Results From
NASA High End Computing (HEC)
WAN File Transfer Experiments/Demonstrations
Super Computing 2013 (SC13)

Bill Fink
Bill.Fink@nasa.gov
Computational And Information Science and Technology Office (CISTO)
High End Computer Networking Team (HECN), Code 606.1
NASA Goddard Space Flight Center
November 17-21, 2013

Overview

- These SC13 Network Research Exhibition (NRE) demonstrations were designed to showcase capabilities for transporting extremely large scale data for petascale scientific research using 100 Gbps long distance Wide Area Networks (WANs) and Local Area Networks (LANs).

- For the SC13 Supercomputing Conference in Denver, a consortium of researchers have implemented a national 100 Gbps optically based network testbed, using ESnet, 100 G access paths, and exchange facilities.

- 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, SC11, and SC12.

Collaborating Organizations

NASA Partners in "Using 100G Network Technology in Support of Petascale Science" Special SC13 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: Brocade

On-site SC13 support from Brocade (Matt Lowe, Wilbur Smith, Peter Vignier and others) and NASA/GSFC (Jarrett Cohen).

NASA HEC WAN File Transfer Experiments/Demonstrations At SC13

Special SC13 Demonstration/Evaluation Experiments

- Use of two custom NASA/HECN Team built network-testing-raid-servers deployed into the LAC/iCAIR booth and at GSFC, capable of:

* >90Gbps bi-directional tcp memory-to-memory data flows over a 100G path

* 115Gbps back-to-back uni-directional disk-to-disk file copies (using 4 40G interfaces and 32 SSDs per server) Nov 2012.

- Demonstrate/Evaluate interoperability between multiple vendor 100G products from Alcatel, Brocade, Ciena, Fujitsu, over SCinet, ESnet, Starlight, and MAX/DRAGON

- Achieved 91Gbps TCP WAN disk-to-disk from SC13 in Denver to NASA Goddard Space Flight Center in Greenbelt, MD

Pre-SC13 Successes and Issues

* Same high performance raid servers that were built for last year's SC12 by HECN were used, although one server was reconfigured to fit into a single chassis.

* Split load between 2 processors (16 cores total) each processor has 1 NIC (uses two 40G ports each) and two raid controllers. Total of 8 4-disk raid arrays (two arrays per raid controller, each array ~1.8Gb/s)

* The two and a half week government shutdown, delayed receipt and testing the Brocade 40G card and network path.

* GSFC to Denver connection established via two 50G limited ESnet paths over ESnet routers via Northern and Southern paths. Just prior to SC13, paths were looped together for testing – 48 Gb/s bi-directional UDP with no drops.

* No travel funds were available for HECN, so a server was sent in a portable rack and the HECN team relied on help from Brocade and others at SC13 and the HECN team established a virtual presence via video conferencing.

SC13 Successes and Issues

* The anticipated Acadia Optronics' 100G NIC was not demoed due to fabrication issues and limited access to testing environment for check out and completion of the driver modifications. They hope to demo their card before the end of this year.

* 100G ESnet path divided into a 50Gbps limited Northern path with a 40ms Round-trip-time (RTT), and a 50Gbps limited Southern path with a 62ms RTT.

* Mellanox driver did not recognize the ports on some of the HotLava NICs (the HECN team established this was due to a firmware issue).

* Server at NASA died and had to be partially rebuilt by the HECN team.

* On the last day, the HECN team upgraded the Mellanox driver and swapped cards around to have ones with newer firmware and upgraded the Operating system from Fedora 17 to Fedora 19.

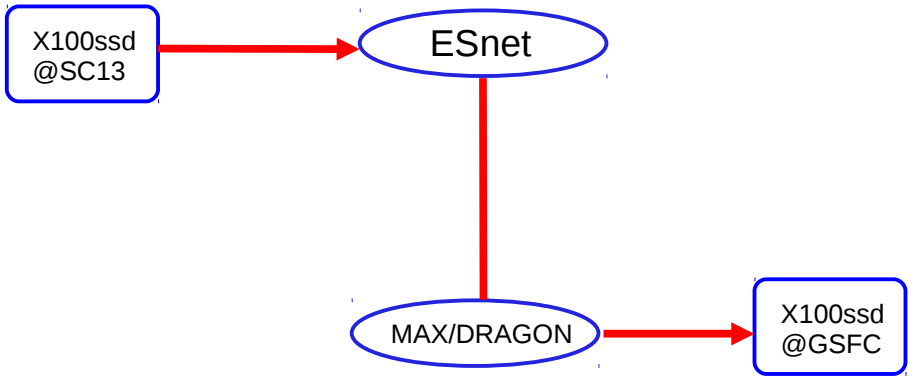
* Obtained 91Gb/s on file transfers from SC13 to Goddard.

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, ESnet, SCinet, Northwestern/iCAIR, UIC/LAC, Starlight, UMD/MAX/DRAGON

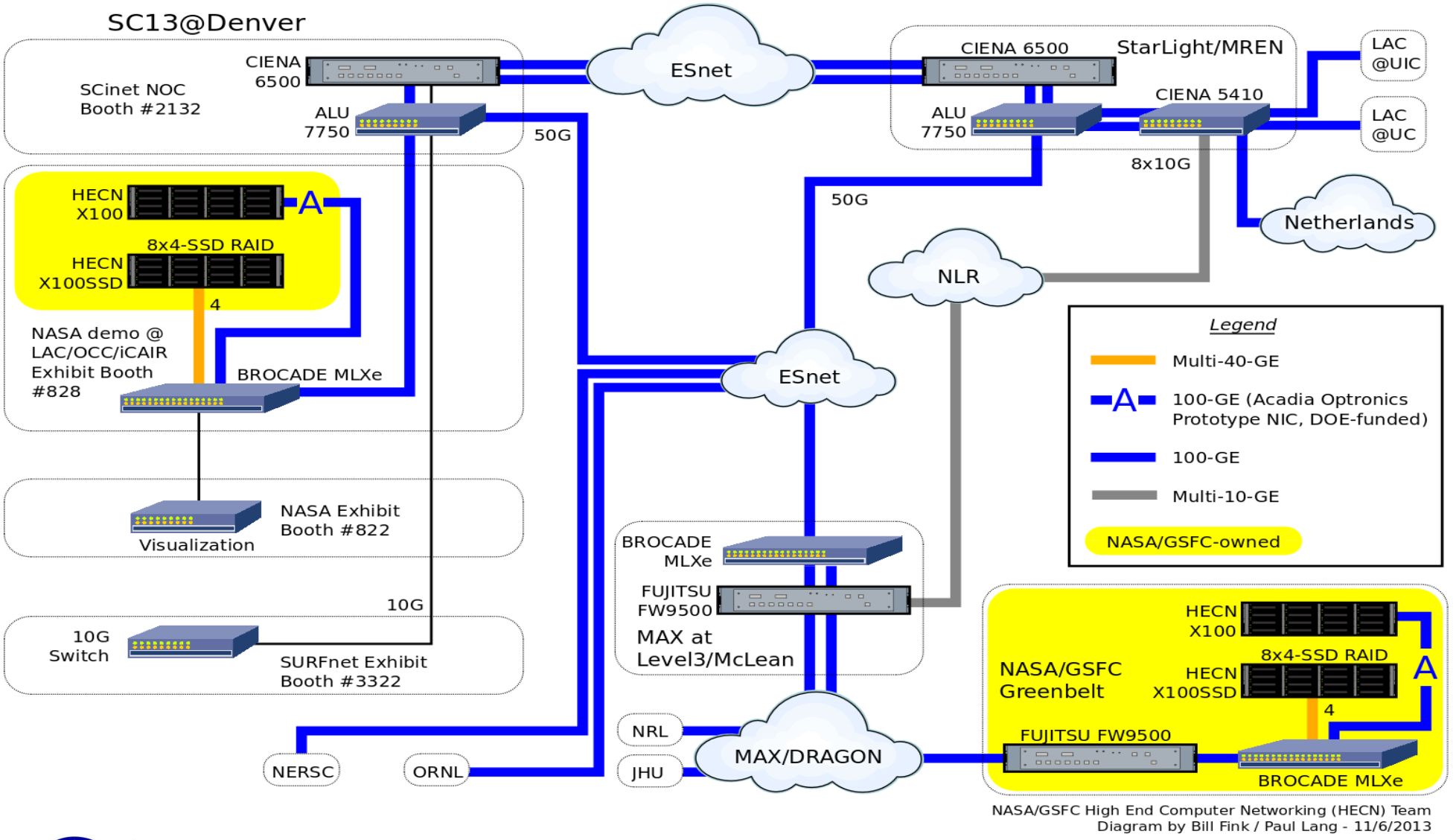
SC13 Demo Summary

Disk-to-Disk 91 Gbps

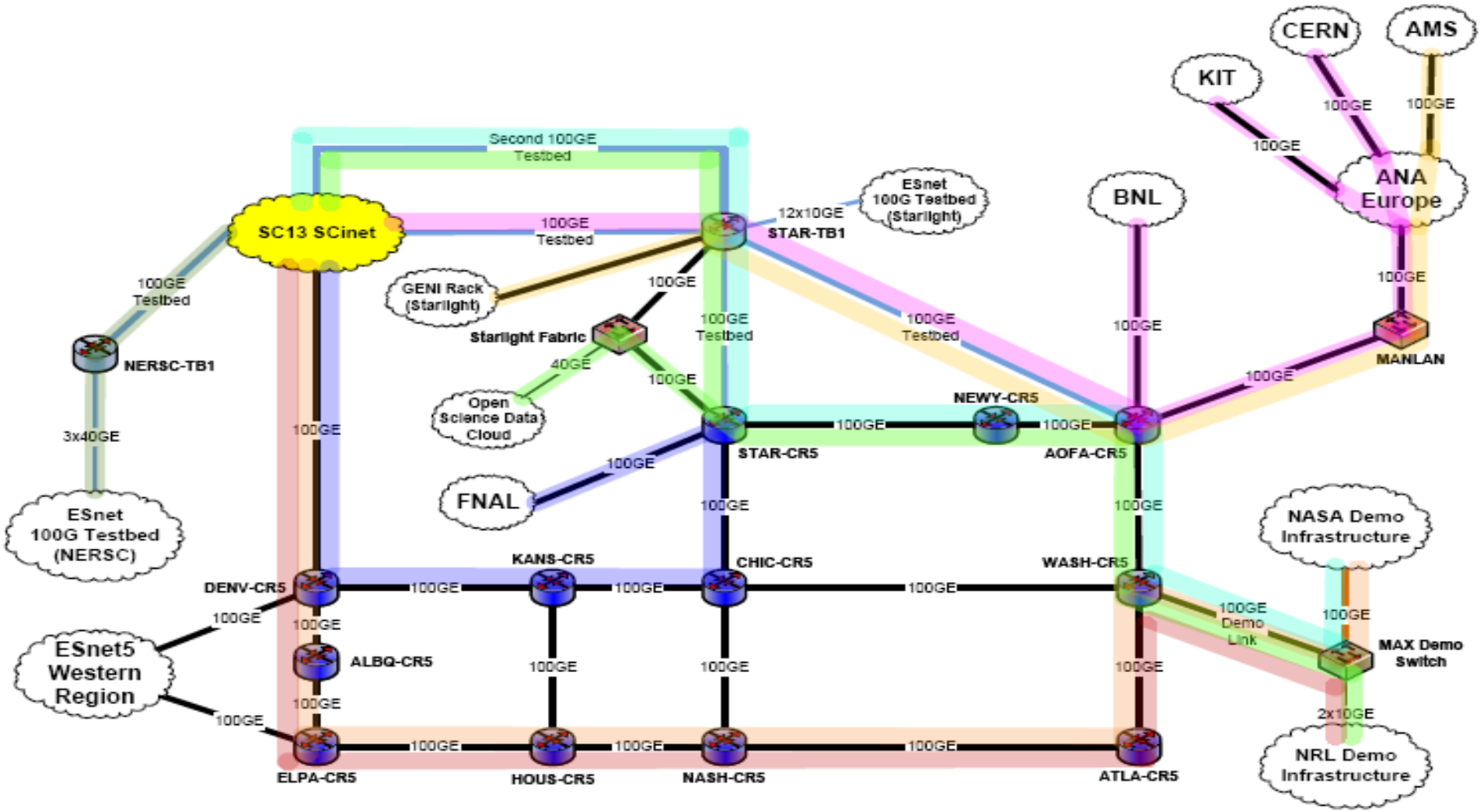


Evaluations/Demonstrations of 100 Gbps Disk-to-Disk WAN File Transfer Performance

An SC13 Collaborative Initiative Among NASA and Several Partners



NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang - 11/6/2013



NRL demo Northern path (20G)	VLANs 1900, 1901
NRL demo Southern path (20G)	VLANs 1902, 1903
NASA demo – production path (50G)	VLAN 1801
NASA demo – testbed path (50G)	VLAN 1800
OpenFlow/SDN demo – ANA path (100G)	VLANs 1921-1929
Caltech demo – ANA path (100G)	VLANs 2602, 2603, 2606, 2607
Caltech demo – FNAL path (60G)	VLANs 2600, 2601
Caltech demo – NERSC TB path (100G)	VLANs 2604, 2605

SC13 demos – ESnet5 map

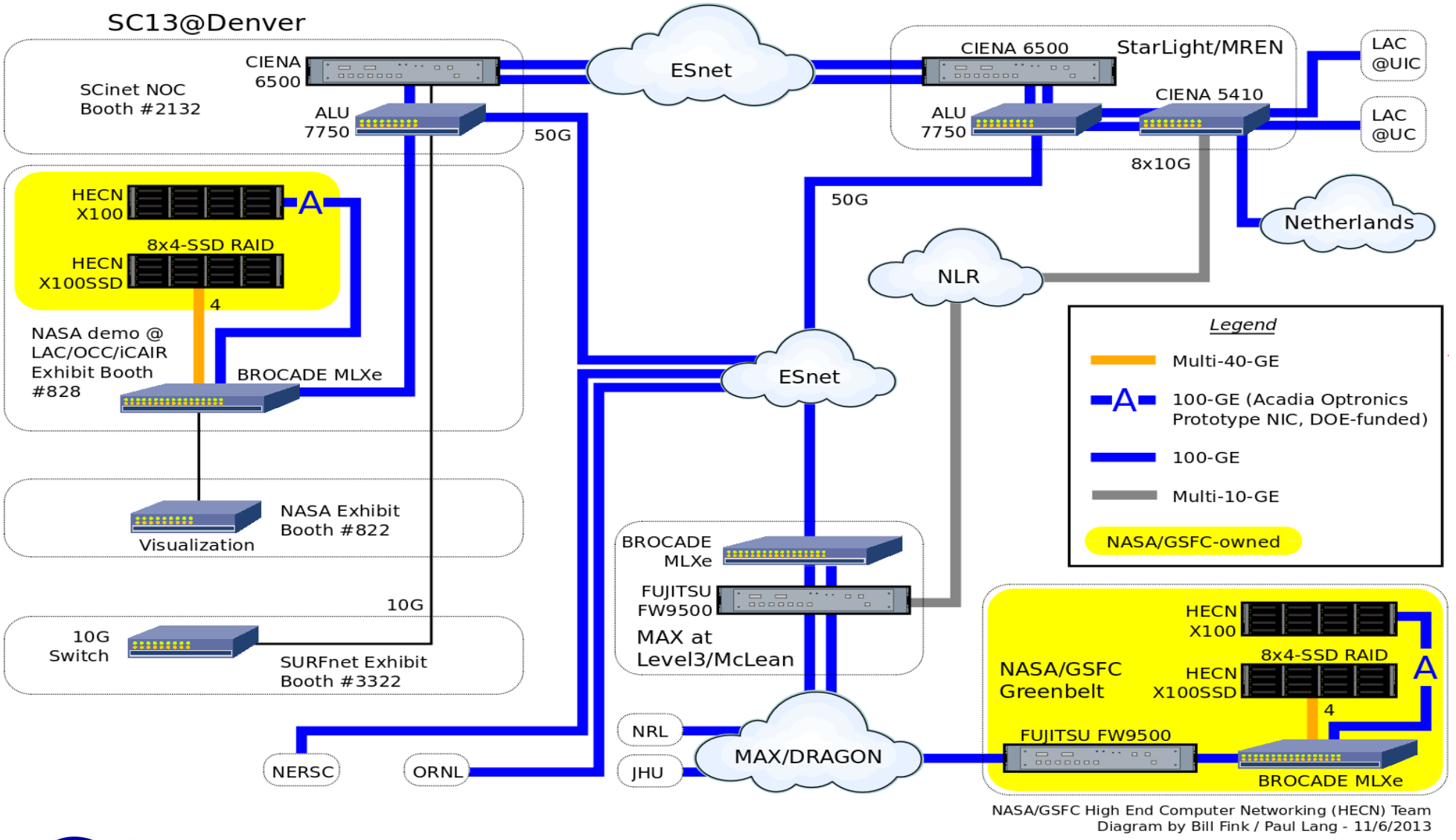
Eli Dart, ESnet 11/14/2013

FILENAME: SC13-DEMOS-V24.VSD

12/5/13

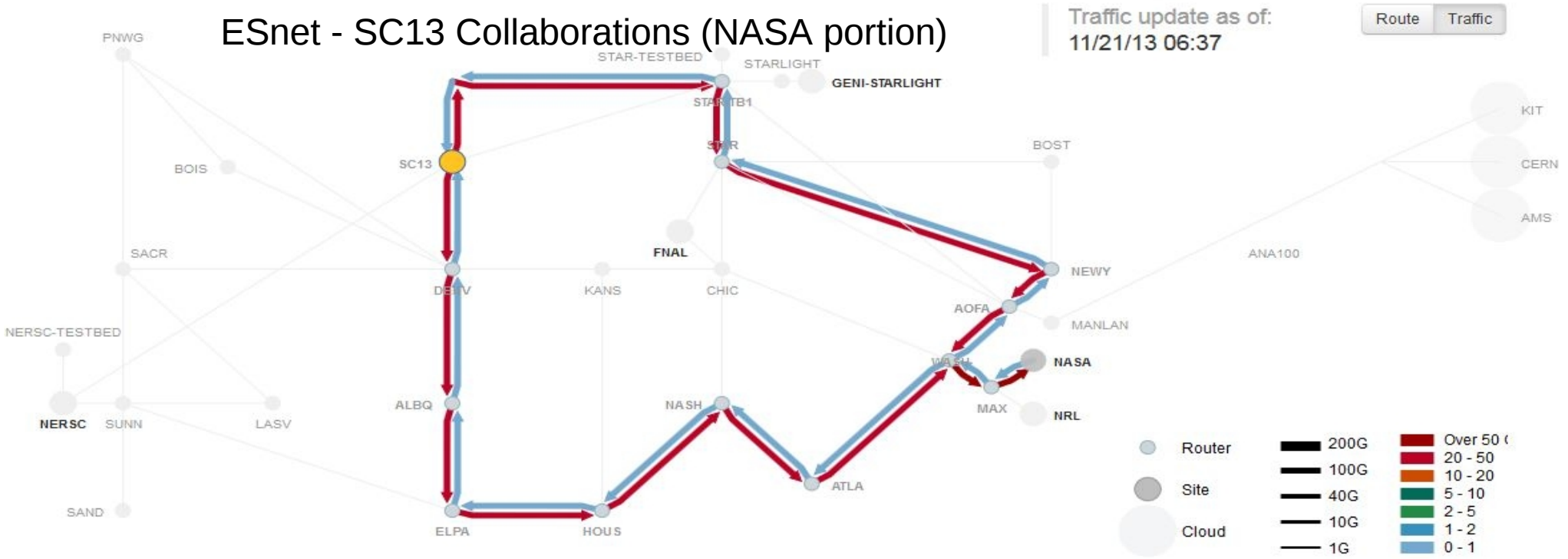
Evaluations/Demonstrations of 100 Gbps Disk-to-Disk WAN File Transfer Performance

An SC13 Collaborative Initiative Among NASA and Several Partners



NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang - 11/6/2013

ESnet - SC13 Collaborations (NASA portion)

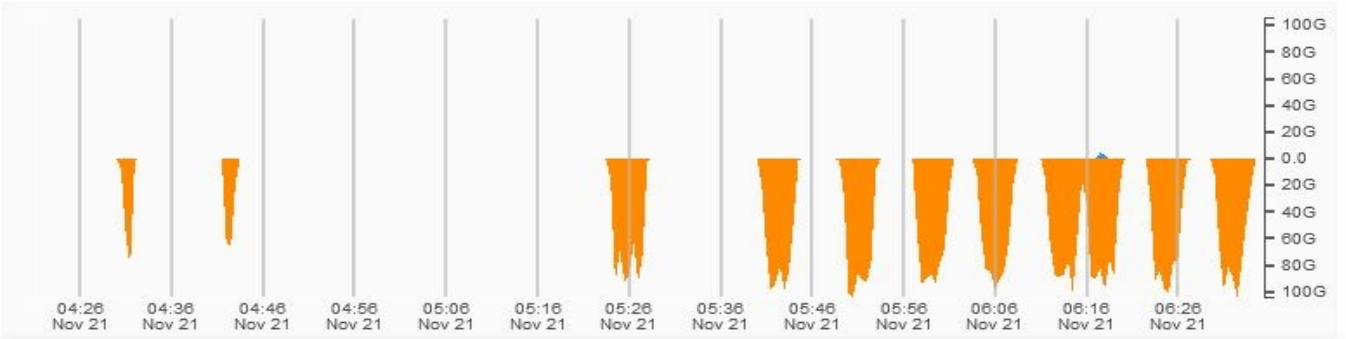


100 Gbps Networks for Next Generation Petascale Science Research and Discovery

Engineers from NASA's Goddard Space Flight Center, in collaboration with the Energy Sciences Network (ESnet), the Starlight International Communications Exchange Facility (Starlight), and the Mid-Atlantic Crossroads (MAX) Gigapop, will demonstrate a 100Gbps disk-to-disk data transfer capability suitable for next-generation Big Data science applications.

Exhibit Booth: 828

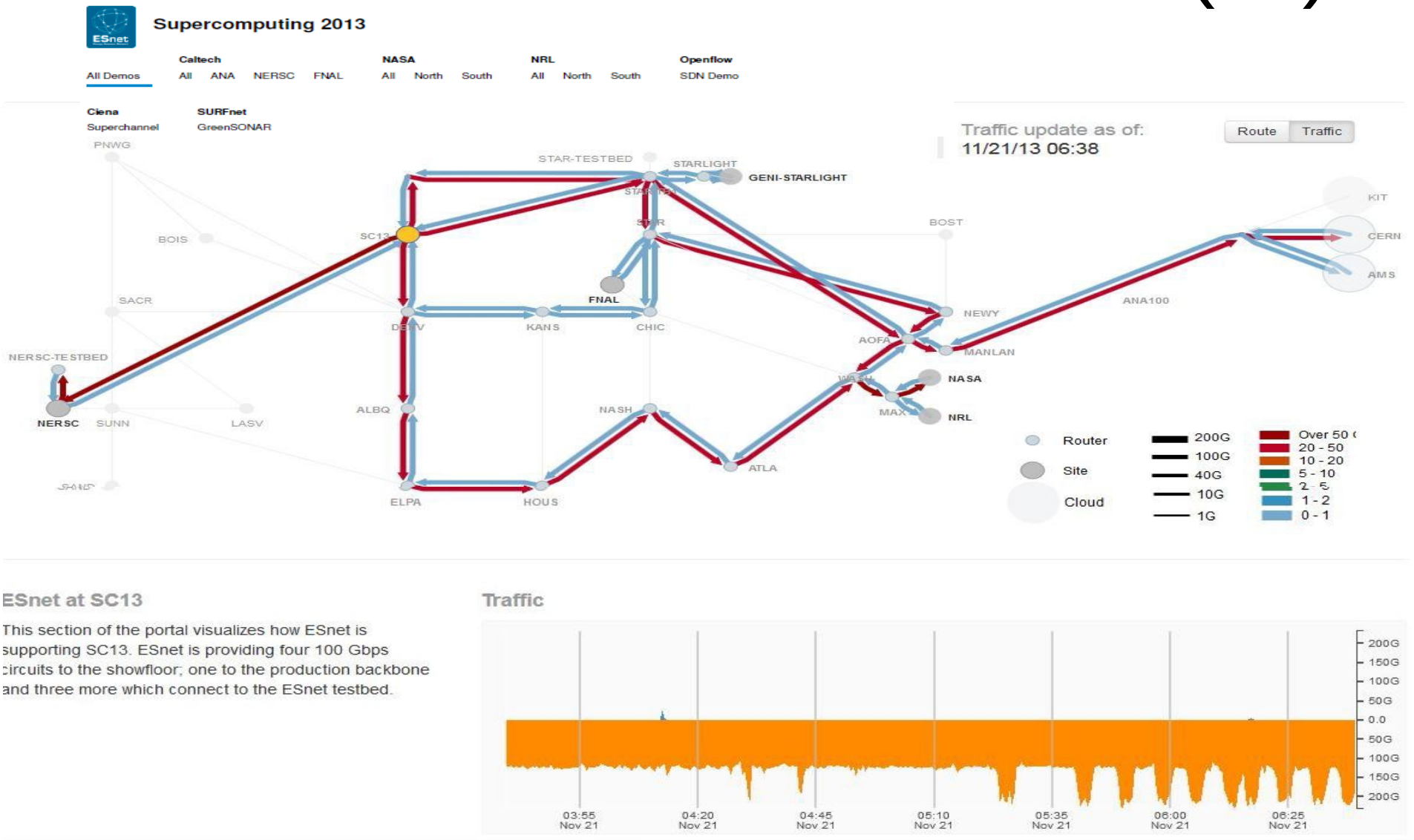
(below: 2TB transfers in 3.5 minutes/each)



ESnet | BERKELEY LAB | U.S. DEPARTMENT OF ENERGY

FAQ
Site Updates

Esnet - SC13 collaborations (all)



ESnet at SC13

This section of the portal visualizes how ESnet is supporting SC13. ESnet is providing four 100 Gbps circuits to the showfloor; one to the production backbone and three more which connect to the ESnet testbed.

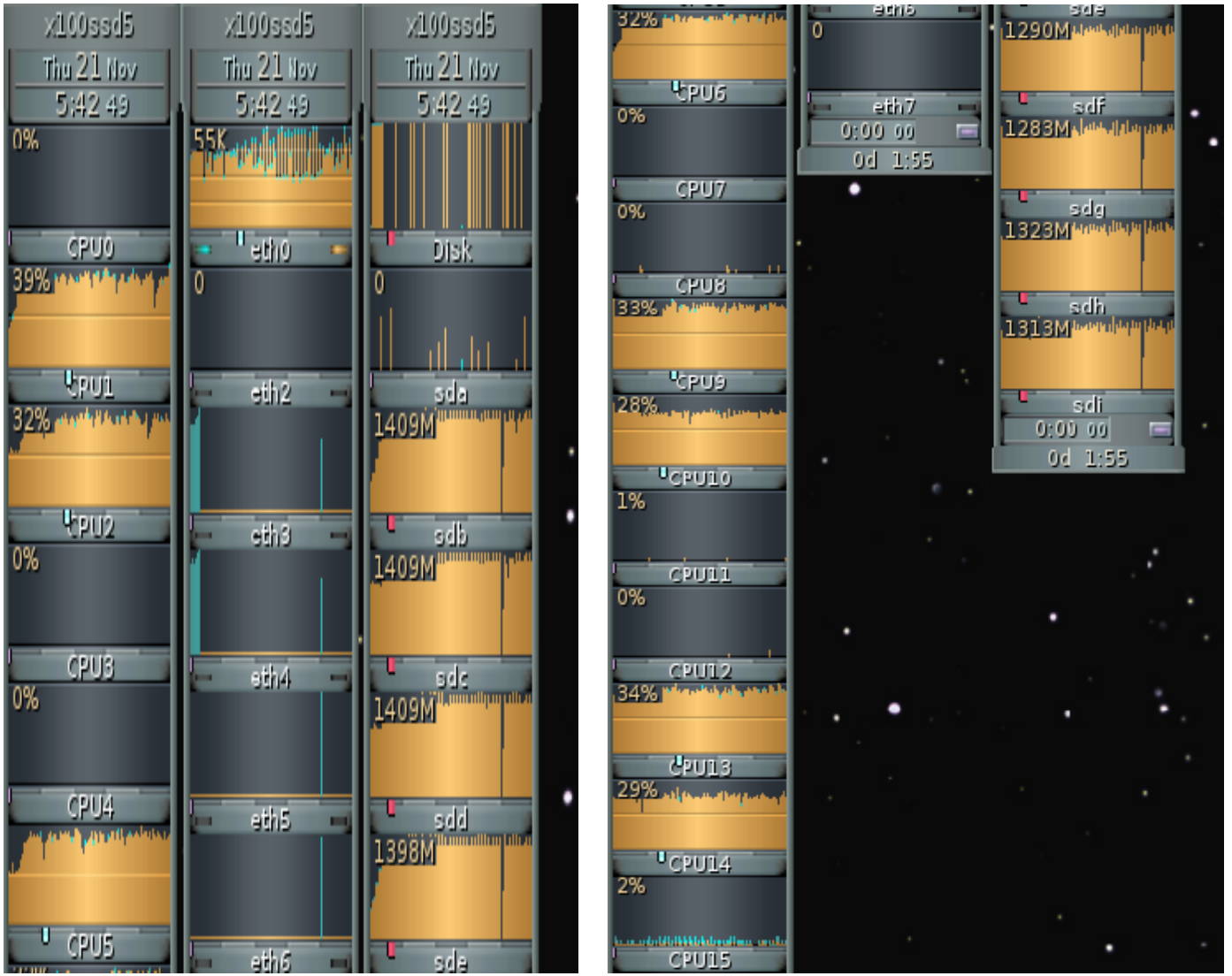
Traffic

ESnet BERKELEY LAB U.S. DEPARTMENT OF ENERGY

FAQ
Site Updates

X100ssd System Monitor (GKrellM)

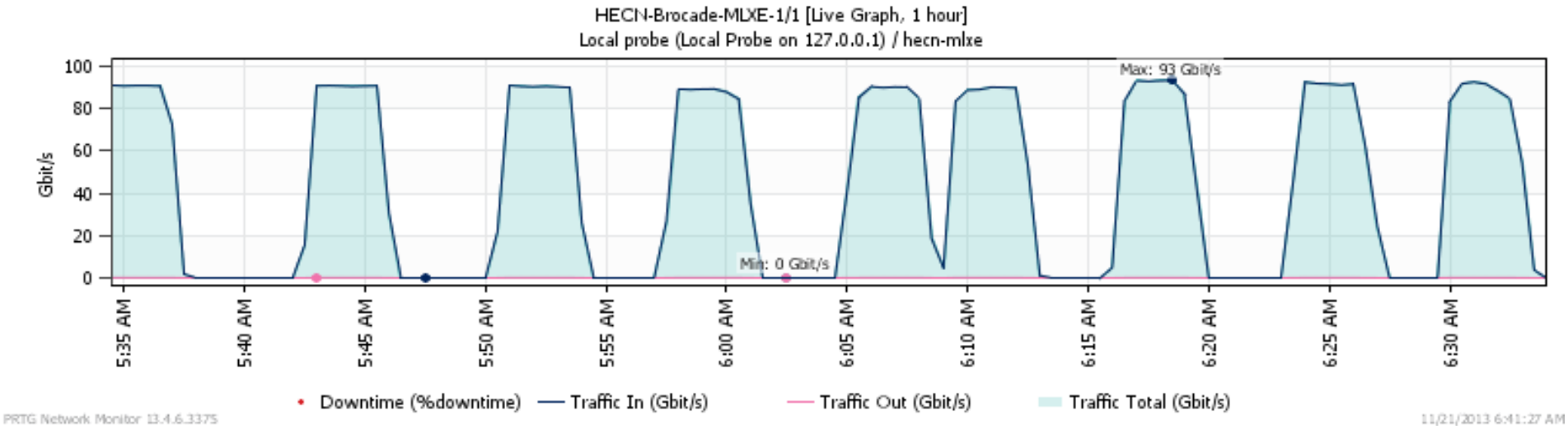
(note: GkrellM not able to display over 2.147GB/s (17.18Gb/s) NIC speeds)



(continued on next column)

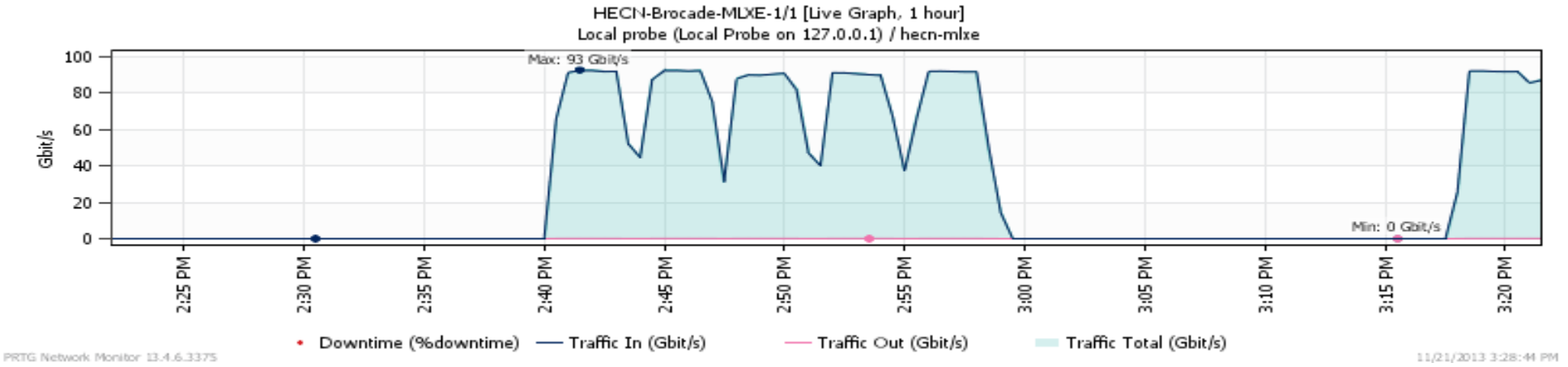
PRTG Graphs of Transfers

(Individual 8 streams of 256GB transfers - 2TB transfers in 3.5 minutes/each)

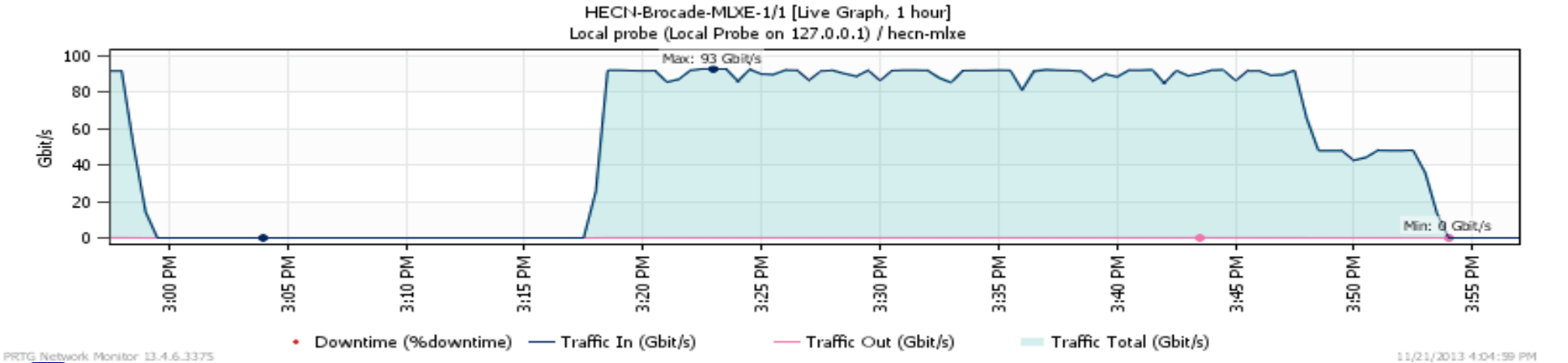


PRTG Graphs of Transfers

(back-to-back transfers of 8 streams of 256GB – total 2TB/each)

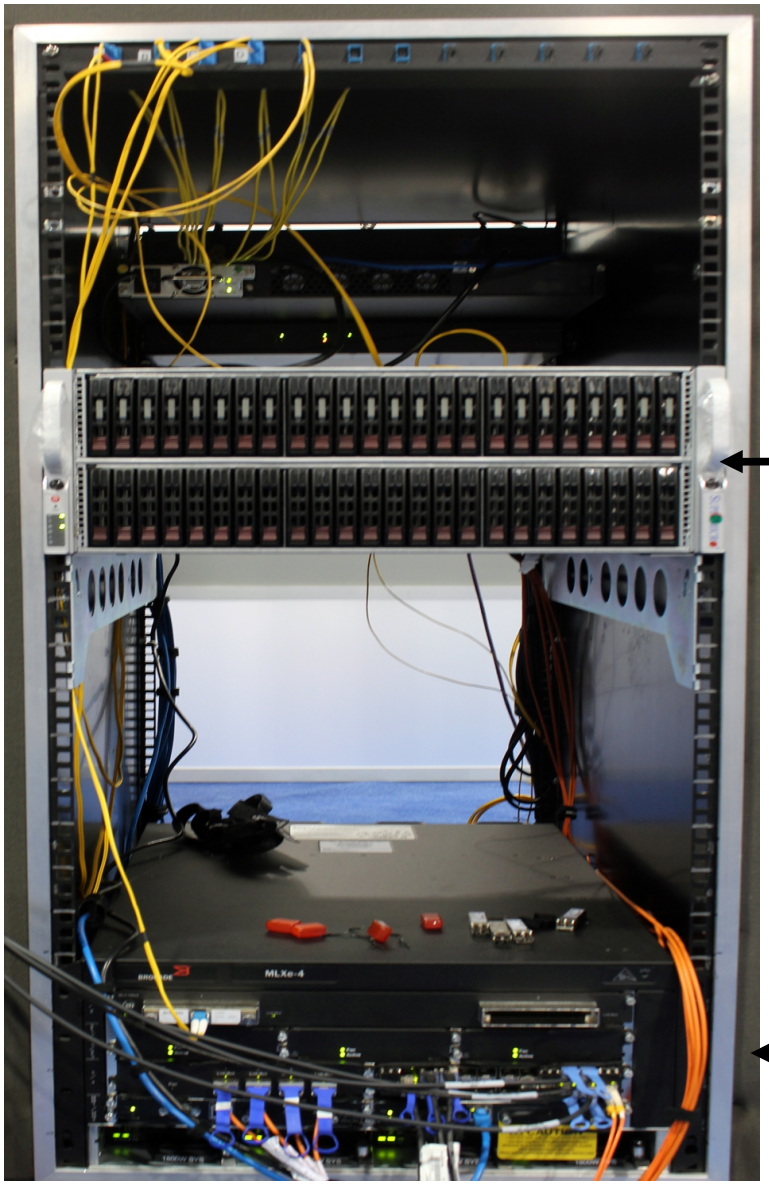


(8 streams of ten 256GB transfers – total 20TB, southern path slower)

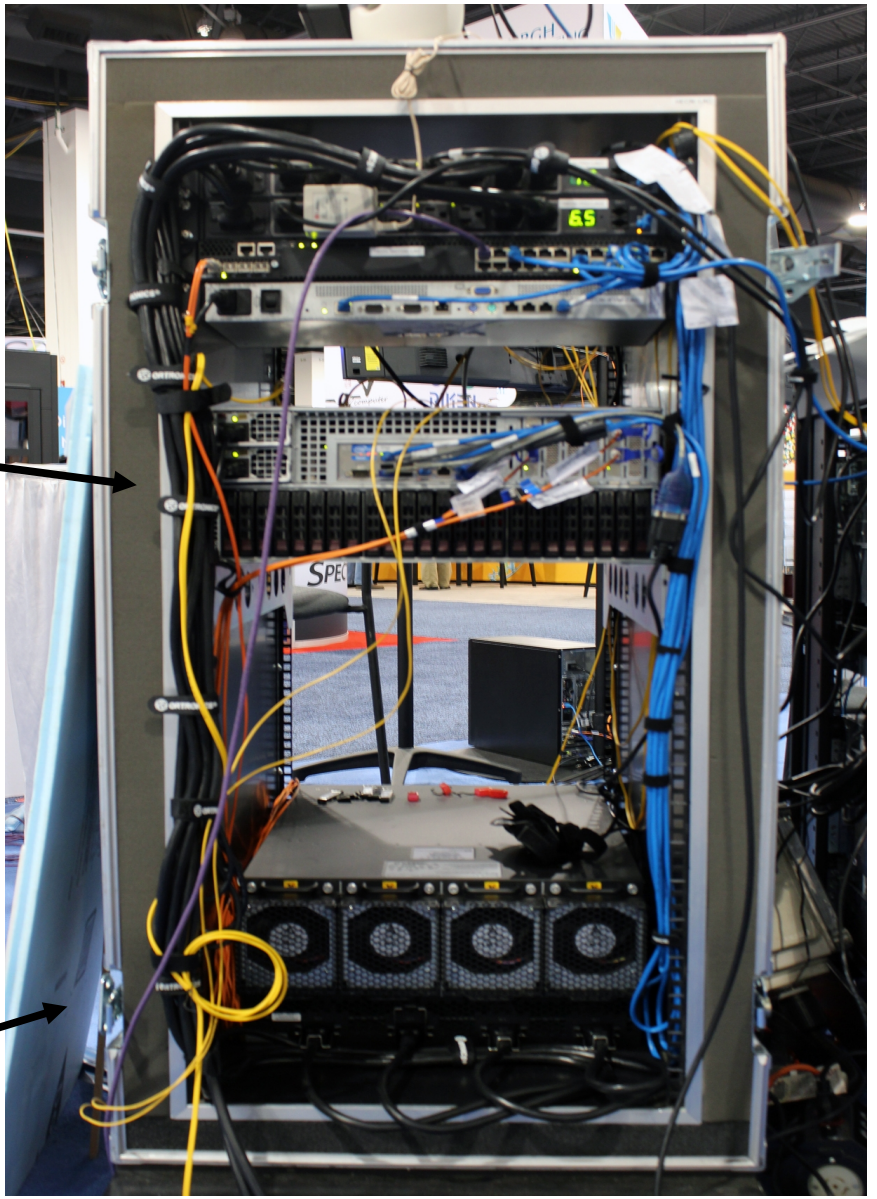


SC13 – Joe Mambretti and Jim Chen, with Paul Lang and Bill Fink on Life Size Screen

Front View of Portable Rack @ SC13



Rear View of Portable Rack @ SC13



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

Possible SC14 Demos

* Acadia Optronics 100G NIC
* Single File Transfer rather than multiple simultaneous file transfers.
* Burst rate adjustment to assist in speed mismatches (100/40G -> 10G, or 100G -> 40G)
* Open-Flow path
* UDP file transfer methods

Chassis modification for SC13

To avoid cooling issues that occured at the SC12 warm showroom floor and to simplify connections (no SAS cables between chassis) at SC13 HECN customized a server to fit into a single chassis.

HECN used a 72 bay 4u Supermicro chassis (417E26-R1400LPB) that we already had which has great air flow. To avoid the noisy fans running full out (too noisy for SC13) or not enough to keep the raid controllers cool, Mike Stefanelli built a temperature sensor driven PWM controller to control a fan based on the raid controller temperature.

HECN needed direct access to the disks (rather than going through a SAS expander), so two of the backplanes were swapped out for ones that did not have the expanders. Two plates that hold the back planes needed to be modified to allow one of the power adapters to plug in.

The chassis had to be modified to allow a path for the additional SAS cables to be to run between the back plane and the raid controllers.

HECN also had to purchase the 2-port version of the 40G HotLava NICs (Part# 2QF3A60A1) to fit this chassis

SC12 X100SSD Parts/Price List

(used same server for SC13, modified one server to fit into one chassis for SC13 showroom)

Description	Model#	Price	Quantity	Subtotals
SM 3u 16bay chassis	836TQ-R800B	$869	1	$869
IcyDock 2.5" adaptor	MB882SP-1S-2B	$17	16	$272
SM front pannel cable	CBL-0084	$3	1	$3
Fan extension cable	12" 3-pin fan extension	$1	1	$1
Red Greatland	18" Slimline SATA adapter	$6	1	$6
SM FDD tray for 2.5"	CP-220-83601-0B	$8	1	$8
Power extension cable	8" 8pin power extension	$8	1	$8
SM 2u 24bay chassis	216A-R900LPB	$869	1	$869
SM motherboard	X9DR3-F	$465	1	$465
Intel Xeon processors	E5-2687W	$1,875	2	$3,750
CPU cooler	Dynatron R14 2U	$46	2	$92
WD 500GB system disk	WD5000BPKT	$87	1	$87
DDR3 ECC Reg (4x8GB)	KVR1600D3D4R11SK4/32G	$279	2	$558
HotLava 3port 40GE NIC	3QF3A60a1	$2,200	2	$4,400
LSI 9271-8i raid	LSI00330	$658	4	$2,632
OCZ Vertex3 SSD	VTX3-25SAT3-120G	$105	32	$3,360
SAS-SATA cable	CBL-SFF8087OCF-10M	$16	4	$64
intSAS-intSAS	1 meter SFF-8087 cable	$20	6	$120
1M ext SAS cable	DMX-8088.8088-01M	$40	4	$160
SAS int/ext adapter	SAS-AD8788-2	$55	2	$110
SAS int/ext adapter	SAS-AD8788-4	$115	1	$115

SM = Supermicro. WD = Western Digital				$17,949
10M 40GE active cable	FCBG414QB1C10 (Finisar)	$400	4	$1,600

				$19,549