently being trialed in the DTN nodes.

Innovation

Current DTN designs are mostly available at the network speed of 10/40/100Gbps. We are working on a solution to use 200Gbps NICs in the server. These NICs connects to the Ethernet switch still using QSFP28; however where each port operates at 100G line rate. These NICs are capable of connecting at 200Gbps using 50G x 4 channels. However, switches with 200Gbps QSFP-DD are still in early development stage. These NICs support RDMA over converged Ethernet, thus supporting NVMeoF offloading for transfer acceleration. We will be demonstrating the use of these NICs to send and receive large distributed datasets stored on the attached NVME storage drives.

HPC and Science Relevance

The lessons learned from these exercises and demonstrations will be widely applicable to any science program trying to achieve transfers of large data flows. LSST, LHC, genomics, medical science and engineering are such an example of science programs.

CPU offloading while transferring data at 200Gbps from a single DTN, when coupled with low latency application interaction with the underlying hardware, provides huge cost savings in terms of freeing CPU time. CPU now can be used for simulations, machine learning and other deep analysis processes.

SCinet and R&E Requirements

As illustrated in the network layout in Figure 1, we require optical links from USC to Caltech and to SCinet. USC also requires VLANs between the Booths shown in the diagram.

We have collaborated with Arista Networks to provide early generation of 400GE switches. These switches will be placed in different booths according to the network layout as seen in Figure 1. We will be providing one switch to SCinet so that it can provide 400GE connections to different booths which are part of the XNET network. This switch will also act as the primary uplink for the USC booth.

400GE link over the OSFP optics is still under continuous development process and is very sensitive to reflections at different patch panels, which may result in loss of link or degraded performance due to line errors. We ask SCinet to provide spliced fiber link end to end for all the 400GE connections.

We need dedicated VLANs between USC and other experiments as indicated below:

Description/Route	VLAN
1. USC Booth – Caltech Booth (via 400GE SCinet switch)	2
2. USC Booth – FIU (via LSST2 Link)	2
3. USC Booth – Clemson Booth (could be L3)	
4. USC Booth – AUTOGOLE	4

Virtual Host for AutoGOLE

1. One Linux VM with proper FQDN name, reverse DNS entry and a valid SSL certificate. This certificate will be added in the iCAIR and NetherLight NSI aggregators.
2. Access from VM to the 400GE switch, so that NSA Agent on the VM can provision VLANs on this switch dynamically. AutoGOLE VLANs needs to be provisioned to StarLight (to reach Kisti, Europe) and PacWave (Australia).

Network Topology

We will work closely with the SCinet, Caltech, Clemson, Starlight and PRP teams to ensure suitable access between external PRP end points, AutoGOLE locations and participants on the exhibition floor. Network topology is shown in the Figure 1.

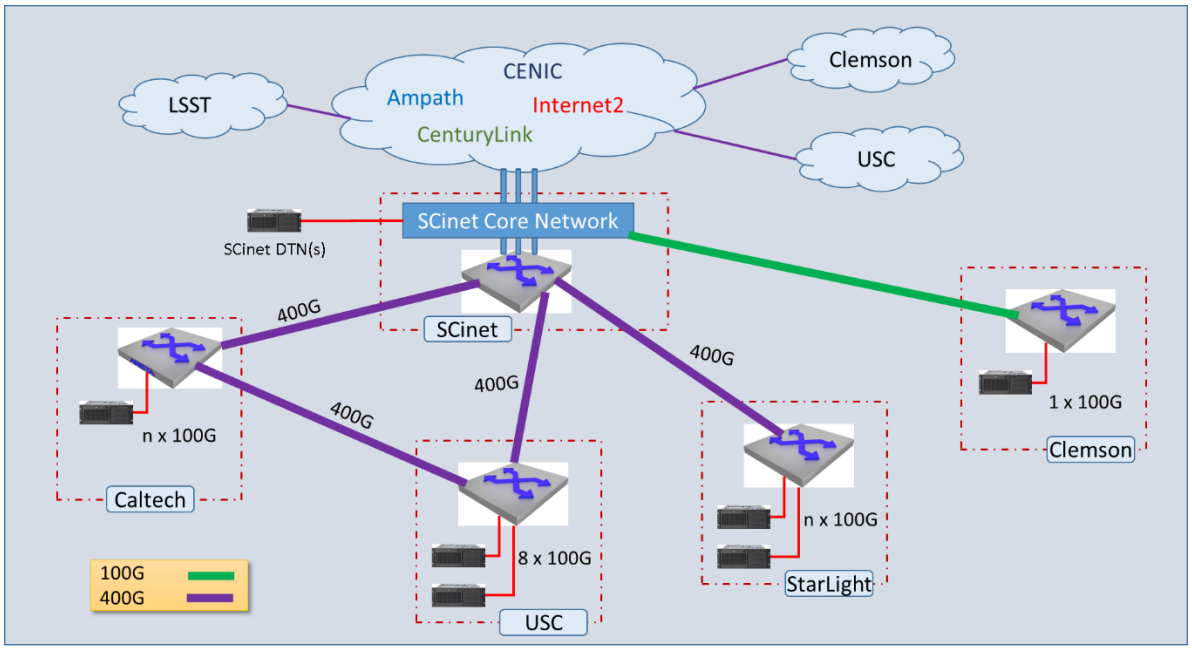

Figure 1: USC Network Topology for SuperComputing 2018

References

[1] Pacific Research Platform. http://prp.ucsd.edu

[2] NVMe Over Fabrics Specifications. https://nvmexpress.org/resources/specifications/