–- DRAFT –--

Results From NASA High End Computing (HEC) WAN File Transfer Experiments/Demonstrations Super Computing 2017 (SC17)

- Bill Fink

- Computational And Information Science and Technology Office (CISTO)
- High End Computer Networking Team (HECN), Code 606.1
- NASA Goddard Space Flight Center

- November 13-16 2017

Overview

- These SC17 Network Research Exhibition (NRE) demonstrations were designed to showcase capabilities for transporting extremely large scale data for petascale scientific research using four 100 Gbps long distance Wide Area Network (WAN) and Local Area Network (LAN) circuits.
- For the SC17 Super Computing Conference in Denver, CO, a consortium of researchers implemented a national optical network testbed consisting of multiple 100 Gbps optical circuits using ESnet and exchange facilities (see attached diagram).
- This testbed is an extension of an existing testbed that was established to develop advanced services and technologies for next generation data- intensive petascale science, under the GSFC High End Computing Program.
- These demonstrations build on earlier efforts related to demonstrations and experiments on a persistent HEC testbed connecting the GSFC and Starlight, and on national WAN and LAN testbeds implemented for SC10 thru SC16.

Collaborating Organizations

- NASA Partners in “Near 400G Disk-to-Disk Network Data Transfers” Special SC16 Demonstration/Evaluation Experiments
- Organizations: Energy Science Network (Esnet), International Center for Advanced Internet Research, Northwestern University (ICAIR), Mid-Atlantic Crossroads, Maryland University (MAX), StarLight International/National Communications Exchange Facility Consortium, Metropolitan Research and Education Network (MREN), Open Cloud Consortium (OCC), Laboratory for Advanced Computing, University of Chicago (LAC), Large Scale Networking Coordinating Group of the Networking and Information Technology Research and Development (NITRD) program.
- Corporations providing loaner equipment include: Arista, Brocade, Ciena, Dell Edgecore
- On-site SC16 support from Brocade (Wilbur Smith).

Pre SC17 Successes and Issues

- Jul 10, to push beyond the CPU limitation of the SC16 demo, we looked into doing RDMA to bypass the CPU. We looked for a server with at least eight PCIe X16 slots and 48 NVMe bays. It looked promising that Supermicro might swap out the SATA backplane for a NVMe backplane in their 4028GR-TR chassis, but didn't happen, so ordered a 2028R-NR48N to cannibalize for its NVMe parts.
- Oct 26, received the Supermicro 2028R-NR48N 48-bay NVMe chassis after 12 week procurement and vendor delay.
- Oct 30, completed modifications to move NMVe backplanes and their power connections into the 4028GR-TR chassis.
- Discovered PCIe negotiation issue with both the AOC-SLG3-4E4T and Chelsio NIC thought to be due to daughter card. RMA'd daughter card but problem persisted. Meanwhile received recommendation to set jumper on SOC-SLG3-4E4T to set it to use X16 PCIe lanes which cleared its negotiation issues. Obtained a total of over 50 GB/s (400 Gb/s) on 24 simultaneous NMVe disk reads.

SC17 Successes and Issues

- Worked through normal setup issues as WAN circuits are connected to the booth.
- Mysterious system hang was tracked down to a bad Chelsio 100G NIC.
- Spent time collecting information for Chelsio on PCIe negotiation issues.
- Had to downgrade Fedora O/S version on all four servers to be able to load Chelsio TCP Offload Engine (TOE) and internet Wide Area RDMA protocol (iWARP) drivers.
- Chelsio configuration file provided for WAN connection needed modification to support WAN. TOE worked well in transmit but not in receive (will do further investigation).

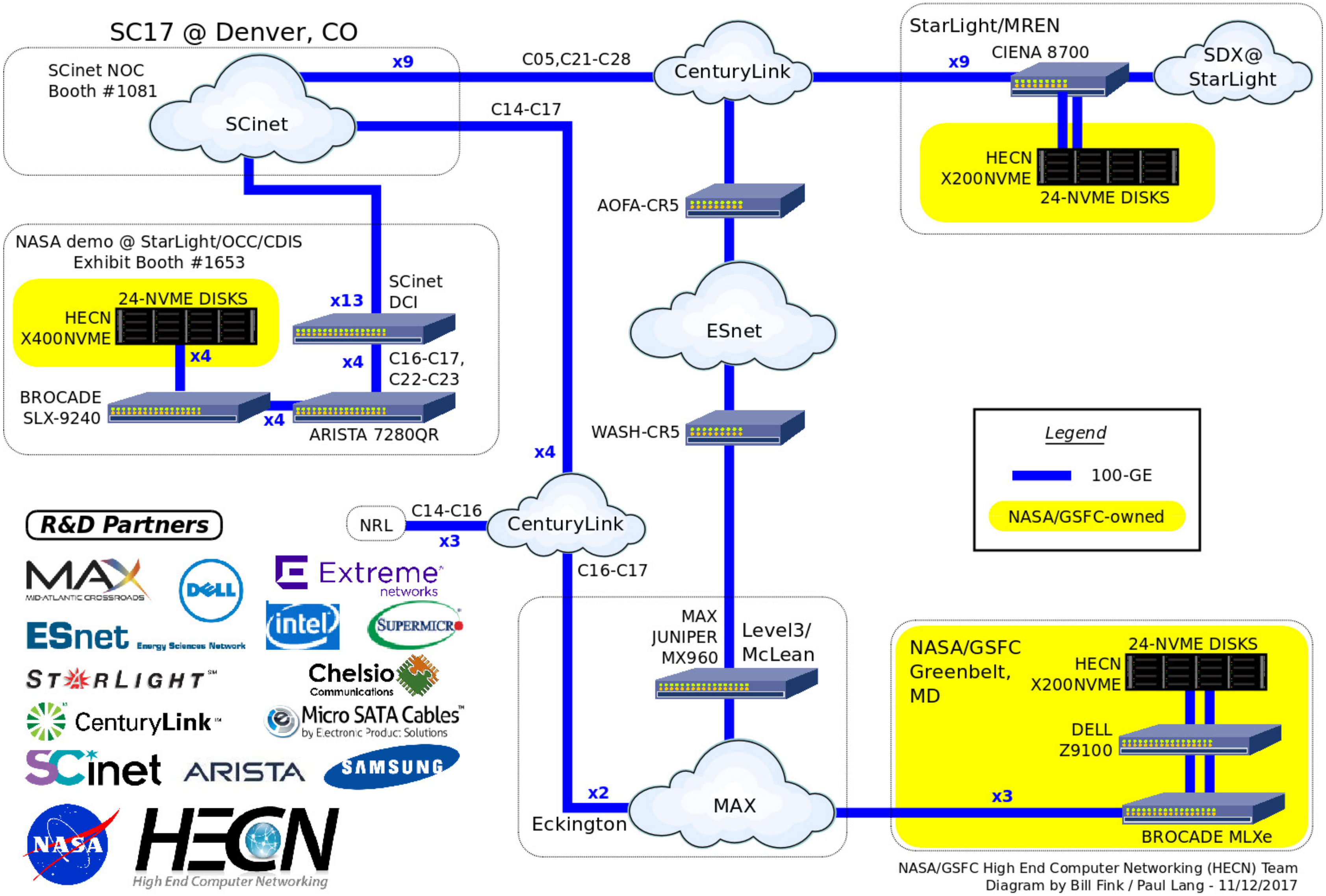
NASA HEC WAN File Transfer Experiments/Demonstrations At SC17

- <u>Special SC17 Demonstration/Evaluation Experiments</u>
- Use of four custom NASA/HECN Team built network-testing-raid-servers, the new 400G NVMe server at SC17 and three of our SC16 200G NVMe servers with Chelsio 100G NICs, deployed into the SC17 LAC/iCAIR booth, Starlight International and National Communications Exchange facility in Chicago, and at GSFC. The 200G servers capable of 181 Gbps non-RDMA back-to-back uni-directional disk-to-disk file copies, using two 100G interfaces and 16 NVMe drives per server. The 400G NVMe server is capable of maxing out the PCIe connection to the NVMe drives and also to the 100G NICs that negotiate their PCIe settings correctly.
- The work on the negotiation and configuration issues prevented running a RDMA demo.
- Plan to continue testing and provide an addendum with updated information.
- Demonstrate/Evaluate interoperability between multiple vendor 100G products from Arista, Brocade, Ciena, Dell/Forece10, Edgecore, Fujitsu, Juniper, over SCinet, Century Link, ESnet, Starlight, and MAX/DRAGON
- Achieved 389 Gbps memory to memory testing, 193+ Gbps from the 400G and 200G servers at SC17 to Starlight and Goddard servers respectively.

SC17

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

An SC17 Collaborative Initiative Among NASA and Several Partners



NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang - 11/12/2017

SC17 – Bill working with Troy from Chelsio, Paul discussing the demo with a visitor to the booth.

High End Computer Networking (HECN) Team

Bill Fink
Acting Project Lead
NASA/GSFC

Paul Lang
Network Engineer
NASA/ADNET Systems

Aruna Muppalla
Network Engineer
NASA/ADNET Systems

Jeff Martz
Network Engineer
NASA/ADNET Systems

Mike Stefanelli
Network Engineer
NASA/ADNET Systems

Pat Gary
(In Memoriam)
43 Years NASA/GSFC

HECN SC17 400G NVME server

(Parts/Price list)

Description	Model	Price	Qty	Subtotal
Eight X16 slot chassis	Supermicro 4028GR-TR	$3,553	1	$3,553
48-bay NVMe chassis	Supermicro 2028R-NR48N	$5,640	1	$5,640
12-core processor	Intel E5-2687W V4	$2,478	2	$4,956
CPU cooler	Dynatron R14	$35	2	$70
8GB 2400MHz DDR4	Kingston ECC 2400 MHz DDR4 memory	$53	8	$424
NVMe M.2 disks	Samsung 512GB 960 Pro	$300	28	$8,400
U.2 to M.2 adapter	MicroSATAcables CASE-944-U2	$40	28	$1,120
Flat Phillips Head screws	M3x4.8mm (Qty 40)	$10	1	$10
PCIe Retimer card	Supermicro AOC-SLG3-4E4T	$160	3	$480
OCuLink cable 85cm	Supermicro CBL-SAST-0820	$48	10	480
Mini-Fit Jr Extraction Tool	Molex 11-03-0044	$21	1	$21
SATA3 system disk	Western Digital Black 1TB 2.5” disk	$65	1	$65
ATA Power Cable	Micro Serial ST-POW/ATA	$4	1	$4
Full Height PCI bracket	Amazon (B01IEGSFN0)	$14	1	$14
100G NICs	Chelsio T62100-CR	$870	4	$3,480

				$28,717

HECN SC17 400G NVME server

(description of modifications)

From the 2028R-NR48N chassis, you will swipe the following parts to put into the 4028GR-TR: two NVMe backplanes, all the cables that were connected to the NVMe backplanes, and the AOC-SLG3-4E4T card from the x16 PCIe slot. The white 8-pin to 4-pin cables you will need to use a pin extractor to swap the yellow and black pins (on the 8-pin side) and trim off the 4 unused pins, then it can be plugged into one of the ten black 8-pin receptacles on the motherboard in the 4028GR-TR. To make the extractor work easier, using a small needle nose pliers slightly bend the very tip of the extractor in towards each other and then slightly spread the tips apart. You will also need to swap out the SAS/SATA backplane in the 4028GR-TR for the NVMe backplanes. This will require, for each backplane, notching the metal bracket that hold the backplanes in place so that the 4-pin power connectors are accessible and notch the tabs on the bracket where it touches a component on the backplane and where it touches one of the OCuLink connectors on the backplane you will also need to add two rivets to strengthen the notched area (see picture for notch and rivet locations). The PCI bracket for the AOC-SLG3-4E4T from the 2028R-48N is low profile (LP), use the PCI bracket from the Amazon (B01IEGSFN0). Set the JP7 jumper on the AOC-SLG3-4E4T cards for X16. (need to find fix for Chelsio PCIe negotiation).

HECN SC17 400G NVME server

(modifications to bracket that holds the backplanes)