Introduction To
NASA High End Computing (HEC)
WAN File Accessing Experiments/Demonstrations
At SC09

Pat Gary
Pat.Gary@nasa.gov
Computational and Information Sciences and Technology Office (CISTO), Code 606
NASA Goddard Space Flight Center
February 19, 2010

Information Supporting NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

NASA HEC WAN File Accessing Team

- GSFC NASA Center for Computational Sciences (NCCS)
  - Dan Duffy/GSFC
    - Hoot Thompson/PTP
    - Kirk Hunter/PTP
- GSFC/NCCS HEC Network Team
  - Pat Gary/GSFC
  - Bill Fink/GSFC
  - Paul Lang/ADNET
    - Aruna Muppalla/ADNET
    - Mike Stefanelli/ADNET
    - Jeff Martz/ADNET
- ARC/NAS Network Team
  - Dave Hartzel/CSC
  - Mark Foster/CSC
  - Kevin Jones/ARC
    - Jeff Becker/CSC
    - Harjot Sidhu/CSC
    - Jason Gunthorpe/Obsidian

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Acknowledgement of Vendor Equipment On Loan

- BlueArc: Titan
- Obsidian Strategics: Longbow E100
- Super Micro: X6DHR
- Vion: HyperStor
- Voltaire: ISR-9024 & ISR2004-534a w/sRB-20210G-65b8

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Identification of GSFC/NCCS Equipment Used

- A++: see following pages
- Arista: 7124 10GE switch/router
- B, B+: see following pages
- Force10: E600i 10GE switch/router
- SMC: 8708 10GE switch

Optimizing Wide-Area File Transfer for 10-Gbps and Beyond

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

- Jointly planned by GSFC's High End Computer Network Team and NCCS' Advanced Development Team, an indication of the wide-area file transfer applications demonstrated and evaluated is shown in the Data Transfer Test Matrix (top figure) and the WAN infrastructure and servers tested are shown in the configuration diagram (bottom figure).

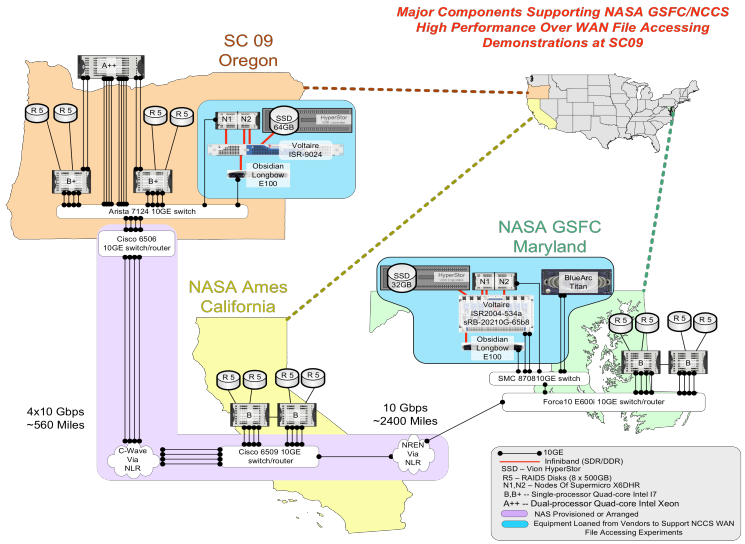
- Demonstration highlights included over 100 gigabits per second (Gbps) uni-directional memory-to-memory data transmissions between in-booth servers, 40-Gbps bi-directional memory-to-memory data transmissions between servers in-booth and at ARC, 10-Gbps disk-to-disk data transfers between in-booth servers, between servers in-booth and at ARC, and between servers in-booth and at GSFC.

POC: Pat Gary, Pat.Gary@nasa.gov, (301) 286-9539, GSFC Computational and Information Sciences and Technology Office

High Performance Wide Area Data Transfer Test Matrix

Tests		Protocols			Connection Points		
		IP	IPoIB	RDMA	GSFC to SC09	ARC to SC09	SC09 Intra-booth
Traditional	bbftp	●	●		● ●		
	scp	●	●		● ●		
	rsync	●	●		● ●		
Experimental	nuttcp	●	●		● ●	● ● ●	● ● ●
	nuttscp	●	●		● ●	● ● ●	● ● ●
	Trperf[1]			●	●		
	Rdma-cp[1]			●	●		
	Rdma-rsync[1]			●	●		
	Xdd[2]	●	●		● ●		
Application	Grid FTP	●	●		● ●		
	iRODS	●	●		● ●		
File Systems	NFS	●	●		● ●		
	NFS Rdma			●	●		
	GPFS	●	●	●	● ● ●		
	Lustre	●	●	●	● ● ●		

[1] Courtesy of Obsidian Research.
[2] End-to-end file transfers supported by the Oak Ridge National Laboratory Extreme Scale System Center and the Department of Defense.



Figures: Data Transfer Test Matrix (Top) and WAN infrastructure and servers tested (bottom) during SC09.

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Objectives of NASA HEC WAN File Accessing Experiments

- Determine optimal ‘tuning parameter” settings to obtain maximum user throughput performance with several traditional and new (or emerging) disk-to-disk file-copying utilities when operating over multi-10Gbps WANs using new state-of-the-art high performance workstations and servers
- Inter-compare throughput findings from traditional versus new file-copying utilities
- As a baseline, determine maximum memory-to-memory throughput performance among the workstations and servers using nuttcp (http://www.nuttcp.org/)
- Are an integral part of GSFC/HEC’s 20, 40 & 100 Gbps Network Testbed Plan

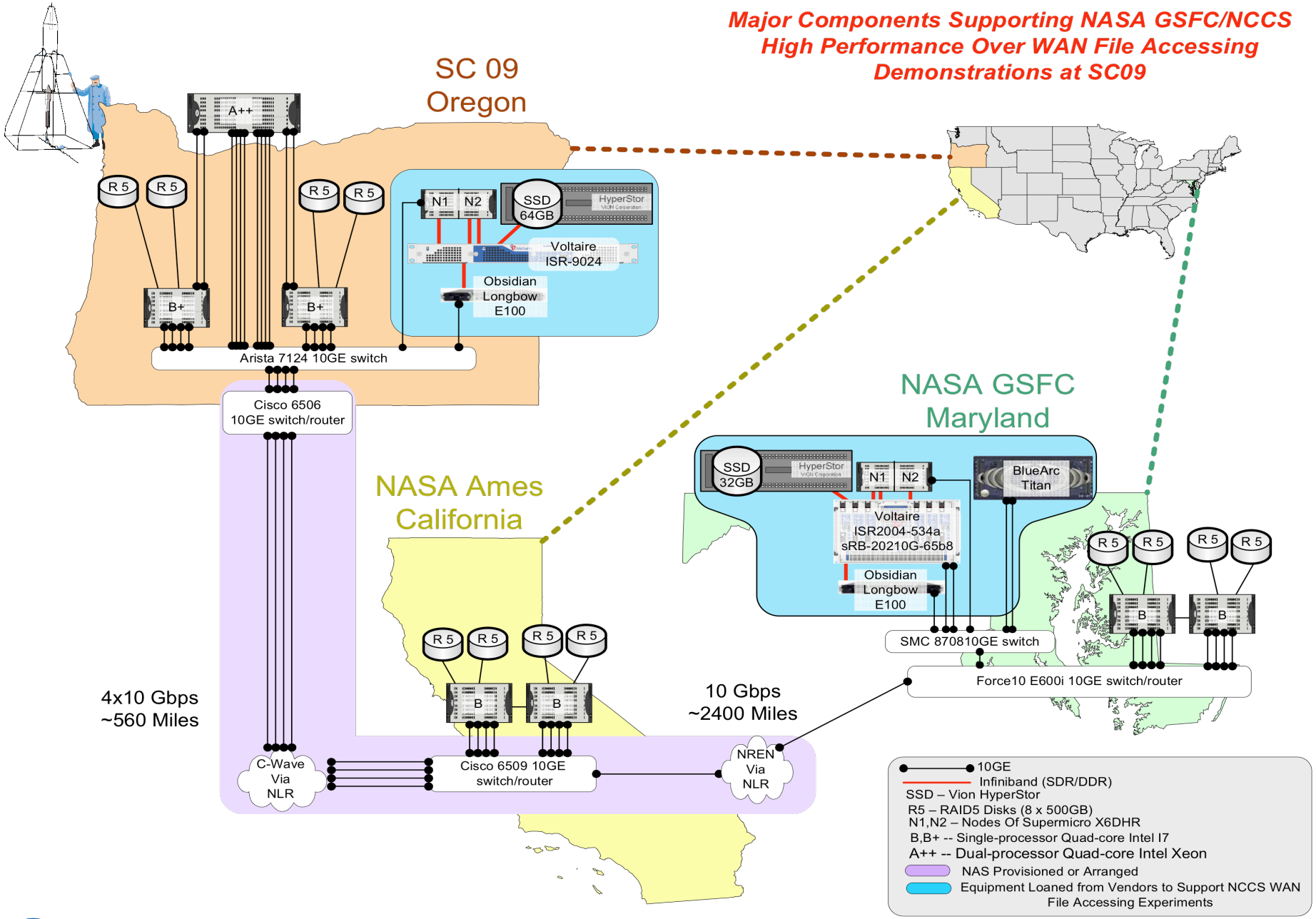
High Performance Wide Area Data Transfer Test Matrix

Tests		Protocols			Connection Points		
		IP	IPoIB	RDMA	GSFC to SC09	ARC to SC09	SC09 Intra-booth
Traditional	bbftp	●	●		● ●		
	scp	●	●		● ●		
	rsync	●	●		● ●		
Experimental	nuttcp	●	●		● ●	● ● ●	● ● ●
	nuttscp	●	●		● ●	● ● ●	● ● ●
	Trperf[1]			●	●		
	Rdma-cp[1]			●	●		
	Rdma-rsync[1]			●	●		
	Xdd[2]	●	●		● ●		
Application	Grid FTP	●	●		● ●		
	iRODS	●	●		● ●		
File Systems	NFS	●	●		● ●		
	NFS Rdma			●	●		
	GPFS	●	●	●	● ● ●		
	Lustre	●	●	●	● ● ●		

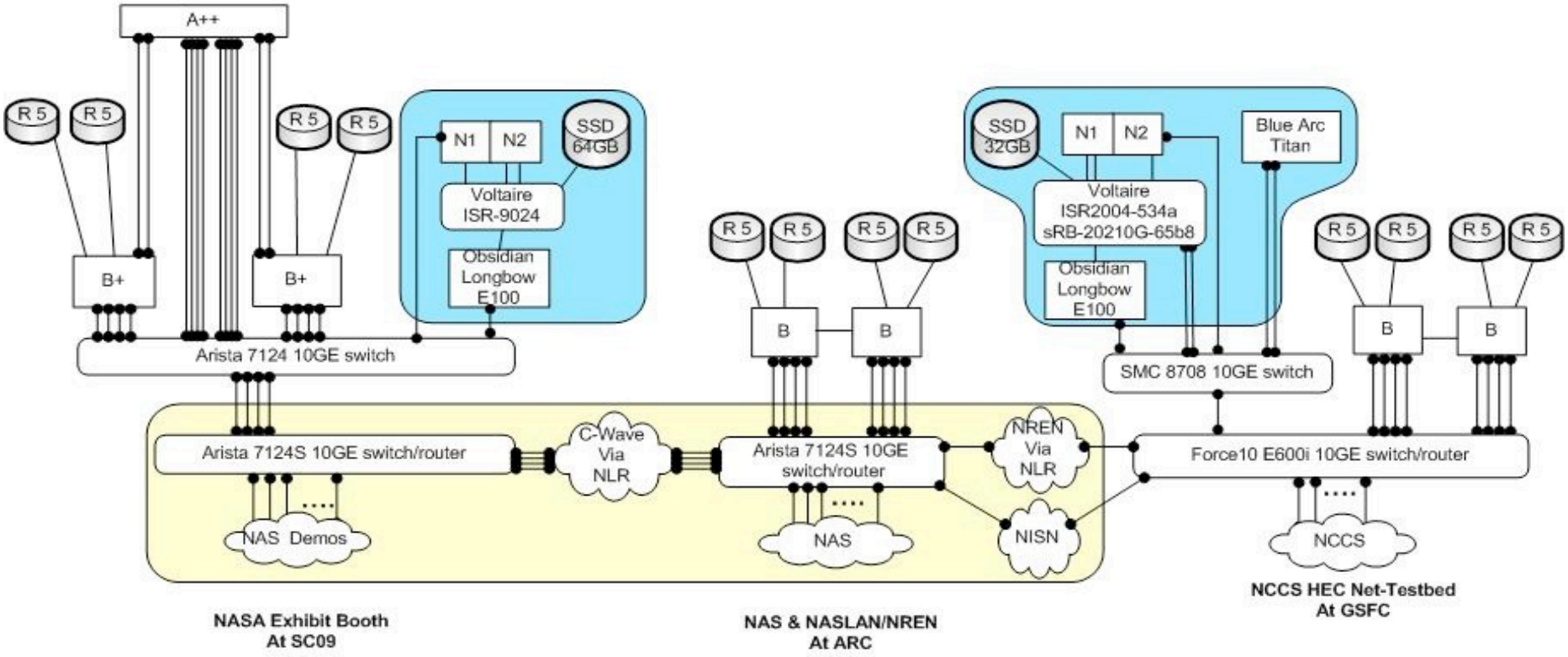
[1] Courtesy of Obsidian Research.

[2] End-to-end file transfers supported by the Oak Ridge National Laboratory Extreme Scale System Center and the Department of Defense.

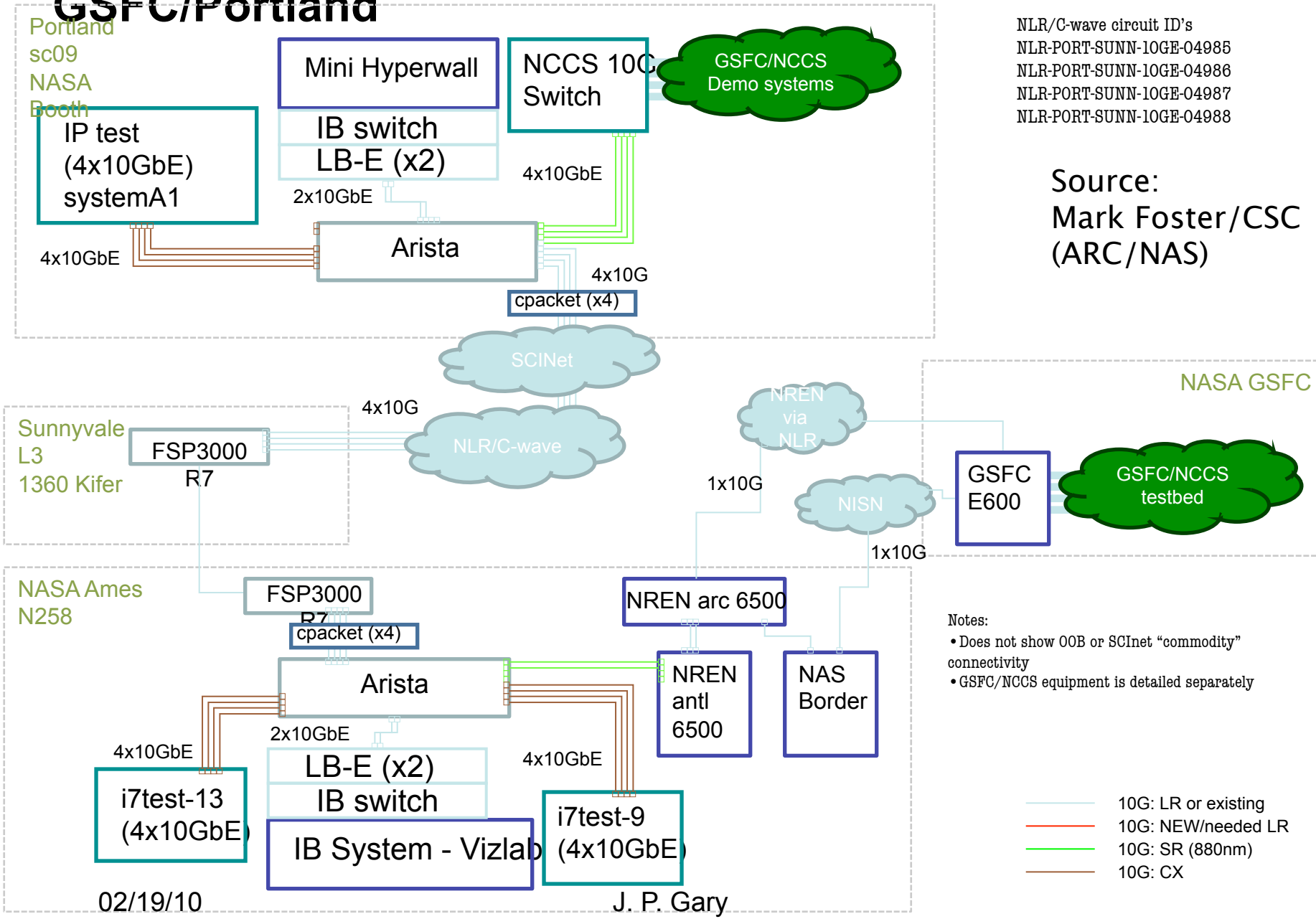
Major Components Supporting NASA GSFC/NCCS High Performance Over WAN File Accessing Demonstrations at SC09



Major Components Supporting NASA GSFC/NCCS High Performance Over WAN File Accessing Demonstrations at SC09



SC09 NASA 20-40Gbps Demos interconnect – ARC/ GSFC/Portland



NLR/C-wave circuit ID's
NLR-PORT-SUNN-10GE-04985
NLR-PORT-SUNN-10GE-04986
NLR-PORT-SUNN-10GE-04987
NLR-PORT-SUNN-10GE-04988

Source:
Mark Foster/CSC
(ARC/NAS)

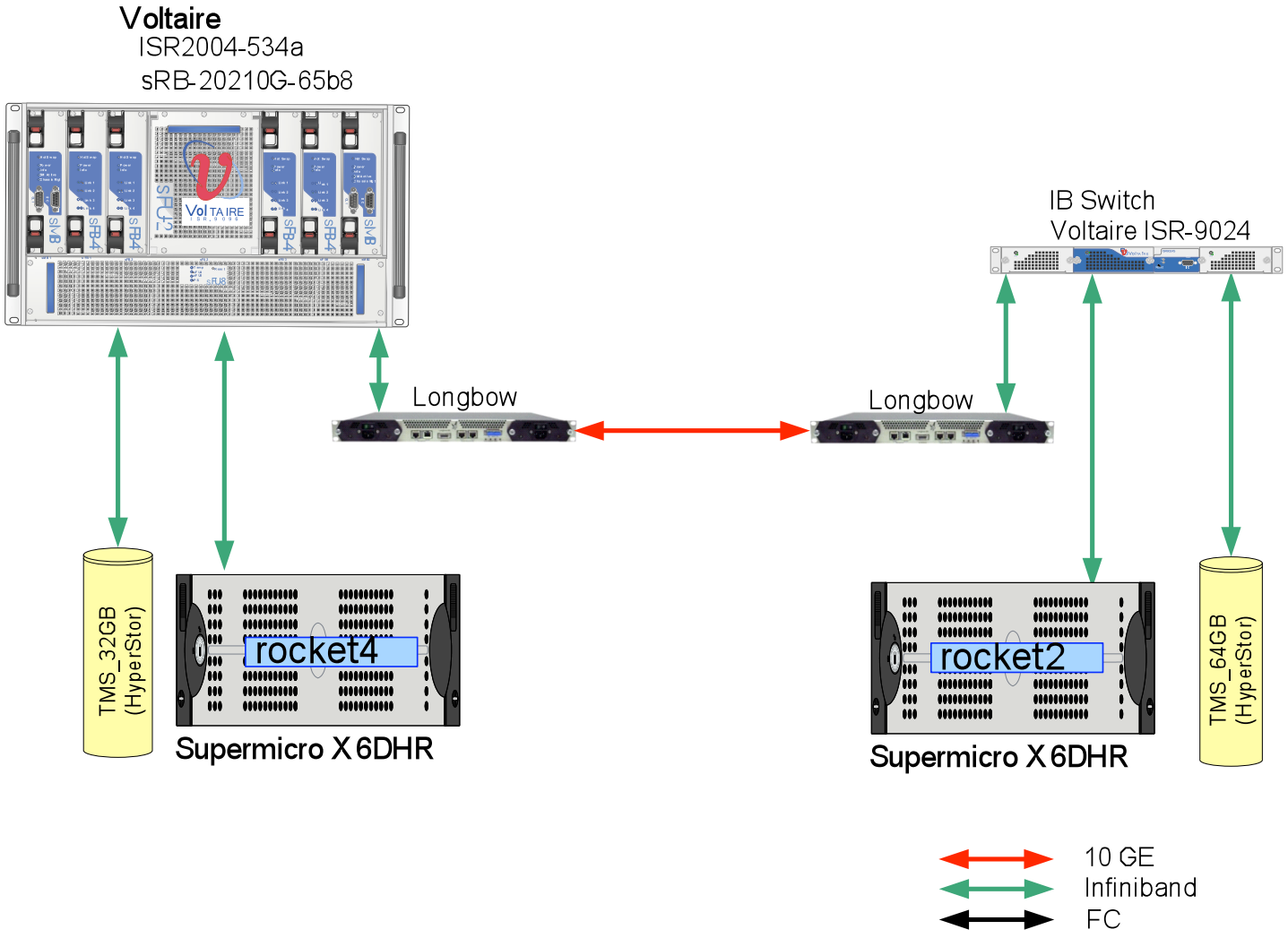
Notes:
- Does not show OOB or SCInet "commodity" connectivity
- GSFC/NCCS equipment is detailed separately

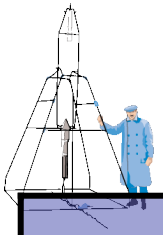
10G: LR or existing
10G: NEW/needed LR
10G: SR (880nm)
10G: CX

V2.5 11/13/09 mf

Pre SC Test Setup

Source: Hoot Thompson/PTP (GSFC/NCCS)





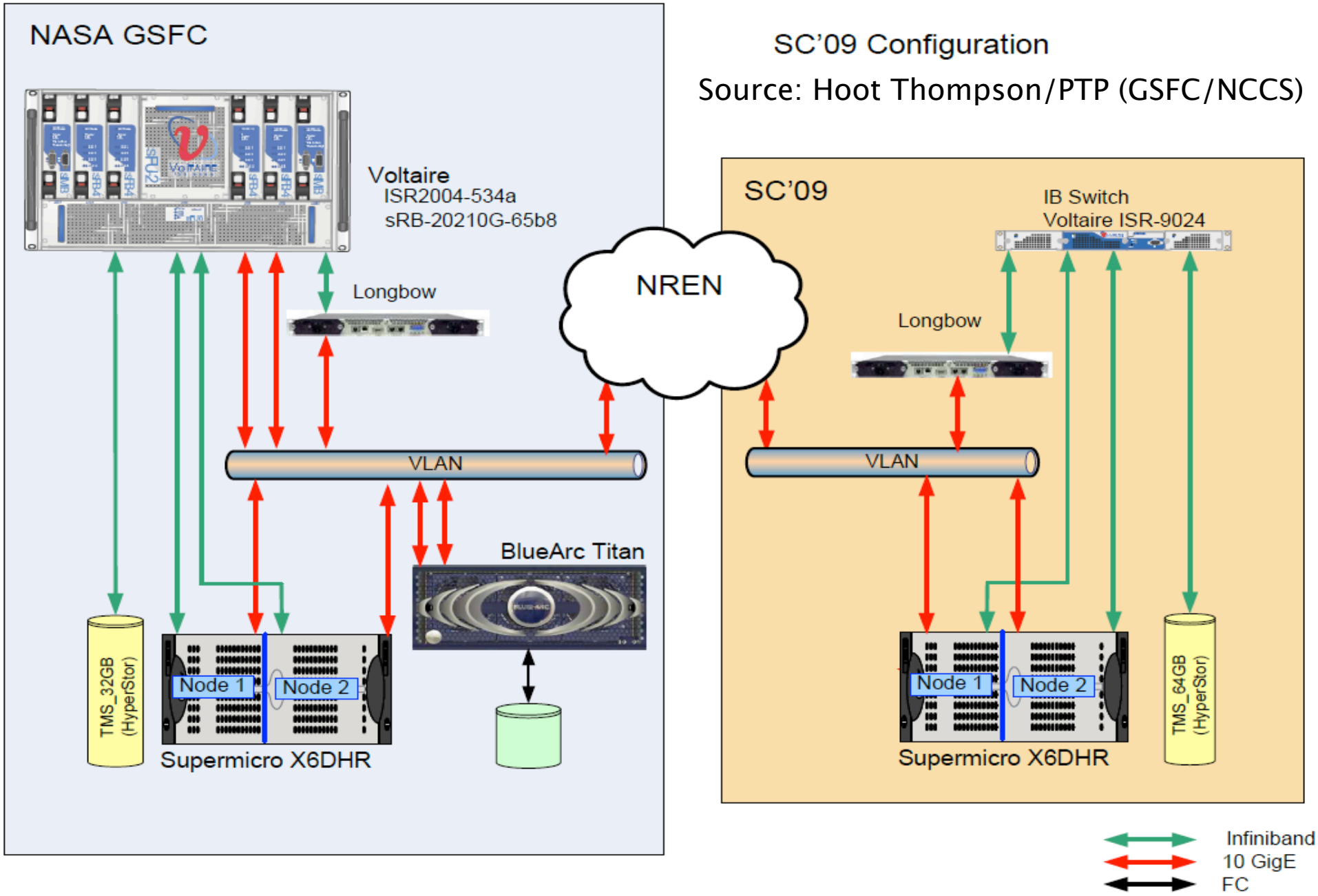
Test Results Pre-3Nov09 (pre-SC09)

Source: Hoot Thompson/PTP (GSFC/NCCS)

Tool	Type	rtt		Comments
		0 msec	100 msec	
nuttcp	Memory ↔ Memory	982 MB/s	920 MB/s	With large rtt, performance builds to peak number
perftest	Memory ↔ Memory	937 MB/s	N/A	rdma_bw test over 10GE NetEffect NICS
rdmacp	Disk ↔ Disk	824 MB/s	~800 MB/s	
bbftp	Disk ↔ Disk	814 MB/s (put) 840 MB/s (get)	33 MB/s (put) 33 MB/s (get)	
iRODS	Disk ↔ Disk	378 MB/s (iput) 379 MB/s (iget)	112 MB/s (iput) 43 MB/s (iget)	
xdd copy	Disk ↔ Disk	981 MB/s (src) 620 MB/s (dest)	493 MB/s (src) 372 MB/s (dest)	Added security related information
dsync	Disk ↔ Disk	N/A	N/A	rdma rsync – just now available
nuttscp	Disk ↔ Disk	577 MB/s	577 MB/s	Default settings
nfs	Disk ↔ Disk	686 MB/s (wrt) 444 MB/s (read)	Not Useful	
nfsrdma	Disk ↔ Disk	319 MB/s (wrt) 326 MB/s (read)	Not Useful	Could not achieve advertised results

SC'09 Configuration

Source: Hoot Thompson/PTP (GSFC/NCCS)



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

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two "B-Systems" (1 of 4) [Source: Bill Fink/GSFC]

- Two simultaneous 64-GB file copies (each file-copy streamed between one RAID5 disk controller hosted on each B-system in a LAN testbed)
  - File copy 1: 5092.5196-Mbps 43% TX 77% RX 0 retrans 0.10ms RTT
  - File copy 2: 5045.3832-Mbps 33% TX 77% RX 0 retrans 0.10ms RTT
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system in a LAN testbed)
  - File copy: 9824.2054-Mbps 58% TX 96% RX 0 retrans 0.10ms RTT
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system in a 40km MAN testbed)
  - File copy: 9402.0330-Mbps 56% TX 98% RX 0 retrans 0.45ms RTT

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two "B-Systems" (2 of 4) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 9548.0962-Mbps 59% TX 97% RX 0 retrans 80.15ms RTT (completed in 57.58 seconds)
    *With receiver B-system over-clocked to 3.4-Ghz instead of 3.2-Ghz
  - [For comparison a 60.16 second memory-to-memory test using nuttcp: 9661.2217-Mbps 26% TX 40% RX 0 retrans 80.14ms RTT]
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system** in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 8931.9535-Mbps 58% TX 97% RX 0 retrans 80.14ms RTT (completed in 61.55 seconds)
    **With receiver B-system clocked normally at 3.2-Ghz

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two "B-Systems" (3 of 4) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 5055.1438-Mbps 31% TX 59% RX 8 retrans 80.15ms RTT (completed in 108.75 seconds)
    *With receiver B-system over-clocked to 3.4-Ghz instead of 3.2-Ghz
  - [For comparison a 30.29 second memory-to-memory test using nuttcp: 5561.7408-Mbps 14% TX 28% RX 4 retrans 80.15ms RTT]
  - Retrans caused by "dropped_bad_crc32" errors at ~10^-6 packet loss rate

Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two "B-Systems" (4 of 4) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km real WAN testbed): GSFC→ARC
  - File copy: 7575.1083-Mbps 47% TX 89% RX 0 retrans 80.58ms RTT (completed in 72.57 seconds)
    *With receiver B-system clocked normally at 3.2-Ghz
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km real WAN testbed): ARC→GSFC
  - File copy: 8284.2127-Mbps 60% TX 95% RX 0 retrans 80.58ms RTT (completed in 66.36 seconds)
    **With receiver B-system clocked normally at 3.2-Ghz

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 .

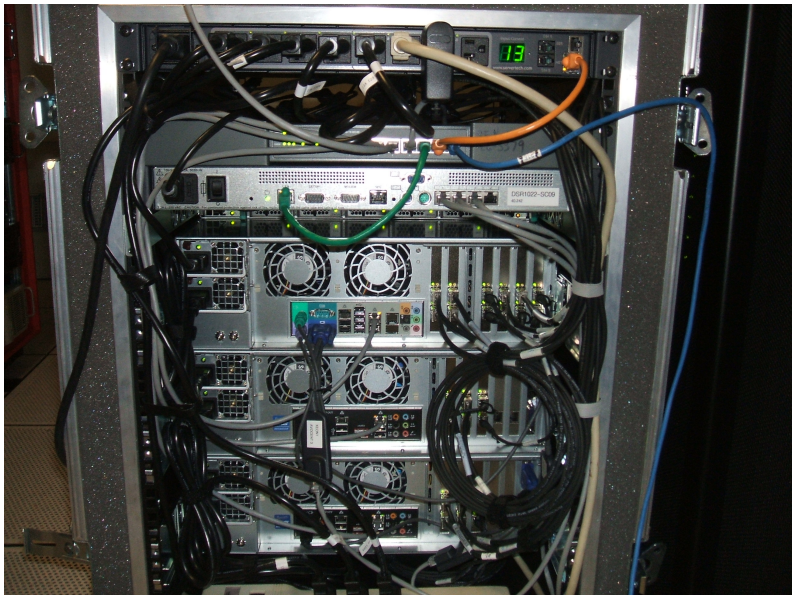


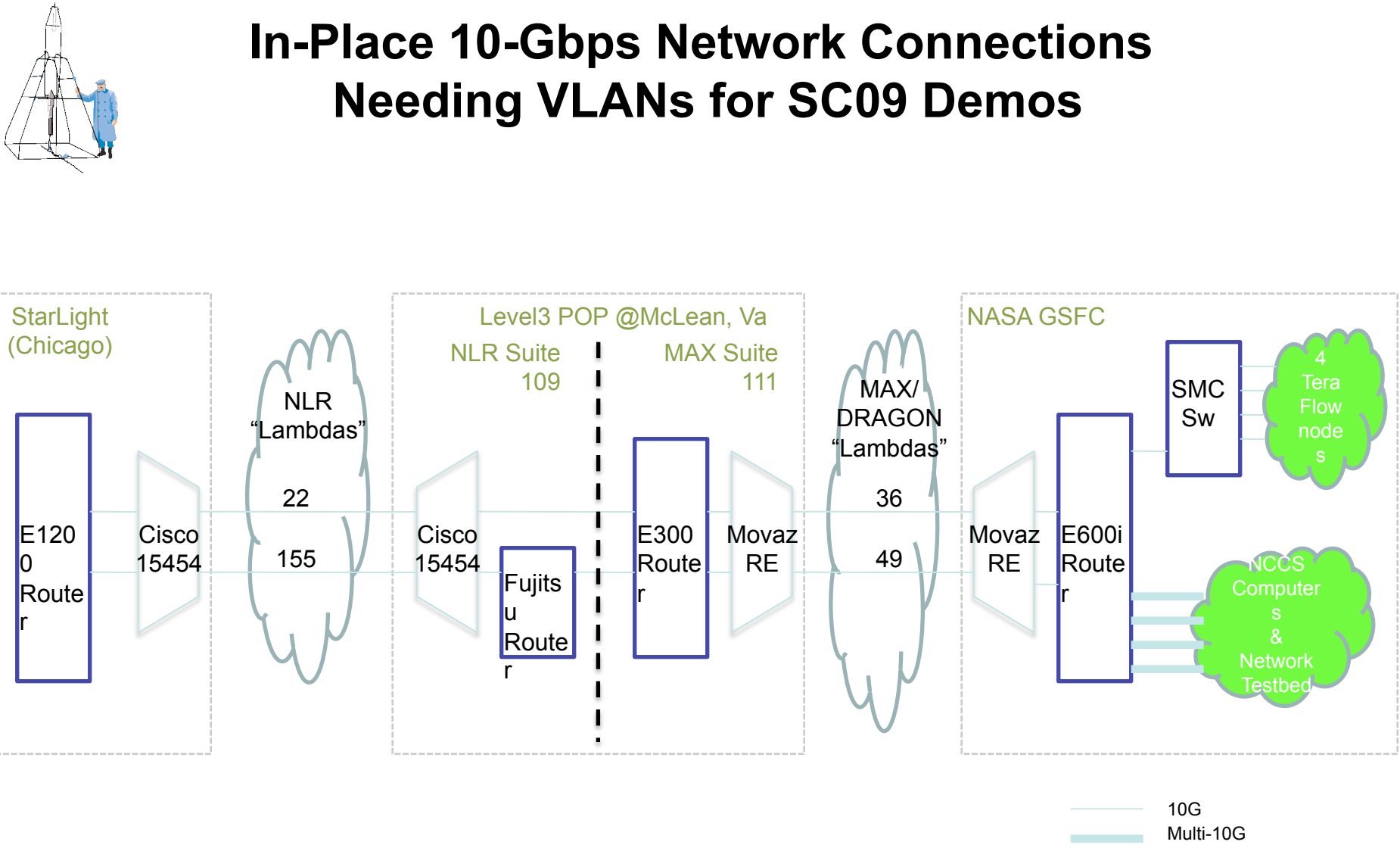
Figure: Xeon and two Core i7 workstations (bottom) interconnected with 10 Gigabit Ethernet switch and management units (top) in a rack for shipping to SC09.

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

Multi-10Gbps WAN Pathways Supporting Joint NASA+NCDM(UIC) Demos During SC09
Including High Performance WAN File Transfers, Distributed Cloud Computing & TeraFlow

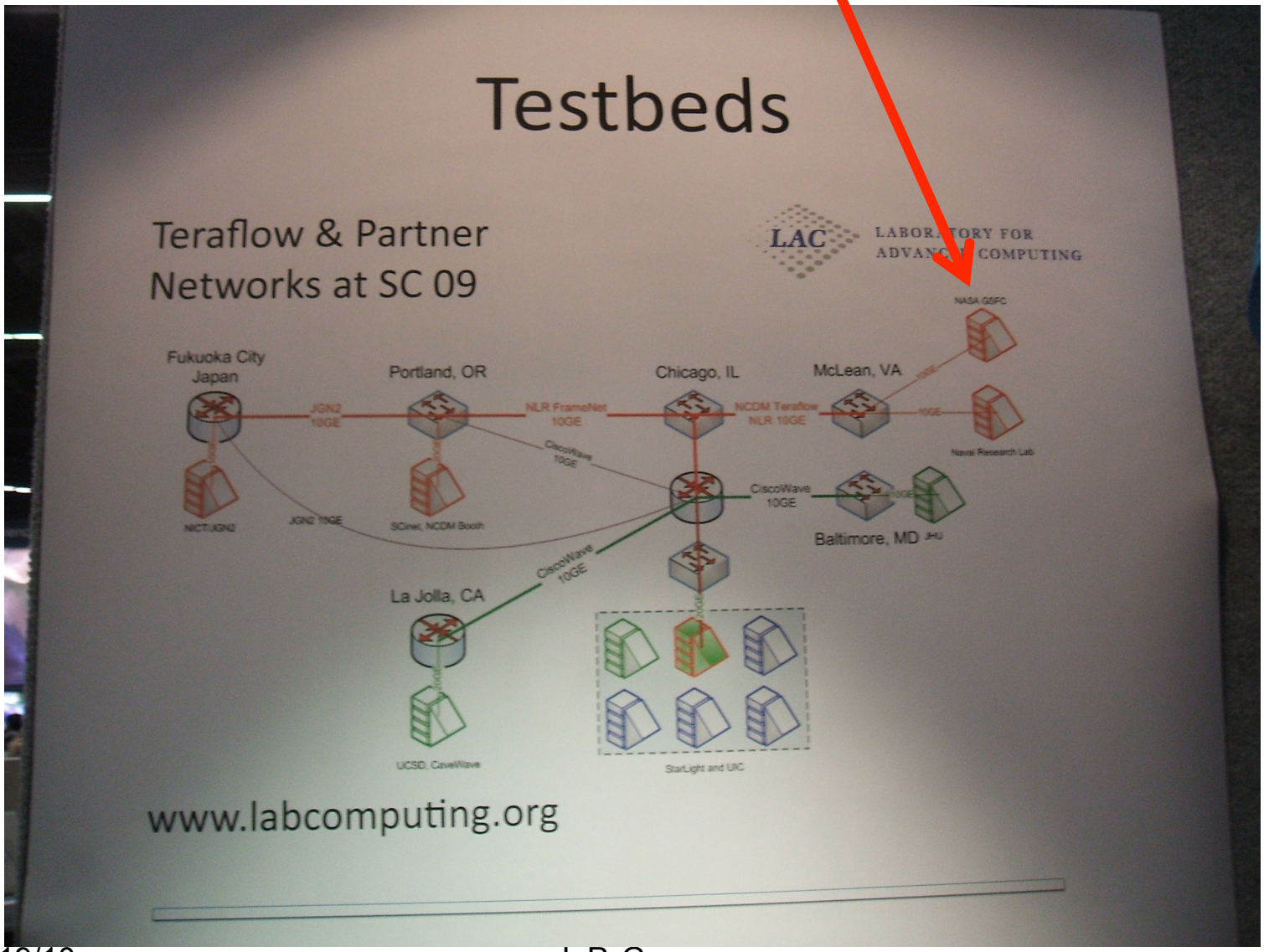
J. P. Gary
11/02/09

In-Place 10-Gbps Network Connections Needing VLANs for SC09 Demos



J. P. Gary
11/02/09

UIC-prepared Diagram Showing GSFC as Part of UIC's Teraflow Testbed and Open Cloud Consortium Initiatives



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

Current Status (1 of 3)

- Our testbed plans are in their early stages of development and are a "work-in-progress"
- We have gotten excellent support from our collaborators, especially the MAX, NAS/NREN & NLR; and we're beginning to pick up quite a bit of steam
  - Completed significant SC09 experiments/demonstrations
  - Two B-systems are to remain deployed at ARC post-SC09
  - Several A-, B-, & C-systems deployed at GSFC
  - Four 10-Gbps HECN-enabled links between GSFC and College Park (CLPK) are ready to deploy
  - Two B-systems are in preparation for deployment to CLPK
  - Four 10-Gbps DRAGON-enabled links between GSFC and McLean (MCLN) are ready to deploy
  - Two B-systems are in preparation for deployment to MCLN

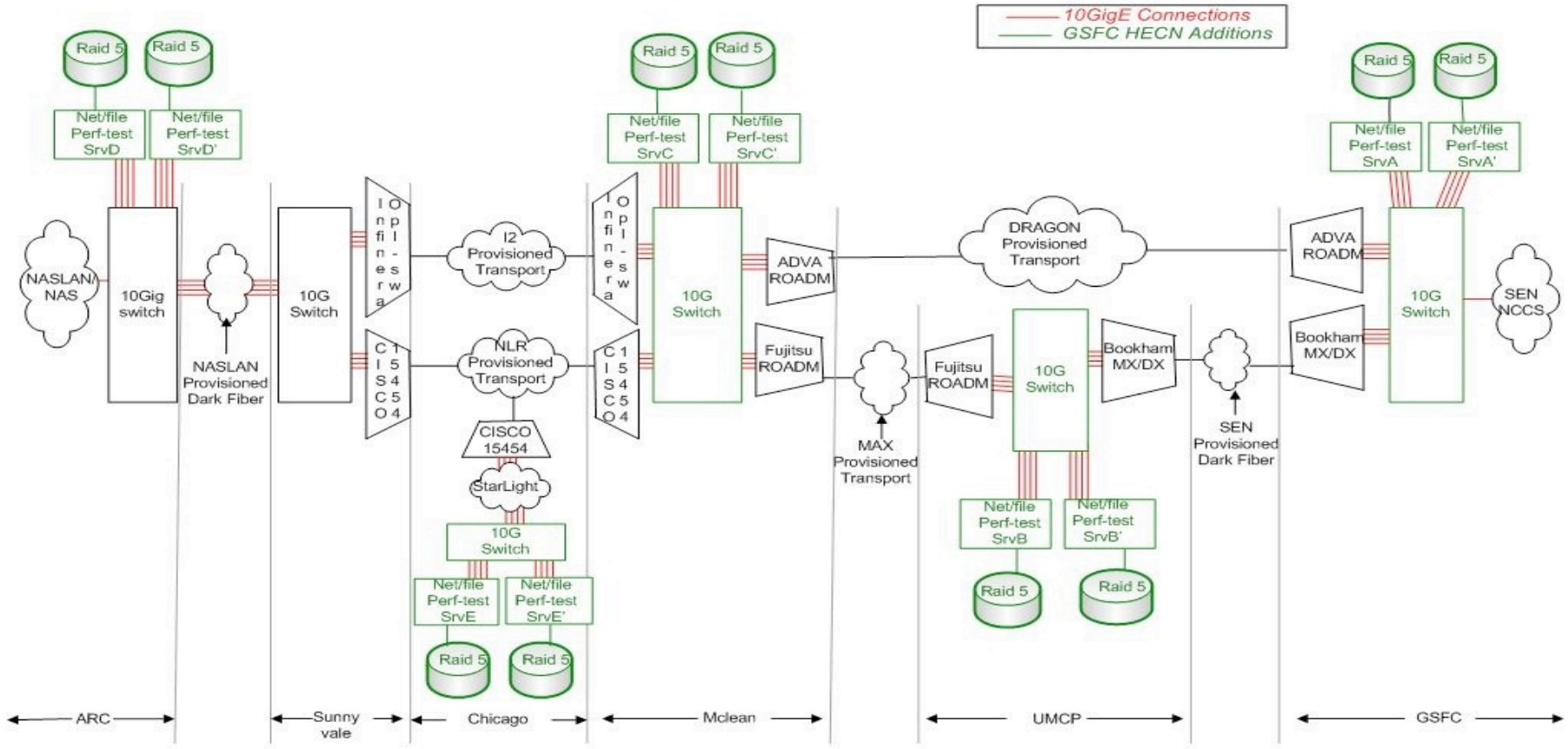
Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Current Status (2 of 3)

- "More steam"
  - Four 10-Gbps NLR-enabled links between MCLN and StarLight (@Chicago) are ready for use
  - Two B-systems are in preparation for deployment to StarLight
  - Expect one 40-Gbps NLR-enabled link between MCLN and StarLight starting in mid-2010
  - Obtained approval to be informal partners in DoE's high performance file accessing testbed and have access to their upcoming 100-Gbps link between ANL (connected via StarLight) and NERSC (connected at Sunnyvale)
  - Expect one 40-Gbps MAX-enabled link between CLPK & MCLN starting in mid-2010, with potential as a 100-Gbps link not long after
- Intra-NASA we'll seek broader sponsorship via an "in-formation" Emerging Network Technology Testbed initiative

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

Current Candidate 40Gbps MAN & WAN Pathways For Use During Early Stages Of Phase 1 20Gbps & Phase 2 40Gbps Testbeds



J.P.Gary/A.Muppalla
6/17/09

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

Current Status (3 of 3)

- In the meantime we offer an open invitation to be involved
  - To extend the base of owning and/or testing network-test workstations like we have
  - To develop and/or test variants to network-test workstations like we have
  - To participate in and/or learn from the various WAN file accessing applications we're investigating
  - To enhance your status vis-à-vis NSF Academic Research Initiative proposals

Optimizing Wide-Area File Transfer for 10-Gbps and Beyond:

Network-Performance Testing, Wide-Area File Systems, and File Transfer Applications at 10, 40 and 100-Gbps Throughput Performance

NASA Research Exhibit at the SC09 Conference, Portland, OR, Nov. 16–19, 2009

Exterior of Oregon Convention Center that hosted SC09.

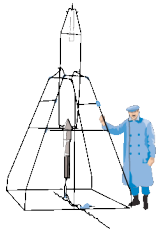
Right front of NASA Research Exhibit at SC09.

Left front of NASA Research Exhibit at SC09.

SCinet booth at SC09.

GSFC's network engineer Bill Fink participates in the live experiments via Lifesize-enabled HD videoconferencing.

Data management and network experts conducting experiments with wide-area file transfer applications.



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Reference Articles & Websites

- "Optimizing Wide-Area File Transfers for 10 Gbps and Beyond"
  - http://www.nas.nasa.gov/SC09/PDF/Datasheets/Gary_OptimizingWide.pdf
- "NASA Successfully Demonstrates Remote High-speed Encrypted InfiniBand Applications Over National LambdaRail"
  - http://www.virtualpressoffice.com/detail.do?contentId=208703&companyId=3273&showId=1215381715818
- "NASA Demos Secure Coast-to-Coast Backup at Full Wire Speed Using Obsidian's New Longbow E100 and DSYNC"
  - http://www.virtualpressoffice.com/publicsiteContentFileAccess?fileContentId=206528&fromOtherPageToDisableHistory=Y&menuName=News&sId=1215381715818&sInfo=Y
- NASA use of NLR during SC09
  - http://www.flickr.com/photos/nationallambdarail/4189002873/