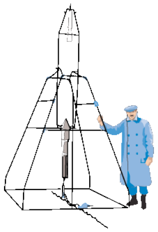


Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Pat Gary/GSFC
Pat.Gary@nasa.gov

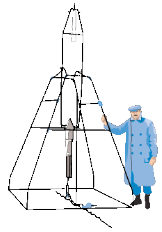
Bill Fink/GSFC
William.E.Fink@nasa.gov

Paul Lang/ADNET
Paul.A.Lang@nasa.gov

Computational and Information Sciences and Technology Office (CISTO), Code 606
NASA Goddard Space Flight Center

August 22, 2010

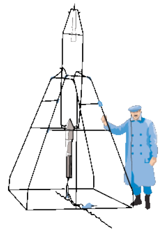
Information for NASA Network Planning



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

Topic Outline

- Problem Statement, Motivation & Background Context
  - NCCS: Increasing capacities/capabilities & “PI distribution”
  - Success of prior >10 Gbps network testbed efforts
  - More 40 & 100 Gbps network technologies on the way
- Value Proposition & Solution Approach
- Multi-Phased Network Testbed Objectives
- Phase 1 Network-Test Workstations
- Early Stages of Phase 1 & 2 Testbeds
  - WAN File Accessing Experiments/Demonstrations at SC09
  - Build out of Phase 1& 2 Testbeds
- Current Status and Next Steps



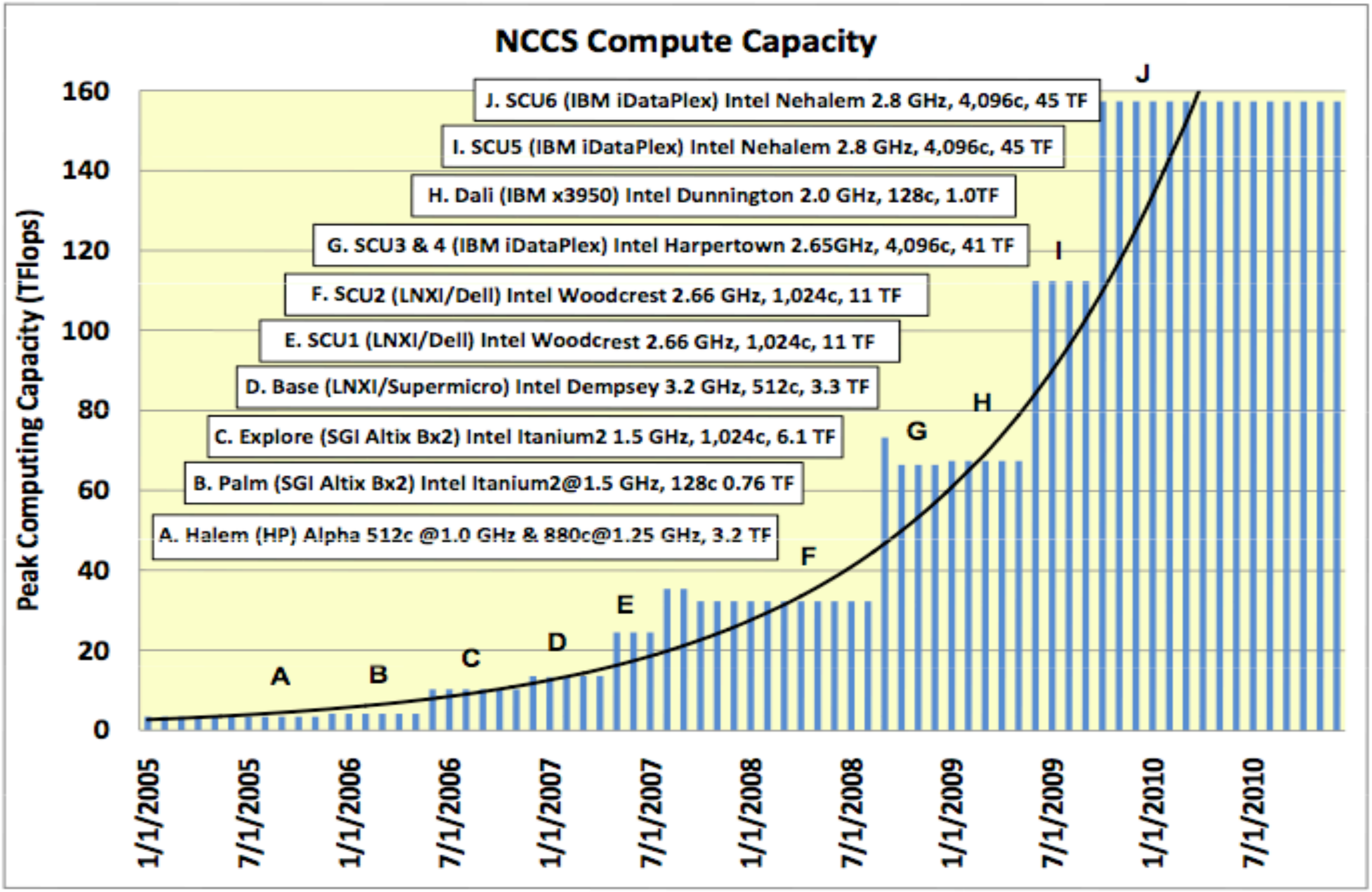
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Problem Statement

- GSFC’s NASA Center for Computational Sciences (NCCS) is increasing its data production/analysis/storage capacities and capabilities

- Higher bandwidth networks can be deployed

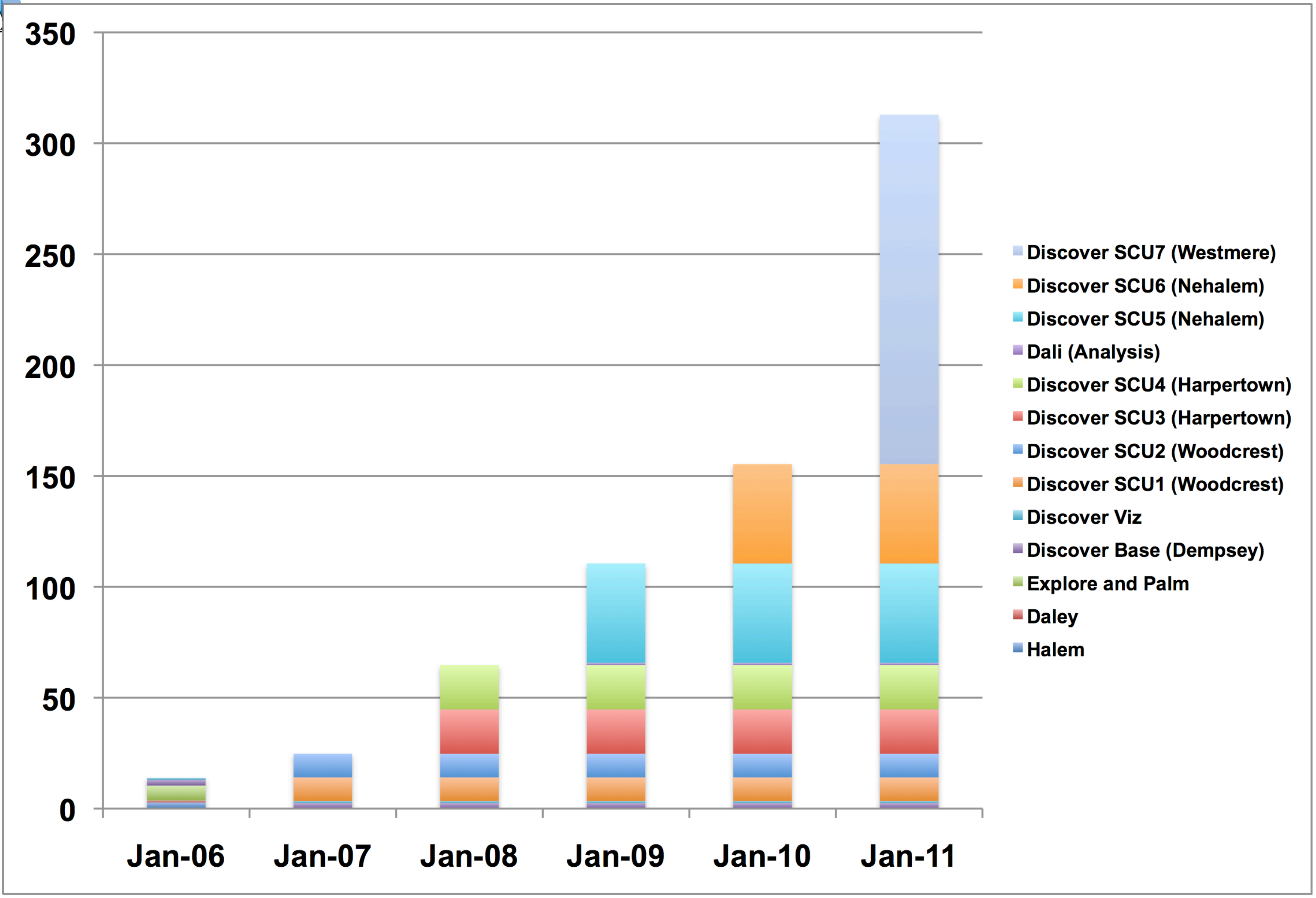
- But something in the combination of our file copying applications, disk I/O subsystems, server/workstation configurations, protocol stack tuning and/or NICs is preventing full use of our higher bandwidth networks



Source: Dan Duffy/GSFC (GSFC/NCCS) & Scott Wallace (CSC) (GSFC/NCCS)

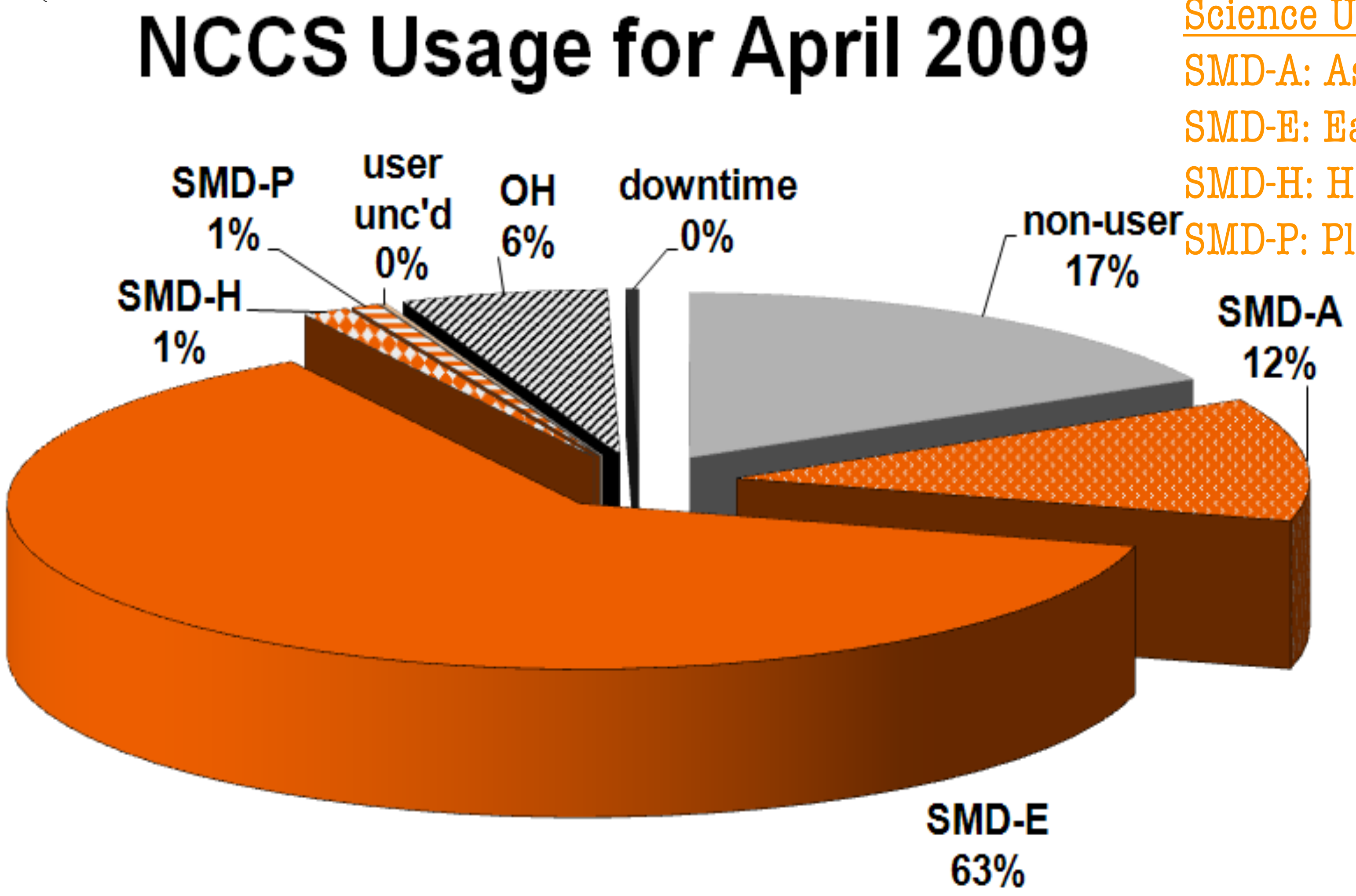
NCCS Peak Computing (TF) Over Time

Source: Dan Duffy (GSFC/NCCS)



08/22/10

Excerpt from HEC Program Monthly Status Report Apr09
Source: Sally Stemwedel/GS&T (GSFC/NCCS)

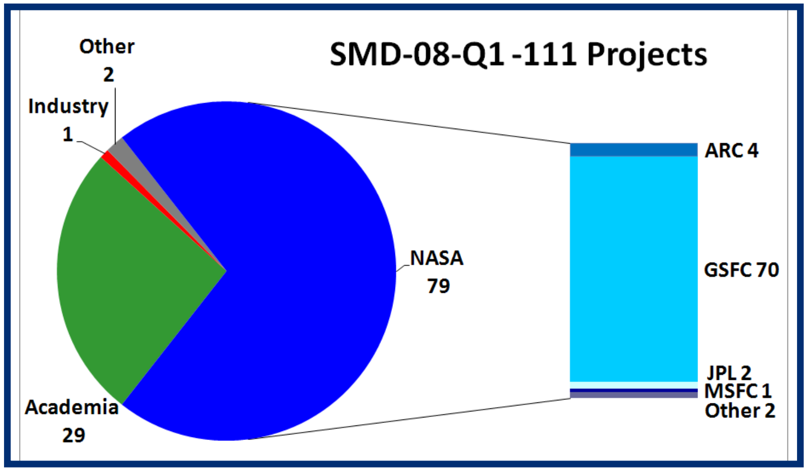


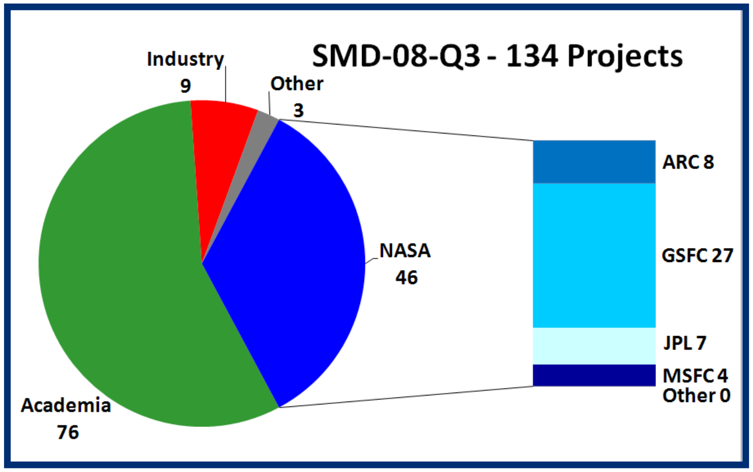
Science Users

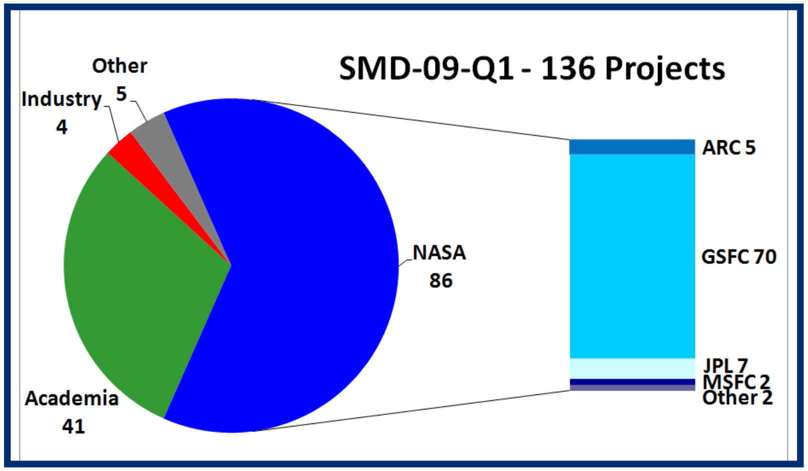
SMD-A: Astrophysics
SMD-E: Earth Science
SMD-H: Heliophysics
SMD-P: Planetary

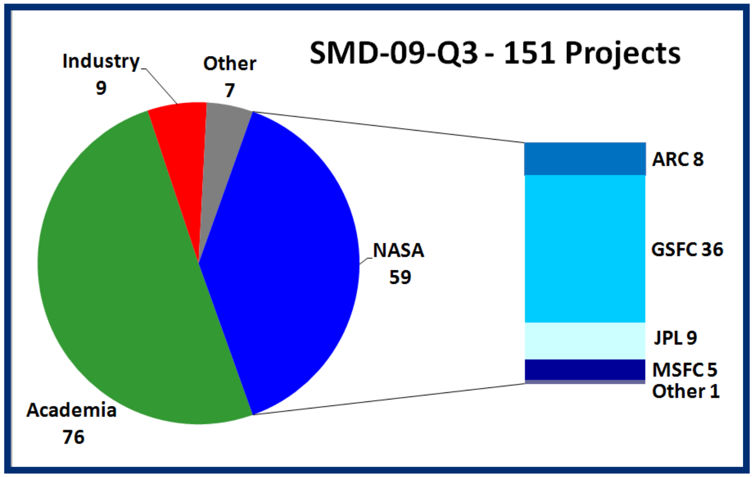
Excerpt from HEC Program Monthly Status Report Apr09
Source: Sally Stemwedel/GS&T (GSFC/NCCS)

SMD PI Organization Distribution









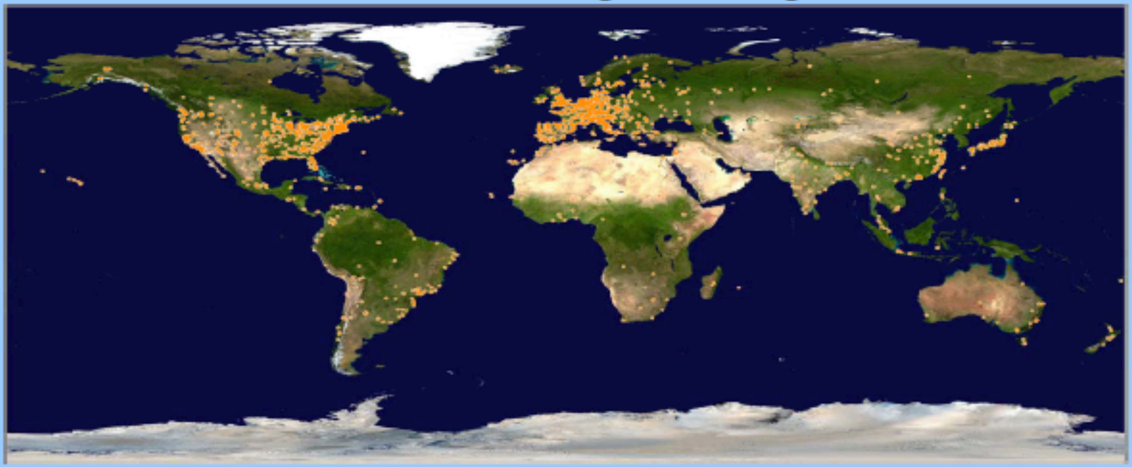
Science Gateway

Earth System Grid

Center for Enabling Technologies

Agencies

● Registered sites



ESG-CET enables Scientific Discovery in Climate Science by providing an international community of over 16,000 registered users with climate simulation data, climate models, analysis and visualization tools, and enabling technologies for a distributed, global science enterprise

ESG turns climate science data into community resources

Data warehouse, search and discovery, access, and reduction

Data used in hundreds of scientific papers

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE

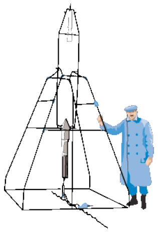
WMO

UNEP

Much of which provided a basis for the 4th Assessment Report of the IPCC

The Nobel Peace Prize 2007

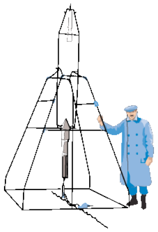
Intergovernmental Panel on Climate Change (IPCC)



Earth System Grid Center for Enabling Technology (ESG-CET) Project [>10,000 users]
Key Multi-Model Data Archive for World Climate Research Program

[Sources: http://esg-pcmdi.llnl.gov/ & http://www.scidacreview.org/0902/html/esg.html]

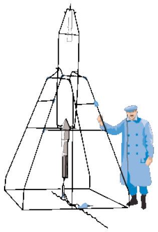
	# of Expr.	# of Models	# of Ctrs.	# of Files	# of TB	# of TB Downloaded
CMIP3 (IPCC AR4) [using models with 140-250 km res.]	12	25	17	~80K	>35	>425
CCSM					~160	>35
+Others thru 2006					~250 total	
CMIP5 (IPCC AR5) [using models with much higher res.]	3x12+	>40	~30			
+Others thru 2011					nx10K total	



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

Topic Outline

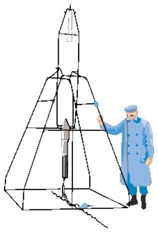
- Problem Statement, Motivation & Background Context
  - NCCS: Increasing capacities/capabilities & “PI distribution”
  - Success of prior >10 Gbps network testbed efforts
  - More 40 & 100 Gbps network technologies on the way
- Value Proposition & Solution Approach
- Multi-Phased Network Testbed Objectives
- Phase 1 Network-Test Workstations
- Early Stages of Phase 1 & 2 Testbeds
  - WAN File Accessing Experiments/Demonstrations at SC09
  - Build out of Phase 1& 2 Testbeds
- Current Status and Next Steps



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

Nuttcp (pronounced as new-t-t-c-p or nut-t-c-p)

- Primary author Bill Fink (william.e.fink@nasa.gov), with Rob Scott (rob@hpcmo.hpc.mil).
- Great follow-on to original ttcp network throughput performance measurement and troubleshooting tool. Highly recommended. One of the best!
- Over 60 examples of use included in Phil Dykstra's noteworthy tutorial for High Performance Data Transfer (at SC0x's).
- Advanced capabilities/features/options still being added to enable more sophisticated use, while retaining ease-of-use defaults.
- At http://www.nuttcp.net & included in perfSONAR's liveCD.

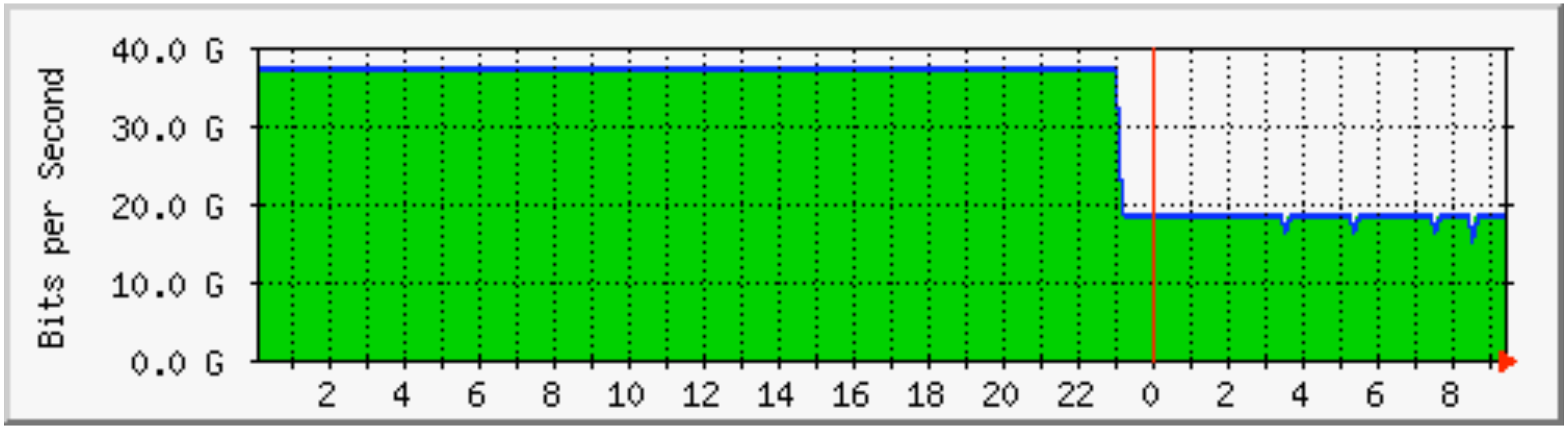


A Sample Traffic Analysis During MAX's 40-Gbps Interface Testing Between CLPK and MCLN

A joint testing effort among Fujitsu, Juniper, MAX & NASA/GSFC

The statistics were last updated **Tuesday, 14 October 2008 at 9:27**

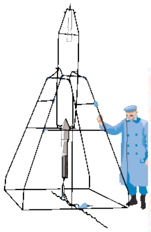
`Daily' Graph (5 Minute Average)



Max In:37.5 Gb/s (94.3%)Average In:31.5 Gb/s (7.8%)Current In:18.7 Gb/s (47.0%)
Max Out:37.6 Gb/s (94.4%)Average Out:31.5 Gb/s (7.8%Current Out:18.7 Gb/s (47.0%)

Snapshot taken during MAX NASA JUNIPER FUJITSU 40G INTEROP Test: One 10G stream from each of two pair of NASA line-rate nuttcp-servers is hard looped between some Juniper T1600 10G interfaces to fill the Juniper OC768c interfaces and Fujitsu 40G optical transponders set up between MAX POP's at College Park, MD and McLean, VA.

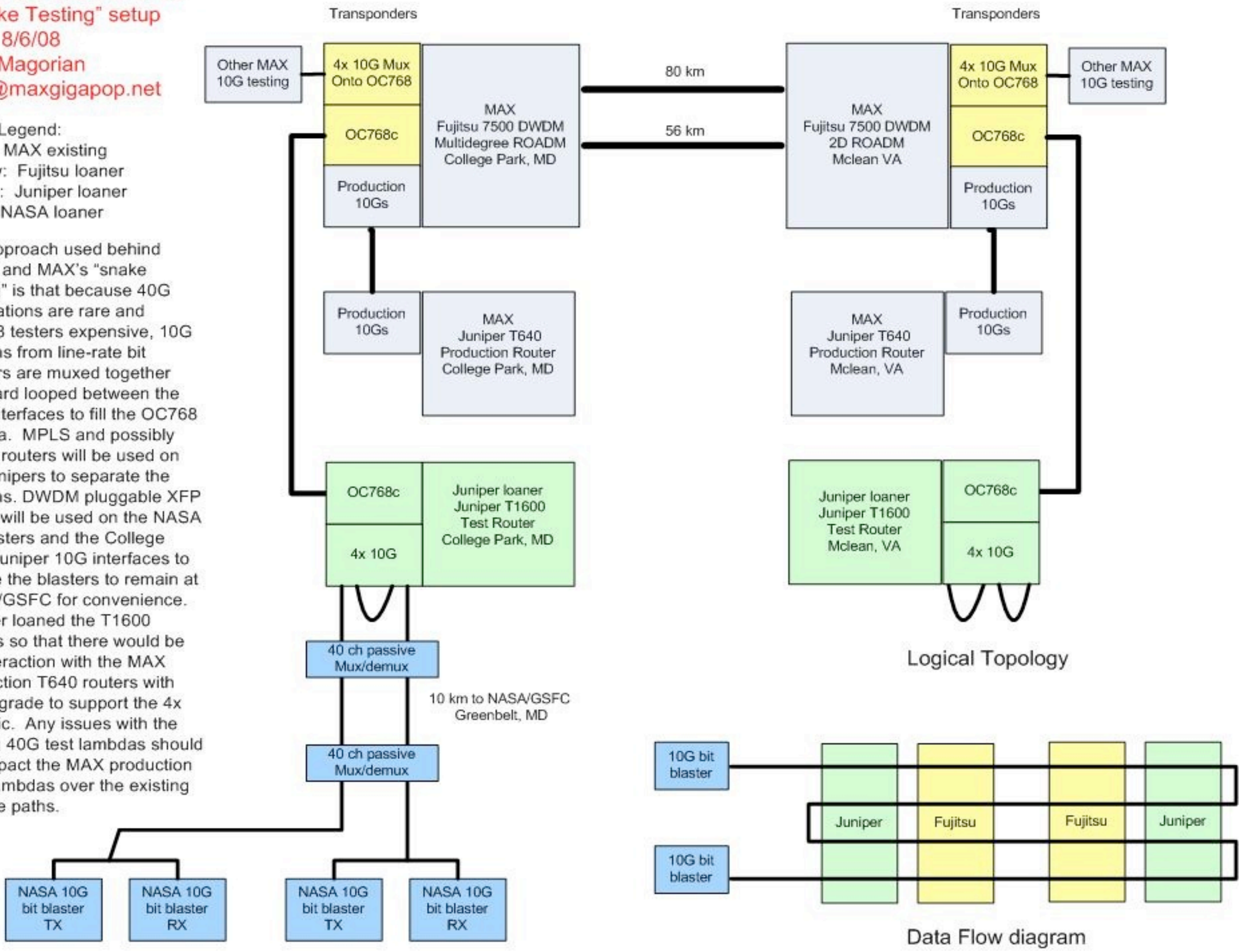
Source: Dan Magorian (MAX)



MAX NASA JUNIPER
FUJITSU 40G INTEROP
"Snake Testing" setup
V1.1 8/6/08
Dan Magorian
40g@maxgigapop.net

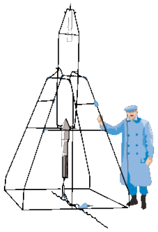
Color Legend:
Grey: MAX existing
Yellow: Fujitsu loaner
Green: Juniper loaner
Blue: NASA loaner

The approach used behind NASA and MAX's "snake testing" is that because 40G applications are rare and OC768 testers expensive, 10G streams from line-rate bit blasters are muxed together and hard looped between the 10G interfaces to fill the OC768 lambda. MPLS and possibly virtual routers will be used on the Junipers to separate the streams. DWDM pluggable XFP optics will be used on the NASA bit blasters and the College Park Juniper 10G interfaces to enable the blasters to remain at NASA/GSFC for convenience. Juniper loaned the T1600 routers so that there would be no interaction with the MAX production T640 routers with the upgrade to support the 4x 10G pic. Any issues with the Fujitsu 40G test lambdas should not impact the MAX production 10G lambdas over the existing diverse paths.



Logical Topology

Data Flow diagram



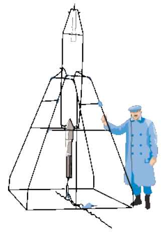
Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

40 Gbps Network Testing Between CLPK and MCLN

- Very successful; very informative
- Summary in “Industry Leaders Collaborate on 40 Gbps Live Network Trial” press release at http://www.businesswire.com/portal/site/google/?ndmViewId=news_view&newsId=20081110005564&newsLang=en

BUT

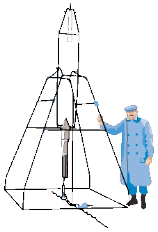
- No file copying application testing
- “Toothpaste-looping” flows, while innovative and sufficient for intervening network testing, are inappropriate for file copy testing
- Need higher performing network-test workstations for investigations focused on application throughput limitations



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

Even 100 Gbps Networks Are Beginning To Emerge

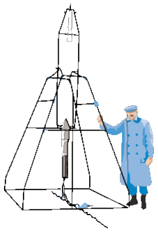
- The Internet2 and the National LambdaRail (NLR) are upgrading their infrastructure to support 40-to-100 Gbps wavelengths
- IEEE P802.3ba Task Force ratified both the 40 and 100 Gbps Ethernet (GE) standards by on July 19, 2010 (http://www.ieee802.org/3/ba/index.html)
- A partnership among ESNet, Internet2, Juniper Networks, Infinera, and Level3 Communications announced the formation of a 100 GE testbed (https://mail.internet2.edu/wws/arc/i2-news/2008-11/msg00000.html)
- http://www.hpcwire.com/offthewire/ESnet-Receives-62M-to-Develop-Worlds-Fastest-Computer-Network-52989552.html?viewAll=y



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

Additional 40G & 100G Emerging Network Technology Testing

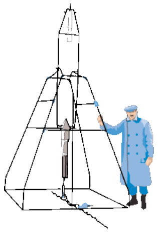
- 40G Cisco transponders in NLR
  - See http://www.nlr.net/release.php?id=52
- 100G Infinera transponders/optical switches, Juniper router interfaces, and Finisar, Opnext & Reflex Photonics CFP transceivers
  - See http://www.networkworld.com/community/node/59599
- Various 40G and 100G WAN trials
  - See http://www.networkworld.com/news/2009/111909-100g-ethernet-cheatsheet.html
- Extreme Networks intros 40GE at under $1,000 per Port
  - See http://investor.extremenetworks.com/releasedetail.cfm?ReleaseID=463163



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

Value Proposition (Time Is Money)

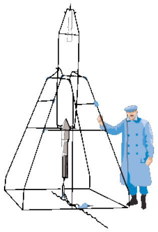
- Time-decreased data flows increase time for other productive work with the data
- Achieving time-efficient data flows over wide areas is notoriously problematic (“un-tuned” read/write/protocols, deficient disk I/O performance, …)
  - Even over 10-Gbps networks, throughput is often only ~10-Mbps, so copying a single 10-GB file typically takes ~17 minutes
- Need to determine 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

Part of the Solution

- Prepare applicable testbeds to identify the bottlenecks and investigate solutions/alternatives
  - Plan multi-phased LAN, MAN & WAN testbeds with 20, 40 & 100 Gbps throughput performance objectives
  - Assemble low-cost high-performance network-test workstations to assess/troubleshoot:
    - New network technologies
    - File copying application tuning
  - Arrange joint collaboration on WAN file-accessing applications with NCCS and NASA Advanced Supercomputing (NAS) Division data management experts to ensure that the network testbed efforts will have at least HEC Program applicability
- Transition findings into production environments



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

Multi-Phased Network Testbed Objectives

- A series of partly overlapping approximately-two-year Phases
- Each producing a demonstrable network capability with end-to-end throughput performance goals targeted respectively at 20, 40 & 100 Gbps
- Each with joint collaboration on WAN file-accessing applications with NCCS and NAS data management experts to ensure that the network testbed efforts will have at least HEC Program applicability

[1] P. Gary, “Plan/Proposal For Initiation of 20, 40 & 100 Gbps Network Technology Testbeds”, 29Jan09 draft proposal for potential submission to the NASA Strategic Investment Business Case (SIBC) Initiative for Networking R&D that is expected to emerge soon within NASA.

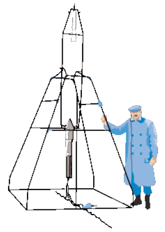
[2] D. Duffy et al, “NASA High End Computing Selected Wide Area Network Testing: Representative Test Plan”, 18Mar09 draft.

Both [1] and [2] were emailed to HEC’s network engineers at NAS and NCCS.

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

Example Throughput Performance Benchmarking

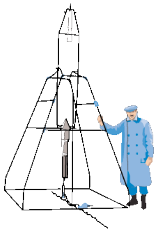
- Several representative test types are described in [2], including ones involving:
  - nuttcp at wire transfer speeds for the much needed performance baseline against which all subsequent performance measurements will be compared
  - scp, bbftp, nfs, iRODS, GridFTP and other file transfer mechanisms that NASA HEC users jobs use at both NAS and the NCCS to move data back and forth over the wide area
  - iSER and iSCSI over RDMA to mount the disks either locally or remotely
  - GPFS and/or Lustre shared file system
- Actual throughput performance results will be widely distributed once they are available



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

Notional Milestone Schedule

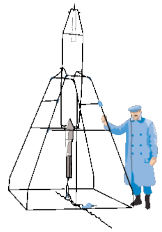
```
                       |    CY09   |   CY10    |    CY11   |   CY12    |
                       JFMAMJJASONDJFMAMJJASONDJFMAMJJASONDJFMAMJJASOND
Phase 0 10 Gbps Testbeds
O LAN & Region/MAN       -------**
O WAN                      ------*
Phase 1 20 Gbps Testbeds
O LAN & Region/MAN            ---------***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



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

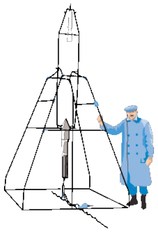
Phase 1 Network-Test Workstation Functional Objectives with Performance Targets (1 of 2)

- “B” (Baseline) systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 40-Gbps unidirectional, 40-Gbps bidirectional (80-Gbps “total”)
  - Secondarily for WAN file copying application throughput evaluations in disk-to-disk testing at up to 10-Gbps unidirectional
- “C” systems:
  - Primarily for WAN file copying application throughput evaluations in disk-to-disk testing at up to 20-Gbps unidirectional
- “A” systems:
  - Primarily for WAN delay emulation at up to 40-Gbps unidirectional, 40-Gbps bidirectional (80-Gbps “total”)
  - Also as firewall at up to 20-Gbps unidirectional, 20-Gbps bidirectional (40-Gbps “total”)



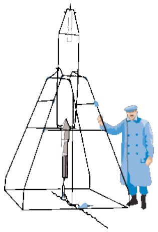
Phase 1 Network-Test Workstation Functional Objectives with Performance Targets (2 of 2)

- “A+” systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 70-Gbps unidirectional, 40-Gbps bidirectional (80-Gbps “total”)
    - Actual performance: On 12Jun09 using eight streams between two A+ systems connected via eight 10GE’s, measured an aggregate performance of 69.2907 Gbps unidirectional, and bidirectional 38.6955 Gbps transmit & 38.5842 Gbps receive (77.2797 Gbps total aggregate)
- “A-” systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 20-Gbps unidirectional, 20-Gbps bidirectional (40-Gbps “total”)



Phase 1.1 Network-Test Workstation Functional Objectives with Performance Targets

- “A++” systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 100-Gbps unidirectional, 50-Gbps bidirectional (100-Gbps “total”)
- Actual performance “in-progress”
  - On 6Aug09 measured an aggregate performance of 100.4637 Gbps in transmits; but currently only up to 56.4703 Gbps in receives
    - Test configuration has each of the two quad-core Xeon processors of one A++ system connected via six 10GE’s to one of two quad-core i7-based A+ systems
    - Twelve streams are generated – one for each of the twelve 10GE connections handled by the one A++ system



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

Approximate Costs (With components acquired via SEWP IV in lot-sizes of 3 - 15, and self assembly) of Phase 1 & 1.1 Network-Test Workstations

“B” System:	~$6.8K
“C” System:	~$9.0K
“A” System:	~$4.6K
“A+” System:	~$6.5K
“A-” System:	~$3.6K
“A++” System:	~$11.1K

- For more detail, contact Paul.Lang@nasa.gov

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.

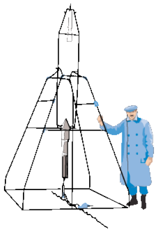
- Each workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with two HighPoint RocketRaid 4320 RAID disk controllers and a Myricom 10 Gigabit Ethernet network interface card in the PCIe Gen2 slots of a Asus P6T6 WS Revolution motherboard. Each RAID controller hosts eight 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



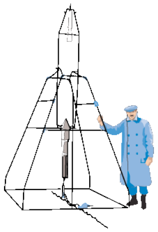
Introduction To
NASA HEC WAN File Accessing
Experiments/Demonstrations At SC09

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

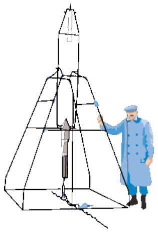
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system in a 40km MAN testbed)
  - File copy: 9402.0330-Mbps 56% TX 98% RX 0 retrans 0.45ms RTT



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]

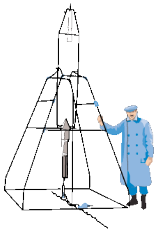
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 9548.0962-Mbps 59% TX 97% RX 0 retrans 80.15ms RTT (completed in 57.58 seconds)
    *With receiver B-system over-clocked to 3.4-Ghz instead 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]

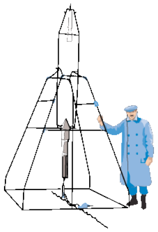
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 5055.1438-Mbps 31% TX 59% RX 8 retrans 80.15ms RTT (completed in 108.75 seconds)
    *With receiver B-system over-clocked to 3.4-Ghz instead 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



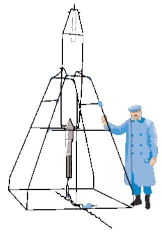
Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Additional 10-Gbps netem-enabled-WAN Sample Test Results (nuttcp-based) Between Two "B-Systems**"* [Source: Bill Fink/GSFC]

*Also see: *Nuttscp Sample Test Results Between Two "B-Systems" (2 of 4)* & *(3 of 4)*

**With receiver B-system over-clocked to 3.4-Ghz instead of 3.2-Ghz

- With Large Receive Offload on the myri10ge driver enabled
  - 30 second test using Linux TCP autotuning:
    6053.8558-Mbps 15% TX 24% RX 0 retrans 80.14ms RTT
  - 30 second test using manually specified 100 MB TCP window:
    6796.0992-Mbps 16% TX 27% RX 0 retrans 80.15ms RTT
- With Large Receive Offload on the myri10ge driver disabled
  - 30 second test using Linux TCP autotuning:
    7029.8505-Mbps 19% TX 29% RX 0 retrans 80.15ms RTT
  - 30 second test using manually specified 100 MB TCP window:
    9442.1071-Mbps 27% TX 39% RX 0 retrans 80.15ms RTT



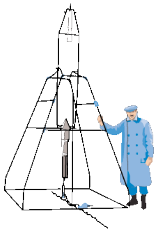
Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Sample 4x10-GigE Bonding Test Results (nuttcp-based) Between Two "B-Systems" in Back-to-Back Direct Connection [Source: Bill Fink/GSFC]

- Kernal L2 load-balanced round-robin bonded interface (aka Link Aggregation)
  - 10 second test:
    31615.4616-Mbps 99% TX 95% RX 31 retrans 0.05ms RTT

- Nuttcp "application bonding" using 4 streams (each across its own 10-GigE path)
  - 10 second test:
    39564.4536-Mbps 81% TX 94% RX 0 retrans 0.11ms RTT

In both cases the use of the "correct" CPU made a significant difference in the achieved network performance. Unfortunately the "correct" CPU did not seem to be deterministic.



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

More 4x10-GigE Bonding Test Results (nuttcp-based) Between Two "B-Systems"* [Source: Bill Fink/GSFC]

- Nuttcp "application bonding" using 4 streams (each across its own 10-GigE path)
  - 10 second test:
    39134.0831-Mbps   99% TX   91% RX   1 = 0+0+1+0 retrans   0.11ms RTT
  - 10 second test:
    39151.9019-Mbps   91% TX   92% RX   1 = 0+0+1+0 retrans   0.11ms RTT
  - 10 second test:
    39318.0384-Mbps   80% TX   90% RX   1 = 0+0+1+0 retrans   0.10ms RTT
  - 10 second test:
    39406.0384-Mbps   79% TX   92% RX   1 = 0+0+1+0 retrans   0.10ms RTT

*Obtained while testing nuttcp-7.1.1's new features for:

- Improved multilink aggregation specification options (e.g., stride & dotted quad)
- Providing summary TCP retrans info for multi-stream TCP (with per-stream info for Linux)
- Allowing local name resolution to occur for third party nuttcp tests if the remote third party host can't resolve the specified test hostname

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.

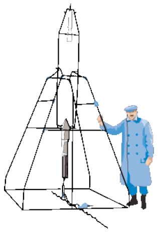
- Each workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with four HighPoint RocketRaid 4320 RAID disk controllers and a Myricom 2-port 10 Gigabit Ethernet network interface card in the PCIe Gen2 slots of a Asus P6T6 WS Revolution motherboard. Each RAID controller 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



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

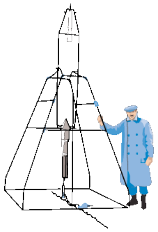
Precursor Tests of "C-Systems" (to show the individual components have the necessary muscle) [Source: Bill Fink/GSFC]

- Disk I/O speeds via dd reads (of=/dev/null) & writes (if=/dev/zero)
  - Read: 68719476736 bytes (69 GB) copied, 25.8791 s, 2.7 GB/s
  - Write: 68719476736 bytes (69 GB) copied, 26.8676 s, 2.6 GB/s

- 2x10-GigE via "nuttcp application bonding"
  - TX: 19805.8537 Mbps 34 %TX 59 %RX 0 retrans 0.11 msRTT
  - RX: 19808.7300 Mbps 39 %TX 53 %RX 0 retrans 0.11 msRTT

Nuttscp Sample Test Results Between Two "C-Systems" (1-of-7) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between four RAID5 disk controllers nested as RAID50 hosted on each C-system in a LAN testbed)
  - Configuration settings:
    - LRO enabled
    - eth2,3 interrupts on CPU0
    - nuttcp application running on CPU1
    - 4xHPT RAID5 interrupts running on CPU2
    - md RAID50 across above
  - Get: 10273.4125 Mbps 52 %TX 99 %RX 0 retrans 0.11 msRTT
  - Put: 10311.2700 Mbps 52 %TX 99 %RX 0 retrans 0.11 msRTT
- Houston, we have a problem! We're definitely not firing on all cylinders. It's obvious what the problem is, namely that the receiver CPU is totally saturated. To go faster is going to require nuttcp using multiple cores in parallel….



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

Nuttscp Sample Test Results Between Two "C-Systems" (2-of-7) [Source: Bill Fink/GSFC]

- One 64-GB file copy similar to "1-of-7" but only one side's RAID50 is real
  - Configuration settings: same as in "1-of-7"
  - Get from RAID50 to /dev/null:
    17324.4416 Mbps 98 %TX 49 %RX 0 retrans 0.11 msRTT
  - Put from /dev/zero to RAID50:
    10129.7218 Mbps 27 %TX 99 %RX 0 retrans 0.11 msRTT
- So, the immediate 20-Gbps challenge is primarily on the write side….

Nuttscp Sample Test Results Between Two “C-Systems” (3-of-7) [Source: Bill Fink/GSFC]

- Two 64-GB file copy (between four RAID5 disk controllers nested as RAID50 hosted on each C-system in a LAN testbed)
  - Configuration settings: same as in “1-of-7” **plus**
    - nuttcp application running on CPU3
  - Put file1:
    7184.8745 Mbps 41 %TX 71 %RX 0 retrans 0.11 msRTT
  - Put file2:
    7082.7940 Mbps 46 %TX 70 %RX 0 retrans 0.11 msRTT
  - Aggregate throughput:
    14267.6685 Mbps
- Better; but there was a lot of disk head contention seeking back and forth between the two files

Nuttscp Sample Test Results Between Two "C-Systems" (4-of-7) [Source: Bill Fink/GSFC]

- A slight variation of "3-of-7", using individual 10-GigE nuttcp streams across individual 10-GigEpaths
  - Put file1:
    7136.7905 Mbps 39 %TX 72 %RX 0 retrans 0.11 msRTT
  - Put file2:
    7123.8836 Mbps 39 %TX 72 %RX 0 retrans 0.11 msRTT
  - Aggregate throughput:
    14260.6741 Mbps
- Basically the same result as "3-of-4"

Nuttscp Sample Test Results Between Two "C-Systems" (5-of-7) [Source: Bill Fink/GSFC]

- Splitting the one RAID50 into two separate RAID50s to avoid the disk head seeking contention
  - Configuration settings:
    - LRO enabled
    - eth2,3 interrupts on CPU0
    - nuttcp s2 application running on CPU1
    - 2xHPT RAID5 interrupts running on CPU2
    - first md RAID50 across above
    - 2xHPT RAID5 interrupts running on CPU2
    - second md RAID50 across above
    - nuttcp s1 application running on CPU3
  - Put file1/s1:
    9318.3251 Mbps 55 %TX 92 %RX 0 retrans 0.11 msRTT
  - Put file2/s2:
    7960.6777 Mbps 47 %TX 79 %RX 0 retrans 0.10 msRTT
  - Aggregate throughput:
    17279.0028 Mbps

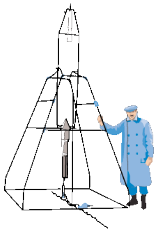
Nuttscp Sample Test Results Between Two "C-Systems" (6-of-7) [Source: Bill Fink/GSFC]

- Similar to "5-of-7" but moving the last 2 HPT RAID5 interrupts to CPU 0, so stream s2could have the same advantage as stream s1
  - Configuration settings:
    - LRO enabled
    - eth2,3 interrupts on CPU0
    - 2xHPT RAID5 interrupts running on CPU0
    - second md RAID50 across above
    - nuttcp s2 application running on CPU1
    - 2xHPT RAID5 interrupts running on CPU2
    - first md RAID50 across above
    - nuttcp s1 application running on CPU3
  - Put file1/s1:

    9161.1181 Mbps 55 %TX 94 %RX 0 retrans 0.11 msRTT
  - Put file2/s2:

    8663.7400 Mbps 52 %TX 89 %RX 0 retrans 0.11 msRTT
  - Aggregate throughput:

    17824.8581 Mbps (90% of maximum 19.8 Gbps)



Nuttscp Sample Test Results Between Two “C-Systems” (7-of-7) [Source: Bill Fink/GSFC]

- We are currently investigating SSD technology, to hopefully double our disk transfer speeds and get us into the 40-Gbps networked disk transfer realm

- But using parallelism of multiple cores and multiple streams is going to be key to going to 40-GigE, 100-GigE, and beyond speeds, since individual cores are not getting significantly faster

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 .

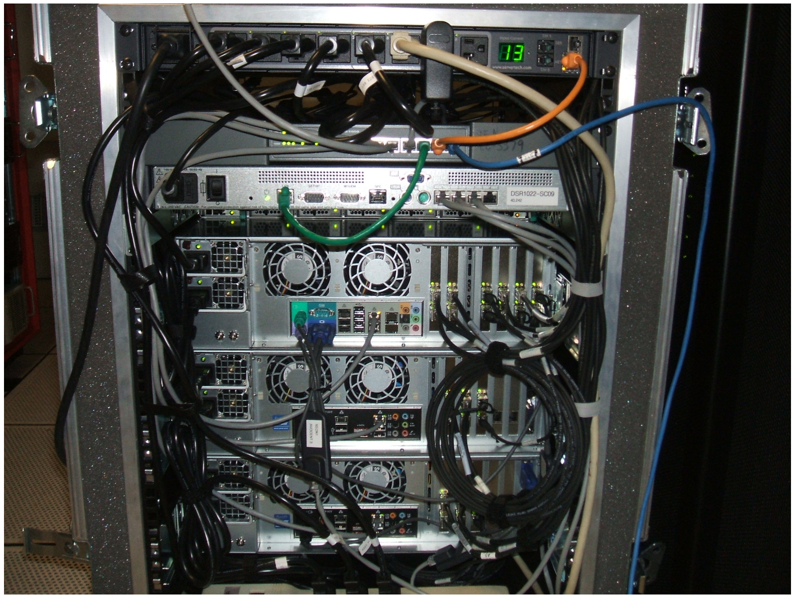
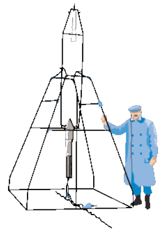
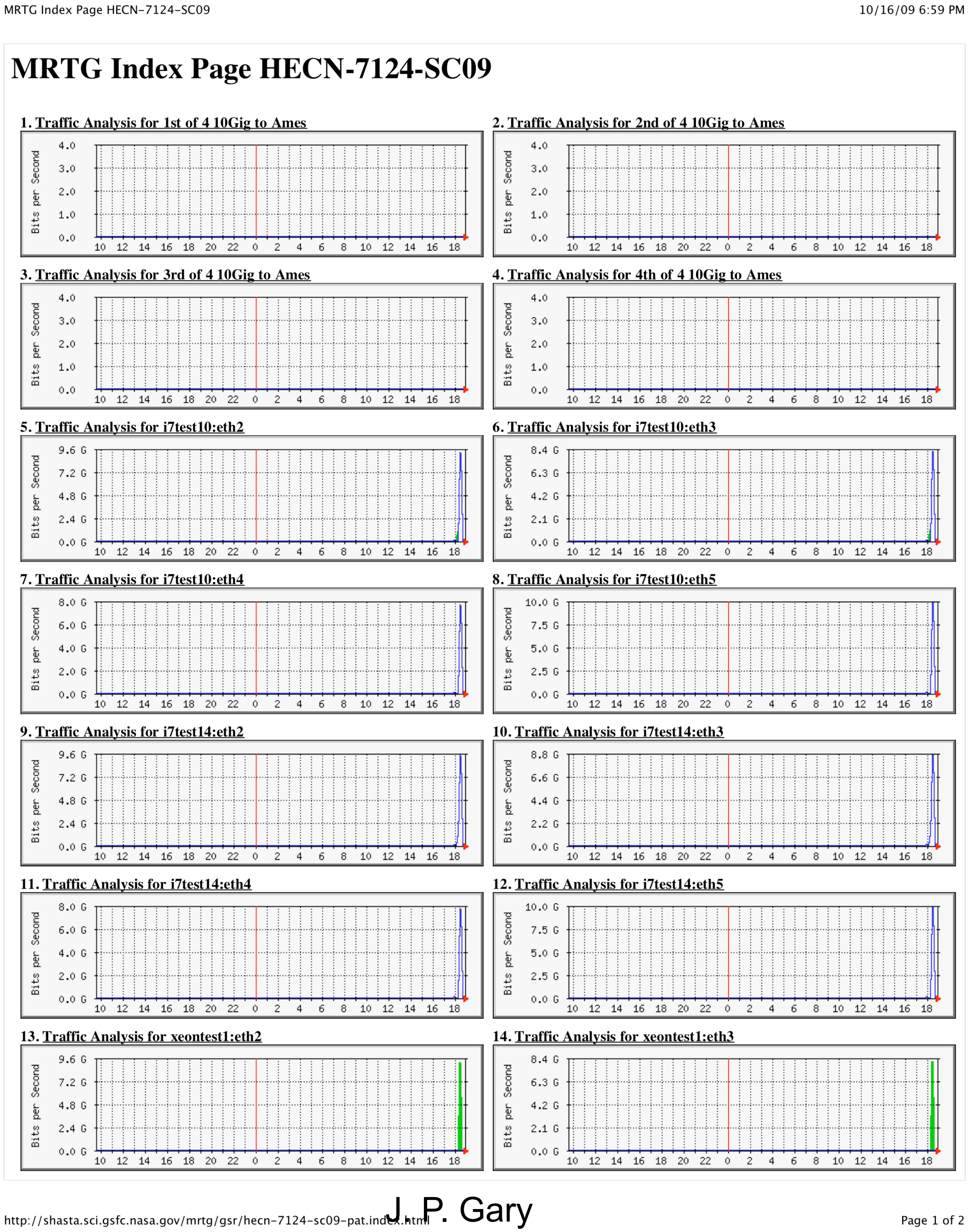


Figure: Xeon and two Core i7 workstations (bottom) interconnected with 10 Gigabit Ethernet switch and management units (top) in a rack for shipping to SC09.

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





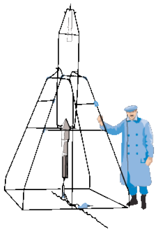
MRTG Index Page HECN-7124-SC09 10/16/09 6:59 PM

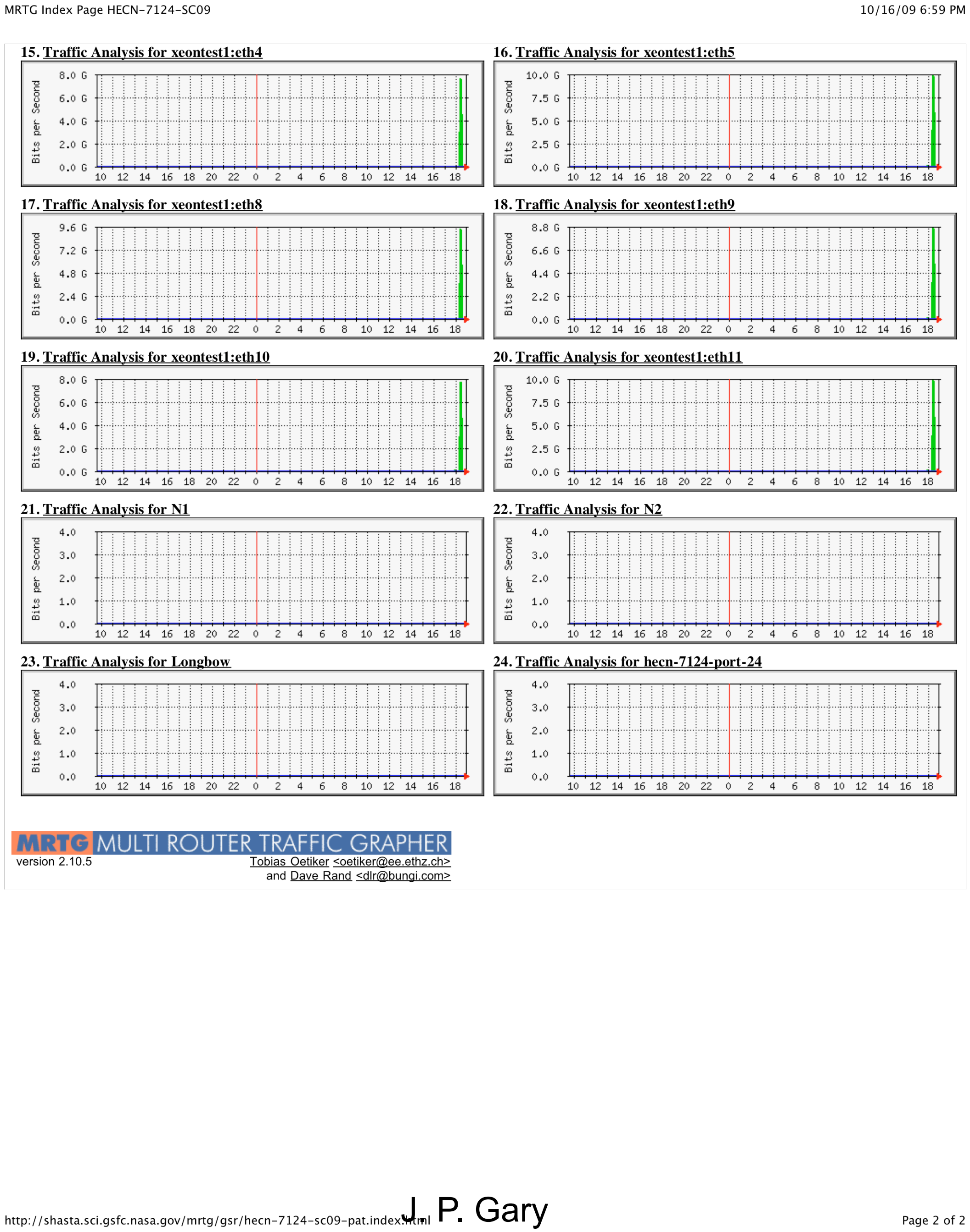
MRTG Index Page HECN-7124-SC09

1. Traffic Analysis for 1st of 4 10Gig to Ames
2. Traffic Analysis for 2nd of 4 10Gig to Ames
3. Traffic Analysis for 3rd of 4 10Gig to Ames
4. Traffic Analysis for 4th of 4 10Gig to Ames
5. Traffic Analysis for i7test10:eth2
6. Traffic Analysis for i7test10:eth3
7. Traffic Analysis for i7test10:eth4
8. Traffic Analysis for i7test10:eth5
9. Traffic Analysis for i7test14:eth2
10. Traffic Analysis for i7test14:eth3
11. Traffic Analysis for i7test14:eth4
12. Traffic Analysis for i7test14:eth5
13. Traffic Analysis for xeontest1:eth2
14. Traffic Analysis for xeontest1:eth3

Bits per Second
4.0 3.0 2.0 1.0 0.0
9.6 G 7.2 G 4.8 G 2.4 G 0.0 G
8.4 G 6.3 G 4.2 G 2.1 G 0.0 G
8.0 G 6.0 G 4.0 G 2.0 G 0.0 G
10.0 G 7.5 G 5.0 G 2.5 G 0.0 G
8.8 G 6.6 G 4.4 G 2.2 G 0.0 G
10 12 14 16 18 20 22 0 2 4 6 8 10 12 14 16 18

http://shasta.sci.gsfc.nasa.gov/mrtg/gsr/hecn-7124-sc09-pat.index.html Page 1 of 2





MRTG Index Page HECN-7124-SC09 10/16/09 6:59 PM

15. Traffic Analysis for xeontest1:eth4

16. Traffic Analysis for xeontest1:eth5

17. Traffic Analysis for xeontest1:eth8

18. Traffic Analysis for xeontest1:eth9

19. Traffic Analysis for xeontest1:eth10

20. Traffic Analysis for xeontest1:eth11

21. Traffic Analysis for N1

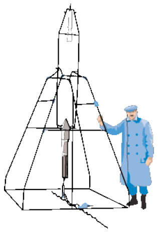
22. Traffic Analysis for N2

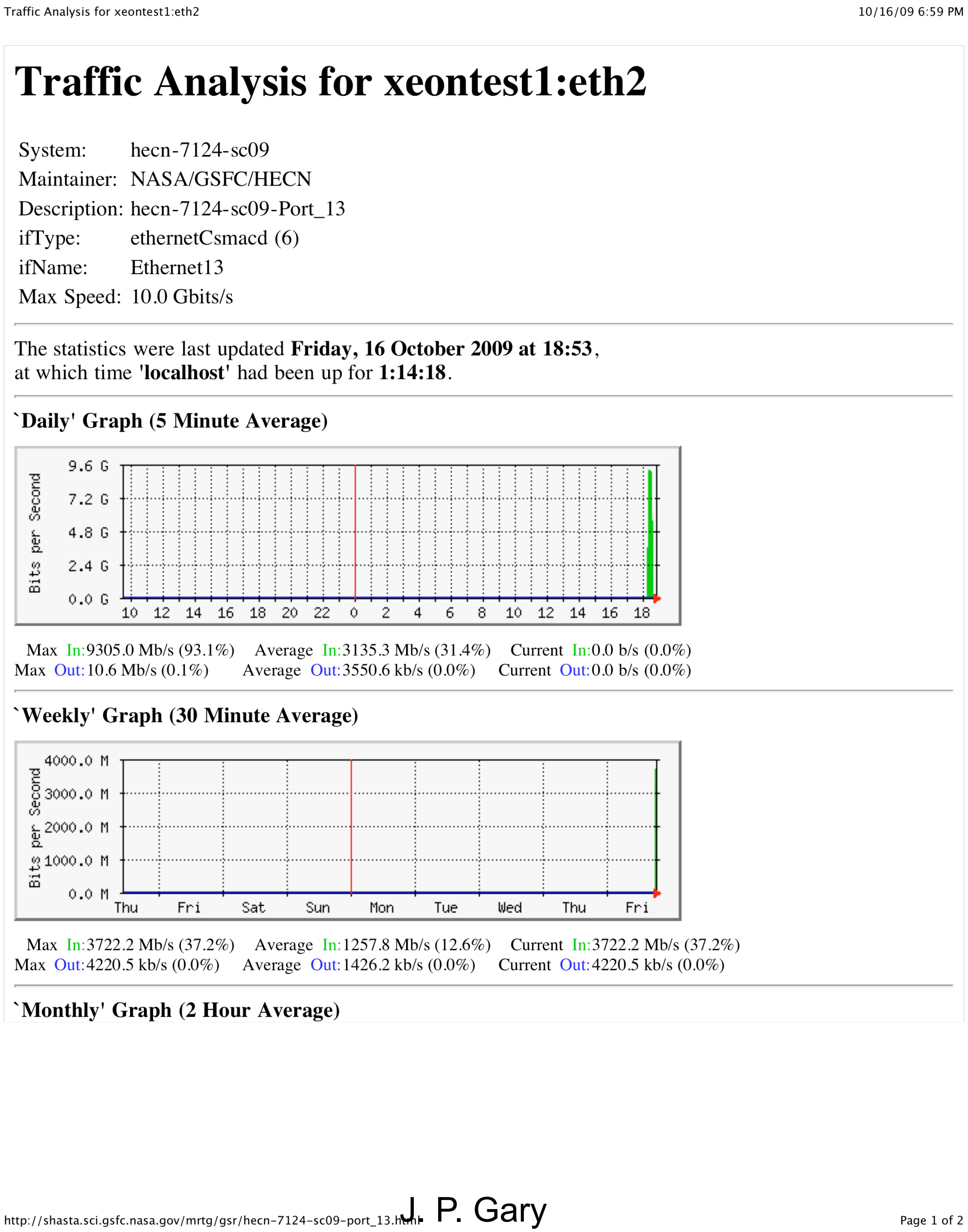
23. Traffic Analysis for Longbow

24. Traffic Analysis for hecn-7124-port-24

MRTG MULTI ROUTER TRAFFIC GRAPHER
version 2.10.5
Tobias Oetiker <oetiker@ee.ethz.ch>
and Dave Rand <dlr@bungi.com>

http://shasta.sci.gsfc.nasa.gov/mrtg/gsr/hecn-7124-sc09-pat.index.html Page 2 of 2





Traffic Analysis for xeontest1:eth2 10/16/09 6:59 PM

Traffic Analysis for xeontest1:eth2

System:	hecn-7124-sc09
Maintainer:	NASA/GSFC/HECN
Description:	hecn-7124-sc09-Port_13
ifType:	ethernetCsmacd (6)
ifName:	Ethernet13
Max Speed:	10.0 Gbits/s

The statistics were last updated **Friday, 16 October 2009 at 18:53**, at which time **'localhost'** had been up for **1:14:18**.

`Daily' Graph (5 Minute Average)

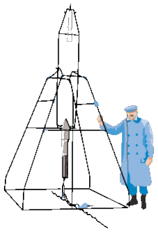
Max In:9305.0 Mb/s (93.1%) Average In:3135.3 Mb/s (31.4%) Current In:0.0 b/s (0.0%)
Max Out:10.6 Mb/s (0.1%) Average Out:3550.6 kb/s (0.0%) Current Out:0.0 b/s (0.0%)

`Weekly' Graph (30 Minute Average)

Max In:3722.2 Mb/s (37.2%) Average In:1257.8 Mb/s (12.6%) Current In:3722.2 Mb/s (37.2%)
Max Out:4220.5 kb/s (0.0%) Average Out:1426.2 kb/s (0.0%) Current Out:4220.5 kb/s (0.0%)

`Monthly' Graph (2 Hour Average)

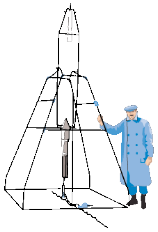
http://shasta.sci.gsfc.nasa.gov/mrtg/gsr/hecn-7124-sc09-port_13.html Page 1 of 2



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

Topic Outline

- Problem Statement, Motivation & Background Context
  - NCCS: Increasing capacities/capabilities & “PI distribution”
  - Success of prior >10 Gbps network testbed efforts
  - More 40 & 100 Gbps network technologies on the way
- Value Proposition & Solution Approach
- Multi-Phased Network Testbed Objectives
- Phase 1 Network-Test Workstations
- → Early Stages of Phase 1 & 2 Testbeds
  - WAN File Accessing Experiments/Demonstrations at SC09
  - Build out of Phase 1& 2 Testbeds
- Current Status and Next Steps



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

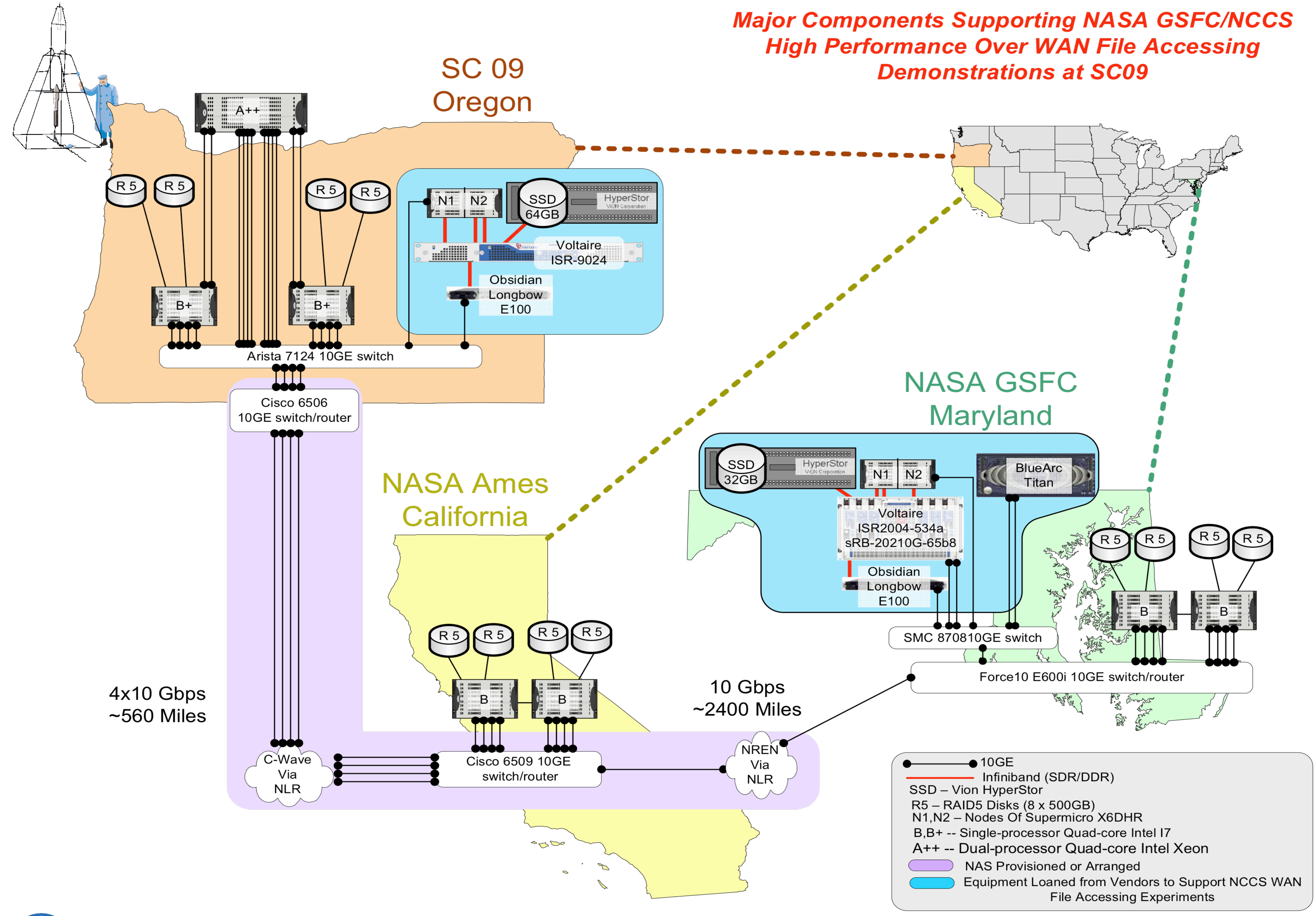
High Performance Wide Area Data Transfer Test Matrix

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	🔵	🟣	🟢	🔵 🟣 🟢		

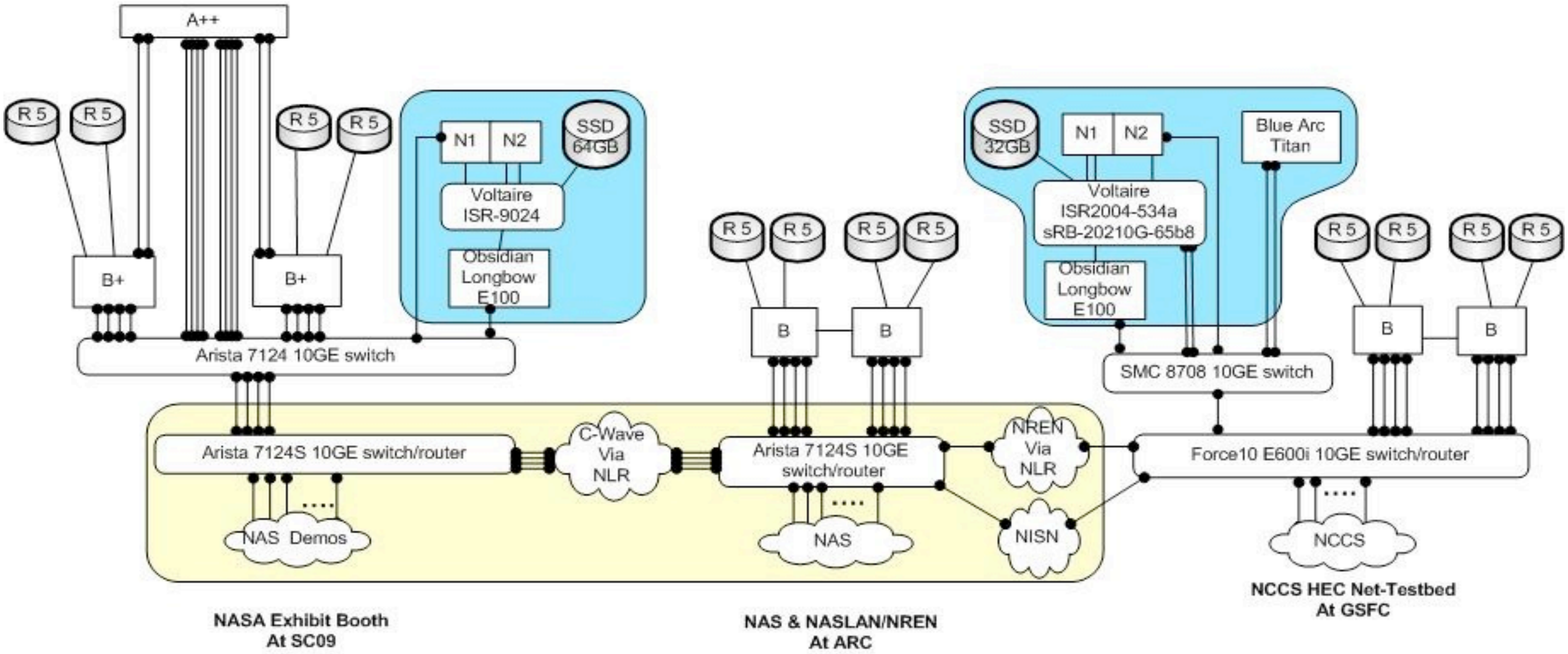
[1] Courtesy of Obsidian Research.

[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



●——● 10GE

—— 10G IB

SSD – Vion HyperStor

R5 – RAID5 Disks (8 x 500GB)

N1,N2 – Nodes Of Supermicro X6DHR

B,B+ -- Single-processor Quad-core Intel I7

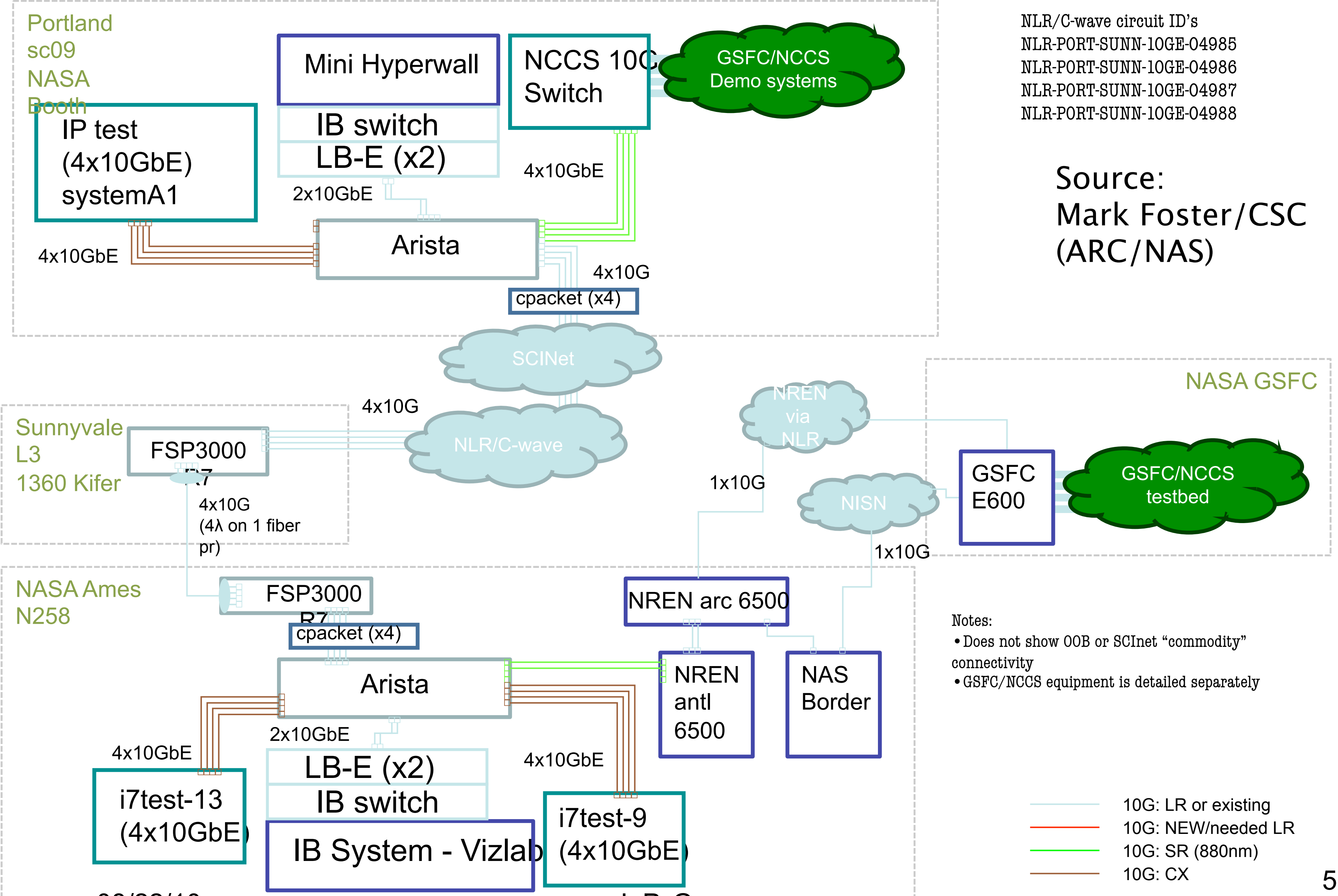
A++ -- Dual-processor Quad-core Intel Xeon

🟨 NAS Provisioned or Arranged

🟦 Equipment Loaned from Vendors to Support NCCS WAN File Accessing Experiments

J.P.Gary
12/7/09

SC09 NASA 20-40Gbps Demos WAN 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

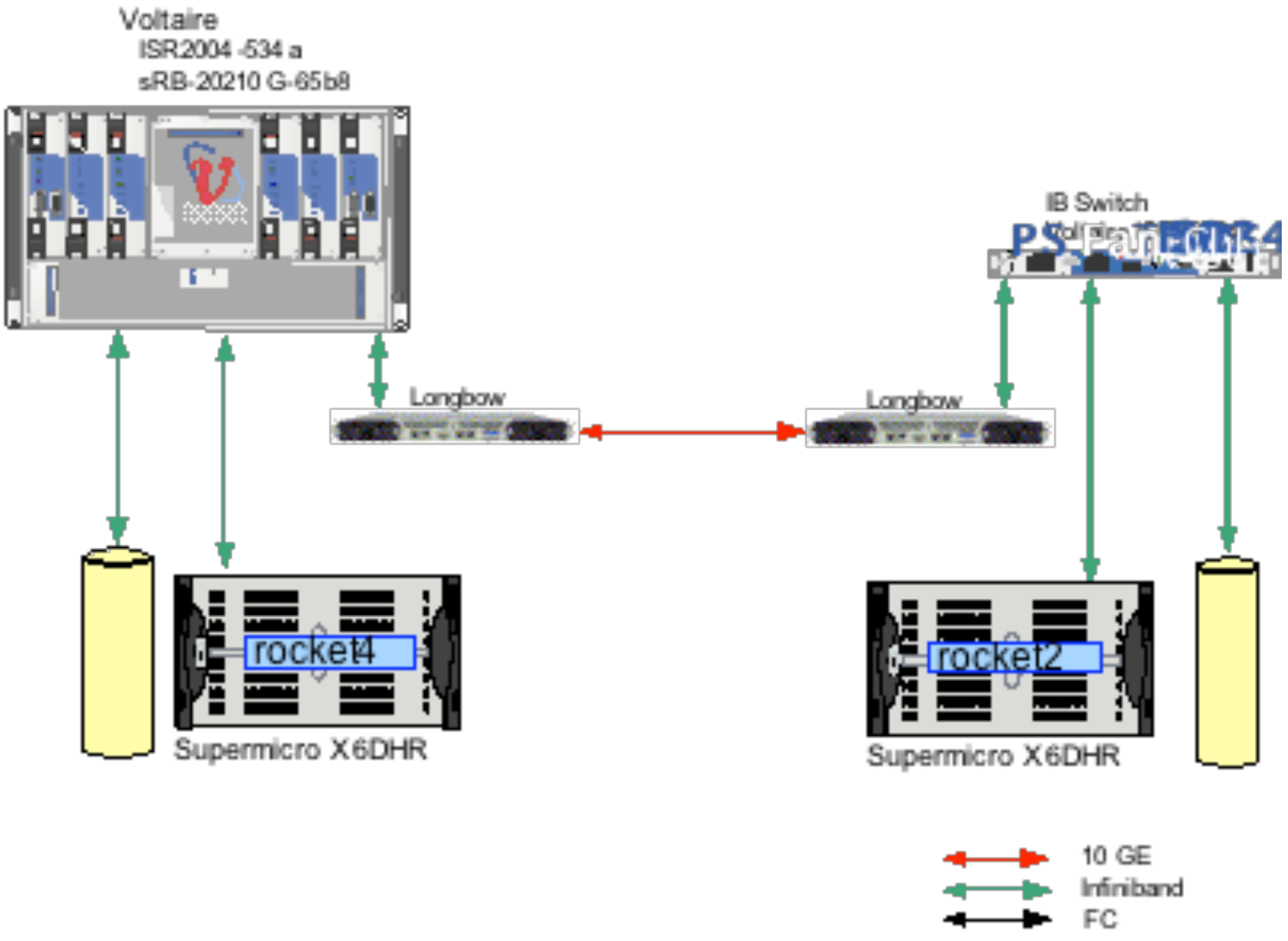
Source:
Mark Foster/CSC
(ARC/NAS)

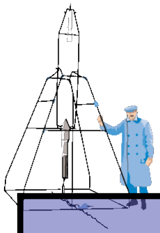
Notes:
- Does not show OOB or SCInet "commodity" connectivity
- GSFC/NCCS equipment is detailed separately

V2.6 04/06/10 mf

Pre SC Test Setup

Source: Hoot Thompson/PTP (GSFC/NCCS)





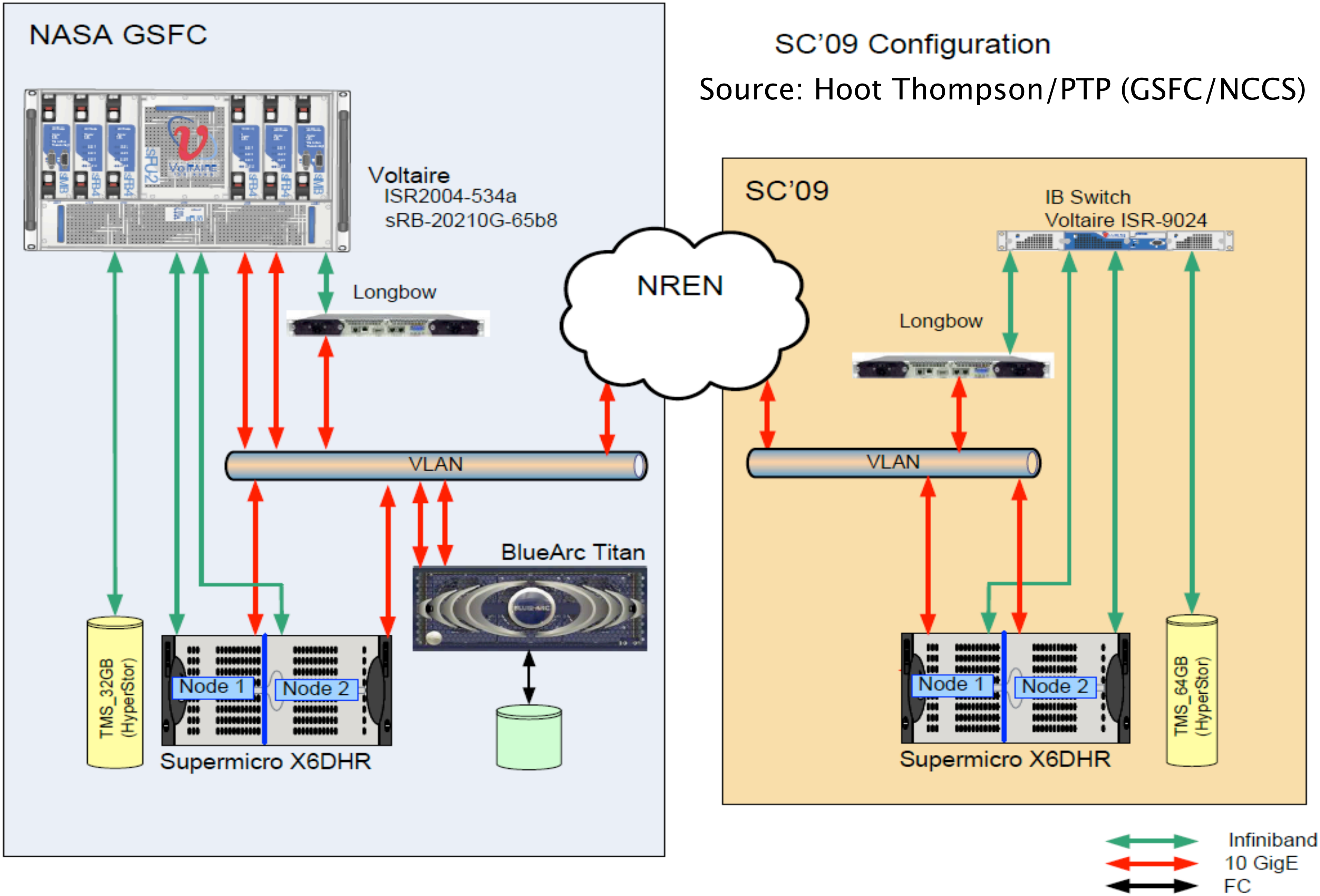
Test Results Pre-3Nov09 (pre-SC09)

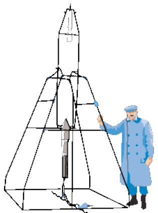
Source: Hoot Thompson/PTP (GSFC/NCCS)

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)





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

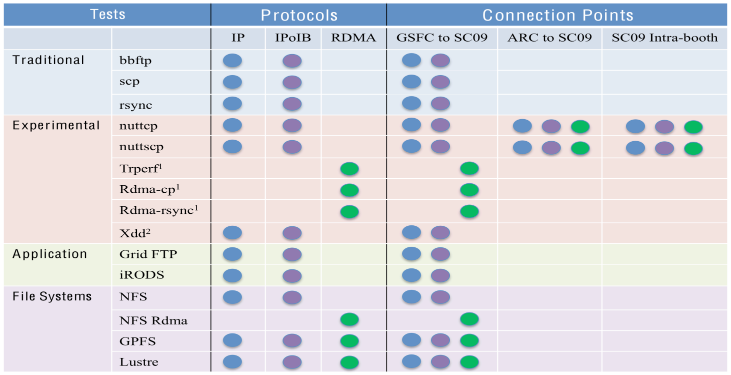
- Demonstrations of network-performance testing, wide-area file systems, and file transfer applications ranging from traditional to experimental were provided in the NASA research exhibit at the SC09 conference, Portland, OR, Nov. 16–19.

- 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

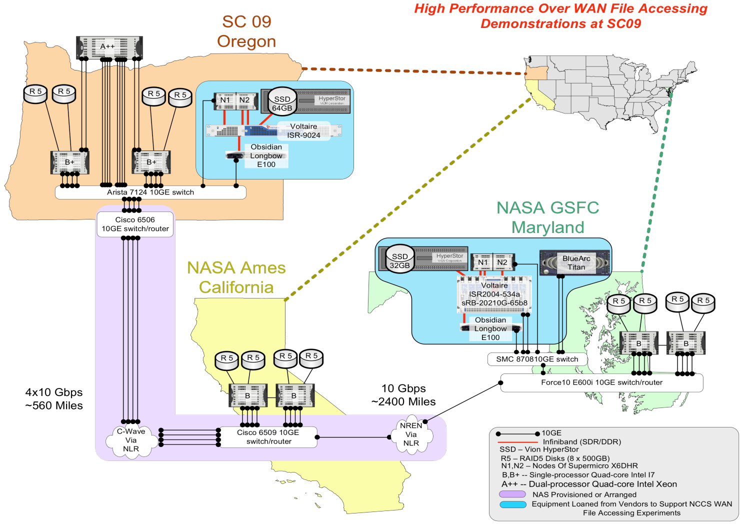
High Performance Wide Area Data Transfer Test Matrix



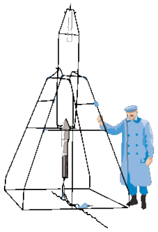
Tests		Protocols			Connection Points		
		IP	IPoIB	RDMA	GSFC to SC09	ARC to SC09	SC09 Intra-booth
Traditional	bbftp	🔵	🟣		🔵 🟣		
	scp	🔵	🟣		🔵 🟣		
	rsync	🔵	🟣		🔵 🟣		
Experimental	nuttcp	🔵	🟣		🔵 🟣	🔵 🟣 🟢	🔵 🟣 🟢
	nuttscp	🔵	🟣		🔵 🟣	🔵 🟣 🟢	🔵 🟣 🟢
	Trperf[1]			🟢	🟢		
	Rdma-cp[1]			🟢	🟢		
	Rdma-rsync[1]			🟢	🟢		
	Xdd[2]	🔵	🟣		🔵 🟣		
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.

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



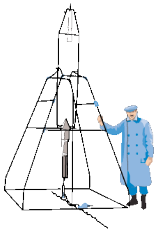
Figures: Data Transfer Test Matrix (Top) and WAN infrastructure and servers tested (bottom) during SC09.



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

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/

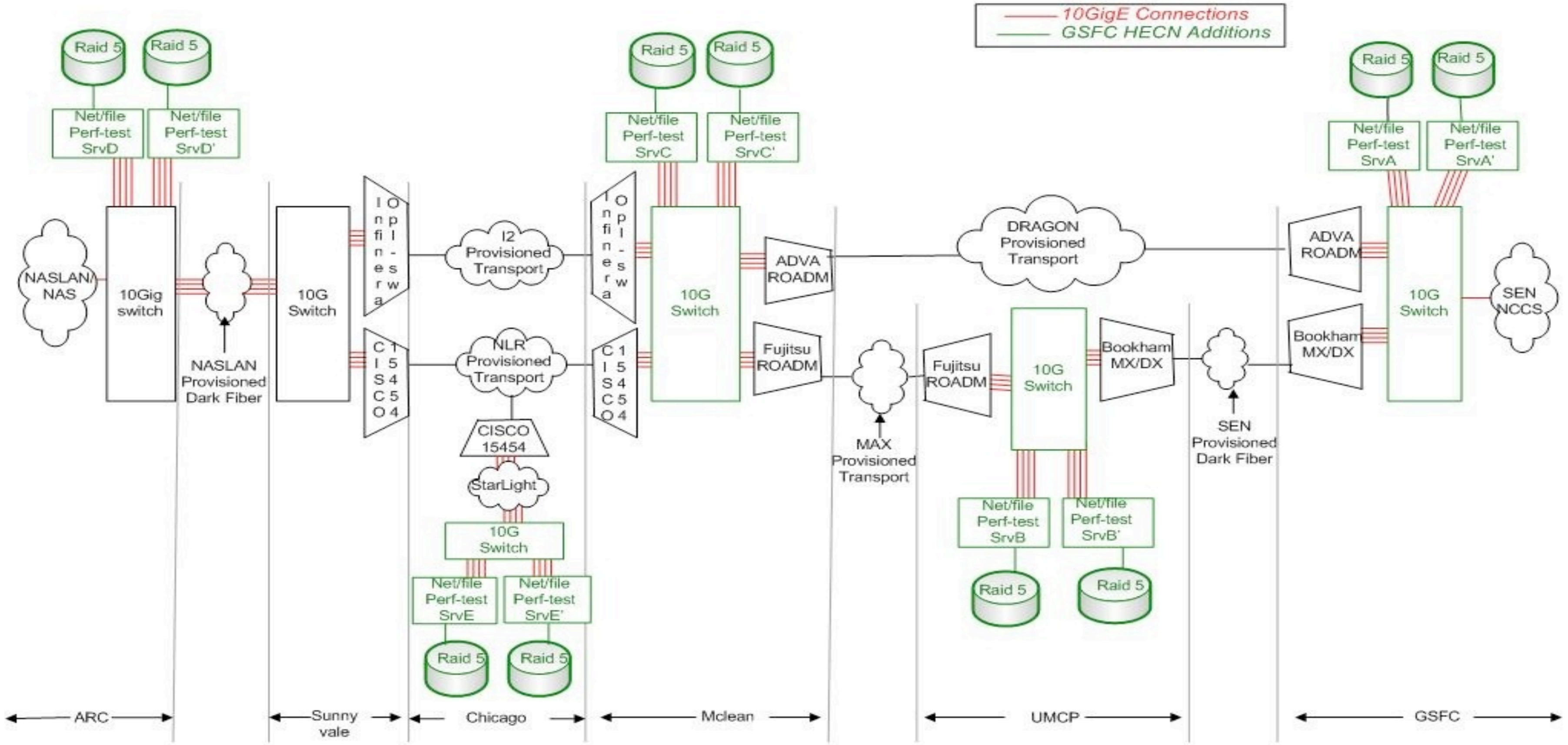


GSFC High End Computer Network (HECN) Project's Research Partners and Collaborators (Partial List)

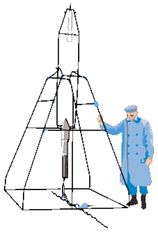
- *DRAGON Project: http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/WebHome*
  - *PI: Jerry Sobieski (UMCP)*
  - *GSFC L-Net on DRAGON network diagram: http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/Network*
- *e-VLBI Project: http://web.haystack.mit.edu/e-vlbi/evlbi.html*
  - *PI: Alan Whitney (MIT/Haystack)*
  - *GSFC L-Net on e-VLBI network diagram: http://cisto.gsfc.nasa.gov/L-Netpdfs/SC04_eVLBI_network.pdf*
- *GLIF: http://www.glif.is/*
  - *Chair: Kees Neggers (SURFnet)*
  - *GLIF network diagrams: http://www.glif.is/publications/#maps*
- *NGC IT Sector: http://www.it.northropgrumman.com/index.html*
  - *PI: Brice Womack (NGC)*
  - *GSFC L-Net on NGC IT Sector Colshire network diagram: http://cisto.gsfc.nasa.gov/L-Netpdfs/DRAGON_NGC_030606.pdf*
- *NLR: http://www.nlr.net/*
  - *CEO: Tom West (NLR)*
  - *NLR network diagram: http://www.nlr.net/infrastructure/*
- *NREN Project: http://www.nren.nasa.gov/*
  - *PM: Ken Freeman (ARC)*
  - *GSFC L-Net/SEN on NREN network diagram: http://cisto.gsfc.nasa.gov/L-Netpdfs/CENIC2006_13_mfoster_excerpts.pdf*
- *OptIPuter Project: http://www.optiputer.net/*
  - *PI: Larry Smarr (UCSD)*
  - *GSFC L-Net on OptIPuter network diagram: http://cisto.gsfc.nasa.gov/L-Netpdfs/SMARR-OptIPuter-AHM-gold.pdf*
- *TeraFlow Testbed Project: http://www.teraflowtestbed.net/*
  - *PI: Robert Grossman (UIC)*
  - *GSFC L-Net on TeraFlow Testbed network diagram: http://www.ncdm.uic.edu/maps/index.jpeg*

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

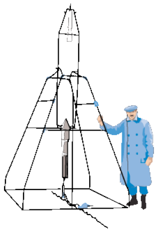


J.P.Gary/A.Muppalla
6/17/09



Key Notes Regarding the Previous Diagram (1 of 2)

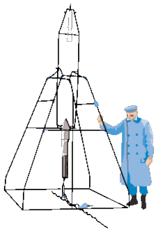
- The diagram's title is "Current Candidate 40-Gbps MAN & WAN Pathways For Use During Early Stages of Phase 1 20-Gbps and Phase 2 40-Gbps Testbeds"; and the referenced Testbeds were identified and/or described in [1] and [2]
- Where different but parallel pathways are shown (such as Internet2 vs NLR, and DRAGON's ADVA vs MAX's Fujitsu MAN pathways to MCLN) only one WAN and one MAN pathway is needed; but we'll have enough ports on our 10-GE switches to accommodate both WAN pathways and both MAN pathways without re-cabling if the opportunities should arise
- In the Early Stages, 4x10-GE link aggregation (LAG) is key (vs later stages if/when 40-GE Media Access Controllers (MACs) become available and affordable)



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

Key Notes Regarding the Previous Diagram (2 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 plan/proposal 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 plan/proposal
  - Costs of 40 and 100 Gbps Layer 1+2+3 router interfaces are likely to be two or more orders of magnitude greater than 40 and 100 Gbps Layer 1+2 Ethernet switch interfaces which are sufficient to enable the needed VLANs



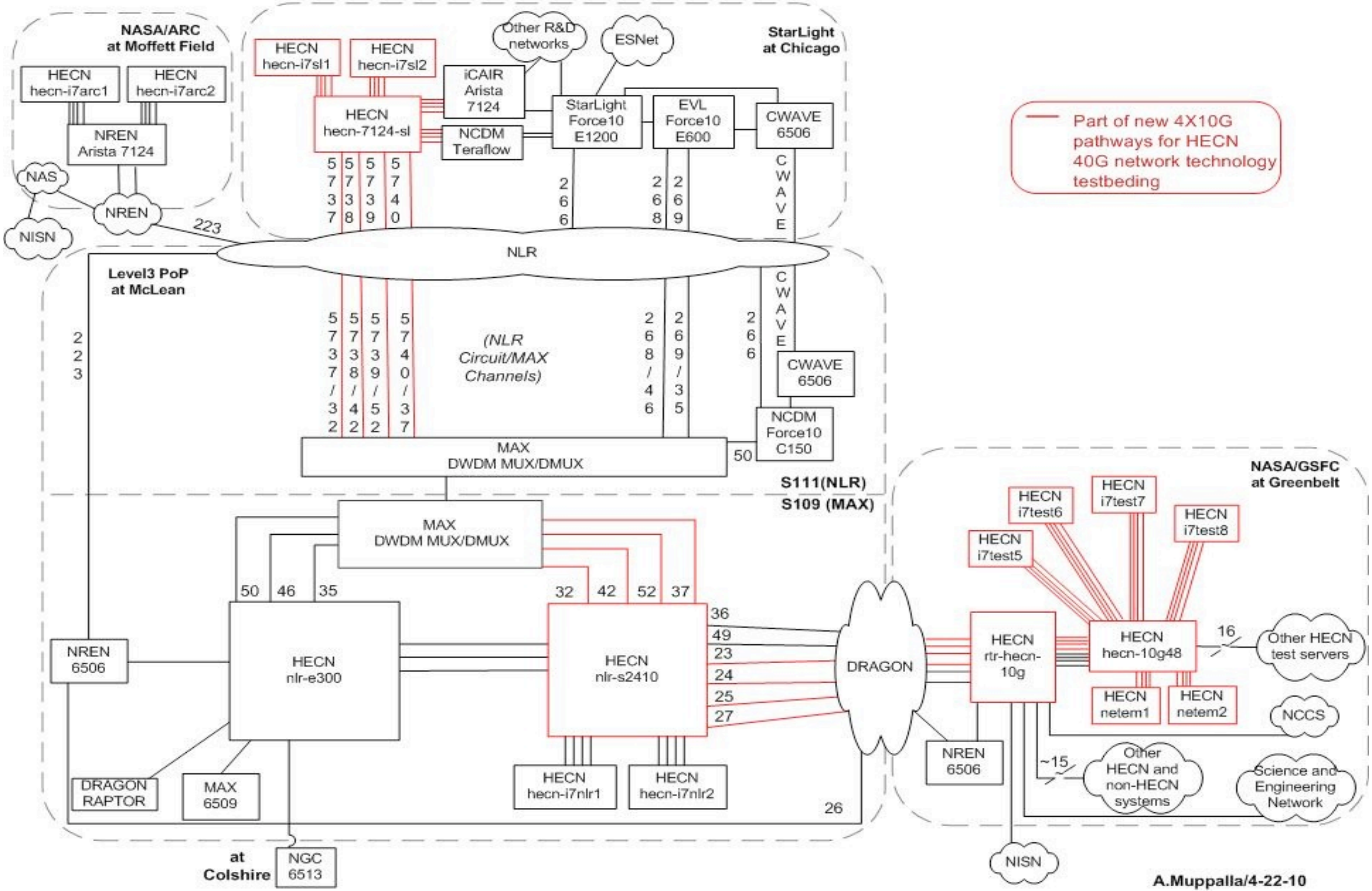
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Current Status (1 of 4)

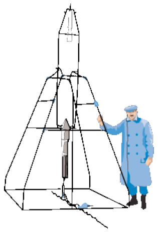
- Our testbed plans are in their early stages of development and are a “work-in-progress”
- We have gotten excellent support from our collaborators, especially the MAX, NAS/NREN & NLR; and we’re beginning to pick up quite a bit of steam
  - Completed significant SC09 experiments/demonstrations
  - Two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

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



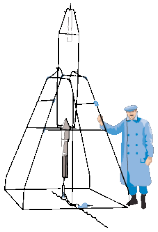
A.Muppalla/4-22-10



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

Current Status (2 of 4)

- “More steam”
  - Four new 10-Gbps HECN-enabled DWDM links between GSFC and College Park (CLPK) are ready to be deployed
  - Two B-systems are in preparation for deployment to CLPK
  - Expect StarLight to install initial 100-Gbps capability in late-2010 per NSF Advanced Research Infrastructure (ARI) Program award of StarLight’s ARI proposal, encouraged by GSFC/HECN’s 20, 40 & 100-Gbps Network Testbeds Plans
  - Expect MAX to install initial 100-Gbps capability in late-2010 per NSF ARI Program award of MAX’s ARI proposal, encouraged by GSFC/HECN’s 20, 40 & 100-Gbps Network Testbeds Plans
    - To include a 100-Gbps MAX-enabled pathway between CLPK & MCLN



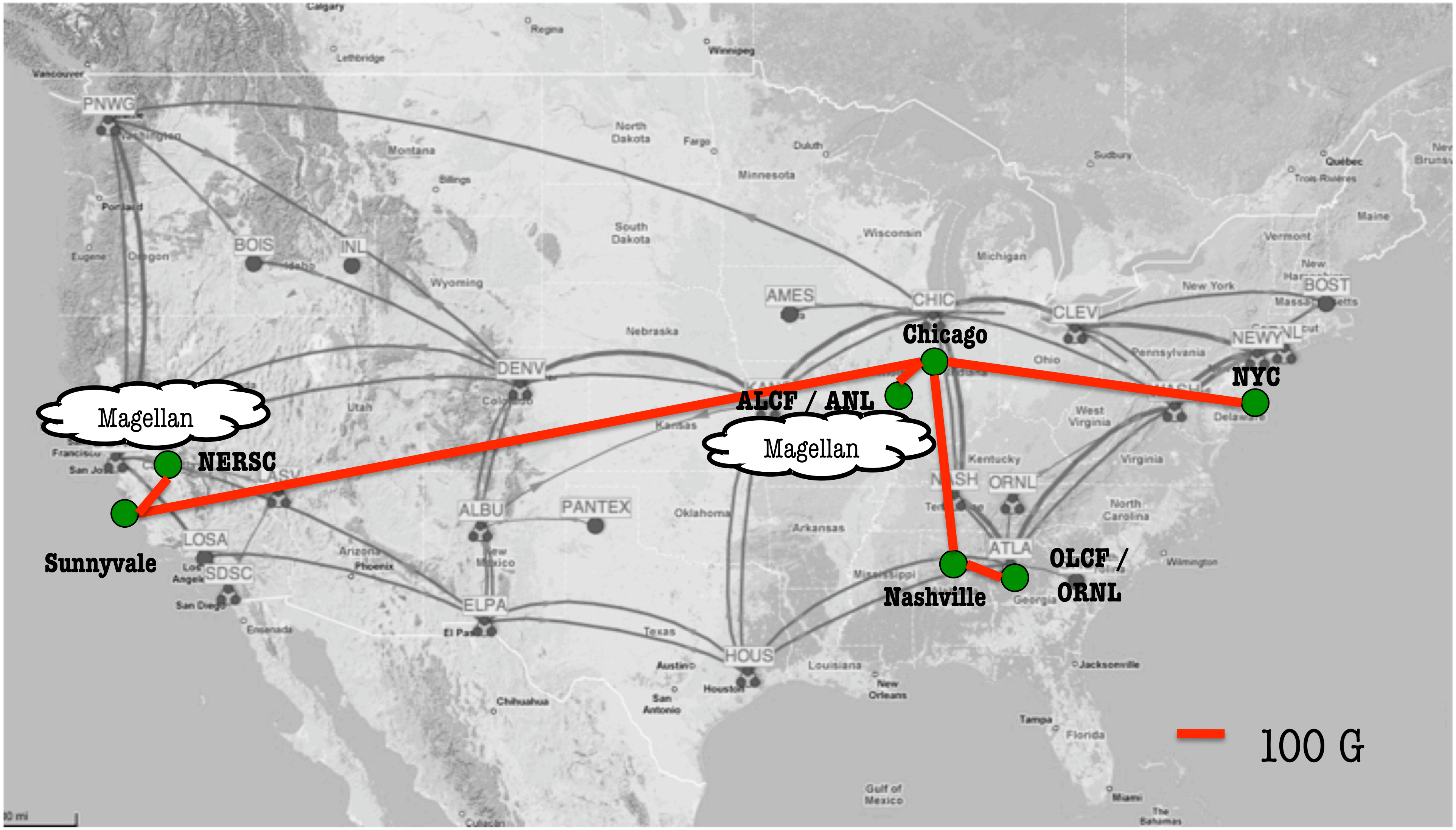
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

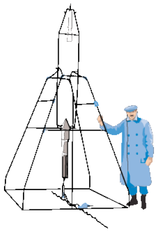
Current Status (3 of 4)

- “Still More steam”
  - Expect one 40-Gbps NLR-enabled muxponder link between MCLN and StarLight starting in 2011
  - Obtained approval to be informal partners in DoE’s high performance file accessing testbed and have access to their upcoming Advanced Network Initiative’s 100-Gbps Prototype Network between ANL (connected via StarLight) and NERSC (connected at Sunnyvale)

Source: Brian Tierney (LBL/ESnet)

Nationwide 100G Prototype Network

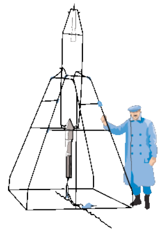




Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

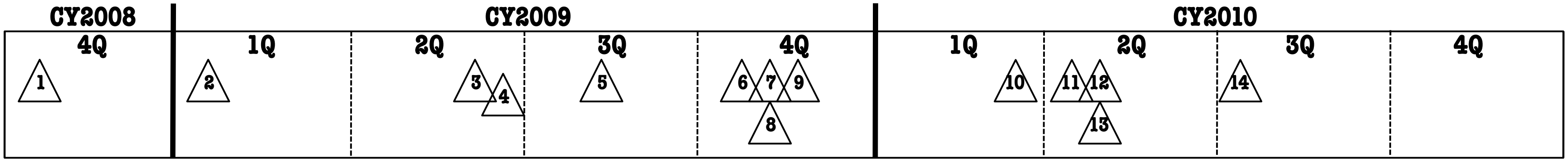
Current Status (4 of 4)

- Intra-NASA we'll seek broader sponsorship via an "in-formation" Emerging Network Technology Testbed initiative
- 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 if you are university based



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

Significant >10G Accomplishments of HECN Team & Partners (1 of 5)

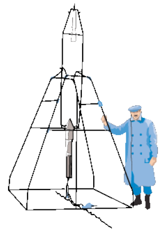


1: 10/14/08 In collaboration with MAX across MAX-provisioned link between College Park and McLean, bi-directionally tested maximum throughput of Fujitsu 40G optical transponders & Juniper OC-768c interfaces on T1600 routers

2: 01/29/09 Created initial drafts of HECN 20, 40 & 100G Network Testbed plans

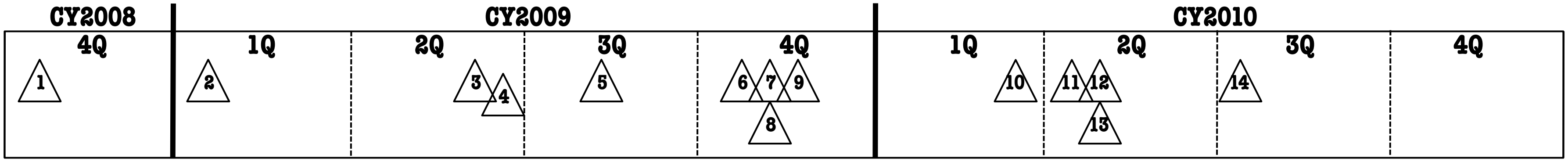
3: 06/12/09 Bench-tested HECN's first >10G net-test-workstations (i7-based) measuring nuttcp-enabled memory-to-memory throughput flows unidirectionally at >69.2G and bi-directionally at >77.2G aggregate

4: 06/24/09 Presented "Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds" at MAX Spring 2009 All Hands Meeting



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

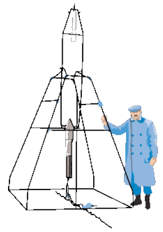
Significant >10G Accomplishments of HECN Team & Partners (2 of 5)



5: 08/06/09 Bench-tested a new HECN net-test-workstation (Xeon-based) measuring nuttcp-enabled memory-to-memory throughput flows unidirectionally (transmit) at >100.4G

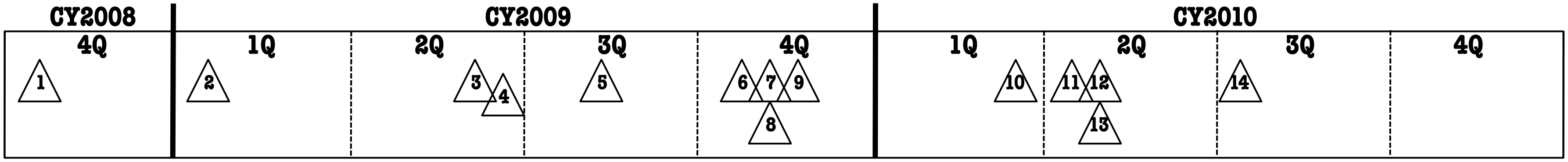
6: 10/30/09 Two HECN net-test-workstations (i7-based) each with 4x10G NIC interfaces, deployed at ARC, used by NREN in prep for SC09 to fully check out new 4x10G links on single fiber pair between ARC and Sunnyvale using ADVA FSP3000 dwdm mux/demuxes

7: 11/02/09 Using two HECN net-test-workstations (i7-based) each with two RAID5 disk controllers nested as RAID50, measured nuttscp-enabled disk-to-disk data throughput unidirectionally at >9.8G in a 0.1ms RTT testbed; on 11/10/09 measured >9.5G in a 80.1ms RTT testbed



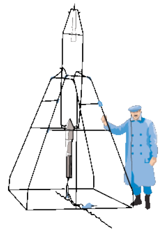
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Significant >10G Accomplishments of HECN Team & Partners (3 of 5)



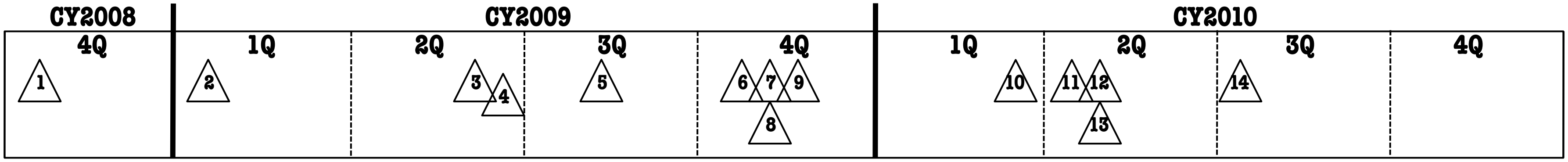
8: 11/02/09 Using two HECN net-test-workstations (i7-based) in nuttcp-enabled memory-to-memory throughput flows, with kernal bonding/ standard link-aggregation among each workstation's 4x10G NIC interfaces measured unidirectionally >31.6G and with nuttcp-enabled "application bonding" measured unidirectionally >39.5G

9: 11/16/09 In the NASA research exhibit at the SC09 conference, Portland, OR, demoed: >100G uni-directional memory-to-memory data throughput between in-booth HECN servers; 40G bi-directional memory-to-memory data throughput between HECN servers in-booth and at ARC across 4x10G NLR/C-Wave links; and 10G disk-to-disk data throughput between in-booth HECN servers, between HECN servers in-booth and at ARC, and between HECN servers in-booth and at GSFC



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

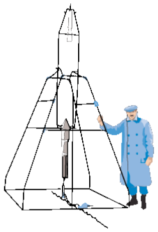
Significant >10G Accomplishments of HECN Team & Partners (4 of 5)



10: 03/24/10 In collaboration with MAX, added 4x10G links across MAX’s DRAGON regional optical network between GSFC and the NASA rack in the Level3 PoP at McLean, and then through MAX’s ADVA dwdm mux/ demuxes to NLR’s Cisco 15454 optical transponder at that PoP

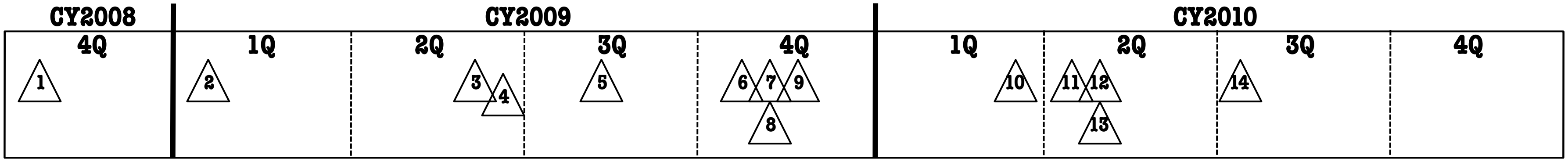
11: 04/07/10 In collaboration with iCAIR, deployed two HECN net-test-workstations (i7-based) at StarLight for 4x10G connections with NLR and 8x10G connections with other R&D networks

12: 04/14/10 In collaboration with NREN, used two HECN net-test-workstations (i7-based) at ARC to measure unidirectional and bidirectional throughput of a Fortinet FortiGate-3810A security gateway with four 10GE ports



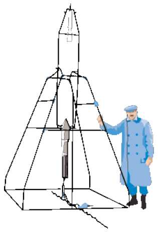
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Significant >10G Accomplishments of HECN Team & Partners (5 of 5)



13: 05/09/10 NLR turned up new 4x10G links between McLean and StarLight for HECN maximum throughput testing; but due to errors that ultimately were isolated to and fixed on HECN equipment at MCLN, HECN did not “accept” the NLR links as active until 08/03/10. In between HECN conducted limited 40G bi-directional memory-to-memory and 10G unidirectional disk-to-disk data throughput between HECN servers at GSFC and StarLight

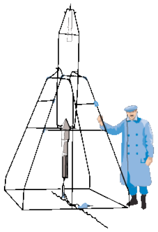
14: 08/14/10 Using two HECN net-test-workstations (i7-based) each with four RAID5 disk controllers nested as RAID50, measured nuttscp-enabled disk-to-disk data throughput unidirectionally at >17.8G (90% of maximum 19.8G) in a 0.1ms RTT testbed



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

Next Steps (Near Term) with Key Challenges

- Complete Phase 1 network-test workstation assembly, bench tests, and deployments - LAN, MAN, WAN
  - Including more comprehensive file-copy application testing and tuning to achieve throughput performance objectives
- Investigation of candidate Phase 2 network-test workstations, refinement of Phase 2 throughput performance objectives, and refinement of Phase 2 LAN, MAN, WAN pathway targets
  - Opportunities/challenges to obtain 40-Gbps disk subsystems at low cost
  - Opportunities/challenges to obtain early 40 & 100 GE-MAC-interfaces in the Phase 2 testbeds
- Leveraging NASA's "in-formation" Emerging Network Technology Testbed initiative



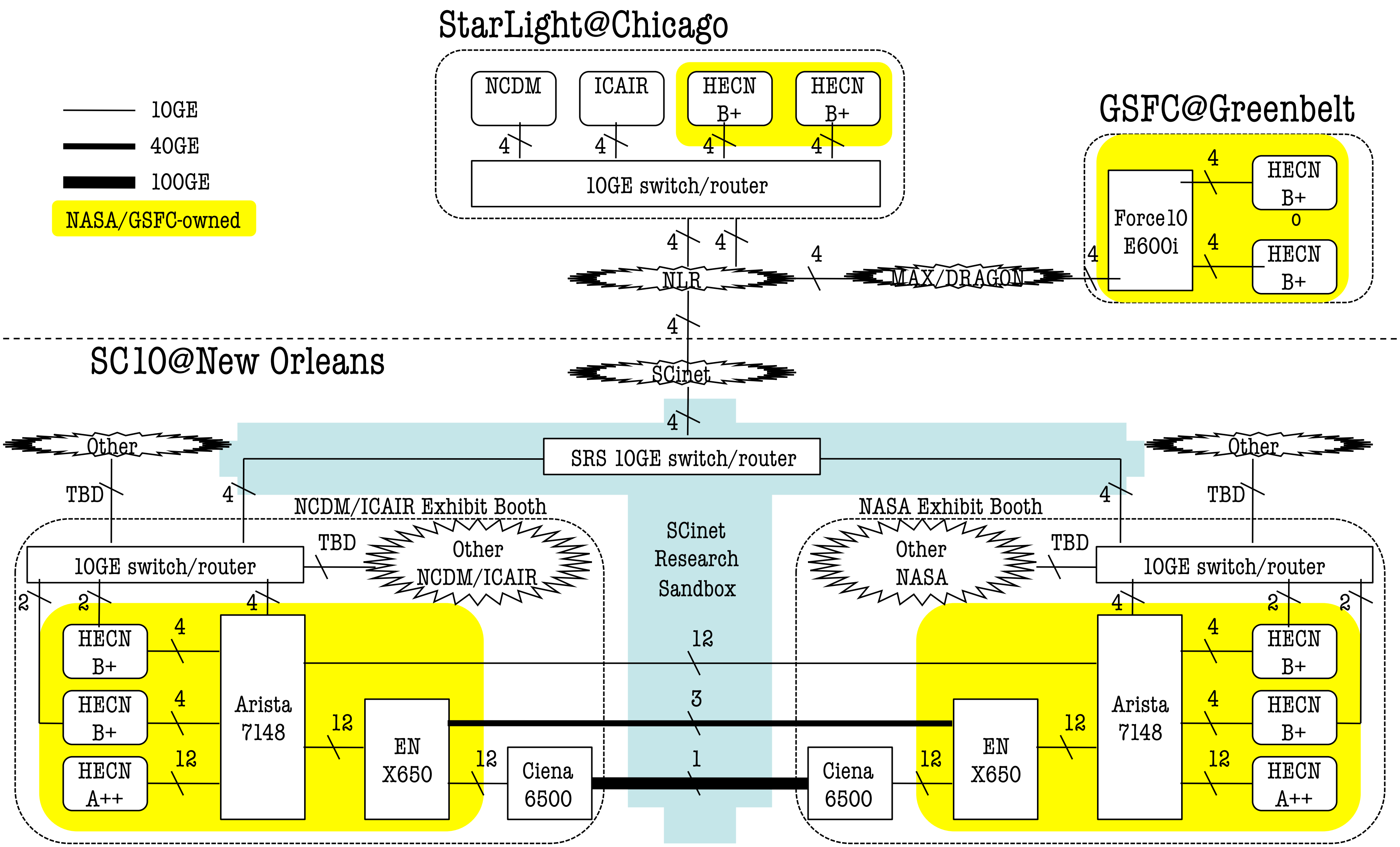
Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Potential SC10 Demonstration/Evaluation Experiments

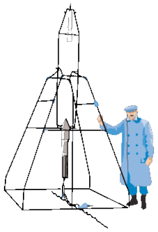
- Deploy a set of HECN's high performance network-testing-workstations and supporting multi-10G/40G switches in both NASA's and NCDM/ICAIR Exhibit Booths, capable of:
  - >100G unilateral memory-to-memory data flows
  - >80G bi-lateral aggregate memory-to-memory data flows
  - >10G unilateral disk-to-disk file copies; likely 20G and possibly 40G using SSDs
- Use SCinet Research Sandbox inter-booth fiber to demonstrate/evaluate different 100G network technology solutions in LAN data flow demonstrations
- Use existing 4x10G pathway across NLR and MAX/DRAGON between GSFC and StarLight, plus a new NLR 40G pathway between StarLight and SC10, to conduct science-oriented WAN data flow demonstrations

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, Northwestern/ICAIR, SCinet & UIC/NCDM



J. P. Gary 8/17/10



Overall Timeline for the HEC 20, 40 &100 Gbps Network Testbeds

```
           | FY10  | FY11  | FY12  | FY13  | FY14  | FY15
4x10G
A. LAN+MAN
   Testing tttttttttttttttt
B. WAN-to-StarLight
   Testing   tttttttttttttttt

1x40G
A. LAN+MAN
   Testing       ttttttttoooooooooooooooo
B. WAN betw ARC+GSFC
   Testing               TTTTTTTTTTTTTTTT
C. WAN betw ARC+GSFC
   Operations                    OOOOOOOOOOOOOOOO

1x100G
A. LAN+MAN
   Testing                       ttttttttoooooooooooooooooo
B. WAN betw ARC+GSFC
   Testing                               TTTTTTTTTTTTTTTT
C. WAN betw ARC+GSFC
   Operations                                    OOOOOOOOOO
```

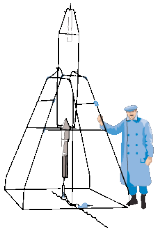
Wherein:

ttt…ttt implies only pre-operational testing use

ooo…ooo implies operational use in LAN+MAN

TTT…TTT implies only pre-operational testing use in WAN

OOO…OOO implies operational use in WAN



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

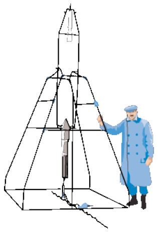
Summary of HEC Test WAN Requirements

- Key SLA Requirements of HEC's WAN Test Network
  - Availability (percent): 80.0
  - Restoral Time: <48 hours
  - Coverage Period: 24x7
  - RTT between sites in CONUS: <100ms
  - Packet Loss: <1E-7
  - Jumbo Frames: Transport of 9000-byte IP MTU jumbo frames without fragmentation
  - Bandwidth between NAS@ARC and NCCS@GSFC:
    - 1Jul11 through 30Jun13: 40Gbps (i.e., 1x40G, not 4x10G or other LAG approaches)
    - 1Jul13 through 30Jun15: 100Gbps (i.e., 1x100G, not 10x10G or other LAG approaches)

- The above SLA requirements apply only to the "TTT…TTT" links
- The SLA requirements of "OOO…OOO" links are similar to "TTT…TTT" links except that "OOO…OOO" links have their Availability (percent) parameters at 99.50 and their Restoral Time parameters at 4 hours; and therefore they likely need to be provisioned from a different supplier than "TTT…TTT" links

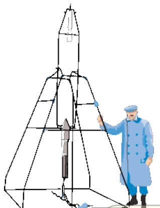
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Q & A



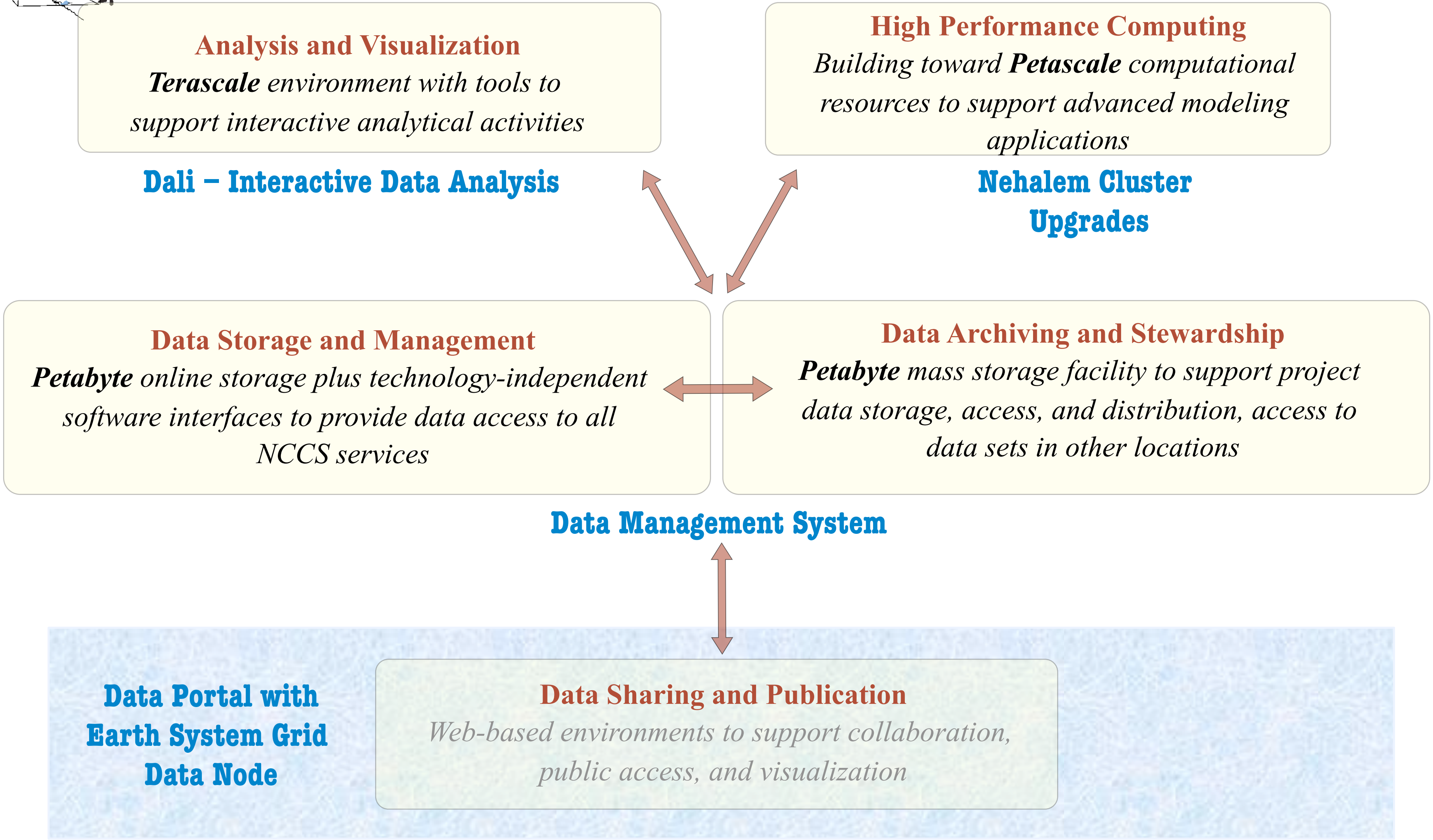
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Backup Slides



Data Centric Notional Architecture

Source: Phil Webster (GSFC/NCCS)



Analysis and Visualization
Terascale *environment with tools to support interactive analytical activities*

Dali – Interactive Data Analysis

High Performance Computing
Building toward ***Petascale*** *computational resources to support advanced modeling applications*

Nehalem Cluster Upgrades

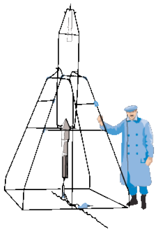
Data Storage and Management
Petabyte *online storage plus technology-independent software interfaces to provide data access to all NCCS services*

Data Archiving and Stewardship
Petabyte *mass storage facility to support project data storage, access, and distribution, access to data sets in other locations*

Data Management System

Data Portal with Earth System Grid Data Node

Data Sharing and Publication
Web-based environments to support collaboration, public access, and visualization

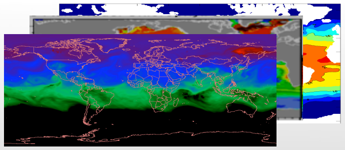


NCCS Data Centric Climate Simulation Environment

Source: Phil Webster (GSFC/NCCS)

Data Sharing and Publication

- Capability to share data & results
- Supports community-based development
- Data distribution and publishing

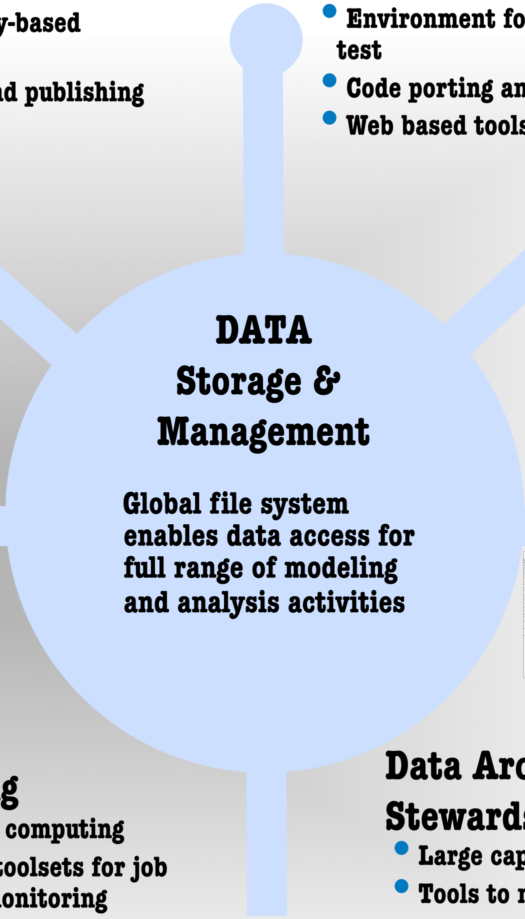


Code Development*

- Code repository for collaboration
- Environment for code development and test
- Code porting and optimization support
- Web based tools

User Services*

- Help Desk
- Account/Allocation support
- Computational science support
- User teleconferences
- Training & tutorials



DATA Storage & Management

Global file system enables data access for full range of modeling and analysis activities

Analysis & Visualization*

- Interactive analysis environment
- Software tools for image display
- Easy access to data archive
- Specialized visualization support

Data Transfer

- Internal high speed interconnects for HPC components
- High-bandwidth to NCCS for GSFC users
- Multi-gigabit network supports on-demand data transfers

HPC Computing

- Large scale HPC computing
- Comprehensive toolsets for job scheduling and monitoring

Data Archival and Stewardship

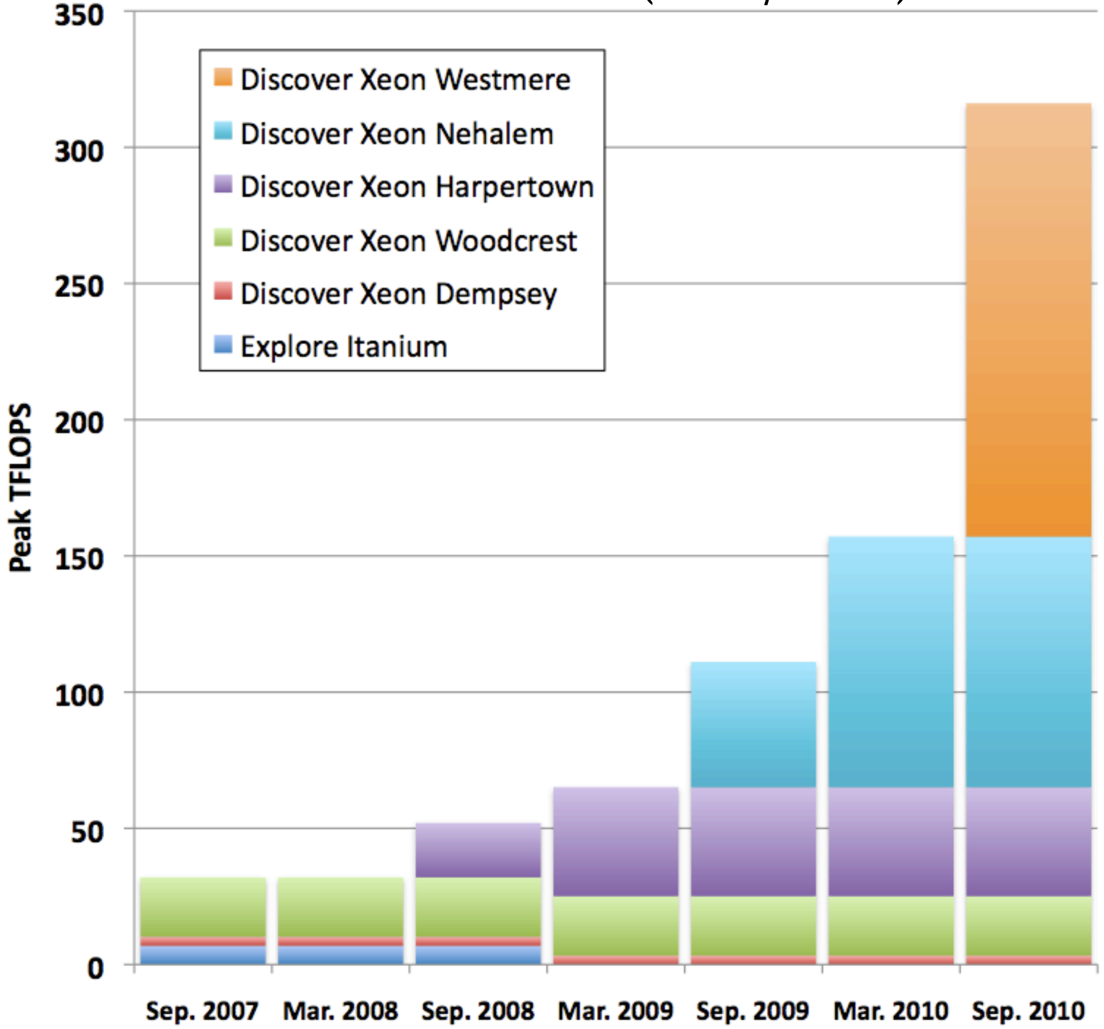
- Large capacity storage
- Tools to manage and protect data
- Data migration support

* Joint effort with SIVO

08/22/10

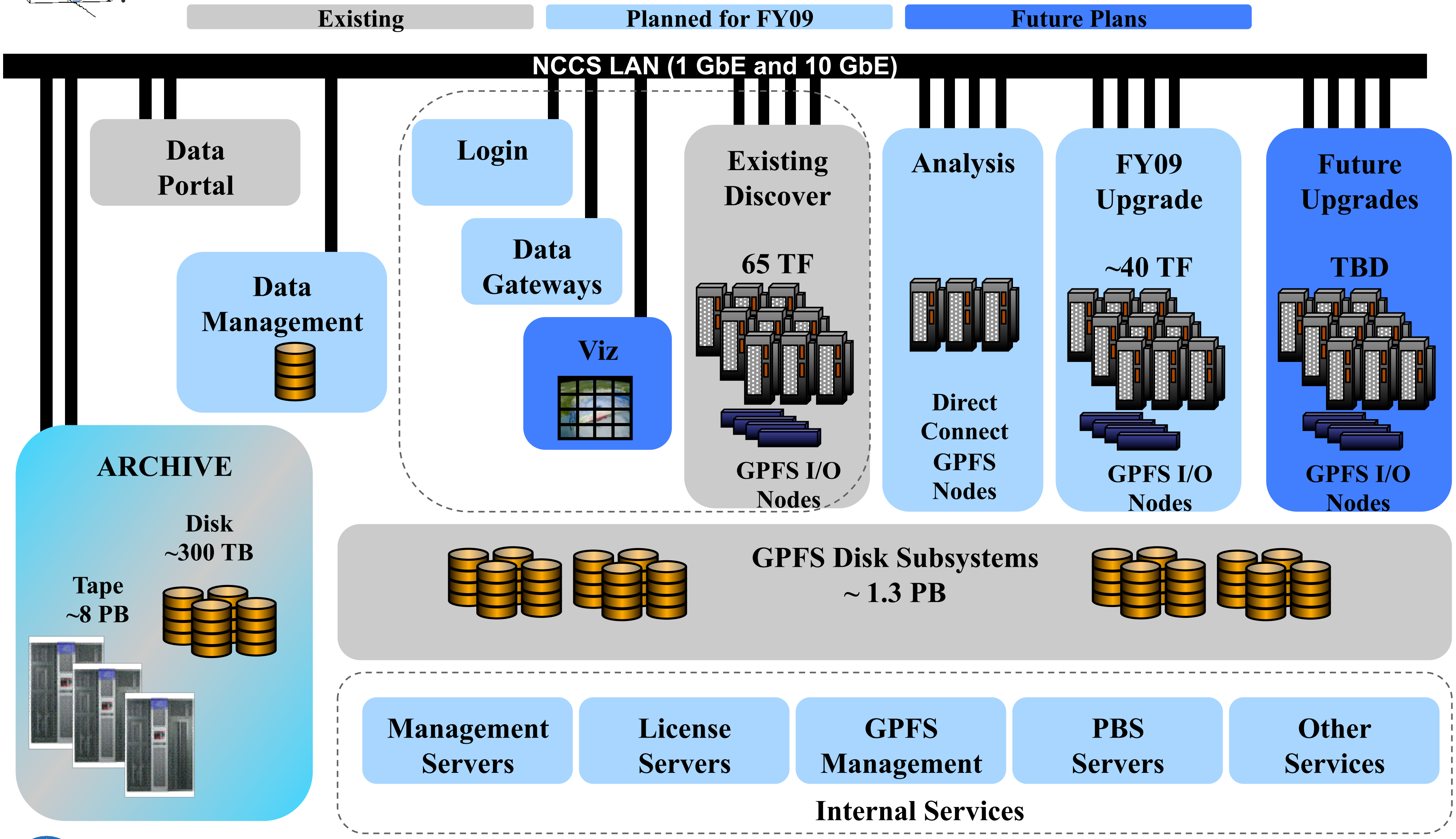
NCCS Compute Capacity Evolution 2007-2010

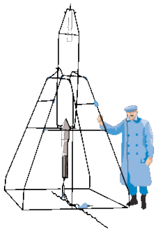
Source: Phil Webster (GSFC/NCCS)



Excerpt from 24Mar09 NCCS User Forum

Representative Architecture



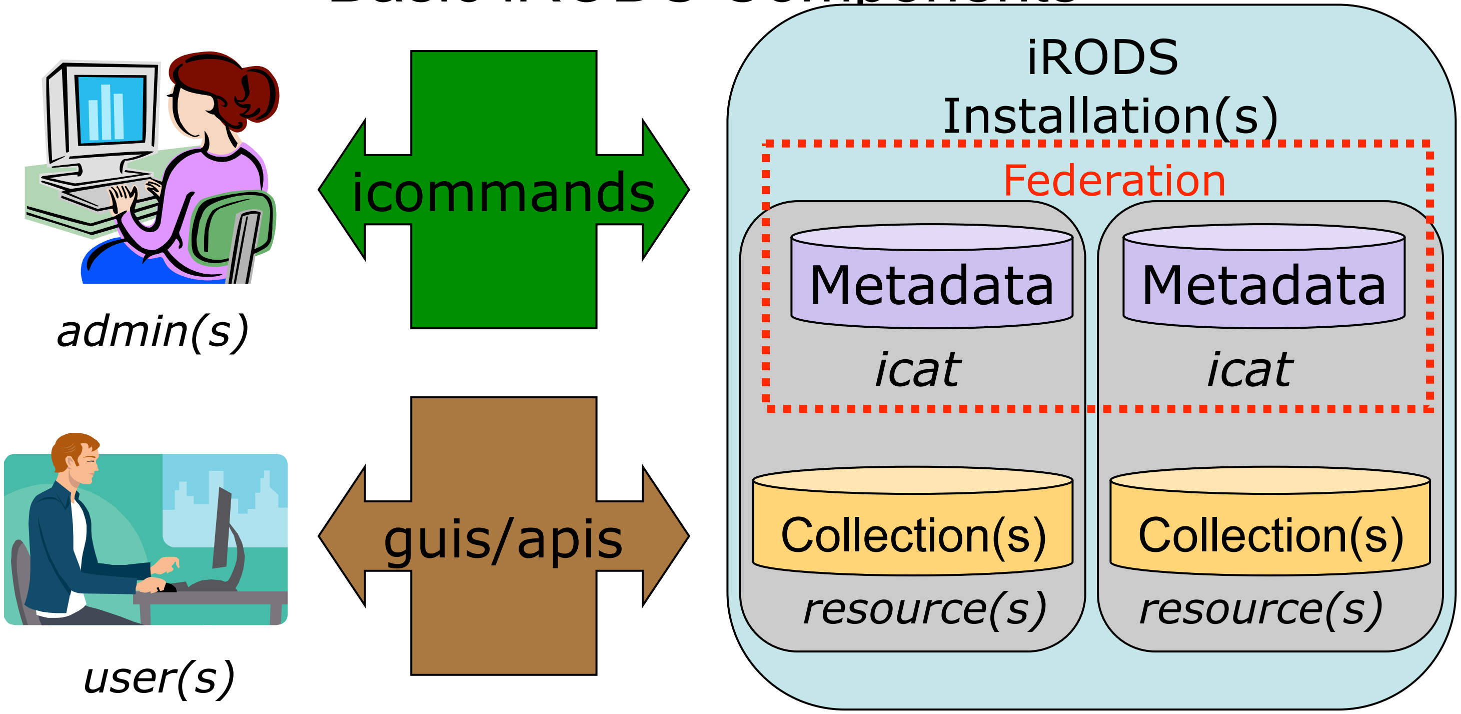


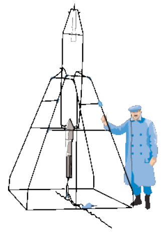
Excerpt from iRODS Update 2009 03 15.ppt
Sources: Dan Duffy/GSFC 606.2(NCCS) & Hoot Thompson/PTP(NCCS)

Integrated Rule-Oriented Data System (iRODS)

•Data grid software system developed by the Data Intensive Cyber Environments (DICE) group (developers of the SRB, the Storage Resource Broker), and collaborators.

Basic iRODS Components

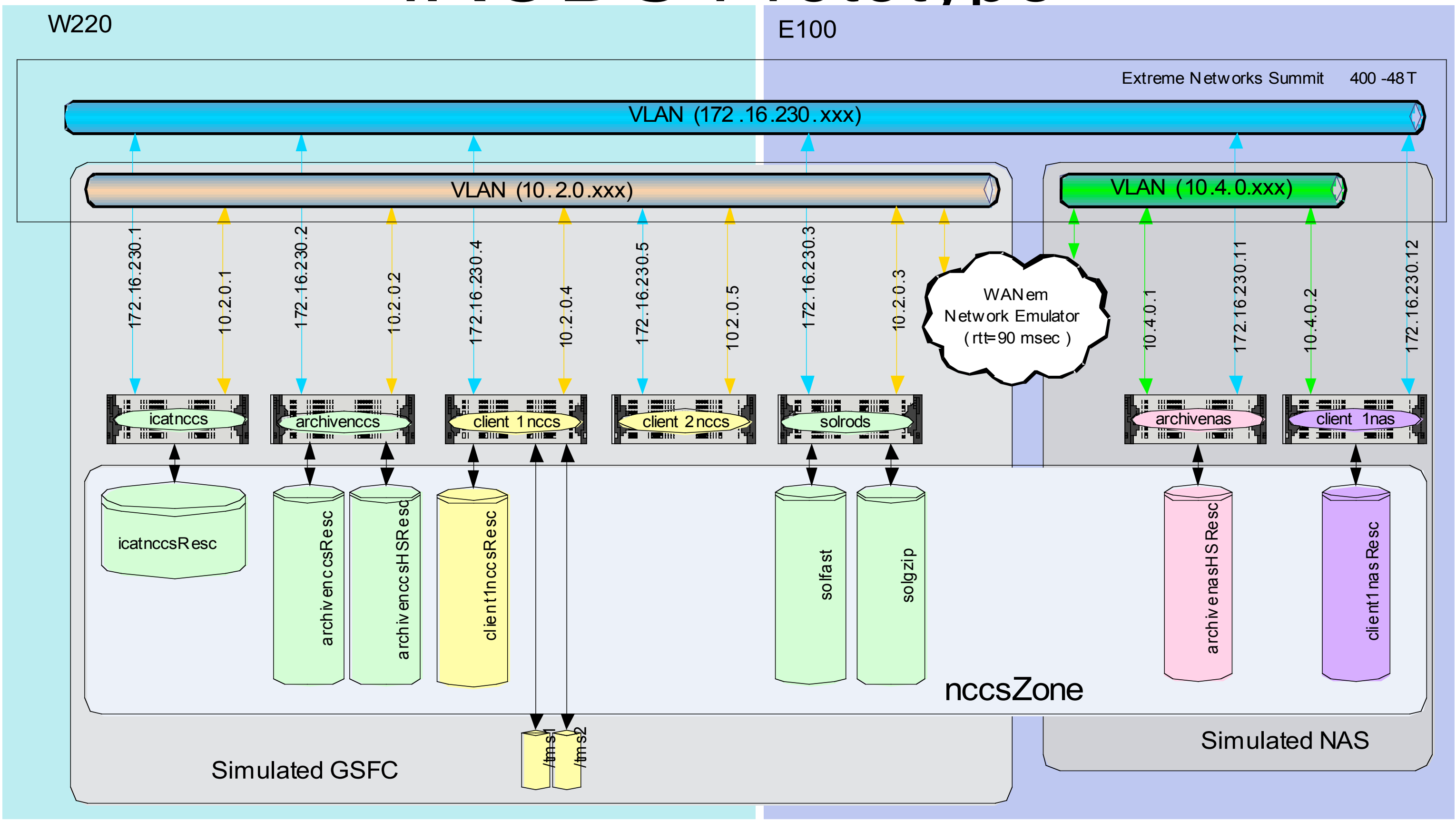




Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

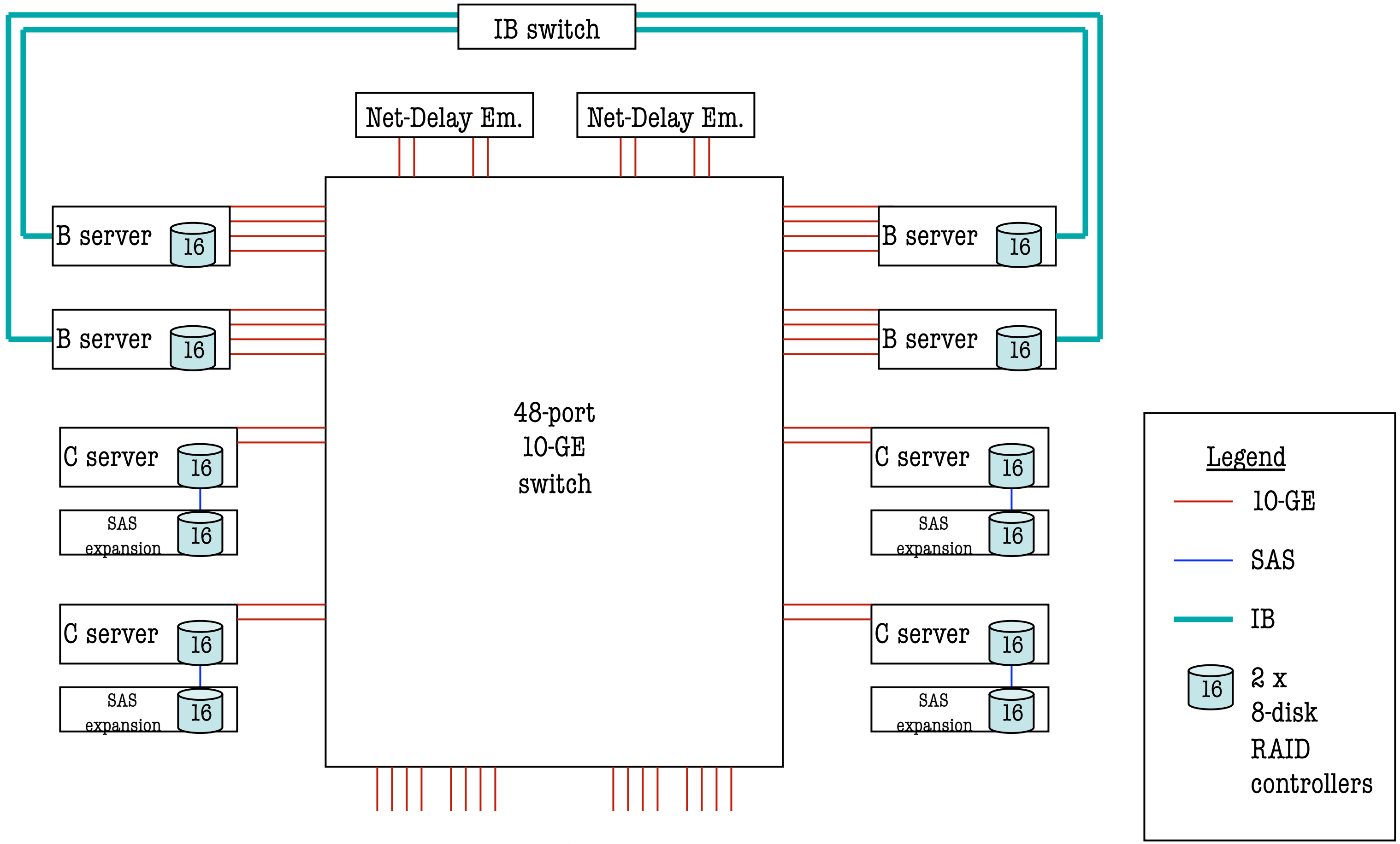
Excerpt from iRODS Update 2009 03 15.ppt
Source: Dan Duffy/GSFC 606.2(NCCS) & Hoot Thompson/PTP(NCCS)

iRODS Prototype



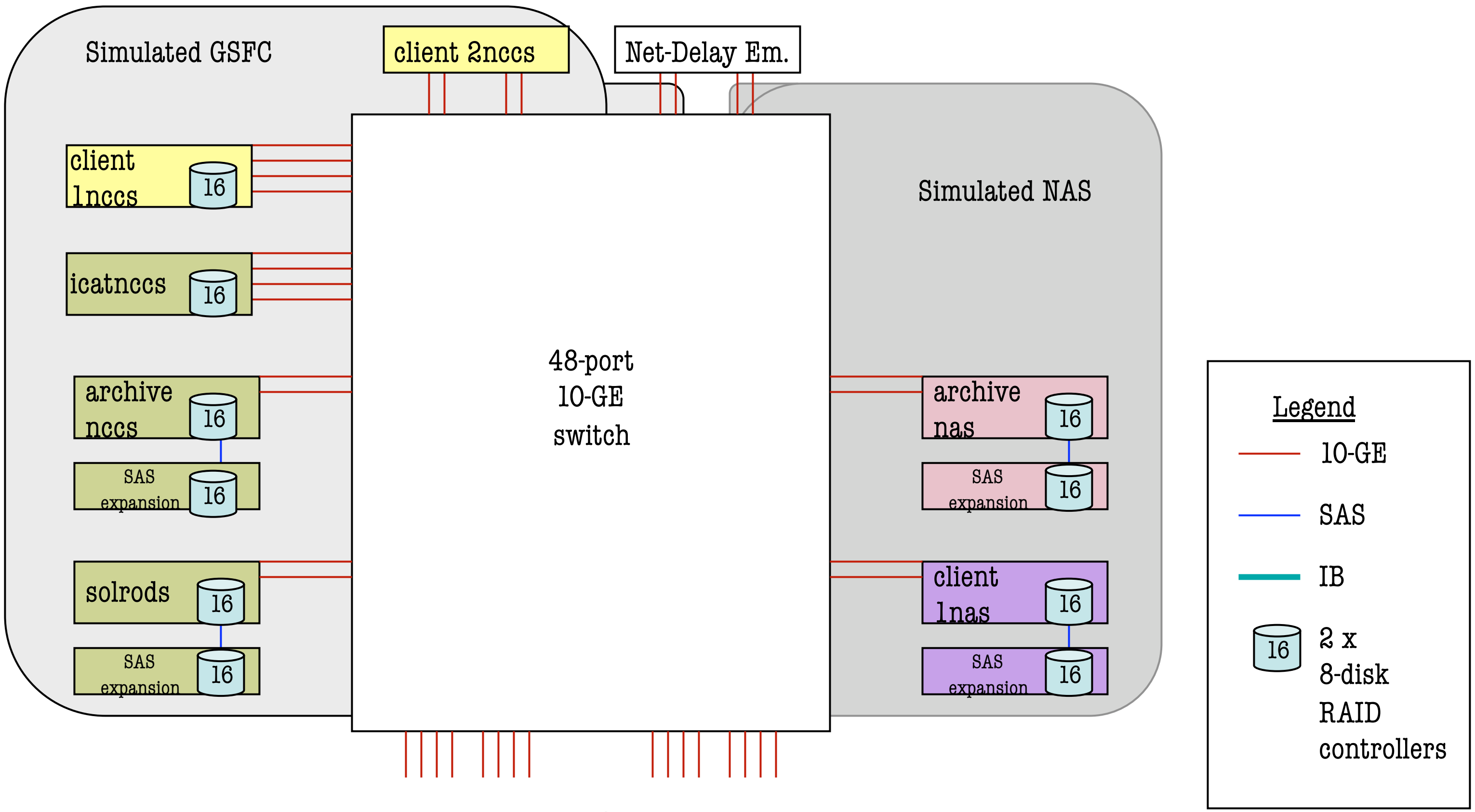
Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Part of the HECN GSFC-local Advanced Networking Testbed



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

Candidate Use of the HECN GSFC-local Advanced Networking Testbed For “Next Generation” NCCS iRODS Prototyping



Simulated GSFC

client 2nccs

Net-Delay Em.

client 1nccs 16

icatnccs 16

archive nccs 16

SAS expansion 16

solrods 16

SAS expansion 16

48-port 10-GE switch

Simulated NAS

archive nas 16

SAS expansion 16

client 1nas 16

SAS expansion 16

Legend

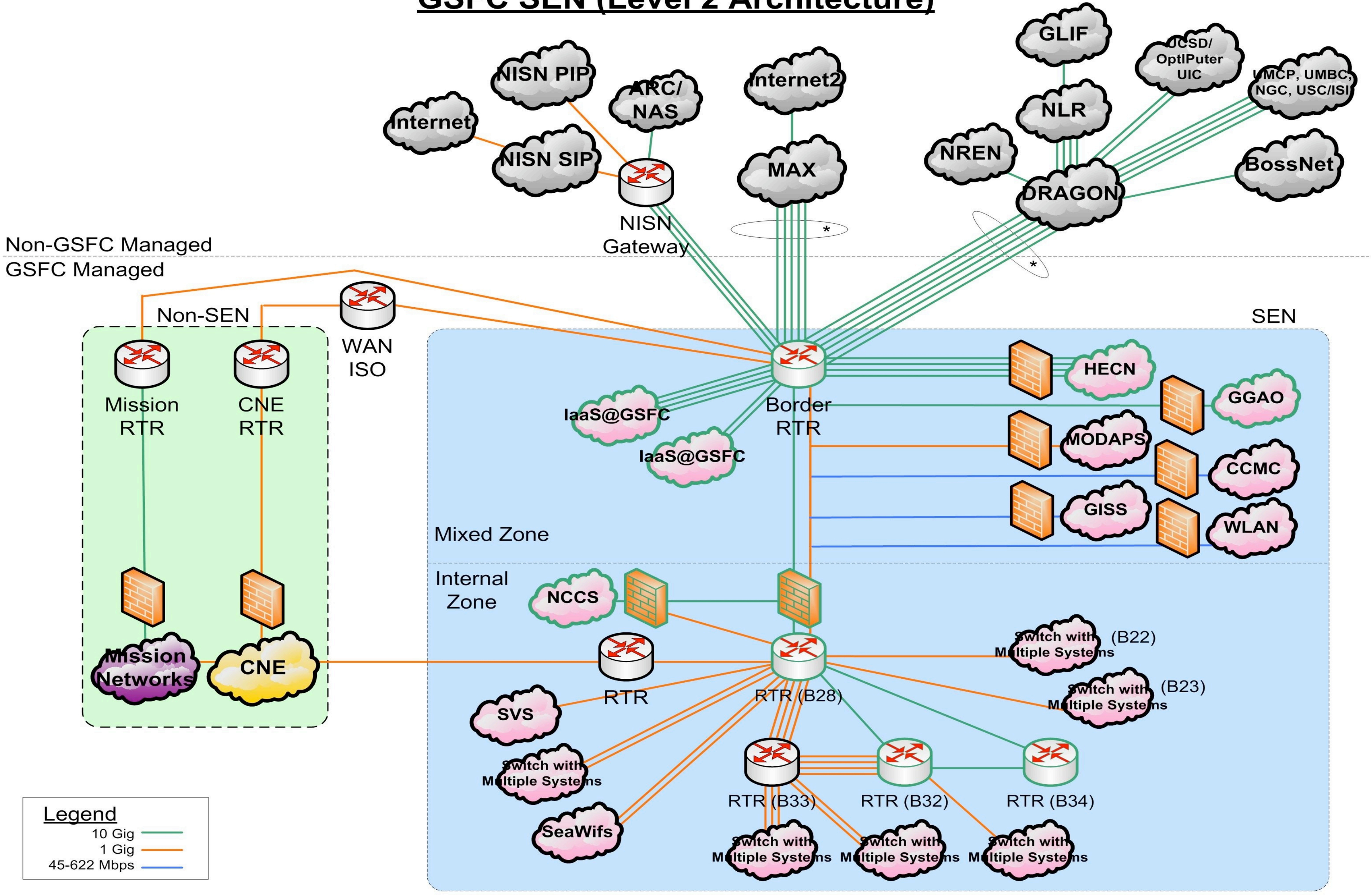
10-GE

SAS

IB

16 2 x 8-disk RAID controllers

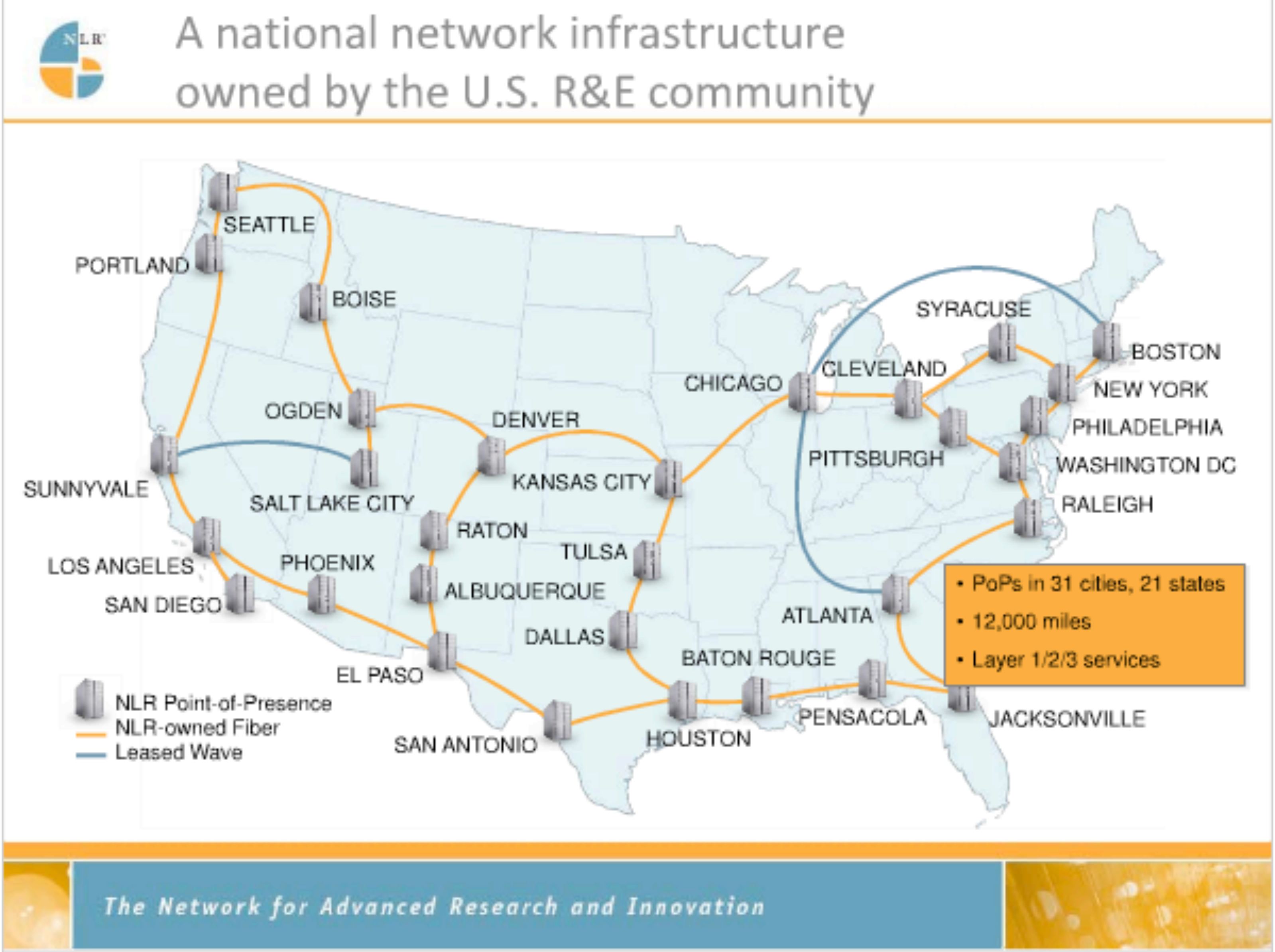
GSFC SEN (Level 2 Architecture)



* One SEN Physical Fiber Pair and Several Different Lambdas

2010-06-15

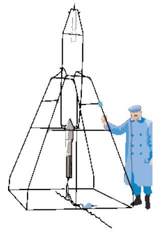
Source: Glenn Ricart (NLR)

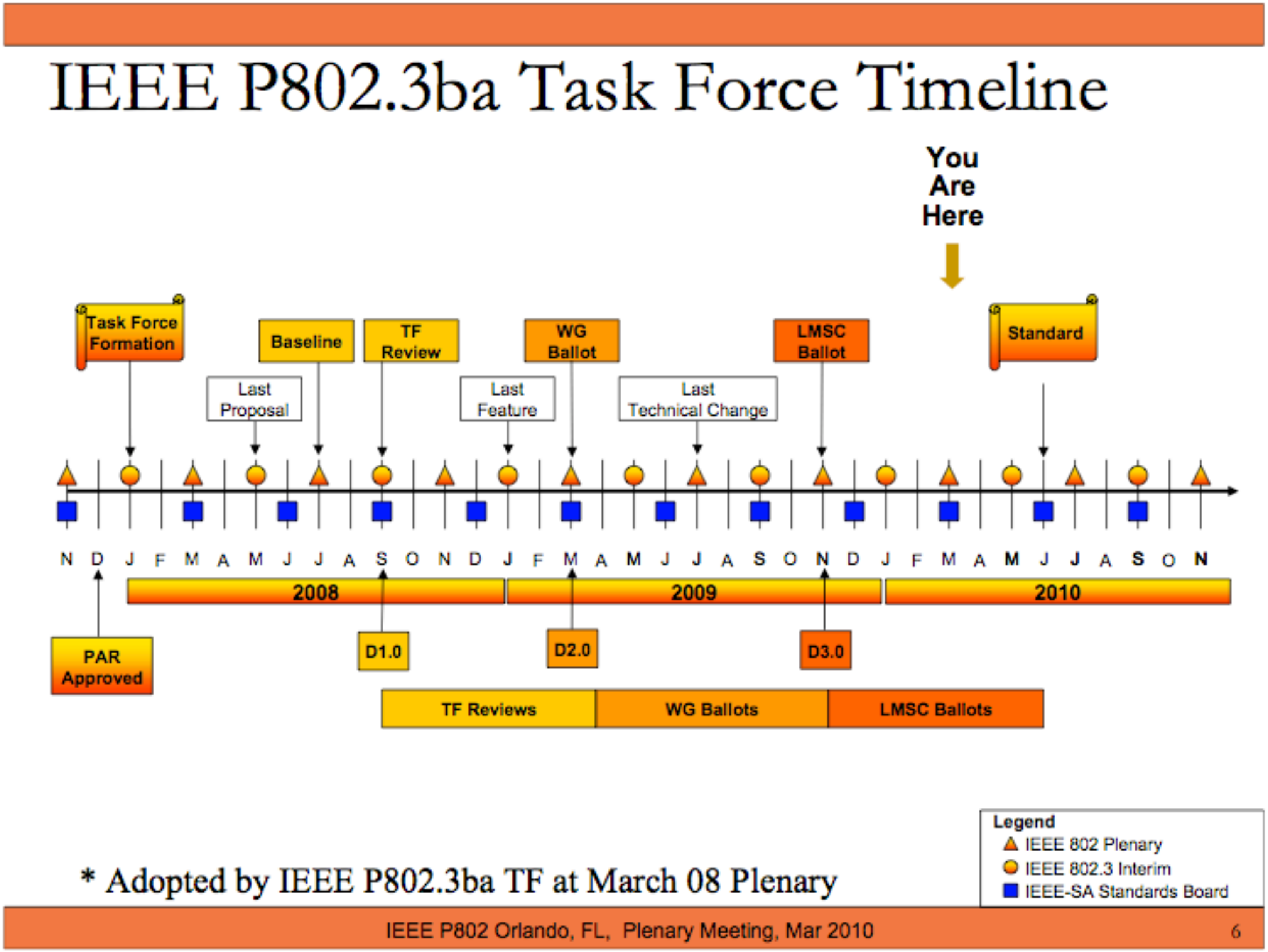


A national network infrastructure owned by the U.S. R&E community

- PoPs in 31 cities, 21 states
- 12,000 miles
- Layer 1/2/3 services

The Network for Advanced Research and Innovation

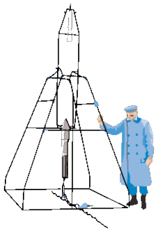




IEEE P802.3ba Task Force Timeline

* Adopted by IEEE P802.3ba TF at March 08 Plenary

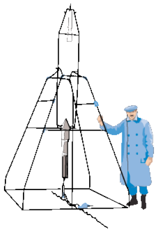
Legend
▲ IEEE 802 Plenary
● IEEE 802.3 Interim
■ IEEE-SA Standards Board



Plan/Proposal For Initiation of 20, 40 & 100 Gbps Network Technology Testbeds

Planned/Proposed Efforts Overview (1 of 4)

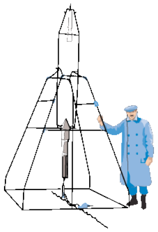
- A multi-phased plan which involves a series of partly overlapping approximately-two-year Phases
- Each Phase focused on producing a demonstrable end-to-end network capability with particular throughput performance goals
  - Matched to then-readily-available technologies
    - Particularly to include dense wave division multiplex (DWDM) optics carrying dynamic circuit network (DCN)-enabled VLANs, switch/routers, NICs, and processors with suitable disk connections
  - Expected respectively to target 20, 40 and 100 Gbps end-to-end throughput performance goals
- Each Phase to include:
  - LAN and Regional/metropolitan area network (MAN) testbed deployments/ demonstrations partly funded from existing HEC-related budgets (assuming existing FY09 levels through FY11)
  - A WAN testbed deployment/demonstration only if additional funding is made available via this proposal
  - Joint collaboration on the iRODS[10] and/or other WAN file-accessing applications with NCCS and NASA Advanced Supercomputing (NAS) Division data management experts to ensure that the network testbed efforts will have at least HEC Program applicability



Plan/Proposal For Initiation of 20, 40 & 100 Gbps Network Technology Testbeds

Planned/Proposed Efforts Overview (2 of 4)

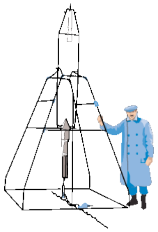
- End-to-end throughput performance goals in each Phase means
  - File-accessing applications, which typically are disk-to-disk, of representative end users will be involved in determining the criterion of success of those Phases
  - Memory-to-memory data flow applications also
    - Oriented to detailing the throughput performance of just the Layer 1 Physical through Layer 4 Transport layers of the end-based testing-computers and the intervening network components
- WAN emulation testing will be included in the LAN testbed deployments/demonstrations
- Regional/MAN and WAN testbed deployments/demonstrations are planned to more appropriately address the full scope of problems that occur when larger round-trip times (RTTs) affect file-accessing applications



Plan/Proposal For Initiation of 20, 40 & 100 Gbps Network Technology Testbeds

Planned/Proposed Efforts Overview (3 of 4)

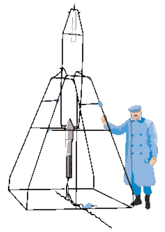
- Use of Layer 1+2 DCN-enabled VLANs versus Layer 1+2+3 full IP routed networks in both the Regional/MAN and WAN testbeds is critical
  - Sufficient to enable more effort to be focused on the primary subjects of this plan/proposal which are the processor interfaces and LAN infrastructure needed on the 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 plan/ proposal
  - Costs of 40 and 100 Gbps Layer 1+2+3 router interfaces are likely to be two or more orders of magnitude greater than 40 and 100 Gbps Layer 1+2 Ethernet switch interfaces which are sufficient to enable the needed VLANs



Plan/Proposal For Initiation of 20, 40 & 100 Gbps Network Technology Testbeds

Planned/Proposed Efforts Overview (4 of 4)

- 40 and 100-GE Regional/MAN and WAN testbeds will include complementing and interconnecting with that of ESnet+Internet2+ Juniper+Infinera+ Level3 and/or other modern-day-NGI networks to the greatest extent possible
- The WAN testbed, which is planned/proposed primarily only between the ARC-based NAS and the GSFC-based NCCS, is critical to enable joint technical interoperability and end-user applicability testing of iRODS and/or other WAN file-accessing applications



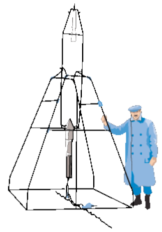
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Phase 1 Network-Test Workstations: Nominal "B" System

- Chassis: Supermicro 836TQ-R800B (3u 16bay 7slot 800W RPS)
- Motherboard: Asus P6T6 WS Revolution (5 PCIe V2 x8)
- Processors: one Intel i7 965 (3.2GHz quad-core Nehalem)
- Memory: Kingston KHX16000D3ULT1K3 (6GB 2000MHz DDR3 CL8)
- System disks: one Western Digital WD2500BEKT (2.5" 250GB)
- NICs: two Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)
- Raid controllers: two HighPoint RocketRaid 4320 (internal, 8 disks each)
- User disks: 16 Western Digital WD5001AALS (500GB)
- IB HCA: one Qlogic QLE7280 (DDR, 8x)
- For more detail, contact Paul.Lang@nasa.gov

Phase 1 Network-Test Workstations: Nominal “C” System

- Nominal “B” (Baseline) System
- Minus:
  - NICs: One Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)
  - IB HCA: one Voltaire (DDR, 8x)
- Plus:
  - Raid controllers: two HighPoint RocketRaid 4322 (external, 8 disks each)
- Plus via SAS-connection:
  - Chassis: one Supermicro 836TQ-R800B (3u 16bay 7slot 800W RPS) with SAS converter/adaptor and cables
  - User disks: 16 Western Digital WD5001AALS (500GB)
- For more detail, contact Paul.Lang@nasa.gov



Phase 1 Network-Test Workstations: Nominal "A" System

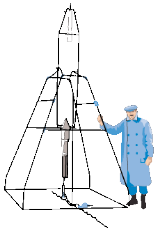
- Nominal "B" (Baseline) System
- Minus:
  - Raid controllers: two HighPoint RocketRaid 4320 (internal, 8 disks each)
  - User disks: 16 Western Digital WD5001AALS (500GB)
  - IB HCA: one Voltaire (DDR, 8x)

- For more detail, contact Paul.Lang@nasa.gov

Phase 1 Network-Test Workstations: "A+" System

- Nominal "A" System
- Plus:
  - NICs: Two Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)

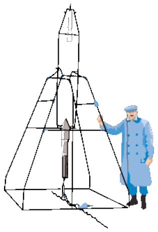
- For more detail, contact Paul.Lang@nasa.gov



Phase 1 Network-Test Workstations: “A-” System

- Nominal “A” System
- Minus:
  - NICs: One Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)

- For more detail, contact Paul.Lang@nasa.gov

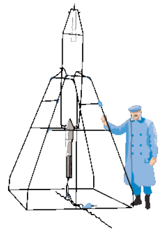


Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Phase 1.1 Network-Test Workstations: “A++” System

- Chassis: Supermicro 836TQ-R800B (3u 16bay 7slot 800W RPS)
- Motherboard: Supermicro X8DAH+-F (6 PCIe V2 (4 x8 & 2 x16))
- Processors: two XEON W5580 (3.2GHz quad-core Nehalem)
- Memory: Kingston KHX16000D3ULT1K3 (6GB 2000MHz DDR3 CL8, running at 1333MHz)
- System disks: one Western Digital WD2500BEKT (2.5” 250GB)
- NICs: six Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)

- For more detail, contact Paul.Lang@nasa.gov



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

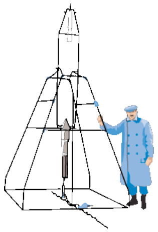
***"A++" Server Approximate Costs** (With components acquired via SEWP IV in lot-sizes of 3 - 15, and self assembly. Source: Paul Lang (ADNET))*

Supermicro 836TQ-R800B 3u 16bay 800W RPS Chassis	$850
Supermicro X8DAH+-F motherborad	$508
Intel W5580 XEON 3.2GHz processor $1669 x 2	$3338
Kingston KHX2000C8D3T1K3 6GB DDR3 2000 CL8 memory x 2	$500
CBL-0084 front pannel cable	$3
12" 3pin fan extension cable	$1
ArkTech slim IDE DVD to SATA adapter	$10
Myri 10G-PCIE2-8B2-2S+E Dual SFP+ NIC $950 x 6	$5700
Dynatron G666 CPU cooler	$35
Western Digital WD2500BEKT 250GB 2.5" system disk	$73
Red Greatland 18" Slimline SATA adapter	$6
Supermicro MCP-220-83601-0B FDD tray for 2.5" disk	$8
eVGA GeForce 8400GS video card	$40
8" 8pin power extension cable	$8
	$11080

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

Example Deployments

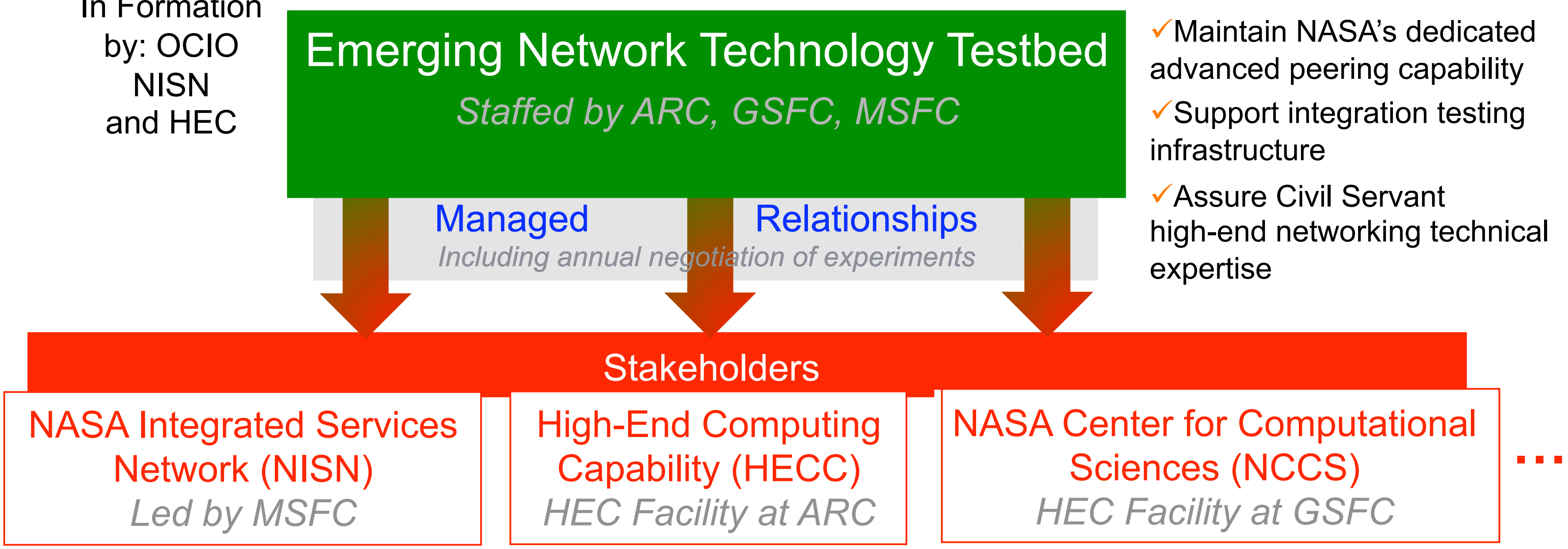
- “B” and “C” Systems will be deployed in the NCCS’ GSFC-local “next generation” iRODS prototyping testbed
- “A” Systems will be deployed as firewalls in the GSFC’s Science and Engineering Network (SEN)
- “A” Systems will be deployed as WAN delay emulators in the HECN Team’s GSFC-local advanced networking testbed
- “A+” Systems will be deployed in the HECN Team’s GSFC-local advanced networking testbed
- “A-” Systems may be deployed to JPL and LaRC for High End Computing Program testing across NISN’s WANX upgrade
- For more detail, contact Pat.Gary@nasa.gov



Source: Jim Fischer/GSFC 606.0

“Emerging Network Technology” Testbed

High-end network technology collaboration within NASA



In Formation by: OCIO NISN and HEC

Emerging Network Technology Testbed

Staffed by ARC, GSFC, MSFC

Managed Relationships

Including annual negotiation of experiments

Stakeholders

NASA Integrated Services Network (NISN)
Led by MSFC

High-End Computing Capability (HECC)
HEC Facility at ARC

NASA Center for Computational Sciences (NCCS)
HEC Facility at GSFC

…

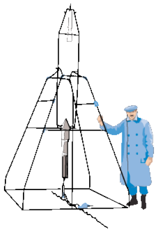
Scope:

✓Prototype advanced commercial networking capabilities in collaboration with stakeholders

✓Maintain NASA’s dedicated advanced peering capability

✓Support integration testing infrastructure

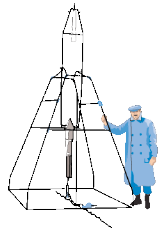
✓Assure Civil Servant high-end networking technical expertise



ENTT Vision

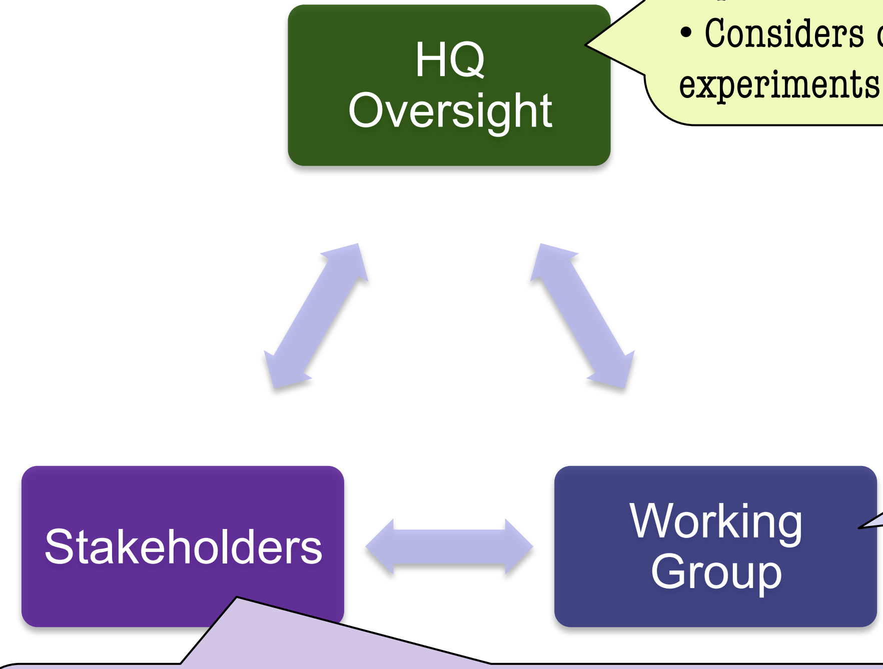
Source: Kevin Jones (ARC)

- Formally establishes an emerging network technology focused program to support NASA’s science and engineering missions
  - Prototypes advanced networking capabilities in collaboration with Stakeholders
  - Supports an innovative network technology testbed between Ames, Goddard and Marshall to address current and future network gaps and challenges
  - Maintains NASA’s dedicated advanced peering capability
  - Assures Civil Servant high-end networking technical expertise
  - Leverages collaboration with multiple government agencies, industry and university partners to conduct technology investigations in a cost effective manner



ENTT Governance

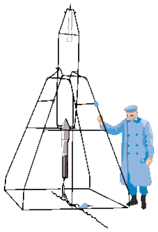
Source: Kevin Jones (ARC)



- Reviews/approves experiments to be negotiated
- Guidelines funds to Working Group for baseline infrastructure and experiments
- Considers optional Stakeholder proposals for co-funding of big experiments

- Engages Stakeholders to identify & prioritize needs/ challenges
- Negotiates win/win experiments with stakeholders including milestones and resource commitment
- Leads experiment execution
- Reports milestone achievement and lessons learned

- Identifies needs/challenges to Working Group
- Participates in experiment negotiations
- Works with HQ Oversight to optionally co-fund big experiments
- Supports experiment execution



NCCS Stakeholder Input to ENTT Formulation

Forces driving NCCS's networking requirements

Plans to enable increased collaboration by GSFC scientists with research partners, including internationals, via productivity enhancing software tools such as the Earth System Grid (ESG)

Paradigm shift in sources of data desired by GSFC scientists from NASA-provided to globally-provided (NASA|global ratio going from 80|20 → 20|80)

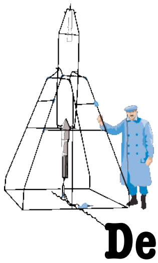
Explosion in data set size due to higher resolution models accommodating larger numbers of data types

Nearterm NCCS networking needs

Significantly enhanced data throughput from NCCS's computational engine (Discover) to NCCS's Internet server (Data Portal)

Support for live access visualization service for remote scientists

Enhanced local analytics involving multiple data sets for ESG users and others



NCCS Stakeholder Input to ENTT Formulation

Derived Network Technology Investigations (prioritized)

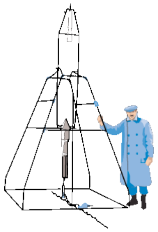
1. Performance measurement tools

2. Network technologies enabling 40G & 100G data flows
   a) Testbed workstations/clusters with enhanced file tranfer capabilities; NCCS' production capabilities will also be used for baselining
   b) Local area Ethernet switches, border routers & Firewalls
   c) Metro and wide area DWDM transport, optical switching &"test" networks that have SLAs very different from those of NASA's production networks

3. QoS approaches, e.g.:
   a) Distributed caching
   b) Application-to-network signaling

4. Alternate network & transport layer protocols, e.g., IPv6, UDT & UTX

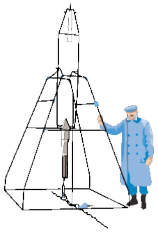
5. More effective collaboration tools involving HiDef video conferencing, whiteboards, etc



Climate Community need for Data Services

Source: Harper Pryor (NCCS/SAIC)

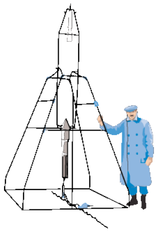
- Climate Scientists need access to a broad range of observational data and model output that resides in research institutions throughout the world. These data collections add up to many petabytes of data.
  - Doing Climate Science:
    - Studying fundamental processes at work in the Earth’s climate system
    - Quantifying the forcings – the factors that affect the Earth’s energy balance and drive climate change
    - Studying the climate record
  - Doing Climate Modeling:
    - Validating the algorithms used to represent physical processes
    - Validating climate models by testing them against available climate data
    - Initializing models



Roadmap of Evolving Capabilities

Source: Harper Pryor (NCCS/SAIC)

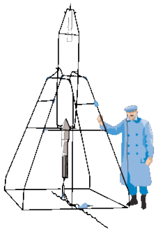
- **Data Portal** – project specific collaborative data sharing services
  - MAP projects with web distribution of data products & visualization data sets
  - NASA Field Campaign with support for flight planning
  - OSSE nature runs and climate data to the international OSSE science teams
  - Cloud Library to distribute GSFC science results from cloud resolving models
- **ESG Data Node** – broad climate community access to model projections
  - IPPC contributions generated by GMAO and GISS
  - Non-IPCC model output to scientific community
  - Provide analytic capabilities unique to the climate data users
- **GES DISC Access** – supporting access to observing system data
  - GES DISC Tools (Giovanni, Mirador, S4P) - data identification, manipulation, and visualization
- **iRODS** – data management layer to provide access to global data sets
  - Access to remote data sets, simulations, and observations without copying data
  - Join federated iRODS community to more easily access and distribute data



Earth System Grid

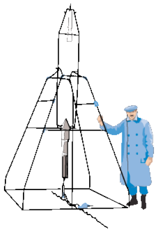
Source: Harper Pryor (NCCS/SAIC)

- Unified infrastructure environment to catalog and widely publish petabytes of distributed climate data so as to make it easily accessible to an international community of potential users – data remain resident within HPC environment where created.
    - Approach:
        - Data node software developed and distributed by PCMDI
        - Adopts standard formats and protocols in use throughout climate community
        - NCCS implementing ESG Data Node into NCCS Data Portal
        - NCCS integrating Live Access Server (LAS); collaborating with LANL to incorporate capabilities for climate data analysis and visualization
    - Capabilities:
        - Secure access
        - Monitor system and service usage
        - Catalog with metadata search capability
        - Transport, aggregation and subsetting
        - Distribute



ESG-CET Project
Major Data Access/Transfer Tools

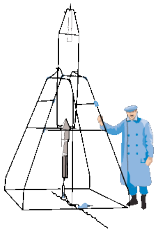
- ESG Gateway/Data Portal (web)
- GridFTP & Reliable File Transfer (RFT) Service
- Live Access Server (LAS)
- Open-source Project for a Network Data Access Protocol (OPeNDAP)
- DataMover (Bulk Data Movement)/Storage Resource Broker
- Data Nodes typically with ”in place” analysis capabilities



Data Management Requirements for Climate

Source: Dan Duffy (GSFC/NCCS)

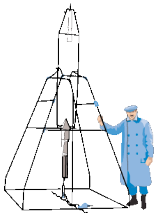
- Easier and high performing mechanism for storing data (long term) when computing in a geographically distributed environment
  - NASA Goddard, NASA Ames, Oak Ridge, and others
- Store and query metadata
- Federate observational and model data sources
- Provide mechanisms to transfer and translate data from its source in one format to its destination in the desired format
- Where did we start?
  - Underlying file transfer mechanisms
  - Open source frameworks for data management
  - User requirements



Objectives for the Wide Area Data Transfer Test Bed

Source: Dan Duffy (GSFC/NCCS)

- Testing of potential wide area data transfer mechanims as the basis for data management infrastructure and/or framework
- Compare traditional versus new file-copying utilities and mechanisms
- Determing optimal tuning parameter settings across wide area networks over multi-10 Gbps links
- As a baseline, determine maximum memory-to-memory throughput performance among the workstations and servers using nuttcp (http://www.nuttcp.org/)
- Basis for building the GSFC/High End Computing 20, 40 and 100 Gbps network testbed



DMS Development Strategy

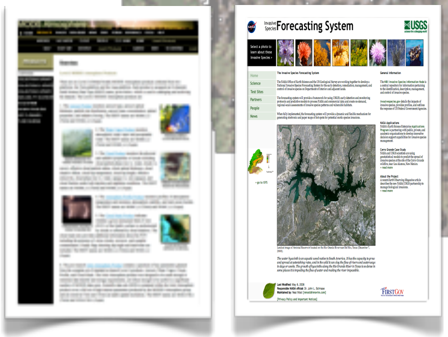
Source: Dan Duffy (GSFC/NCCS)

A progressive, three-part development strategy that systematically achieves the project's goals and produces the defined deliverables ...

Part 1 – Production MODIS Data Service

Part 2 – Prototype IPCC Data Service

Part 3 – Production IPCC Data Service



Part 1 – Production MODIS Data Service

Develop basic iRODS expertise in the controlled setting of a production, mission-oriented observational data system and tech-friendly collaborators and adopters ...

Part 2 – Prototype IPCC Data Service

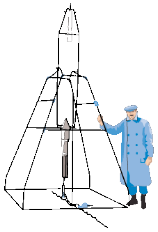
Refine and apply iRODS expertise in the controlled setting of a prototype climate simulation analysis environment ...

Part 3 – Production IPCC Data Service

Translate Part 1 and Part 2 experiences into a comprehensive plan for integrating iRODS technology into NCCS operations, delivering combined observational and simulation data services, connecting to the the Earth System Grid, etc. ...

The plan ...

- Builds on NCC's past investments and experiences with data grid technologies, including SRB.
- Leverages new technological advances (iRODS).
- Lays out an imcremental, risk-controlled strategy of introducing a data grid.

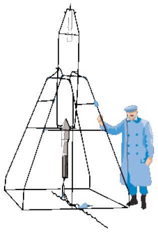


NCCS Stakeholder Input to ENTT Formulation

Future NCCS Network Technology Needs: Frameworks

- The network technology needs include:
  - Not only wide area network infrastructures within NASA and with NASA external partners
  - But also local area network components such as user-managed firewalls, switches and routers
  - And even network aspect of facility-based and/or end-user computers and software application mods to enable more effective data transfers to enhance science research

- The network technology needs change with time, such as indicated in the draft timeline of NCCS’ 4x10G, 1x40G & 1x100G testbedding interests



NCCS Stakeholder Input to ENTT Formulation

Future NCCS Network Technology Needs: Derived

- Testbed* workstations/clusters with 40G & 100G data flow capabilities
  - NICs: initially via 4x10G link aggregation, but native asap
  - Motherboards with multiple PCIe G2 & G3 interfaces
  - Four and more cores per processor: Nehalem architecure or better
  - Very fast but inexpensive disks: likely SSD based
  - Various data transfer applications for benchmarking and, where possible, tuning

  *NCCS' production capabilities will also be used for baselining

- Local area network components
  - Ethernet switches and border routers; multiple 10G, 40G & 100G interfaces
  - Firewalls: multiple 10G, 40G & 100G interfaces

- Metro and/or wide area network components
  - DWDM optical transport and switching
  - Note: a significant amount of end-to-end data flow testing can and should be done over "test" networks that have SLAs very different from those of NASA's production networks

GSFC SEN+HECN Summary Information

GSFC High End Computer Network (HECN) Team

Pat Gary/GSFC

Bill Fink/GSFC

Paul Lang/ADNET

Aruna Muppalla/ADNET

Jeff Martz/ADNET

Mike Stefanelli/ADNET

NETWORK BOTTLENECKS