SC10 Demonstration
"Using 100 Gbps Network Technology in Support of Petascale Science"

J. Patrick (Pat) Gary
Network Projects Leader
Networks and Information Technology Security Group (Code 606.1)
Computational and Information Sciences and Technology Office
NASA Goddard Space Flight Center
Pat.Gary@nasa.gov
March 2, 2011

Presentation for 10th Annual International ON*VECTOR Photonics Workshop, 2Mar11

10 Years of NCCS Highlights

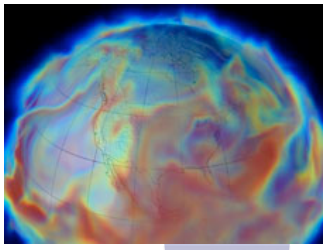
Source: Dan Duffy (GSFC/NASA Center for Climate Simulation (NCCS))

NCCS User Forum Sep. 14, 2010

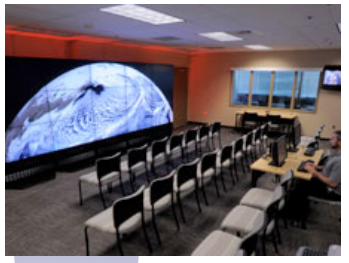
Halem – HP/Alpha Cluster
18th fastest computer in the world

Large SGI Shared Memory Systems at GSFC

Discover – First production IB commodity cluster in NASA HEC



Introduction of Dali analysis environment, interactive access to large data sets



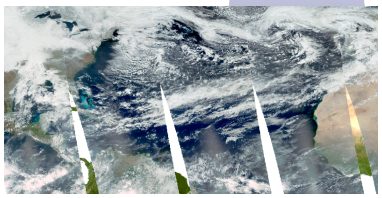
Climate Data Theater
New tool for scientific analysis and visualization

2000 **2005** **2010**

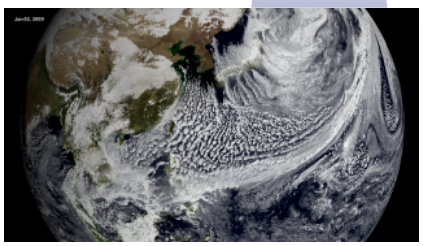
Support for the IPCC AR4 modeling runs by GISS – Nobel Prize

10 GbE NCCS LAN
10 GbE between HEC centers

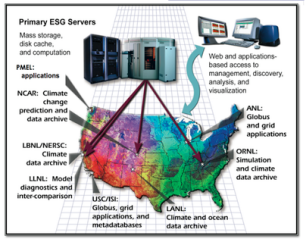
Introduction of the Data Portal System and Services
WMS, MAP, Cloud Library



Discover upgrades supports 3 KM global resolution model – highest resolution ever run

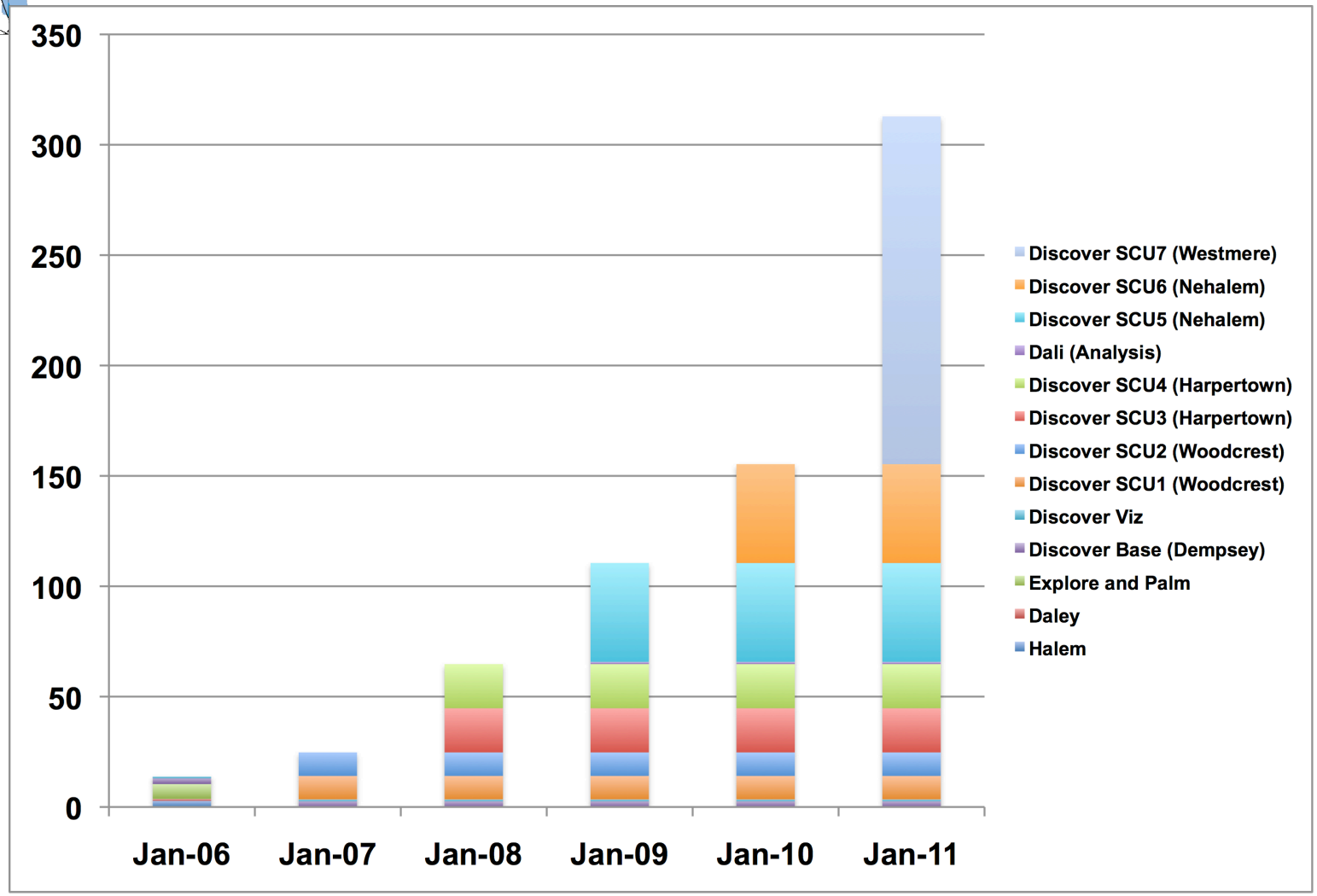


Advancement of Data Services with the Data Management System (DMS) and the Earth Systems Grid (ESG)



NCCS Peak Computing (TF) Over Time

Source: Dan Duffy (GSFC/NCCS)



03/02/11

GSFC High End Computer Network (HECN) Team

From left: Pat Gary, Bill Fink, Paul Lang, Aruna Muppalla, Jeff Martz, Mike Stefanelli

Science and Engineering Network (SEN)

Typically enabling 1-10 gigabit per second (Gbps) user connections, the SEN is a non-mission-dedicated high-end computer network at GSFC serving GSFC projects/users who have computer network performance requirements greater than those baselined for GSFC's general-use campus-wide Center Network Environment (CNE)

http://science.gsfc.nasa.gov/606.1/SENuser.html

20, 40 & 100 Gbps Network Testbed Plans & Accomplishments to Date

http://science.gsfc.nasa.gov/606.1/docs/HECN_10G_Testbeds_082210.pdf

Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Problem Statement

- GSFC's NASA Center for Climate Simulation (NCCS) is increasing its data production/analysis/storage capabilities and its accessing of other large remote data
- Higher bandwidth networks can be deployed
- But bottlenecks in the combination of our file copying applications, disk I/O subsystems, server/workstation configurations, protocol stack tuning and/or NICs is preventing full use of our higher bandwidth networks
- Need to determine server/workstation configurations, data transfer utilities and protocols that enable higher throughput, especially given the emergence of 40- to 100-Gbps networks

Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Notional Milestone Schedule

```
                          |    CY09    |    CY10     |     CY11    |    CY12     |
                          JFMAMJJASONDJFMAMJJASONDJFMAMJJASONDJFMAMJJASOND
Phase 0 10-Gbps Testbeds
O LAN & Region/MAN       -------**
O WAN                      ------*
Phase 1 20-Gbps Testbeds
O LAN & Region/MAN            ---------***
O WAN                              --------***
Phase 2 40-Gbps Testbeds
O LAN & Region/MAN                          ---------***
O WAN                                          --------***
Phase 3 100-Gbps Testbeds
O LAN & Region/MAN                                     ---------***
O WAN                                                     --------***
```

Legend for Milestone Schedule

------ Planning and acquisition subphase

****** I&T plus demo subphase

Test Matrix for Optimizing Wide-Area File Transfer

Tool	Type
aspera fasp**	Disk ↔ Disk
nuttscp	Disk ↔ Disk
GridFTP	Disk ↔ Disk
RocketStream**	Disk ↔ Disk
iRODS	Disk ↔ Disk
bbFTP	Disk ↔ Disk
FDT	Disk ↔ Disk
HPN-SCP	Disk ↔ Disk
dsync*	Disk ↔ Disk
xdd	Disk ↔ Disk
rsync	Disk ↔ Disk
FTP	Disk ↔ Disk
gpfs	Mem/Disk ↔ Disk
pNFS	Mem/Disk ↔ Disk
NFS	Mem/Disk ↔ Disk
NFS-RDMA*	Mem/Disk ↔ Disk
nuttcp	Mem ↔ Mem

*RDMA options – IB/Obsidian, iWARP, RoCE

**Commercial product

- Desired measurements: disk-to-disk file-copying-throughput performance (in Gbps), plotted against different file-sizes and different conditions
- Key single-file-sizes in GBs: 16, 32, 64, 128
- Primary different conditions:
  - File-copying-applications, e.g., GridFTP, bbFTP, nuttscp, ...
    - Both well-established and experimental/emerging ones
  - Key round trip times (RTTs) in milliseconds: 0, 15, 90, 180
    - Corresponding very roughly to LAN, large MAN/RON, trans-USA/WAN, trans-Atlantic
- Secondary different conditions, when time permits:
  - Several-file-sizes yet with a constant 256 GB total volume: 16@16, 8@32, 4@64, 2@128
  - Many-file-sizes yet with a constant total volume
  - Cases with packet loss, corruption, etc
  - Real tests
    - @10Gbps: GSFC-GSFC (RTT=0); GSFC-StarLight (RTT=~17); GSFC-ARC (RTT=~87)
    - @20Gbps: GSFC-GSFC; GSFC-StarLight
    - @40Gbps: GSFC-GSFC; GSFC-SC10 (RTT=~35)

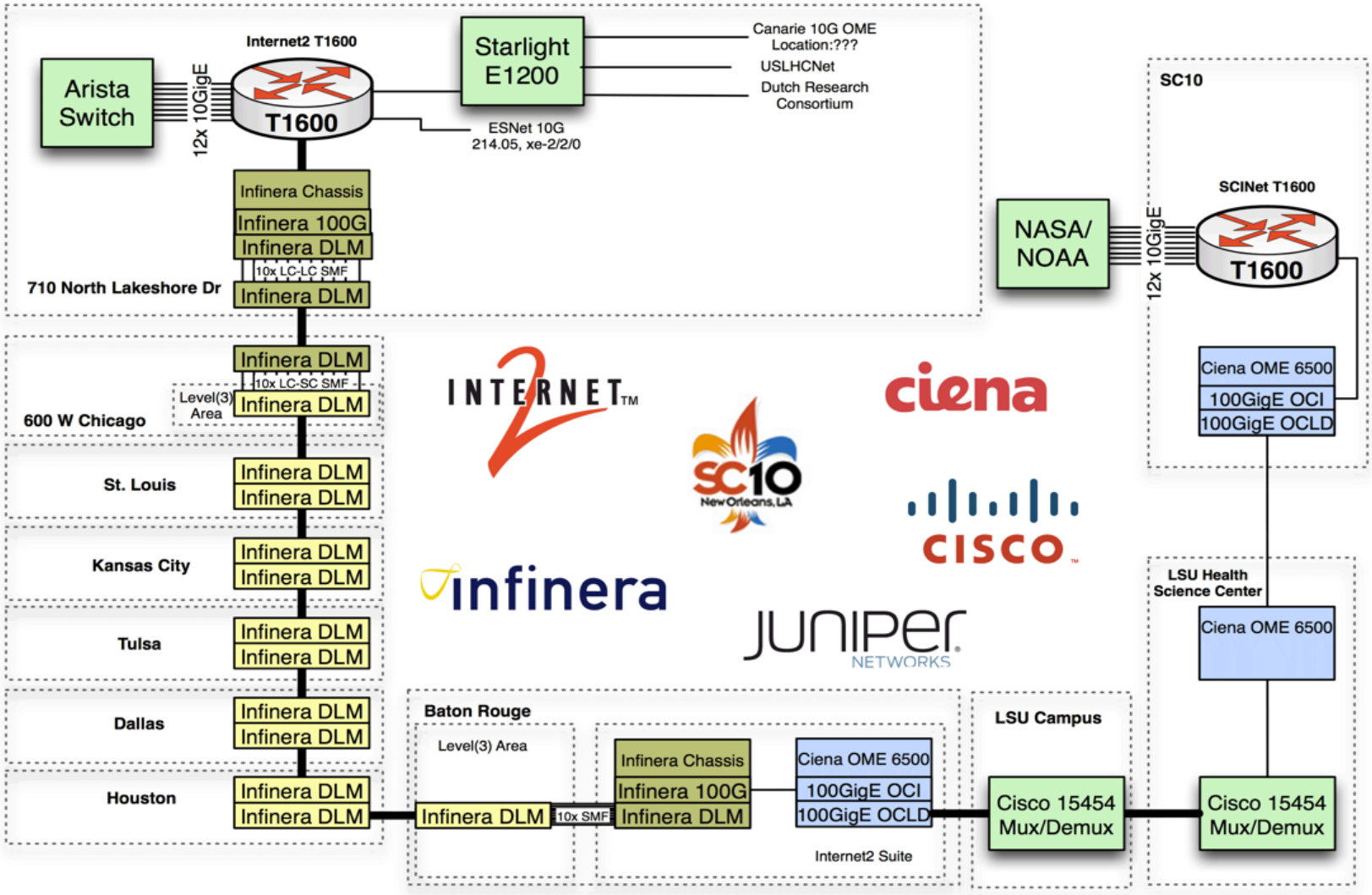
SC10 Demonstration "Using 100 Gbps Network Technology in Support of Petascale Science"

WAN Connectivity for SC10's SCinet

- National LambdaRail (NLR): Eight 10-Gbps links, from StarLight:
  - L2 FrameNet Wave 1_Western Route
  - L2 FrameNet Wave 2_Eastern Route
  - L2 FrameNet_STAR EXPRESS (STAR to NEW ORLEANS)
  - L3_PacketNet Wave 1_Western Route (HOUS)
  - L3 PacketNet Wave 2_Eastern Route (ATLA)
  - Cisco CWAVE West (LOSA)
  - Cisco CWAVE East (JACK)
  - L1 WaveNet_STAR EXPRESS (STAR to NEW ORLEANS) [for NASA]
- ESnet: Four 10-Gbps links, including one each supporting three projects using the SCinet Research Sandbox
  - http://www.lbl.gov/cs/SC10/demos.html
- Internet2: Four "normal" 10-Gbps links (two to the IP Network, one to the ION Infrastructure & one for NOAA) plus a special 1x100G pathway for their Multi-Vendor 100GigE Demo Between Chicago and SC10

Internet2's SC2010 Multi-Vendor 100G Demonstration

Source: Chris Robb/Internet2



SC10 Demonstration
"Using 100 Gbps Network Technology in Support of Petascale Science"

Special SC10 Demonstration/Evaluation Experiments

- Use a set of the NASA/HECN Team's network-testing-workstations deployed into both the NCDM-LAC/iCAIR and NASA Exhibit Booths, capable of:
  - >100-Gbps uni-directional memory-to-memory data flows
  - >80-Gbps aggregate-bidirectional memory-to-memory data flows
  - ~40-Gbps uni-directional disk-to-disk file copies (using SSDs)
- Demonstrate/evaluate different vendor-provided 40G/100G network technology solutions with full-duplex 40G and 100G LAN data flows across SCinet Research Sandbox inter-booth fiber
- Use existing 4x10G dedicated pathway across NLR and MAX/ DRAGON between GSFC and StarLight, plus a mix of 8 other 10G NLR+Cisco-provisioned pathways and a 1x100G Internet2-provisioned pathway between StarLight and SC10, to conduct science-oriented WAN data flow demonstrations

SC10 Demonstration
"Using 100 Gbps Network Technology in Support of Petascale Science"

NASA HEC WAN File Accessing Team

- GSFC NASA Center for Climate Simulation (NCCS)
  - Dan Duffy/GSFC
    - Hoot Thompson/PTP
    - Kirk Hunter/PTP
- GSFC/NCCS HEC Network (HECN) Team
  - Pat Gary/GSFC
  - Bill Fink/GSFC
    - Paul Lang/ADNET
    - Jeff Martz/ADNET
    - Aruna Muppalla/ADNET
    - Mike Stefanelli/ADNET
- ARC/CIO Network Team
  - Kevin Jones/ARC
  - Hugh LaMaster/ARC
    - Dave Hartzel/CSC
    - Mark Foster/CSC
    - Matt Mountz/CSC

SC10 Demonstration "Using 100 Gbps Network Technology in Support of Petascale Science"

NASA "Infrastructure-Owner" Partners in "Using 100G Network Technology …" Special Demos

- International Center for Advanced Internet Research (iCAIR), PI: Dr. Joe Mambretti/Northwestern University
- Internet2, POC: Chris Robb/Internet2
- Laboratory for Advanced Computing (LAC), PI: Dr. Bob Grossman/UIC
- Mid-Atlantic Crossroads (MAX), PM: Peter O'Neil/UMCP
- National LambdaRail (NLR), POC: Bonnie Hurst/NLR
- National Oceanic and Atmospheric Administration (NOAA), POC: Jerry Janssen
- SCinet Research Sandbox (SRS), Chair: Rodney Wilson/Ciena

SC10 Demonstration "Using 100 Gbps Network Technology in Support of Petascale Science"

Acknowledgement of Vendor Equipment On Loan (1 of 2)

- Arista: Two 7148SX 48-port 10-GE switches
- Ciena: Two Optical Multiservice Edge 6500 units each with 100-Gbps transport and 10x10Gbps-to-1x100Gbps muxponder interfaces
- Cisco: Two CRS-3 routers each with one 100-GE and 14x10-GE interfaces, plus use of a third CRS-3 with two 100-GE interfaces
- ColorChip: Two DragonFly 40G-LR (up to 10km) QSFP transceivers (beta)
- cPacket: Two cVu 320G 32-port 10-GE traffic monitoring switches
- Extreme Networks: Two VIM3-40G4X 4-port 40-GE modules (beta, for Summit X650 10-GE switches)

SC10 Demonstration "Using 100 Gbps Network Technology in Support of Petascale Science"

Acknowledgement of Vendor Equipment On Loan (2 of 2)

- Fusion-io: Two Octal cards (SSDs on PCIe Gen2 x16)
- HP: Two ProLiant DL580 G7 servers with two 2x10-GE NICs and one QDR IB HCA
- Panduit: Two CN1 Net-Access Switch Cabinets with accessories

"Indirectly On Loan" (via Internet2)

- Juniper: Two T1600 routers each with one 100-GE and 12x10-GE interfaces

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10

Demo Summary

12x10Gbps* between: HECN XSSD++ + HECN A+ + HECN A+ ⟷ HECN XSSD++ + HECN A+ + HECN A+

1x40Gbps full-duplex across: Extreme X650 ⟷ (w/ColorChip LR's) ⟷ Extreme X650

1x100Gbps full-duplex across: Ciena 6500 ⟷ Ciena 6500

1x100Gbps full-duplex across: Cisco CRS-3 ⟷ Cisco CRS-3 ⟷ Cisco CRS-3

1x100Gbps full-duplex across: Juniper T1600 ⟷ Internet2 ⟷ Juniper T1600

8x10Gbps full-duplex across: Wksts @SC10 ⟷ NLR ⟷ Wksts @StarLight

40Gbps disk-to-disk between: HECN XSSD++ ⟷ HECN XSSD++

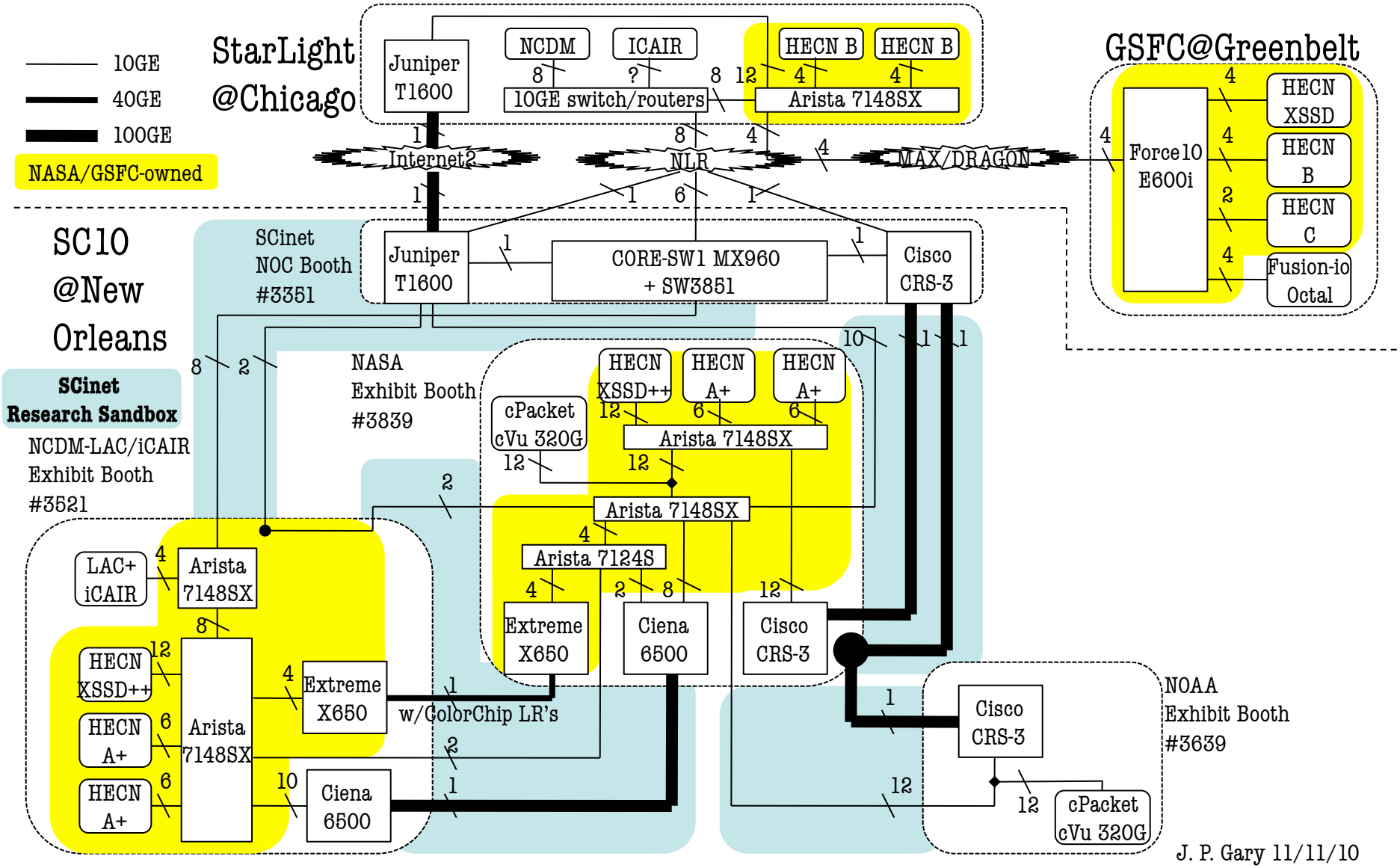
40Gbps disk-to-disk across: XSSD++ @SC10 ⟷ Internet2 + NLR + MAX/DRAGON ⟷ XSSD @GSFC

***bi-directionally**

J. P. Gary 11/01/10

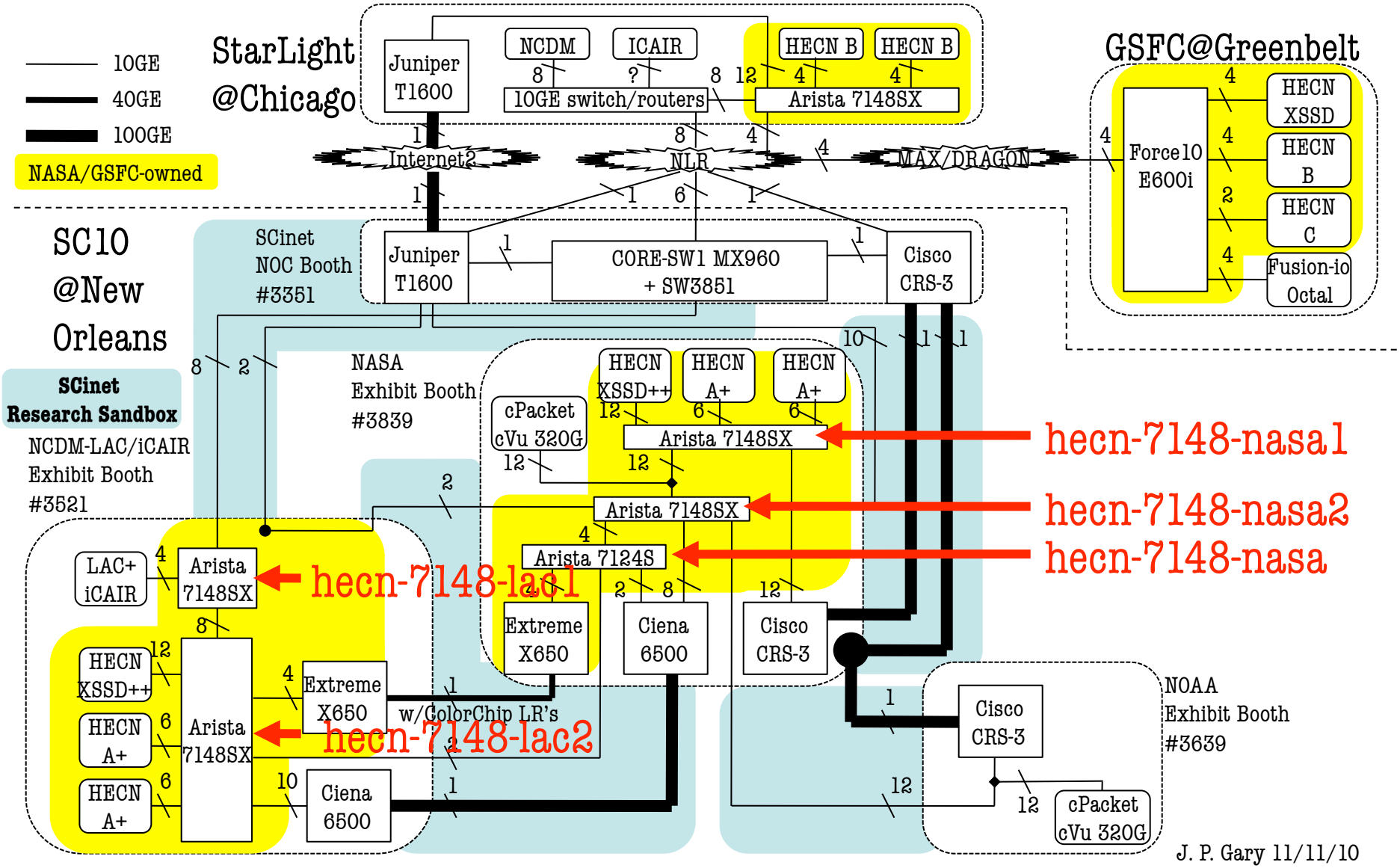
Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10



Using 100G Network Technology in Support of Petascale Science

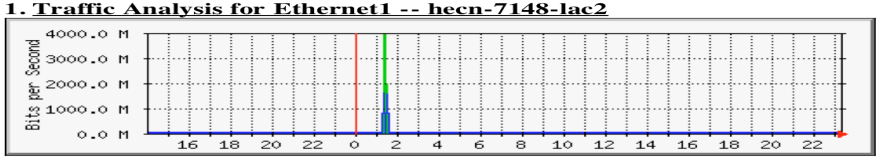
A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10



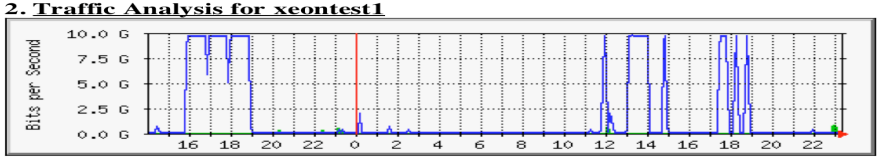
MRTG Index Page for hecn-7148-lac2

Nov 17, 2010, 10:28 PM CT

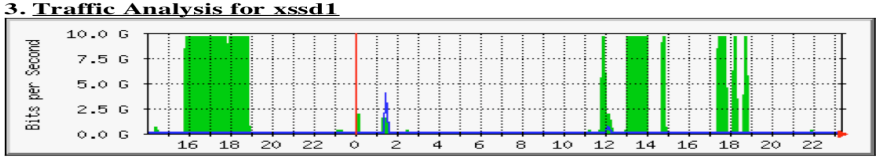
1. Traffic Analysis for Ethernet1 -- hecn-7148-lac2



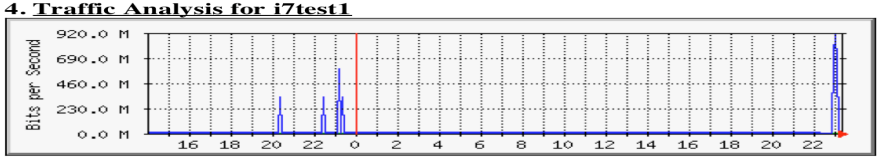
2. Traffic Analysis for xeontest1



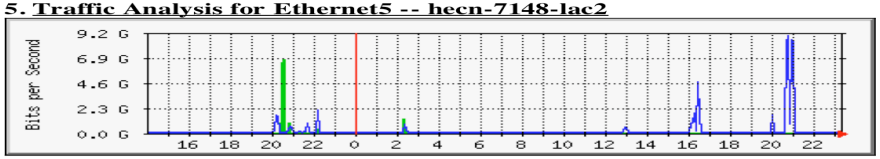
3. Traffic Analysis for xssd1



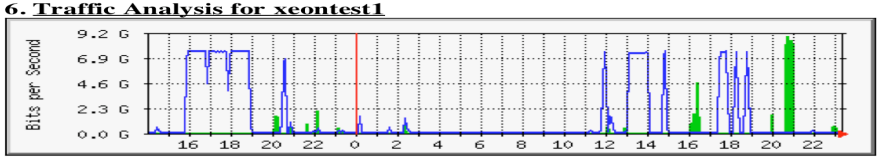
4. Traffic Analysis for i7test1



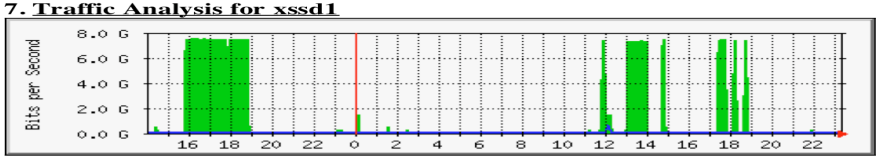
5. Traffic Analysis for Ethernet5 -- hecn-7148-lac2



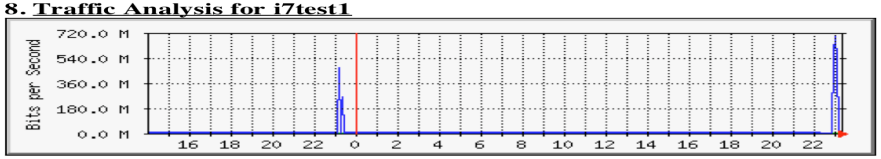
6. Traffic Analysis for xeontest1



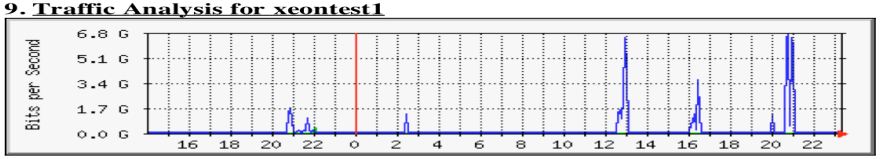
7. Traffic Analysis for xssd1



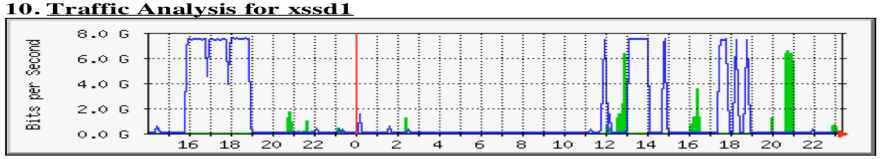
8. Traffic Analysis for i7test1



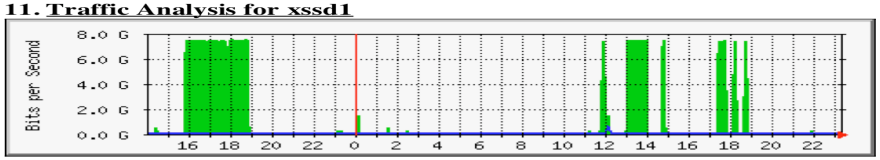
9. Traffic Analysis for xeontest1



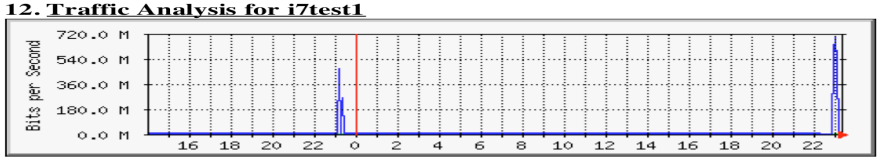
10. Traffic Analysis for xssd1



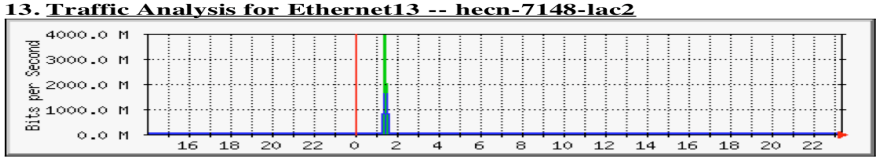
11. Traffic Analysis for xssd1



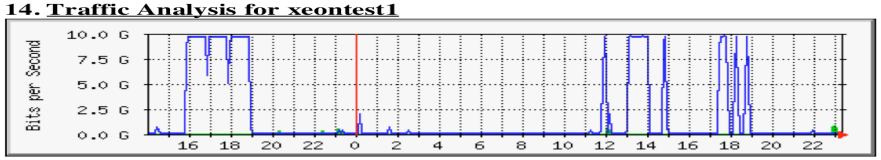
12. Traffic Analysis for i7test1



13. Traffic Analysis for Ethernet13 -- hecn-7148-lac2



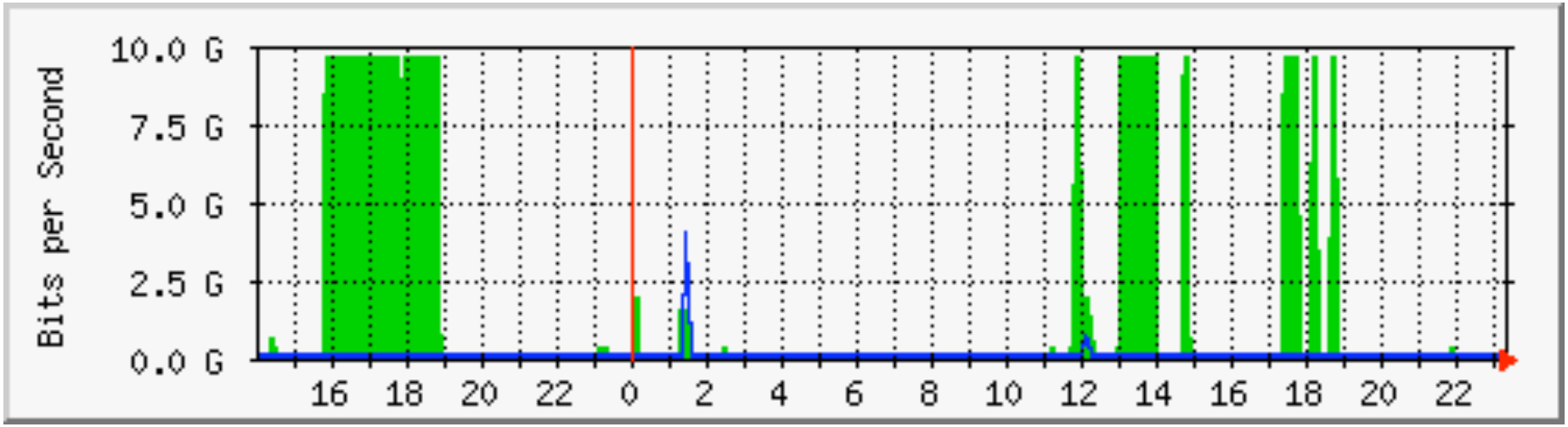
14. Traffic Analysis for xeontest1



A Sample MRTG-Generated Traffic Analysis From NASA's hecn-7148-lac2 10-GE Switch In NCDM's Exhibit Booth During SC10 Of NASA Workstation XSSD1's 10-GE Network Interface #1

The statistics were last updated **Wednesday, 17 November 2010 at 22:28**

`Daily' Graph (5 Minute Average)

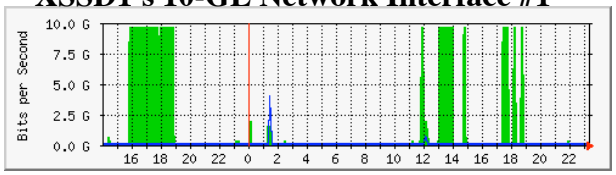


Sample MRTG-Generated Traffic Analyses From NASA's hecn-7148-lac2 10-GE Switch In NCDM's Exhibit Booth During SC10 Of NASA Workstation XSSD1's 10-GE Network Interfaces

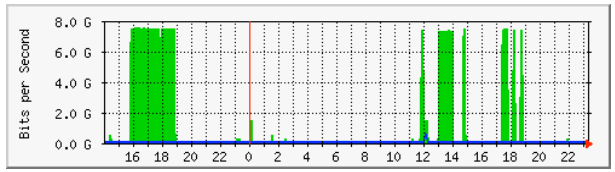
The statistics were last updated **Wednesday, 17 November 2010 at 22:28**

`Daily' Graph (5 Minute Average)

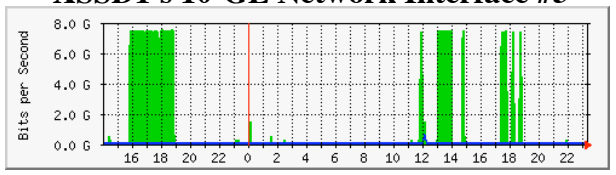
XSSD1's 10-GE Network Interface #1



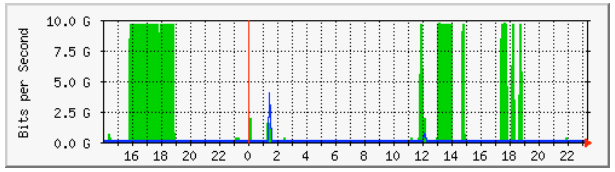
XSSD1's 10-GE Network Interface #2



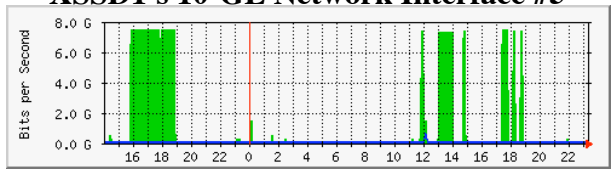
XSSD1's 10-GE Network Interface #3



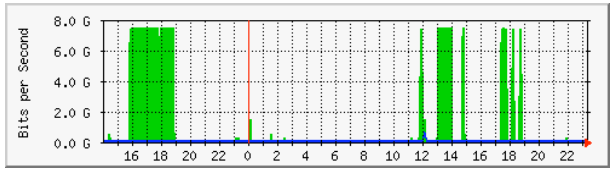
XSSD1's 10-GE Network Interface #4



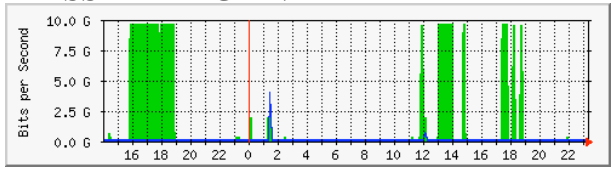
XSSD1's 10-GE Network Interface #5



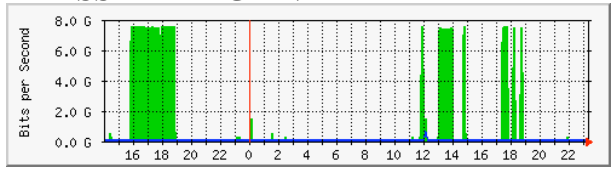
XSSD1's 10-GE Network Interface #6



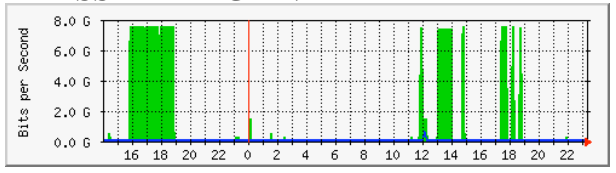
XSSD1's 10-GE Network Interface #7



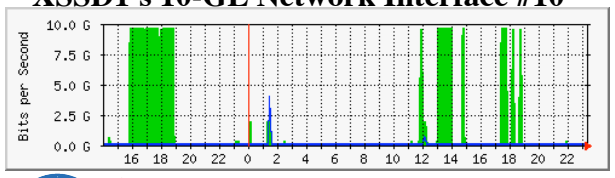
XSSD1's 10-GE Network Interface #8



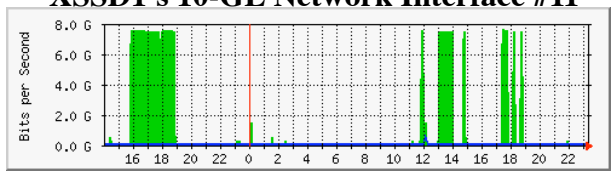
XSSD1's 10-GE Network Interface #9



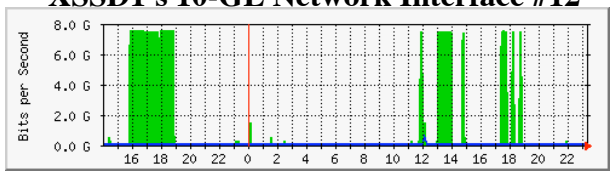
XSSD1's 10-GE Network Interface #10



XSSD1's 10-GE Network Interface #11



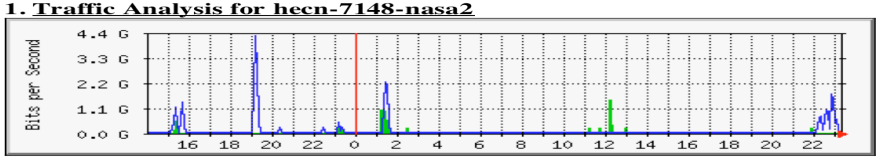
XSSD1's 10-GE Network Interface #12



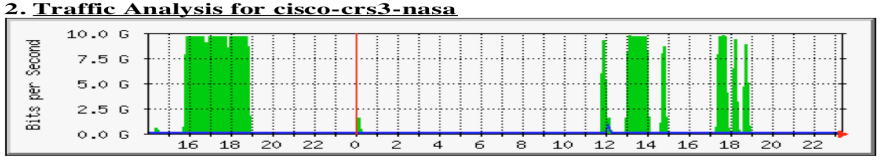
MRTG Index Page hecn-7148-nasa1

Nov 17, 2010, 10:23 PM CT

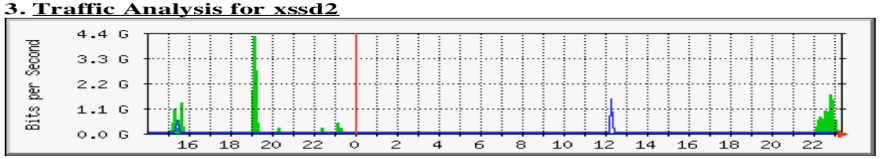
1. Traffic Analysis for hecn-7148-nasa2



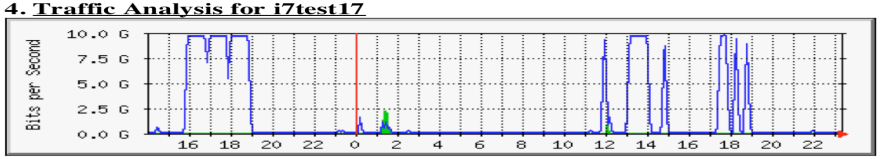
2. Traffic Analysis for cisco-crs3-nasa



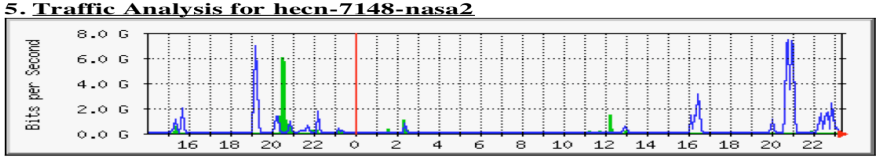
3. Traffic Analysis for xssd2



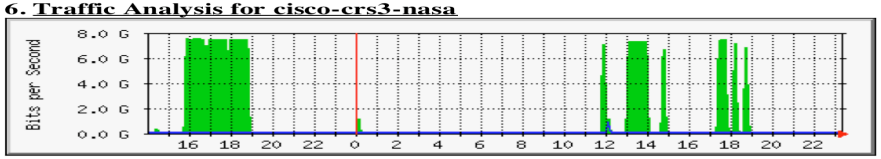
4. Traffic Analysis for i7test17



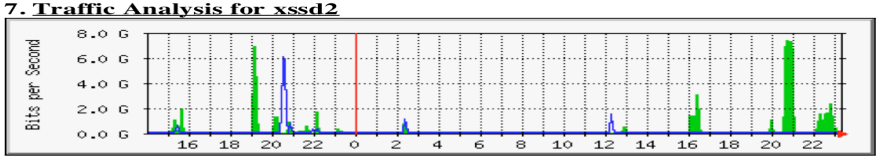
5. Traffic Analysis for hecn-7148-nasa2



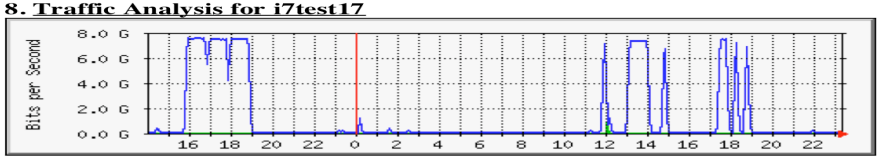
6. Traffic Analysis for cisco-crs3-nasa



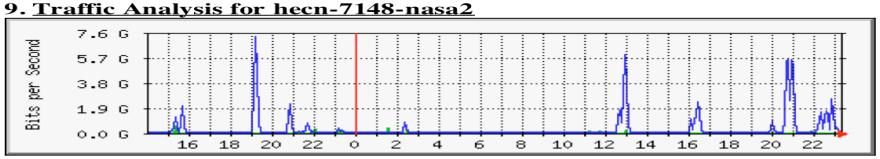
7. Traffic Analysis for xssd2



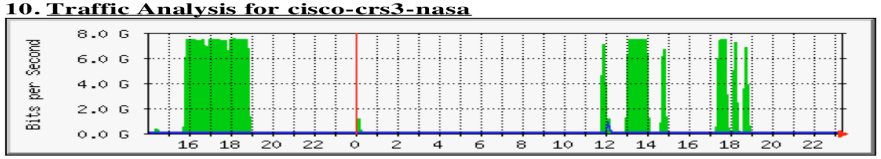
8. Traffic Analysis for i7test17



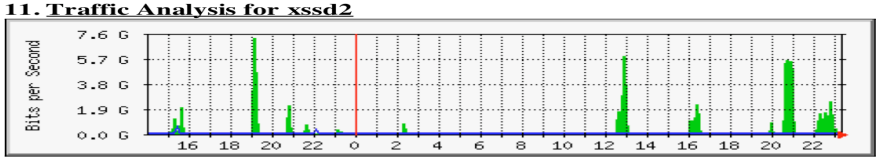
9. Traffic Analysis for hecn-7148-nasa2



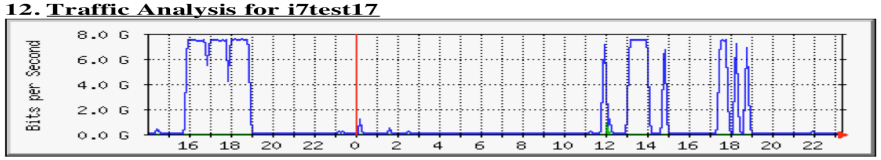
10. Traffic Analysis for cisco-crs3-nasa



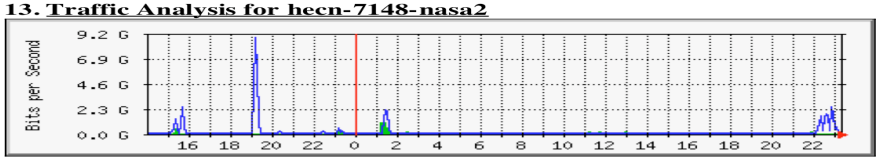
11. Traffic Analysis for xssd2



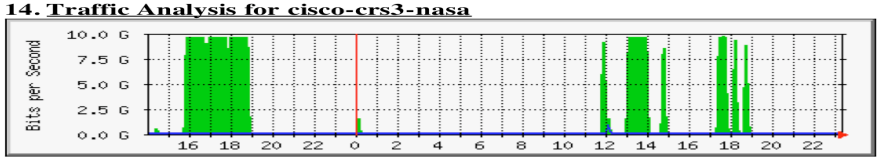
12. Traffic Analysis for i7test17



13. Traffic Analysis for hecn-7148-nasa2



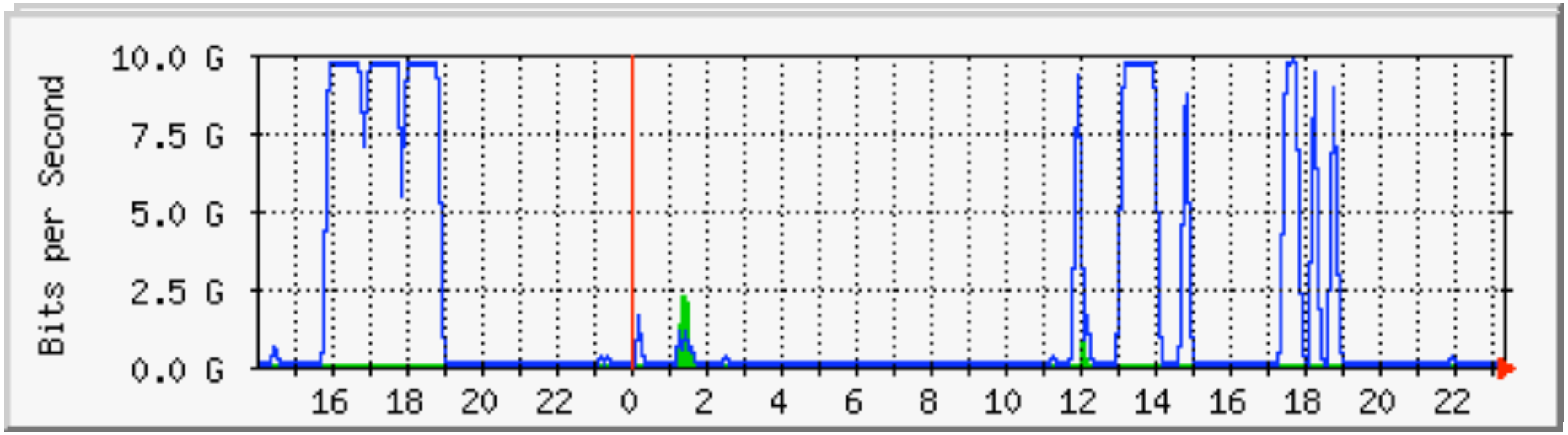
14. Traffic Analysis for cisco-crs3-nasa



A Sample MRTG-Generated Traffic Analysis From NASA's hecn-7148-nasa1 10-GE Switch In NASA's Exhibit Booth During SC10 Of NASA Workstation i7test17's 10-GE Network Interface #1

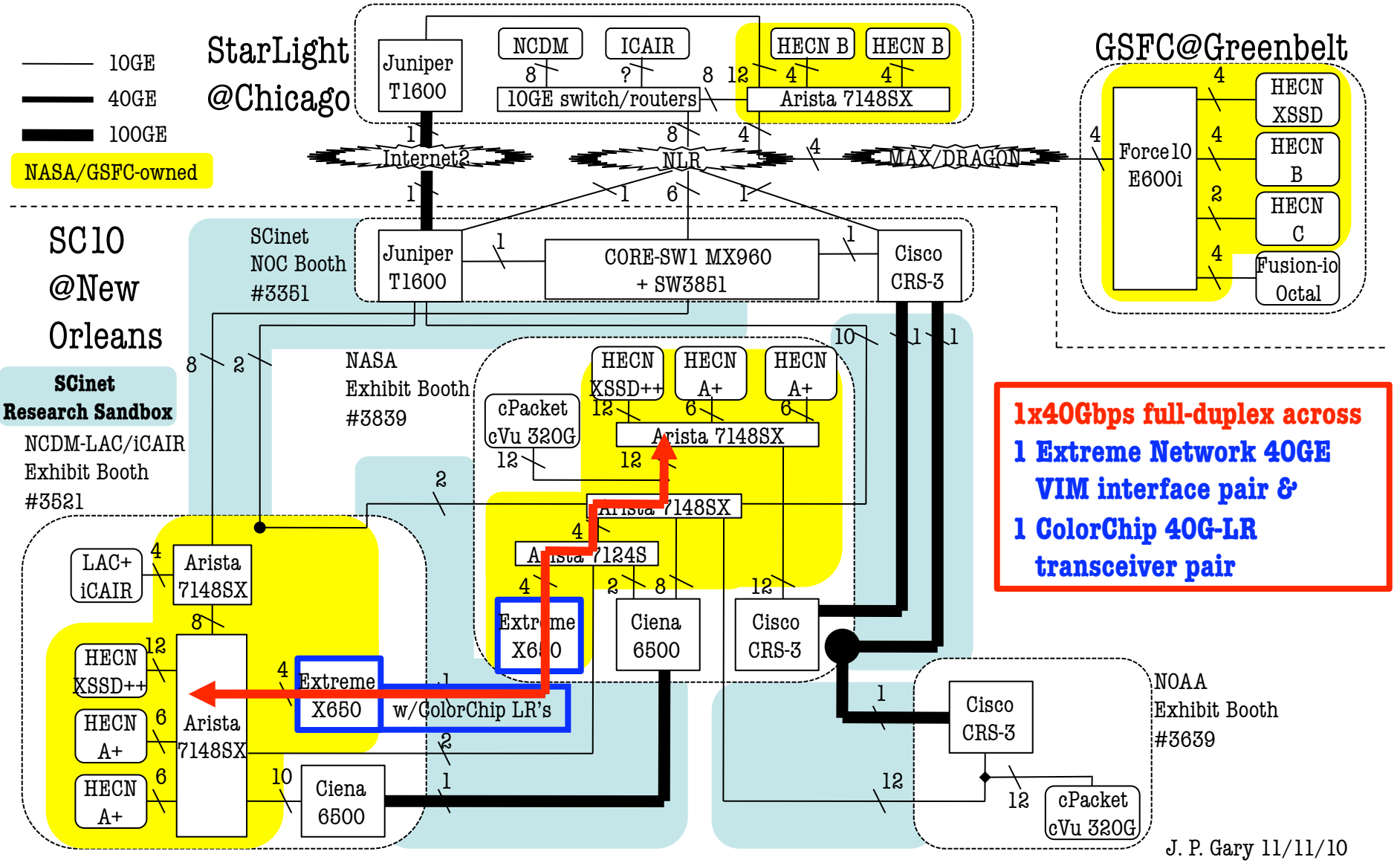
The statistics were last updated **Wednesday, 17 November 2010 at 22:23**

`Daily' Graph (5 Minute Average)



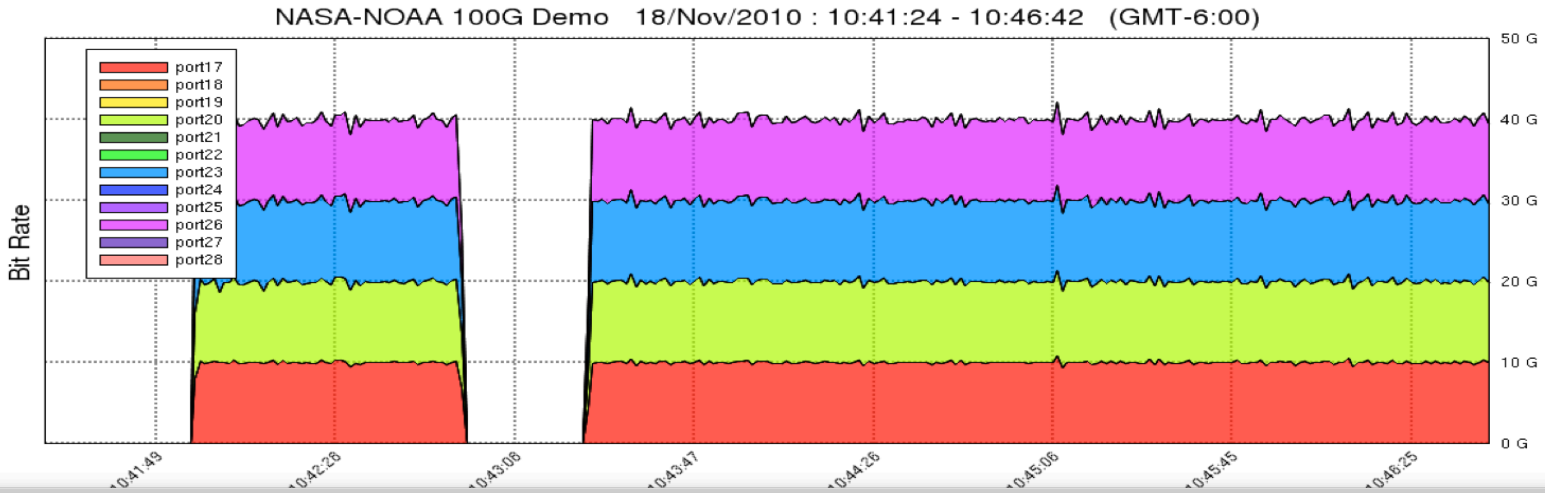
Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10



http://140.221.146.100:8100/live/dev__overview_140.221.148.30__nasa_graph_tx.html

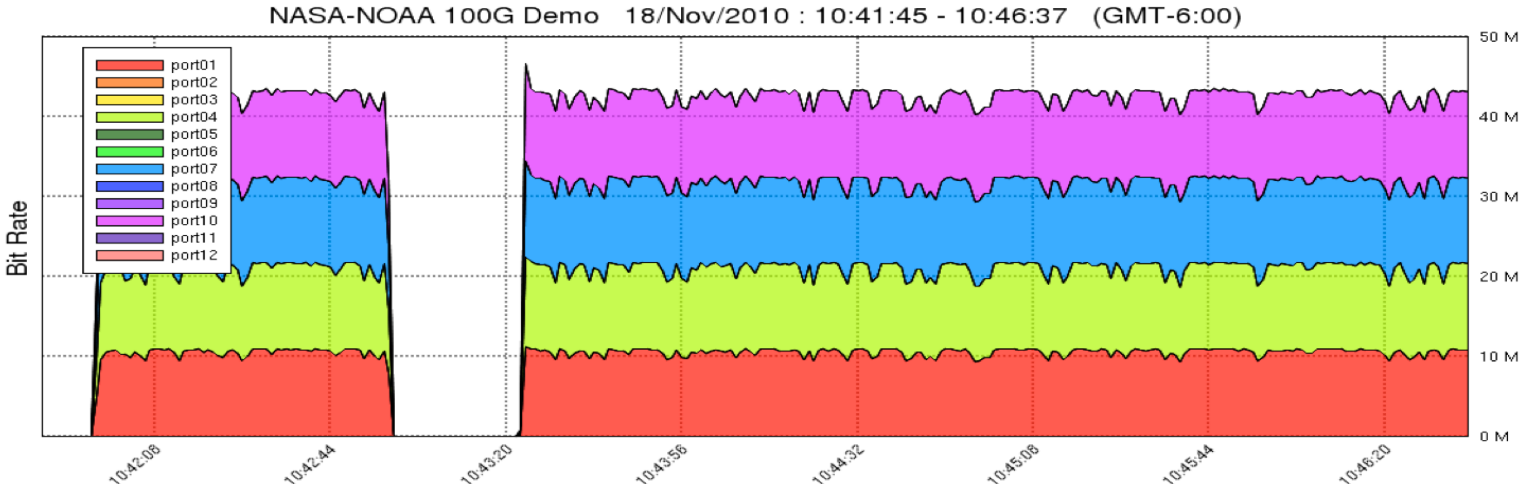
SC10 100G DEMO



Visualization Tools

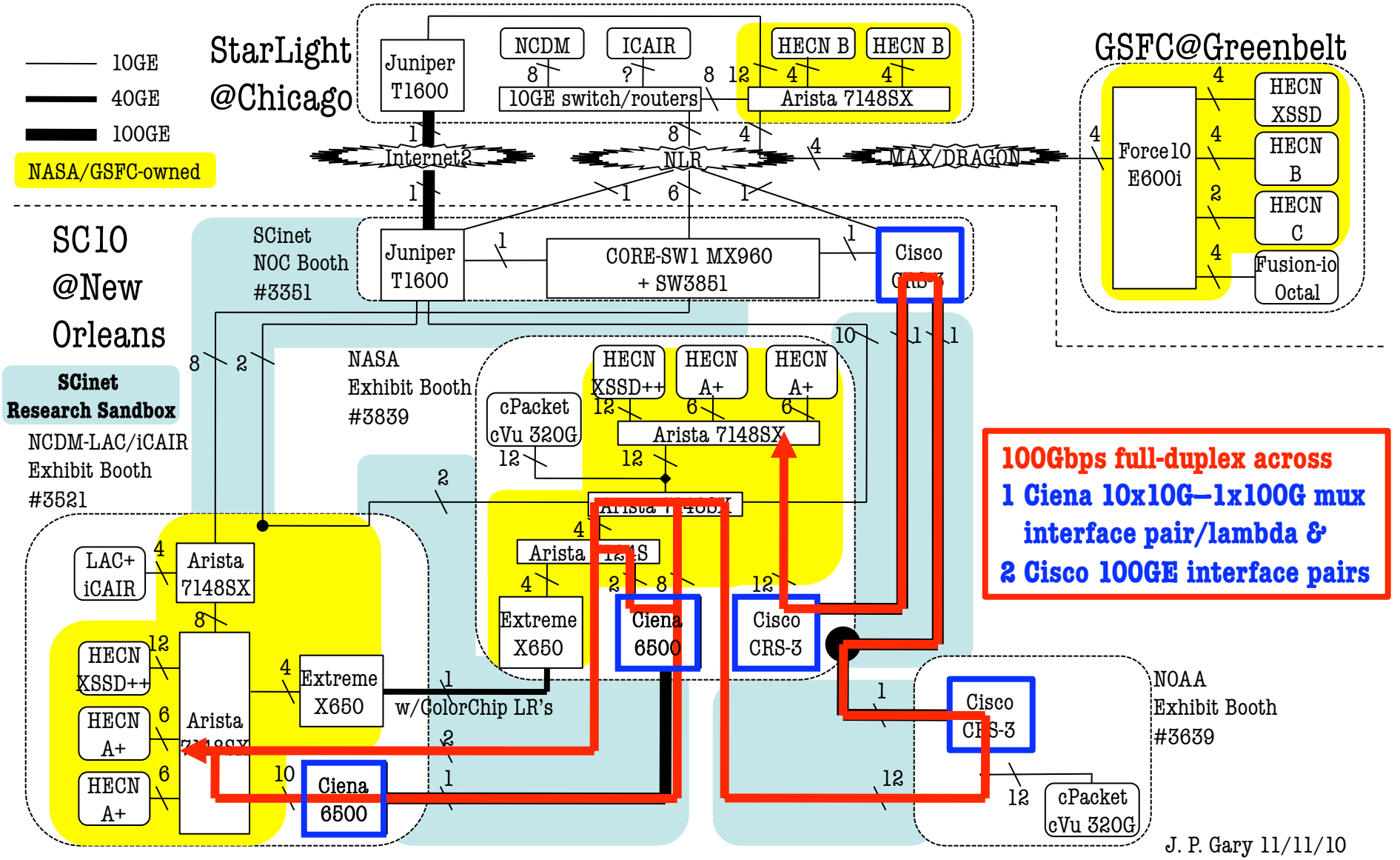
140.221.146.100:8100/live/dev__140.221.148.30.csv__nasa_graph2.html

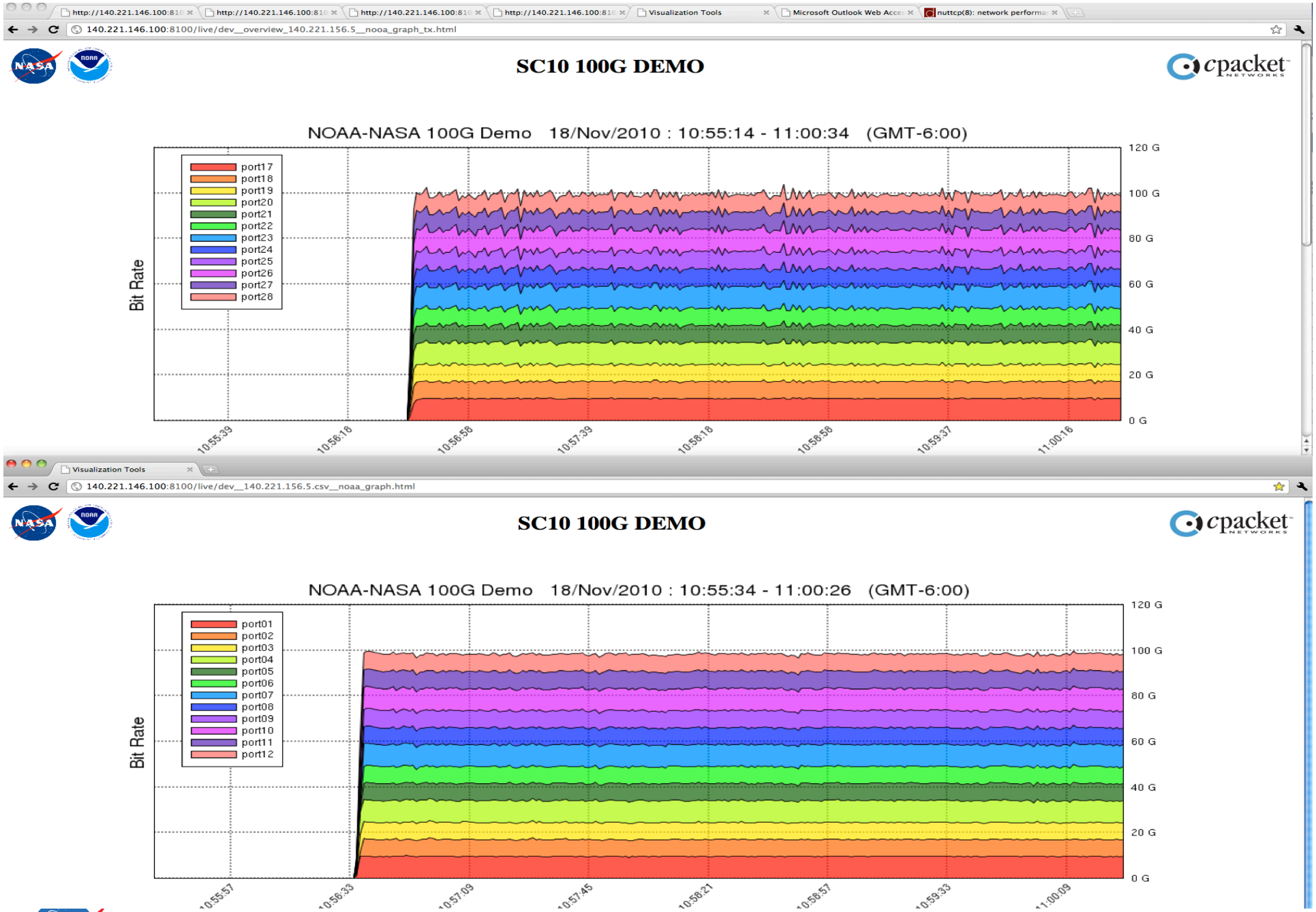
SC10 100G DEMO



Using 100G Network Technology in Support of Petascale Science

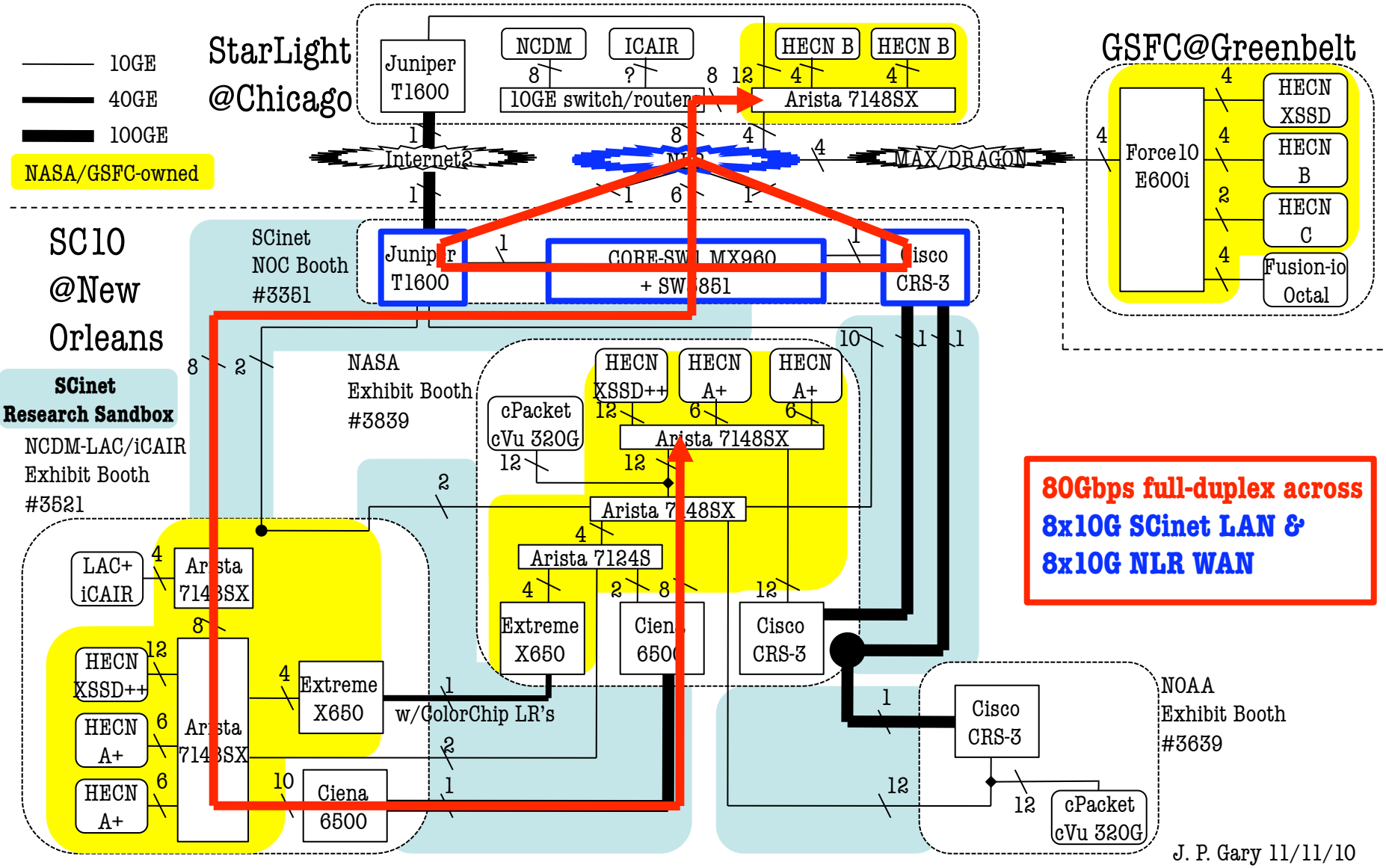
A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10

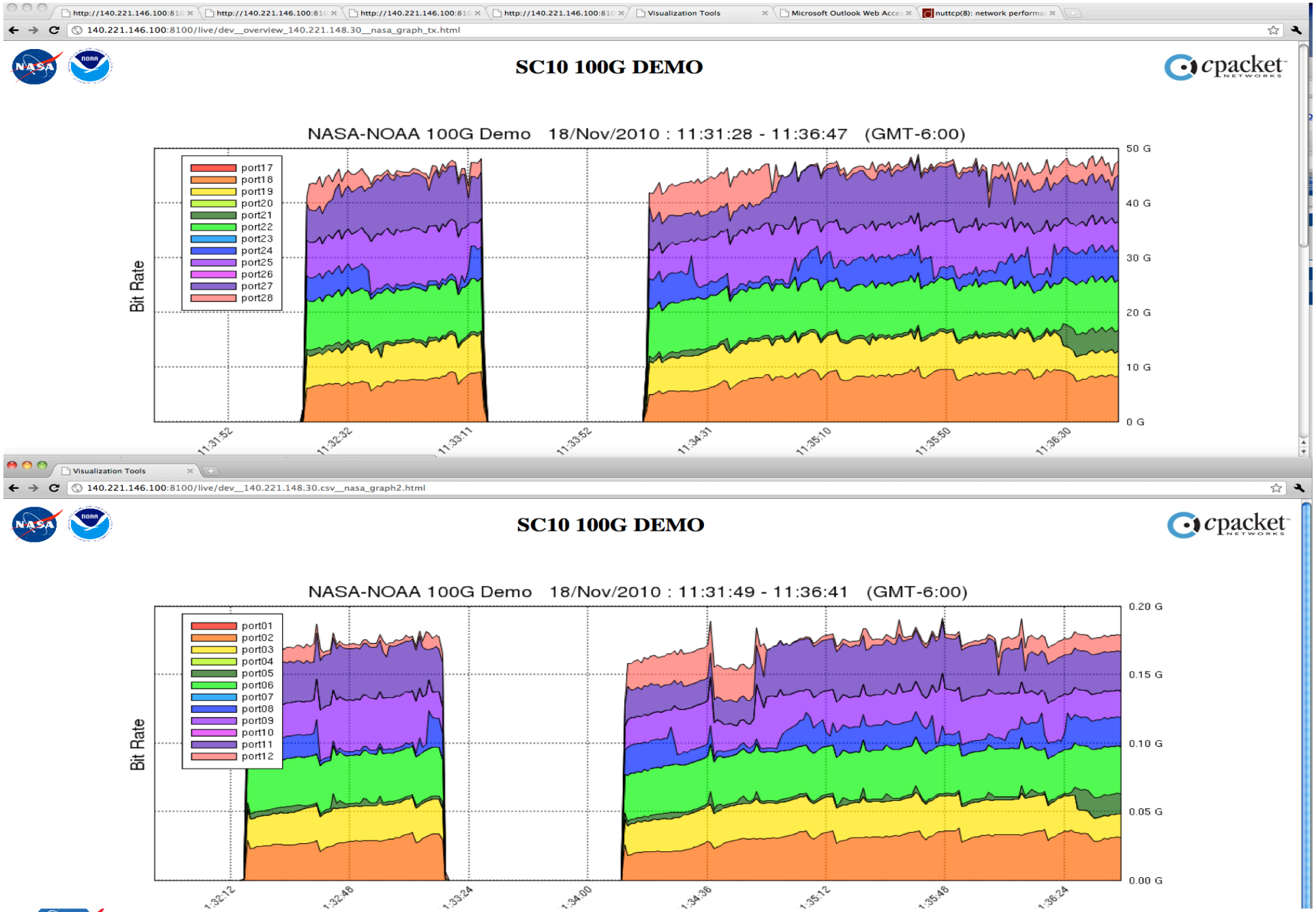



http://140.221.146.100:8100/live/dev__overview_140.221.156.5__nooa_graph_tx.html
SC10 100G DEMO
cpacket NETWORKS
NOAA-NASA 100G Demo 18/Nov/2010 : 10:55:14 - 11:00:34 (GMT-6:00)
Bit Rate
port17
port18
port19
port20
port21
port22
port23
port24
port25
port26
port27
port28
120 G
100 G
80 G
60 G
40 G
20 G
0 G
10:55:39
10:56:18
10:56:58
10:57:39
10:58:18
10:58:58
10:59:37
11:00:16
Visualization Tools
http://140.221.146.100:8100/live/dev__140.221.156.5.csv__noaa_graph.html
SC10 100G DEMO
cpacket NETWORKS
NOAA-NASA 100G Demo 18/Nov/2010 : 10:55:34 - 11:00:26 (GMT-6:00)
Bit Rate
port01
port02
port03
port04
port05
port06
port07
port08
port09
port10
port11
port12
120 G
100 G
80 G
60 G
40 G
20 G
0 G
10:55:57
10:56:33
10:57:09
10:57:45
10:58:21
10:58:57
10:59:33
11:00:09

Using 100G Network Technology in Support of Petascale Science

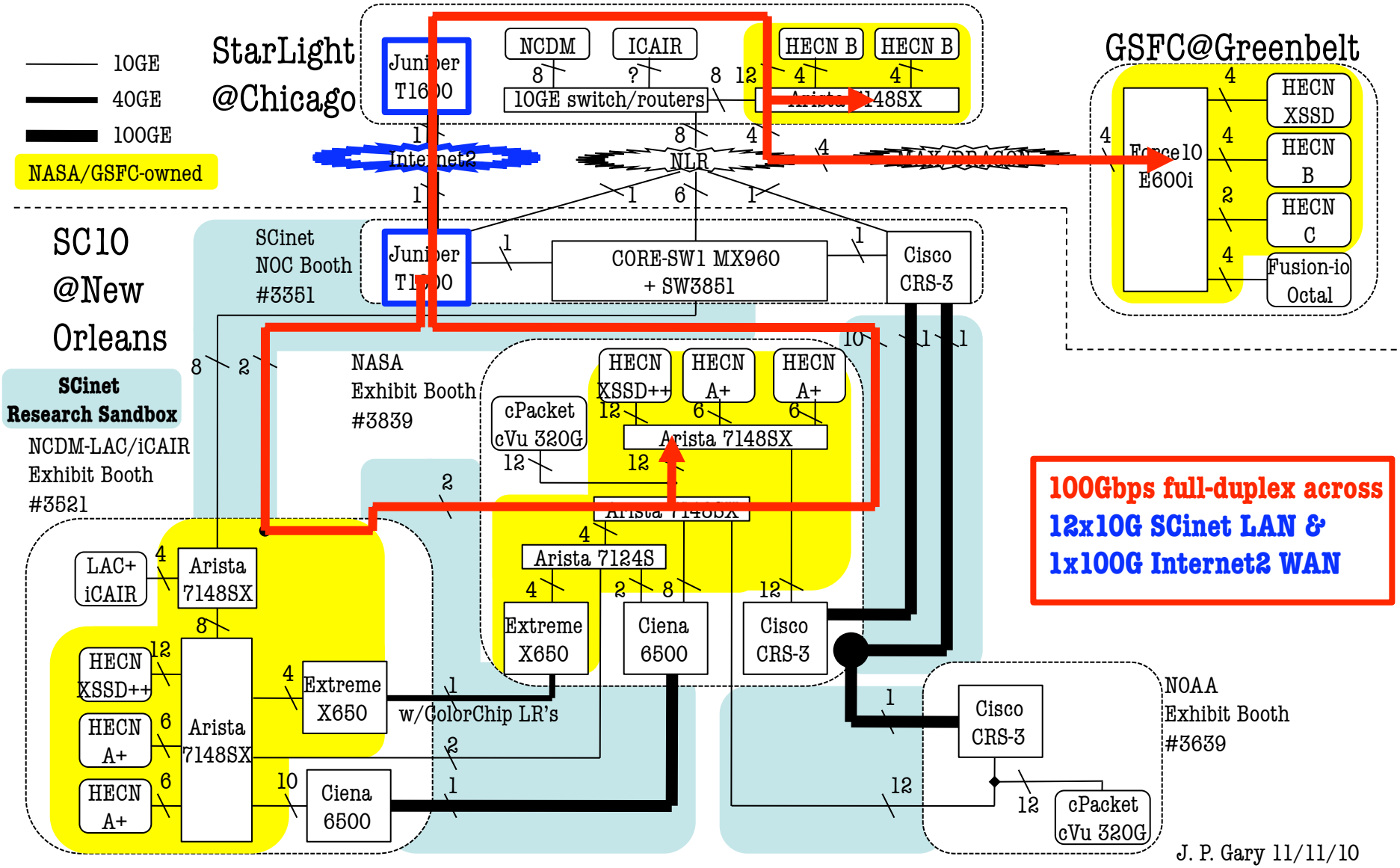
A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10

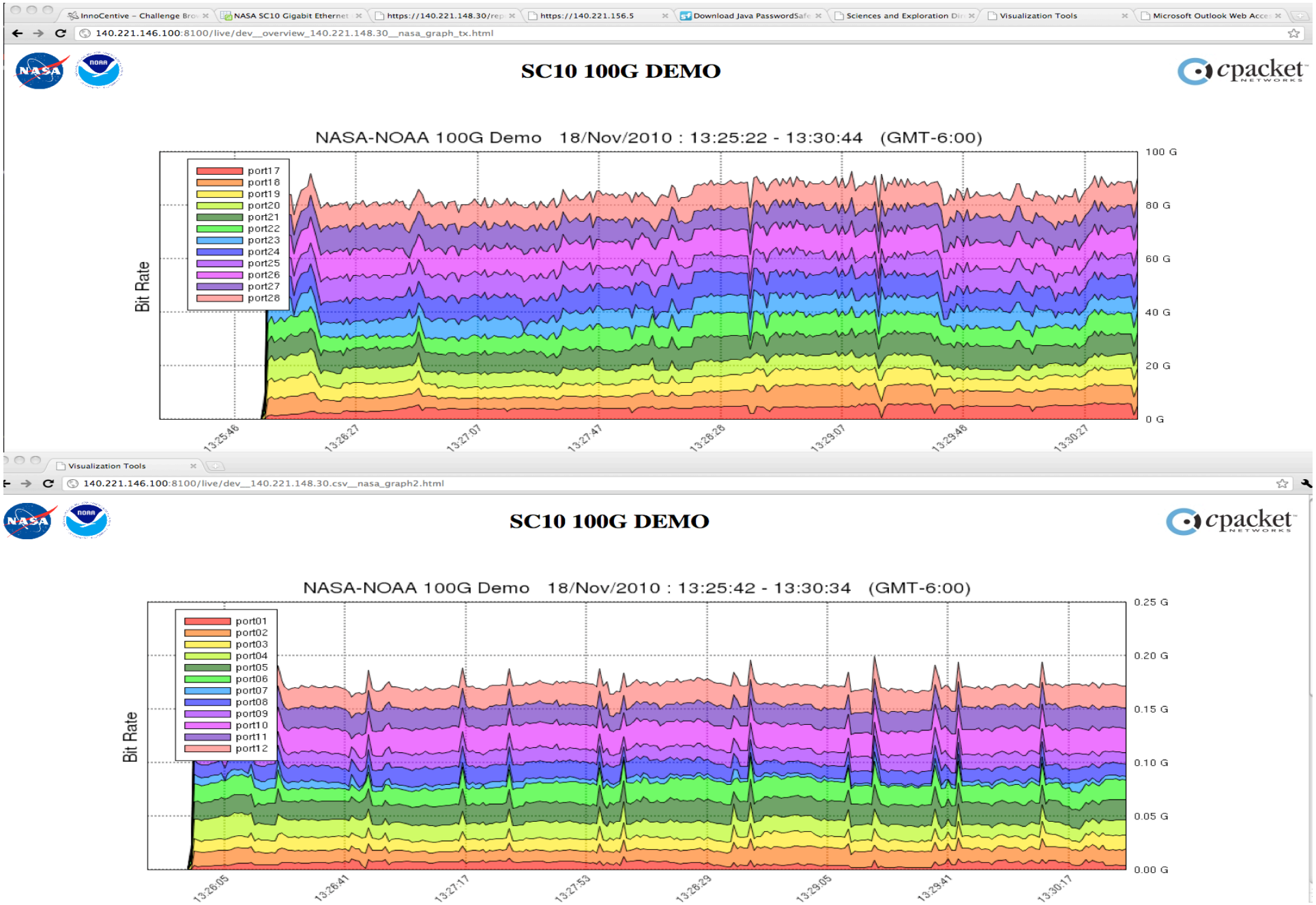



http://140.221.146.100:8100/live/dev__overview_140.221.148.30__nasa_graph_tx.html
SC10 100G DEMO
cpacket NETWORKS
NASA-NOAA 100G Demo 18/Nov/2010 : 11:31:28 - 11:36:47 (GMT-6:00)
Bit Rate
port17
port18
port19
port20
port21
port22
port23
port24
port25
port26
port27
port28
50 G
40 G
30 G
20 G
10 G
0 G
11:31:52
11:32:32
11:33:11
11:33:52
11:34:31
11:35:10
11:35:50
11:36:30
Visualization Tools
http://140.221.146.100:8100/live/dev__140.221.148.30.csv__nasa_graph2.html
SC10 100G DEMO
cpacket NETWORKS
NASA-NOAA 100G Demo 18/Nov/2010 : 11:31:49 - 11:36:41 (GMT-6:00)
Bit Rate
port01
port02
port03
port04
port05
port06
port07
port08
port09
port10
port11
port12
0.20 G
0.15 G
0.10 G
0.05 G
0.00 G
11:32:12
11:32:48
11:33:24
11:34:00
11:34:36
11:35:12
11:35:48
11:36:24

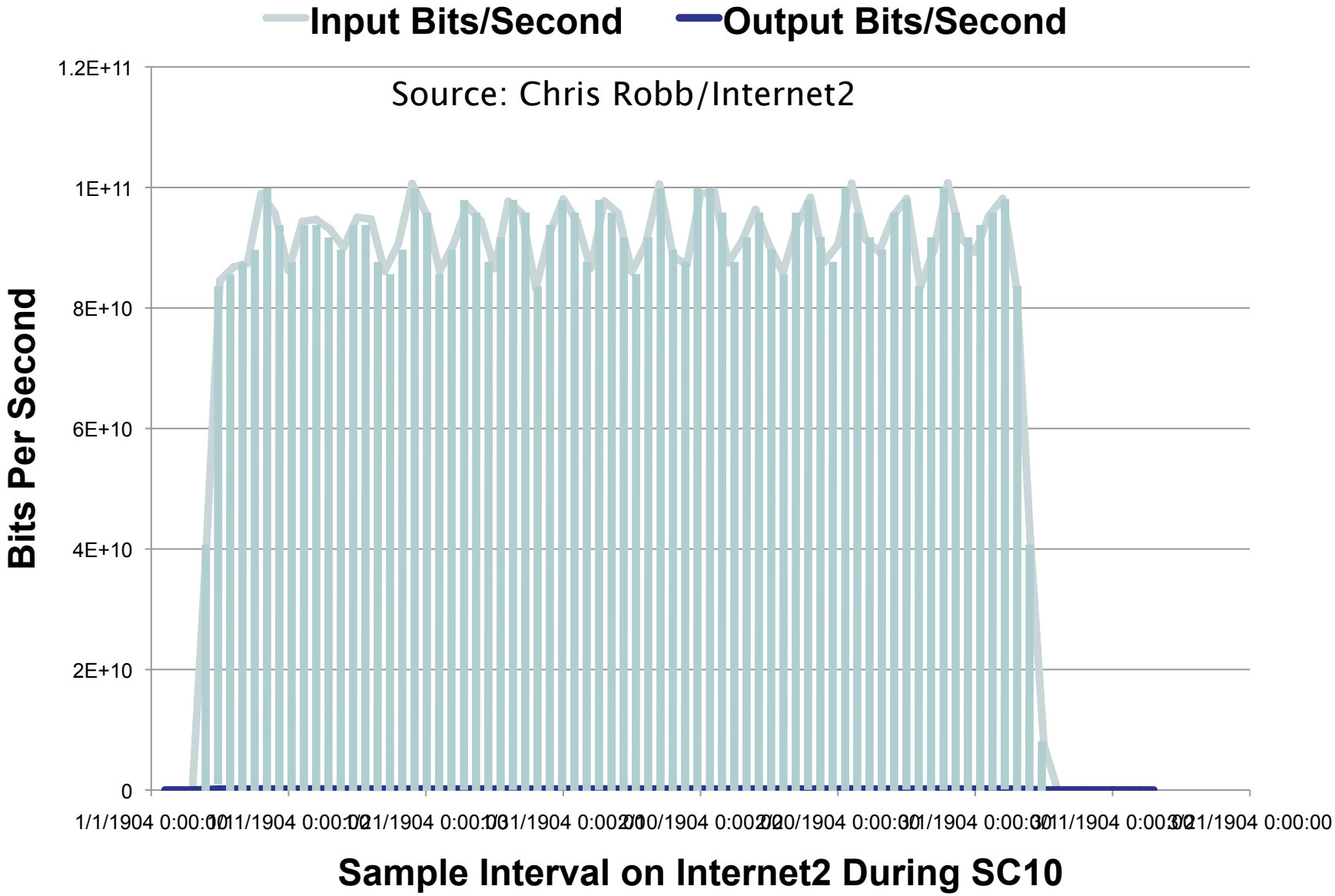
Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10



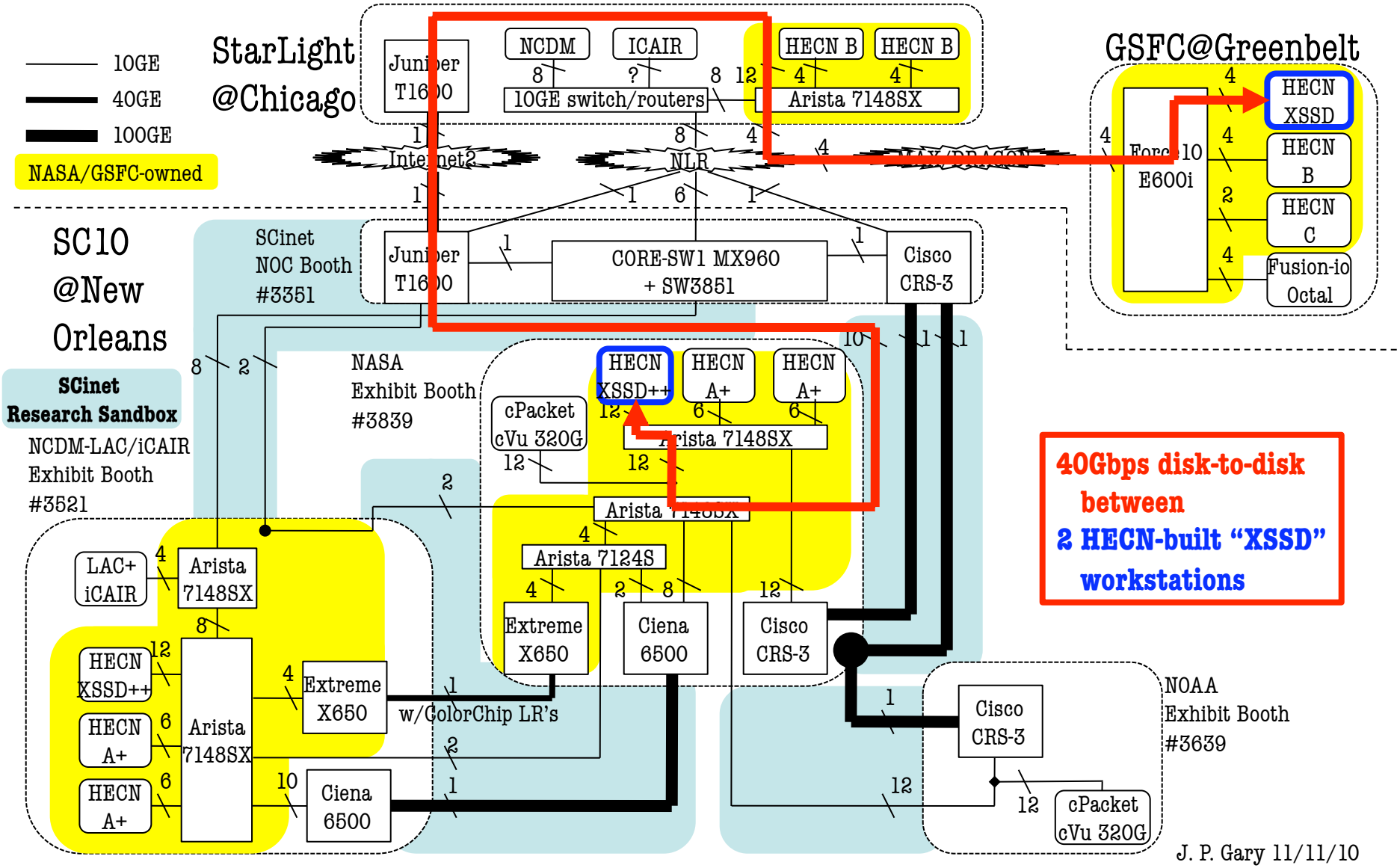

InnoCentive – Challenge Bro
NASA SC10 Gigabit Ethernet
https://140.221.148.30/rep
https://140.221.156.5
Download Java PasswordSafe
Sciences and Exploration Dir
Visualization Tools
Microsoft Outlook Web Acces
140.221.146.100:8100/live/dev__overview_140.221.148.30__nasa_graph_tx.html
SC10 100G DEMO
cpacket NETWORKS
NASA-NOAA 100G Demo 18/Nov/2010 : 13:25:22 - 13:30:44 (GMT-6:00)
Bit Rate
port17
port18
port19
port20
port21
port22
port23
port24
port25
port26
port27
port28
100 G
80 G
60 G
40 G
20 G
0 G
13:25:46
13:26:27
13:27:07
13:27:47
13:28:28
13:29:07
13:29:48
13:30:27
Visualization Tools
140.221.146.100:8100/live/dev__140.221.148.30.csv__nasa_graph2.html
SC10 100G DEMO
cpacket NETWORKS
NASA-NOAA 100G Demo 18/Nov/2010 : 13:25:42 - 13:30:34 (GMT-6:00)
Bit Rate
port01
port02
port03
port04
port05
port06
port07
port08
port09
port10
port11
port12
0.25 G
0.20 G
0.15 G
0.10 G
0.05 G
0.00 G
13:26:05
13:26:41
13:27:17
13:27:53
13:28:29
13:29:05
13:29:41
13:30:17

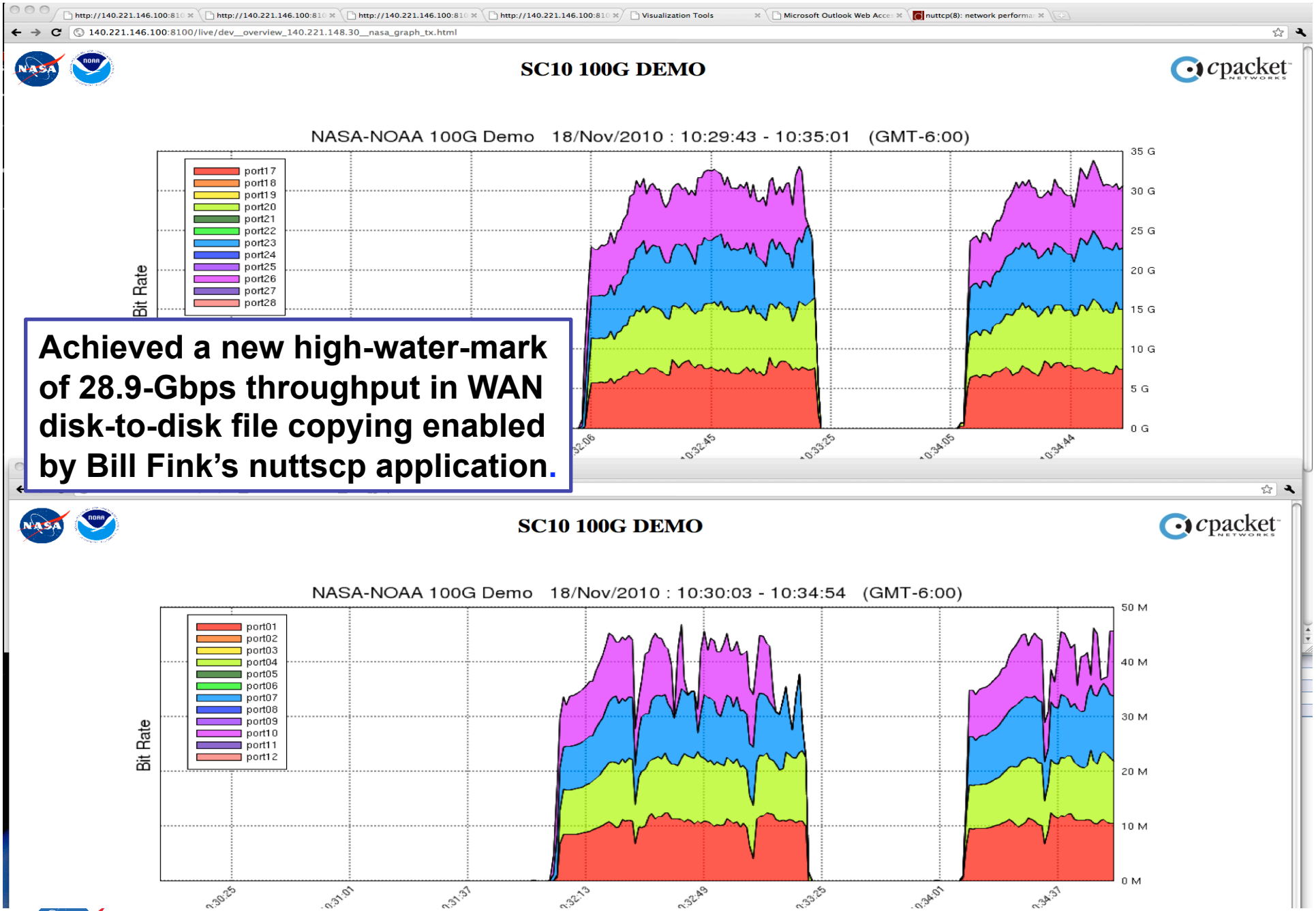
Internet2 100-Gbps Network Sample Interval During SC10
Showing NASA-Generated Input Traffic At 100-Gbps for 10-Minute Test Duration



Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10





Achieved a new high-water-mark of 28.9-Gbps throughput in WAN disk-to-disk file copying enabled by Bill Fink's nuttscp application.

SC10 Demonstration
"Using 100 Gbps Network Technology in Support of Petascale Science"

General Significance of This SC10 Demo

- "Existence proof" of emergent 40G and 100G products
  - Independent throughput performance testing by NASA and its "infrastructure-owner" partners in a semi-public forum
  - All of the vendor products individually operated at their advertised 40G or 100G wire-speeds
  - Direct or "indirect" interoperability was sustained in a complex mixed vendor system
- Modern workstation potential demonstrated, to its limits
  - >100-Gbps aggregate transmitted across PCIe G2 backplane (used two x16 6x10GE NICs, typically averaged ~8.7-Gbps on 12x10GEs)
  - Essentially saturated distributed processing of two 3.3-GHz Nehalem quad-cores
  - ~30-Gbps disk-to-disk over WAN achieved with only two controllers each with 8 SSDs, with limited "tuning" (50-Gbps may be possible with "hero" tuning, or next-gen components)

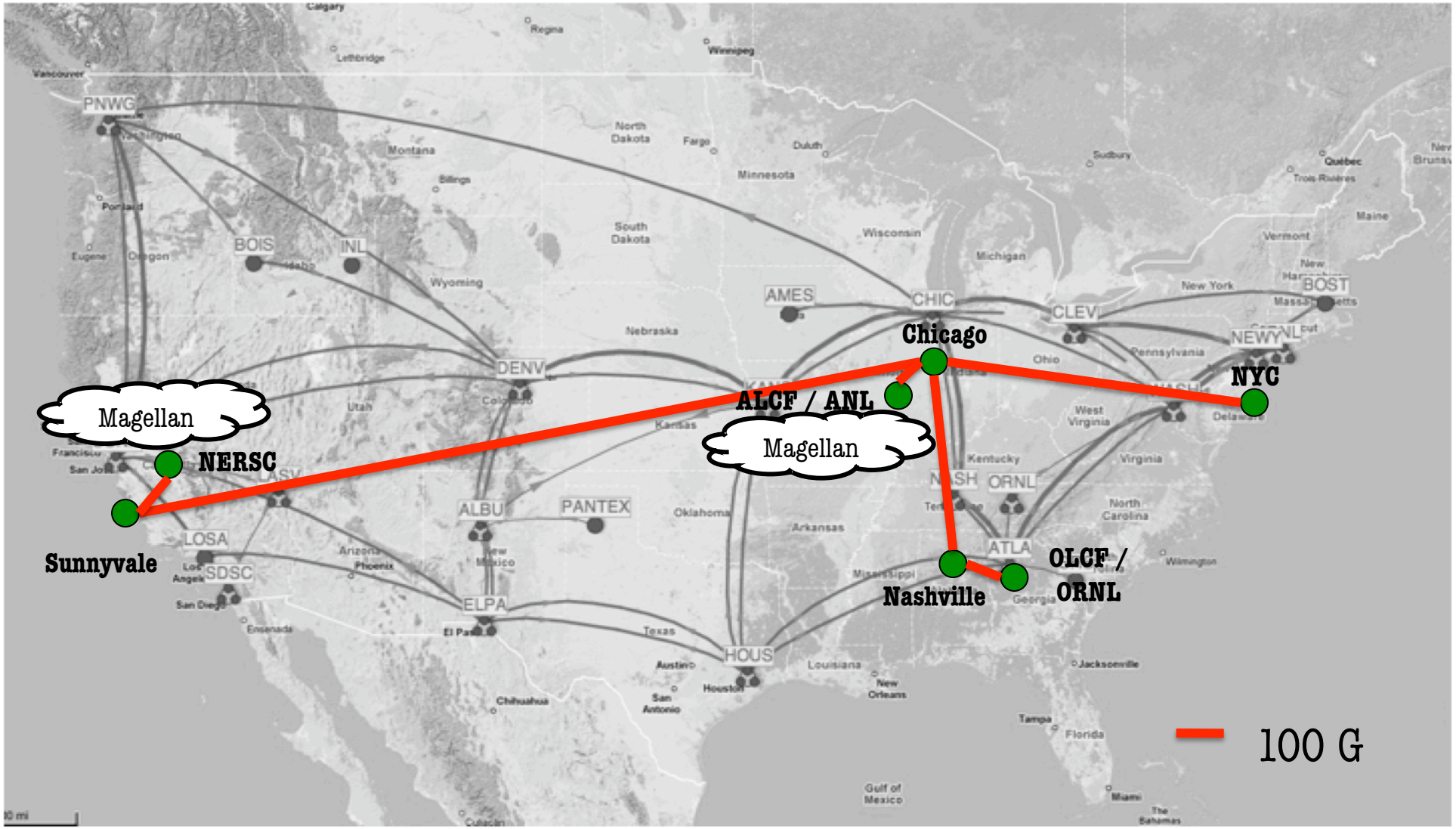
SC10 Demonstration "Using 100 Gbps Network Technology in Support of Petascale Science"

Preparing Future Plans

- Need to share and apply knowledge gained in achieving greater than the Phase 1 Goal of GSFC High End Computing 20, 40 &100 Gbps Network Testbeds initiative: 20-Gbps WAN disk-to-disk user-throughput
  - NASA: NCCS climate simulation data flows for IPCC AR5
  - NOAA: Extreme data flows between GFDL and ORNL
  - DoE: Assist in assessing throughput performance of DoE Advanced Network Initiative (ANI)'s 100G Prototype Network
- Need to refine approaches to achieving the respective Phase 2 & 3 Goals of GSFC High End Computing 20, 40 &100 Gbps Network Testbeds initiative: 40- & 100-Gbps WAN disk-to-disk user-throughput
  - File copying applications
  - Workstations with NICs and disk controllers using PCIe Gen 3
  - WAN test infrastructure

Source: Brian Tierney (LBL/ESnet)

Nationwide 100G Prototype Network



SC10 Demonstration "Using 100 Gbps Network Technology in Support of Petascale Science"

Preparing Future Plans – Near Term

- Replace HECN's two "B" net-test workstations at StarLight
  - Upgrade to two XSSDs: each >100G memory-to-memory, ~30G disk-to-disk
  - Assist DoE with checkout of their ANI 100G Prototype Network
  - Assist StarLight with checkout of their NSF-awarded 100G-capable upgrades in part of StarLight's infrastructure
- Upgrade DRAGON's lambda pathways between GSFC and McLean in cooperation with the MAX
  - Replace ADVA-based 10-Gbps DWDM
  - Assist MAX with checkout of their NSF-awarded 100G-capable upgrades in part of MAX's infrastructure
- Leverage Internet2's and/or NLR's 100G pathways between McLean and StarLight
  - https://lists.internet2.edu/sympa/arc/i2-news/2010-11/msg00003.html
  - http://www.nlr.net/release.php?id=62

SC10 Demonstration
"Using 100 Gbps Network Technology in Support of Petascale Science"

Reference URL Summary

- This presentation is soon to be available
  - http://science.gsfc.nasa.gov/606.1/docs/SC10_HECN-demo_ON_VECTOR_030211.pdf

- HECN Team Demo at SC10 in Context of 20, 40 &100 Gbps Network Testbeds
  - http://science.gsfc.nasa.gov/606.1/docs/SC10_HECN-demo-summary_011011.pdf

SC10 Demonstration
"Using 100 Gbps Network Technology in Support of Petascale Science"

Q & A

Thank You!

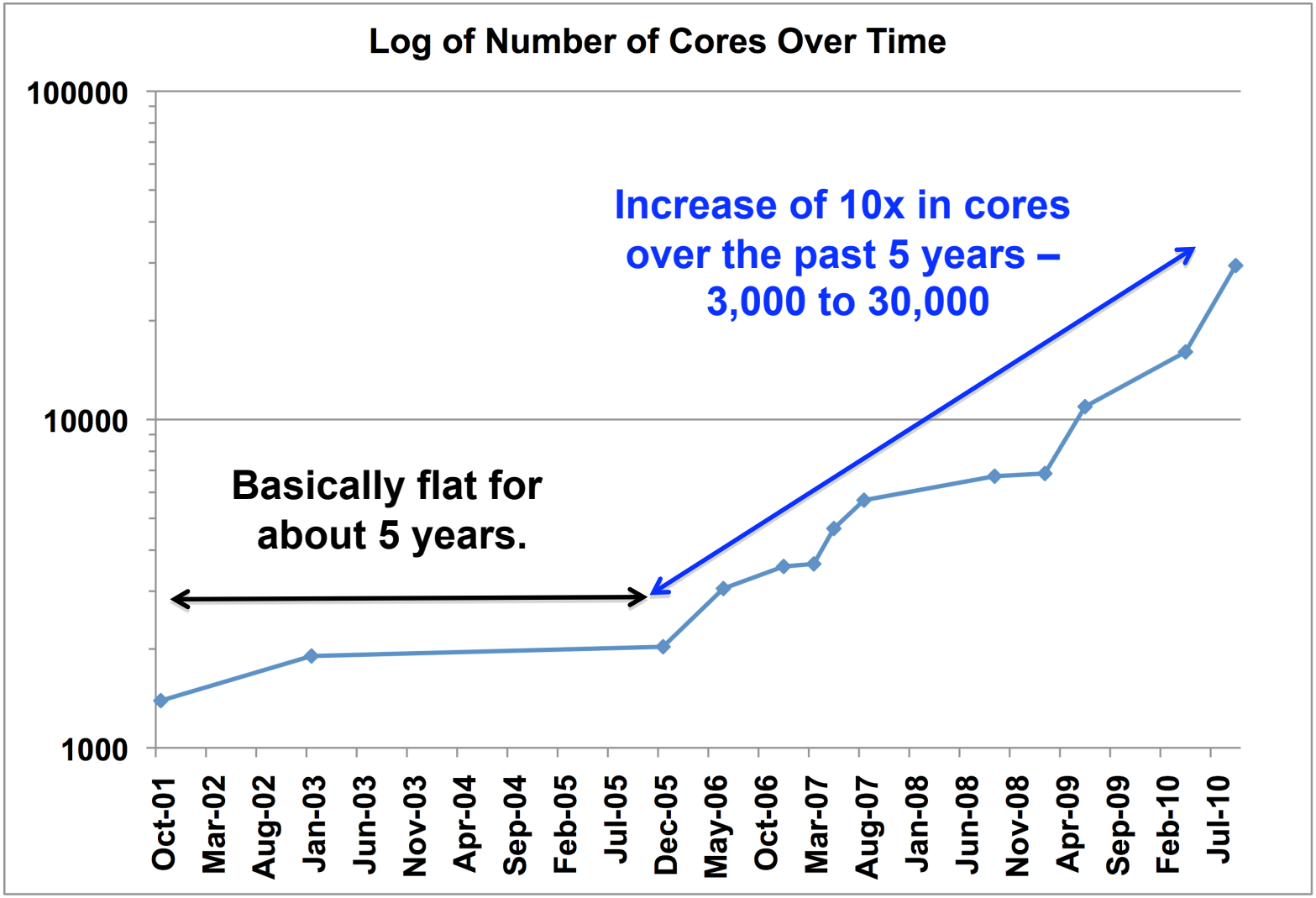
SC10 Demonstration
"Using 100 Gbps Network Technology in Support of Petascale Science"

Backup Slides

Growth of Computing Cores Over the Last 10 Years

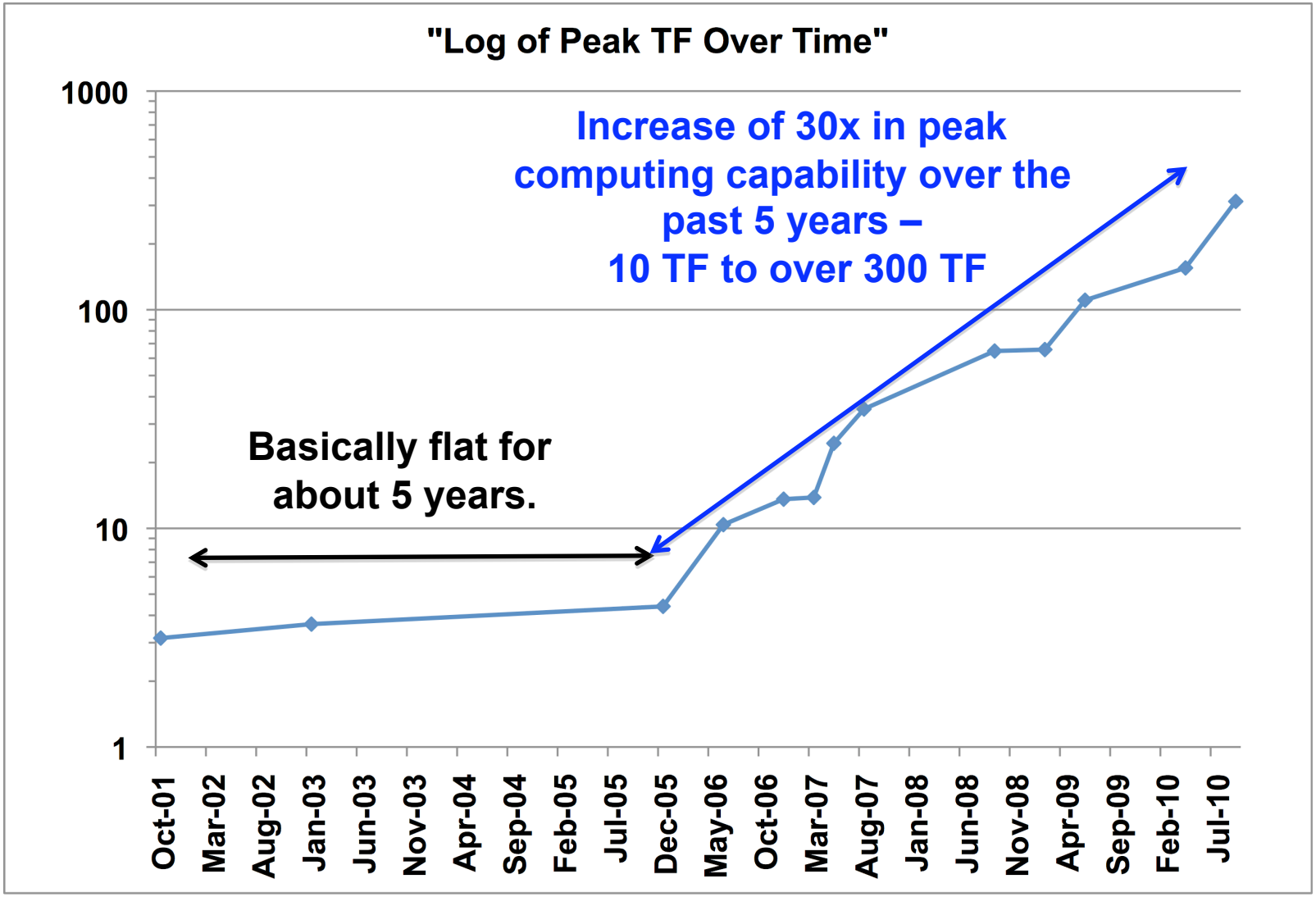
Source: Dan Duffy (GSFC/NASA Center for Climate Simulation (NCCS))

NCCS User Forum Sep. 14, 2010



Growth of Peak Computing Capability Over the Last 10 Years

Source: Dan Duffy (GSFC/NCCS), NCCS User Forum Sep. 14, 2010



Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Key Considerations (1 of 2)

- Use of Layer 1+2 DCN-enabled VLANs versus Layer 1+2+3 full IP routed networks in both the Regional/MAN and WAN testbeds is critical
  - Sufficient to enable more effort to be focused on the primary subjects of this effort which are the processor interfaces and LAN infrastructure needed on the ends of the intervening links
  - Core IP routing issues (while otherwise interesting with many R&D challenges remaining) are not the primary subject of this effort
  - Costs of 40 and 100 Gbps Layer 3 router interfaces are likely to be two or more orders of magnitude greater than 40 and 100 Gbps Layer 2 Ethernet switch interfaces which are sufficient to enable the needed VLANs

Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Key Considerations (2 of 2)

- "Suitable" near-typical server/workstations used by the science community are essential to the acceptance of the testbed findings
- HECN Team has chosen to iteratively specify and then assemble net-test server/workstations to achieve the disk-to-disk throughput performance goals of the respective Phases
  - Lowest cost approach
  - Update specs and assemble new server/workstations to overcome newly discovered bottlenecks
  - Memory-to-memory throughput tests significantly help to calibrate the infrastructure's wire-speed

10-Gbps Disk-to-Disk File Copies Achieved Via Workstations Costing Less Than $7,000

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network Team specified and assembled workstations that individually costs less than $7,000 and are capable of over 10 gigabits per second (Gbps) disk-to-disk file copying.
- Each workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with two HighPoint RocketRaid 4320 RAID disk controllers and a Myricom 10 Gigabit Ethernet network interface card in the PCIe Gen2 slots of a Asus P6T6 WS Revolution motherboard. Each RAID controller hosts eight Western Digital WD5001AALS 500-gigabyte disks.
- Over 10-Gbps disk-to-disk file-copying throughput between two of the workstations was measured using the nuttscp (www.nuttcp.net) file copying tool.
- Demonstrations of these workstations supporting network-performance testing, wide-area file systems, and file transfer applications ranging from traditional to experimental are planned in the NASA research exhibit at the SC09 conference, Portland, OR, November 16–19 .

***Figure*:** Two Core i7 workstations interconnected via 10 Gigabit Ethernet in test configuration prior to shipping to SC09.

POC: Bill Fink, William.E.Fink@nasa.gov, (301) 286-7924, GSFC Computational and Information Sciences and Technology Office

17.8-Gbps Disk-to-Disk File Copies Achieved Via Workstations Costing Less Than $9,000

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network Team specified and assembled workstations that individually costs less than $9,000 and are capable of over 17.8 gigabits per second (Gbps) disk-to-disk file copying.
- Each workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with four HighPoint RocketRaid 4320 RAID disk controllers and a Myricom 2-port 10 Gigabit Ethernet network interface card in the PCIe Gen2 slots of a Asus P6T6 WS Revolution motherboard. Each RAID controller hosts eight Western Digital WD5001AALS 500-gigabyte disks.
- Over 17.8-Gbps disk-to-disk file-copying throughput between two of the workstations was measured using the nuttscp (www.nuttcp.net) file copying tool.
- While SSD technology is next to be investigated, parallelism of multiple cores and multiple streams is likely to be key to going to 40-Gbps and beyond, since individual cores are not getting significantly faster.

***Figure*:** Right case houses Core i7 cores, DDR3 memory, NIC, two "internal" controllers each with eight disks and two "external" controllers; left case houses sixteen SAS-connected disks.

POC: Bill Fink, William.E.Fink@nasa.gov, (301) 286-7924, GSFC Computational and Information Sciences and Technology Office

100 Gigabits per Second Transmissions Achieved Via A Single Workstation

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network (HECN) Team specified and assembled a workstation that costs less than $11,000 and is capable of over 100 gigabits per second (Gbps) data transmission – 10 times the transmission speed of most high end computers.

- The workstation consists of a 3.2-GHz dual-processor (quad core) Intel Xeon W5580 (Nehalem) with six Myricom dual-port 10-Gigabit Ethernet network interface cards in the PCIe Gen2 slots of a Supermicro X8DAH+-F motherboard.

- Over 100-Gbps aggregate-throughput transmissions from the Xeon-workstation to two Intel Core i7 workstations (also specified and assembled by the HECN Team) were measured using the nuttcp (www.nuttcp.net) network-performance testing tool.

- Demonstrations of these workstations supporting network-performance testing, wide-area file systems, and file transfer applications ranging from traditional to experimental are planned in the NASA research exhibit at the SC09 conference, Portland, OR, Nov. 16–19 .

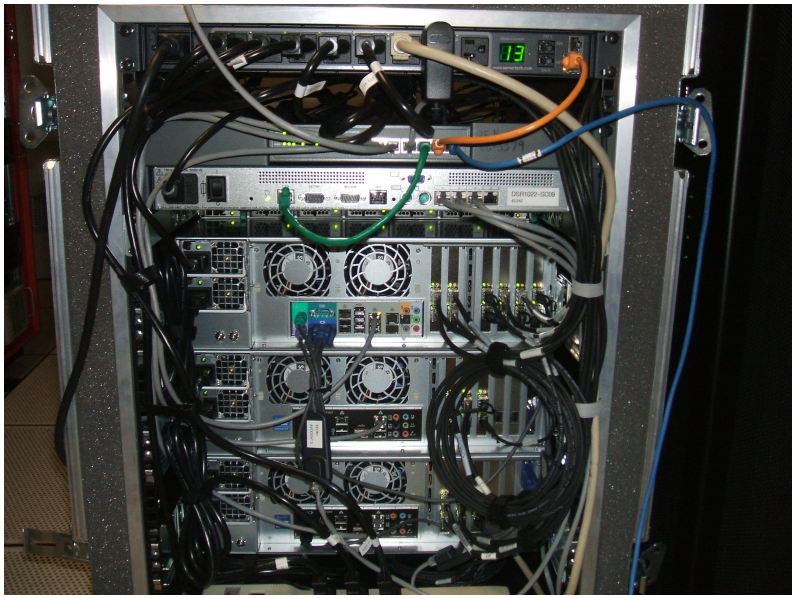


Figure: Xeon and two Core i7 workstations (bottom) interconnected with 10 Gigabit Ethernet switch and management units (top) in a rack for shipping to SC09.

POC: Bill Fink, William.E.Fink@nasa.gov, (301) 286-7924, GSFC Computational and Information Sciences and Technology Office

Aggregate 55+ Gigabits per Second (Gbps) Transmits, 52+ Gbps Receives and 75+ Gbps Bi-Directional Transmissions Achieved Via A Single Workstation With a Single 6x10-Gigabit Ethernet Network Interface Card

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network (HECN) Team tested a HotLava six-port 10-Gigabit Ethernet network interface card (NIC) in a HECN Team-assembled workstation that costs less than $ 6,800 with the NIC and achieved aggregate 55+ Gbps transmits, 52+ Gbps receives and 75+ Gbps bi-directional memory-to-memory data transmissions.

- The workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with one HotLava NIC in one PCIe Gen2 x16 slot of an Asus P6T6 WS Revolution motherboard.

- Transmissions between the above workstation and two other HECN Team-assembled Intel Core i7 workstations with other NICs were measured using the nuttcp (www.nuttcp.net) network-performance testing tool.

- Demonstrations of similar workstations supporting 100 Gbps network testing and near-40 Gbps file transfer applications are planned in the NASA research exhibit at the SC10 conference, New Orleans, LA, Nov. 15–18.

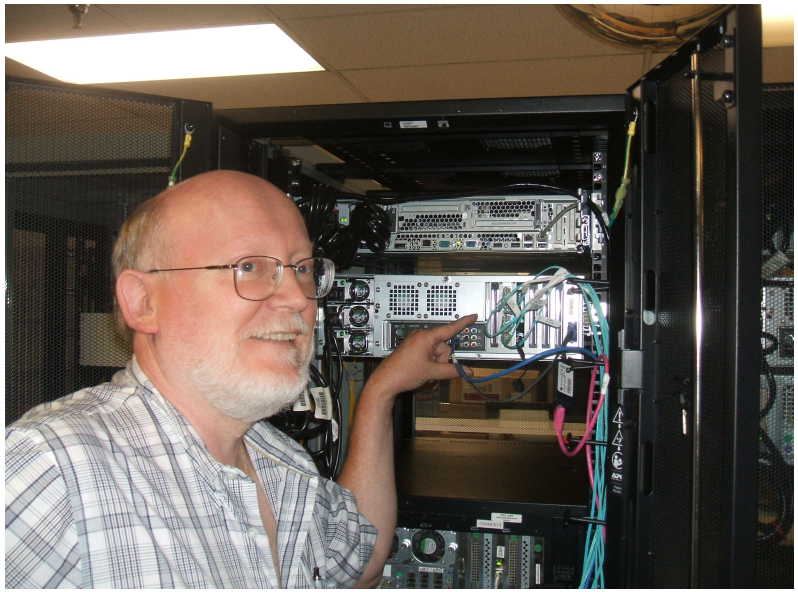


***Figure*:** Bill Fink, author of nuttcp and the throughput performance tests, pointing to the 6x10GE HotLava NIC in the HECN Team's Intel Core i7 based workstation.

POC: Bill Fink, Bill.Fink@nasa.gov,
GSFC Computational and Information Sciences and Technology Office

Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

"XSSD1++" Server Approximate Costs *(With components acquired via SEWP IV in lot-sizes of 3 - 15, and self assembly. Source: Paul Lang (ADNET))* ***(1 of 2)***

- Supermicro 836TQ-R800B 3u 16bay 800W RPS chassis $828
- Supermicro X8DAH+-F motherboard $506
- Intel X5590 4-core 3.3GHz Xeon proc. $1,612 X 2 = $3,224
- 3X2GB 1600 MHz DDR3 memory $227 X 2 = $454
- CBL-0084 front pannel cable $3
- 12" 3pin fan extension cable $1
- ArkTech slim IDE DVD to SATA adapter $10
- HotLava Tanbora 6xSFP+ NIC 6ST2A30A1F1 $1,401 X 2 = $2,802
- Dynatron G666 CPU cooler $35 X 2 = $70
- 2.5" SATA system disk (WD2500BEKT 250GB) $60
- Red Greatland 18" Slimline SATA adapter $6
- Supermicro MCP-220-83601-0B FDD tray for 2.5" disk $8
- PCIe Video card eVGA GeForce 8400GS $40
- 8" 8pin power extension cable $8

Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

"XSSD1++" Server Approximate Costs *(With components acquired via SEWP IV in lot-sizes of 3 - 15, and self assembly. Source: Paul Lang (ADNET))* ***(2 of 2)***

- LSI MegaRAID 9261-8i Raid Controllers $460
- LSI MegaRAID 9280-8e Raid Controllers $660 X 3 = $1,920
- Supermicro 216-R900LPB chassis 2u, 24x2.5"bay $891
- OCZ Vertex 2 EX 50GB SLC SSD $837 x 32 = $26,784
- 2.5" to 3.5" adapter (IcyDock MB882SP-1S-2B) $12 x 8 = $96
- Dual SFF-8087/SFF-8088 (CoolDrives 36Hx2-26TX2) $49 x 3 = $147
- SAS to 4SATA cable (3ware CBL-SFF8087OCF-06M) $16
- SAS-8888-05m .5m SFF-8088 SAS cable $42 x 3 = $126

$38,456

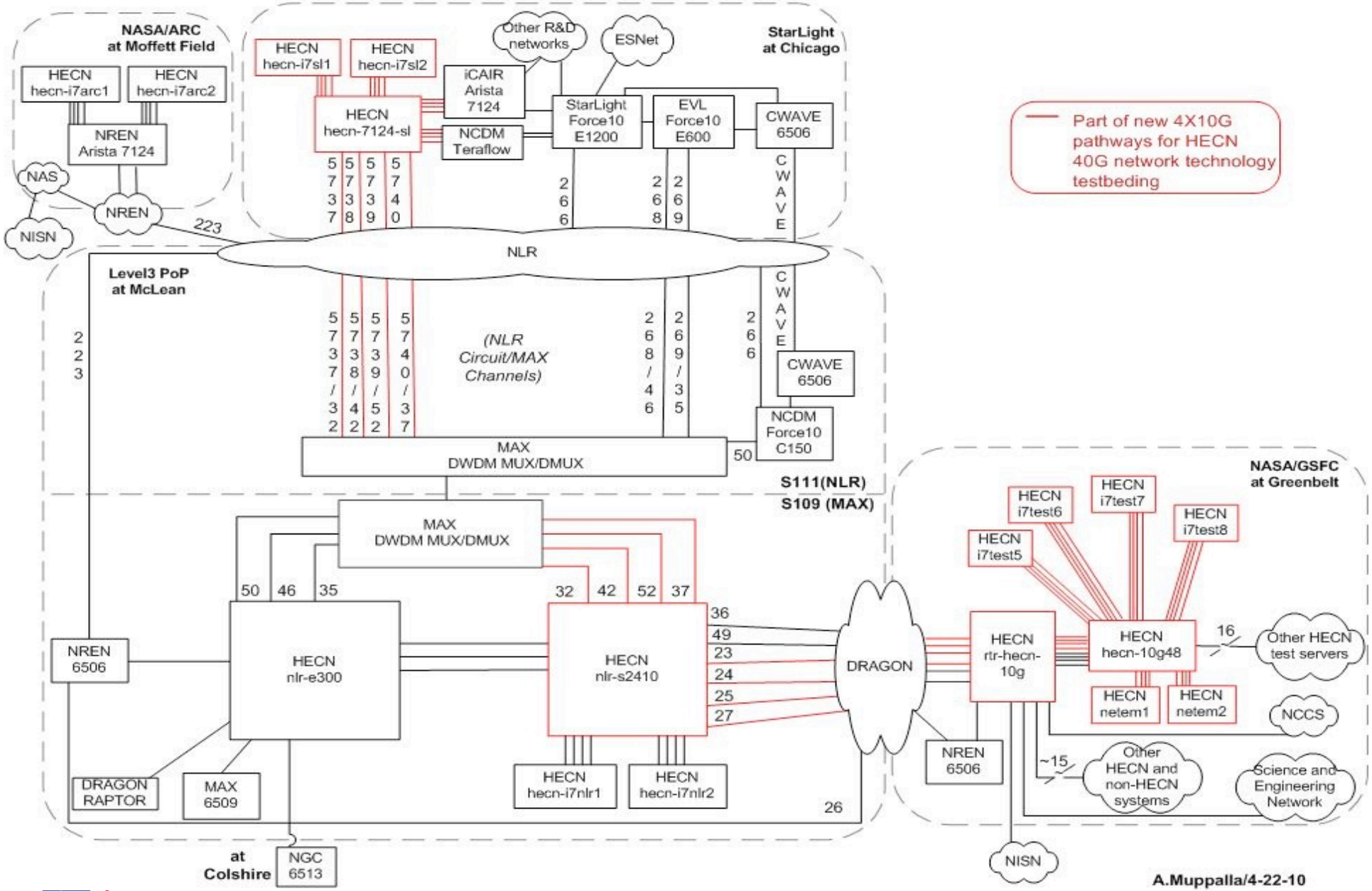
SC10 Demonstration "Using 100 Gbps Network Technology in Support of Petascale Science"

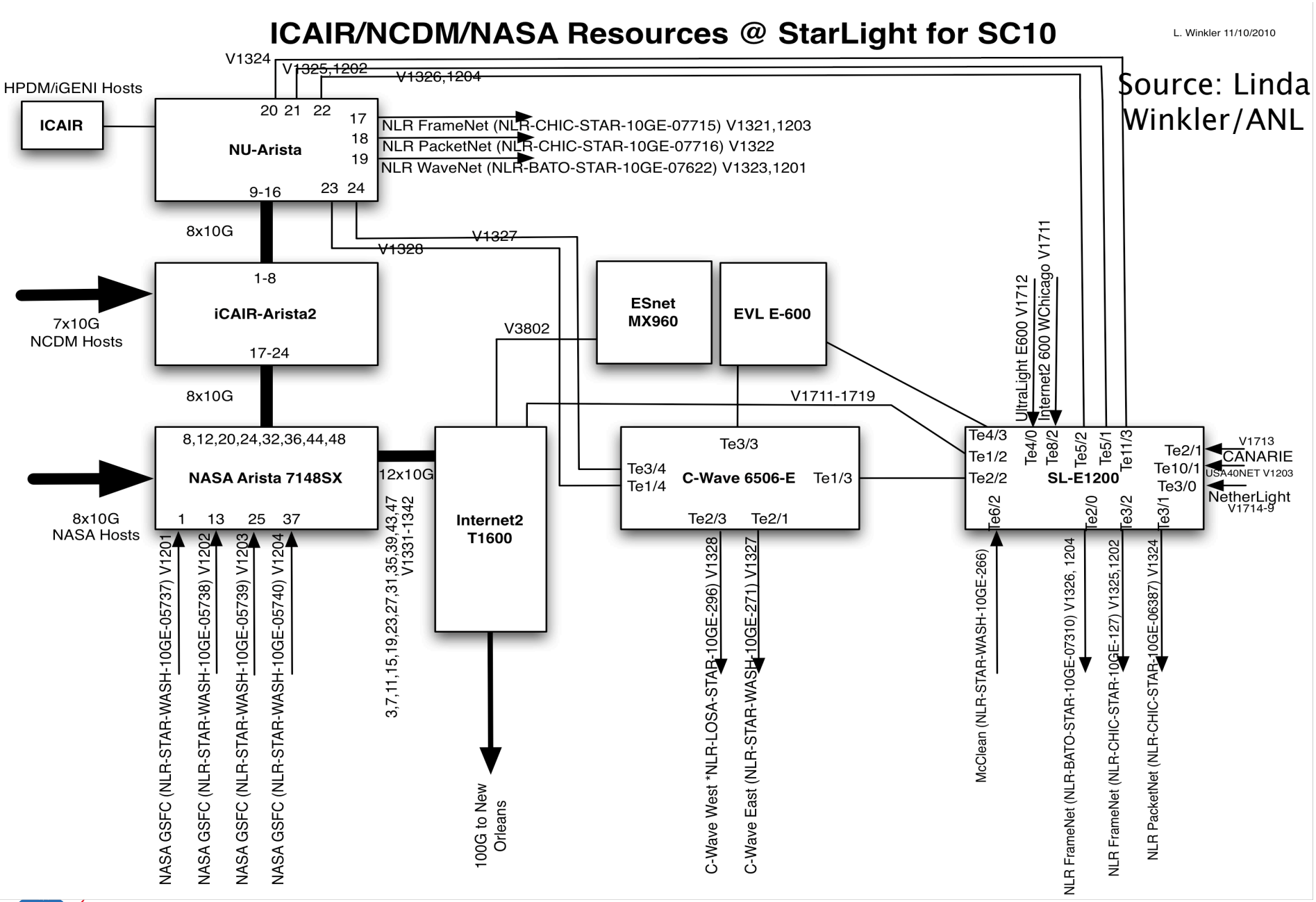
SC10's SCinet Research Sandbox Opportunity

- SC10
  - 23rd annual international Supercomputing 2010 (SC10) conference on high-performance computing, networking, storage and analysis, New Orleans, Nov. 13-18, 2010
- SCinet
  - Provisioned each year for the short duration of the conference
  - One of the most powerful and advanced networks in the world: ~300-Gbps LAN capacity
- SCinet Research Sandbox
  - For network researchers: network monitoring, performance optimization, power / thermal research, network security…
  - http://sc10.supercomputing.org/?pg=scinetsandboxprojects.html

GSFC/High End Computer Network (HECN) and Partners 10GE and 10G Lambda Connections Through McLean

Note: The non-GSFC/HECN systems shown typically have other connections that are not shown in this diagram, as the focus is primarily GSFC/HECN connections

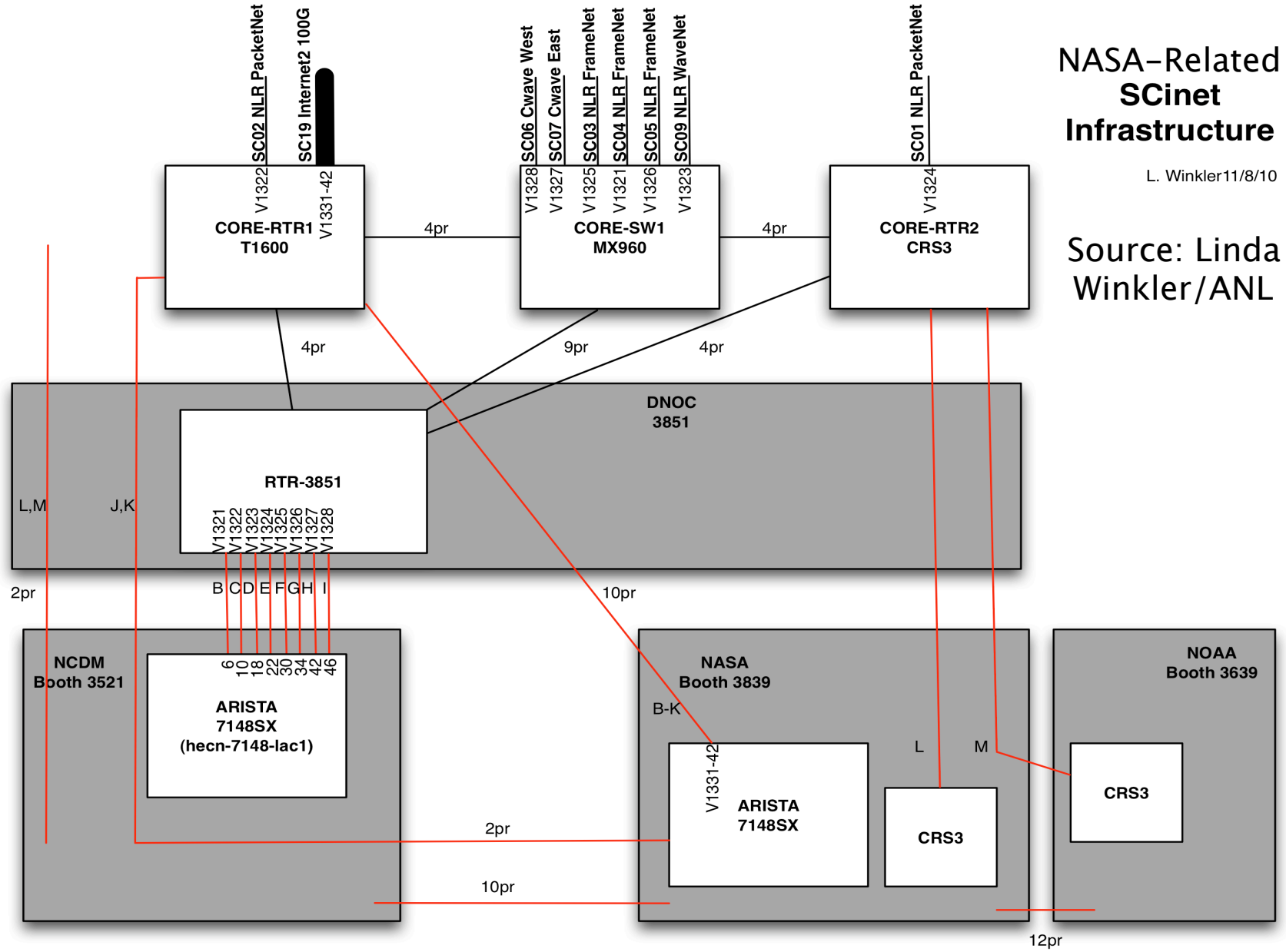



ICAIR/NCDM/NASA Resources @ StarLight for SC10
L. Winkler 11/10/2010
Source: Linda Winkler/ANL
HPDM/iGENI Hosts
ICAIR
NU-Arista
20 21 22
17
18
19
9-16
23 24
V1324
V1325,1202
V1326,1204
NLR FrameNet (NLR-CHIC-STAR-10GE-07715) V1321,1203
NLR PacketNet (NLR-CHIC-STAR-10GE-07716) V1322
NLR WaveNet (NLR-BATO-STAR-10GE-07622) V1323,1201
8x10G
V1327
V1328
1-8
iCAIR-Arista2
17-24
7x10G
NCDM Hosts
ESnet
MX960
EVL E-600
V3802
V1711-1719
UltraLight E600 V1712
Internet2 600 WChicago V1711
8x10G
8,12,20,24,32,36,44,48
NASA Arista 7148SX
1 13 25 37
8x10G
NASA Hosts
12x10G
Internet2
T1600
Te3/3
Te3/4
Te1/4
C-Wave 6506-E
Te1/3
Te2/3 Te2/1
Te4/3
Te1/2
Te2/2
Te4/0
Te8/2
Te5/2
Te5/1
Te11/3
SL-E1200
Te2/1
Te10/1
Te3/0
V1713
CANARIE
USA40NET V1203
NetherLight
V1714-9
Te6/2
Te2/0
Te3/2
Te3/1
NASA GSFC (NLR-STAR-WASH-10GE-05737) V1201
NASA GSFC (NLR-STAR-WASH-10GE-05738) V1202
NASA GSFC (NLR-STAR-WASH-10GE-05739) V1203
NASA GSFC (NLR-STAR-WASH-10GE-05740) V1204
3,7,11,15,19,23,27,31,35,39,43,47
V1331-1342
100G to New Orleans
C-Wave West *NLR-LOSA-STAR-10GE-296) V1328
C-Wave East (NLR-STAR-WASH-10GE-271) V1327
McClean (NLR-STAR-WASH-10GE-266)
NLR FrameNet (NLR-BATO-STAR-10GE-07310) V1326, 1204
NLR FrameNet (NLR-CHIC-STAR-10GE-127) V1325,1202
NLR PacketNet (NLR-CHIC-STAR-10GE-06387) V1324

NASA-Related SCinet Infrastructure

L. Winkler11/8/10

Source: Linda Winkler/ANL




NETWORK
BOTTLENECKS