"Demonstrations of 400 Gbps Disk-to-Disk WAN File Transfers using iWARP and NVMe-oF"

Bill Fink
NASA Goddard
bill.fink@nasa.gov

Paul Lang
NASA Goddard
paul.lang@nasa.gov

Abstract

NASA requires the processing and exchange of ever increasing vast amounts of scientific data, so NASA networks must scale up to ever increasing speeds, with 100 Gigabit per second (Gbps) networks being the current challenge. However it is not sufficient to simply have 100 Gbps network pipes, since normal data transfer rates would not even fill a 1 Gbps pipe. The NASA Goddard High End Computer Networking (HECN) team will demonstrate systems and techniques to achieve near 400G line-rate disk-to-disk data transfers between a high performance 4x100G NVMe Server at SC18 to or from a pair of high performance 2x100G NVMe servers across two national wide area 4x100G network paths, by utilizing NVMe-oF (NVME over Fabrics) and iWARP (Internet Wide Area RDMA Protocol) to transfer the data between the servers' NVMe drives.

I. Overview

To achieve near 400G line-rate disk-to-disk network data transfers, all the various components in the end-to-end path, including NVMe drives, 100G NICs, network switches, and links, must work in unison to avoid any bottlenecks that would stall the data transfer pipeline. The NASA Goddard HECN team builds custom high performance NVMe servers, utilizing motherboards that have the necessary PCIe Gen3 slots and lanes, the latest Xeon processors with high speed DDR4 memory, 3500/2100 MB/s read/write speed NVMe drives, and Chelsio 100G NICs. Data transfer performance and system CPU utilization from tests with the NVMe-oF/iWARP configuration will be compared with that from tests using more traditional TCP/IP network data transfer methods. This will demonstrate the feasibility of transferring large amounts of data at 400G rates across a wide area network utilizing NVMe-oF/iWARP via Chelsio 100G NICs and Samsung NVMe drives to or from a high performance NVMe server with four 100G NICs.

II. Innovation

NASA and many other organizations have a need to support the processing and exchange of rapidly growing vast amounts of science data, thus requiring the ability to transport extremely large scale datasets for petascale scientific research using 100G local area and wide area networks. The experience gained through the development of the high performance NVMe servers via the innovative integration of the various system and network hardware and software components allows us to actually achieve 400G network data transfers, and to pass the knowledge gained on to others with similar requirements.

III. HPC and Science Relevance

The experience and knowledge gained from the development of the high performance NVMe servers is shared with the network R&D and HPC science communities, so they can also achieve greatly improved network data transfer rates, as needed to support their mission data requirements.

IV. SCinet and R&E Requirements

- The HECN demo requires a layer2 4x100G SCinet connection to the StarLight booth (#2851).
- The HECN demo requires a layer2 4x100G network path across the MAX R&D infrastructure and CenturyLink 100G circuits between NASA Goddard and SC18, and also a layer2 2x100G network path from a server at StarLight to SC18.

V. Network Topology

The HECN demo requires a layer2 4x100G network path between NASA Goddard and the StarLight booth at SC18 via the MAX R&D network and CenturyLink, and also a layer2 2x100G network path between StarLight and the StarLight booth at SC18 as depicted in the diagram below.

SC18

Demonstrations of 400 Gbps Disk-to-Disk WAN File Transfers using iWARP and NVMe-oF

An SC18 Collaborative Initiative Among NASA and Several Partners