ESnet updates: The ANI Project and more

Brian Tierney, ESnet,

Open Science Grid Storage Forum, Sept 21, 2010

Talk Overview

ANI Project

- 100G prototype network
- Testbed

New public “Disk I/O performance” test hosts

ESnet’s network knowledge base: fasterdata.es.net

perfSONAR update

ANI: Advanced Network Initiative

Project Start Date: September, 2009

Funded by ARRA for 3 years

Designed, built, and operated by ESnet staff

3 ARRA "Advanced Network Initiative" (ANI) projects in the DOE

- ANI 100G Prototype
- ANI Network Testbed
- 4 ANI research projects

DOE’s Advanced Networking Initiative

ANI Project scope ($66.8M):

- Build an end-to-end 100 Gbps prototype network
  - Handle proliferating data needs between the three DOE supercomputing facilities and NYC international exchange point
- Build a network testbed facility for researchers and industry
  - Includes $5M in network research that will use the testbed facility

Magellan:

- Separate DOE-funded ($32.8M) nationwide scientific mid-range distributed computing and data analysis testbed to explore whether cloud computing can help meet the overwhelming demand for scientific computing
- NERSC / LBNL & ALCF / ANL configured with multiple 10’s of teraflops and multiple petabytes of storage, as well as appropriate cloud software

ANI Project Goals

Prototype network:

- Accelerate the deployment of 100 Gbps technologies
- Build a persistent infrastructure that will transition to the production network ~2012
  - Key step toward DOE's vision of a 1-Terabit network linking DOE supercomputing centers and experimental facilities
- Not for production traffic, not routed to the general internet

Testbed:

- Build an experimental network research environment at sufficient scale to usefully test experimental approaches to next generation networks
  - Funded for 3 years, then roll into the ESnet program
  - Breakable, reserveable, configurable, resettable
  - Enable R&D at speeds up to 100 Gbps

ANI 100G Technology Evaluation

Most devices are not designed with any consideration of the nature of R&E traffic – therefore, we must ensure that appropriate features are present and devices have necessary capabilities

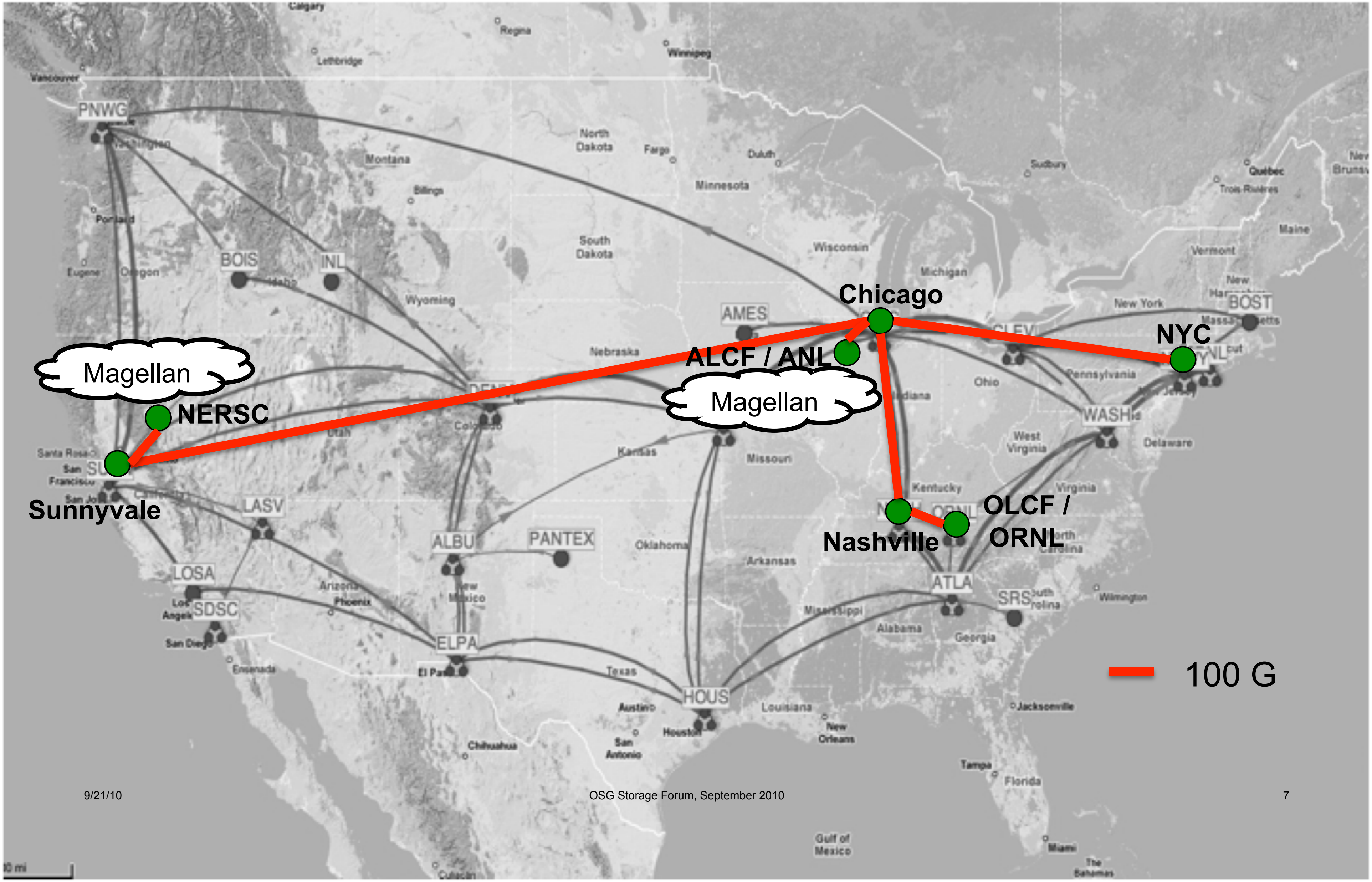
Goals (besides testing basic functionality):

- Test unusual/corner-case circumstances to find weaknesses
- Stress key aspects of device capabilities important for ESnet services

Many tests conducted on multiple vendor alpha-version routers, examples:

- Protocols (BGP, OSPF, ISIS, etc)
- ACL behavior/performance
- QoS behavior
- Raw throughput
- Counters, statistics, etc

Nationwide 100G Prototype Network



ANI Prototype Network Status Update

RFP issued June 15, asking for:

- 100G Service between MANLAN, ORNL, ANL, and NERSC
- 20-year dark fiber IRU for nationwide footprint
  - Northern route, southern route, SF Bay Area MAN, Chicago area MAN

Responses due Aug 20

Proposals are currently being evaluated

Tentative Schedule for future:

- make a decision on 100G RFP by Oct 1
- DOE approval and contract negotiation: Oct-Dec
- Contract Awarded: December
- 100G Router RFP will be issued early October
- Prototype network ready to begin testing: September, 2011

Testbed Overview

Progression

- Start out as a tabletop testbed
- Move to Long Island MAN when dark fiber is available
- Extend to WAN when 100 Gbps available

Capabilities

- Ability to support end-to-end networking, middleware and application experiments, including interoperability testing of multi-vendor 100 Gbps network components
- Researchers get “root” access to all devices
- Use Virtual Machine technology to support custom environments
- Detailed monitoring so researchers will have access to all possible monitoring data

Sample Projects

Examples of the types of projects the current testbed will support include the following:

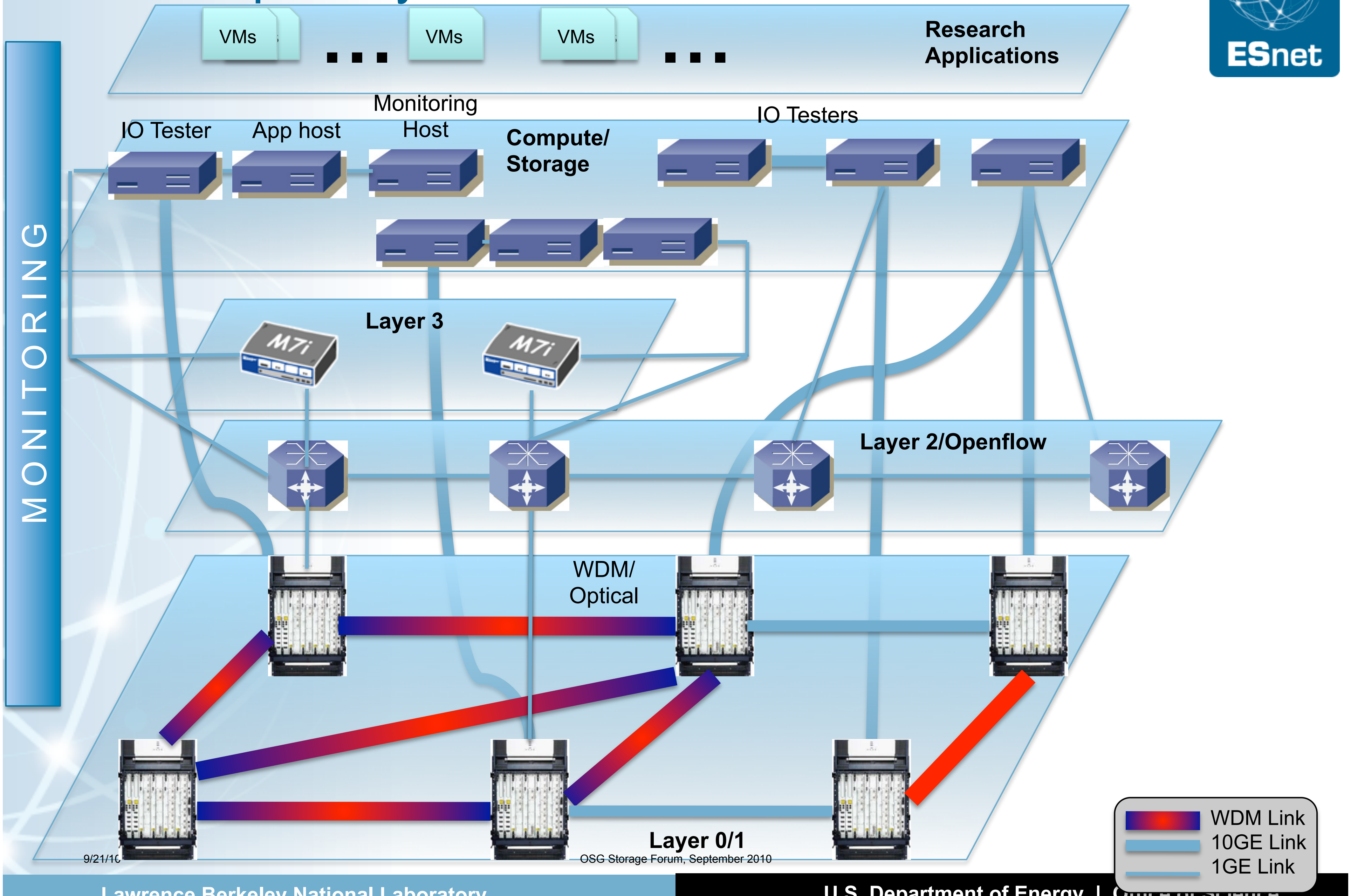
- Path computation algorithms that incorporate information about hybrid layer 1, 2 and 3 paths, and support 'cut-through' routing
- ***New transport protocols for high speed networks***
- Protection and recovery algorithms
- Automatic classification of large bulk data flows
- New routing protocols
- New network management techniques
- Novel packet processing algorithms
- ***High-throughput middleware and applications research***

Network Testbed Components

Table Network Testbed consists of:

- 6 DWDM devices (Layer 0-1)
- 4 Layer 2 switches supporting Openflow
- 2 Layer 3 Routers
- Test and measurement hosts
  - Virtual Machine based test environment
  - 4x10G test hosts initially
    - Eventually 40G and 100G from Acadia 100G NIC project
- This configuration will evolve over time

Tabletop: A layered view



Testbed Status

Tabletop Testbed available for researchers to log in as of late June.

- researchers are logging in, configuring VMs, running tests, etc.
- can reserve testbed components using Google calendar.

User documentation mostly complete:

- https://sites.google.com/a/lbl.gov/ani-testbed-user-guide/

Per-project Monitoring set up:

- https://tb-webdav-1.es.net/ganglia/

Testbed-support@es.net email list is quite active

A few remaining tasks to be done: e.g.: web interface to claim reserved resources

For Phase 2: RFP for the Long Island dark fiber ring has been signed and construction has started.

Testbed Access

Proposal process to gain access described at:

https://sites.google.com/a/lbl.gov/ani-testbed/

Currently there are 4 DOE-funded projects that have access to the testbed

- 3 more are waiting for 40G capability

First round of proposals are due Oct 1

- Accepted proposals announced Dec 10, 2011

Proposal review committee members:

- DOE Lab: Phil DeMar, FNAL; Les Cottrell, SLAC
- University: Ben Yoo, UC Davis
- Industry: Bikash Koley (Google); David Richardson (Amazon); Steve Wolff (Cisco); Wes Doonan, Adva
- International: Cees De Laat, U Amsterdam; Mauro Camponelli, GARR; Tomohiro Kudoh, AIST
- Other: Jerry Sobiesky, Nordunet; Kevin Thompson, NSF

New ESnet I/O Performance Testing Service

New ESnet Diagnostic Tool: 10 Gbps IO Tester

16 disk raid array: capable of > 10 Gbps host to host, disk to disk

Runs ***anonymous*** read-only GridFTP – no keys needed

Accessible to anyone on any R&E network worldwide

1 deployed on now (lbl-diskpt1.es.net)

- 2 more (anl-diskpt1 and bnl-diskpt1) being deployed next week

Already used to debug many problems

Available for Supercomputing demo’s

See: http://fasterdata.es.net/disk_pt.html

ESnet I/O Testers

Security Model

- Runs in a Linux ‘jail’
- Anonymous Access to a set of read-only data sets:
  - 1G, 10G, 50G, and 100G data sets
- Temporary write access possible

Other tools under consideration

- bbcp, Phoebus, BestMan

Sample Commands

```
# copy 10G file using 4 parallel streams

globus-url-copy -vb -p 4 ftp://lbl-diskpt1.es.net:2811/data1/10G.dat file:///tmp/test.out

# memory to memory: /dev/zero to to /dev/null

globus-url-copy -vb -p 4 -len 10G ftp://lbl-diskpt1.es.net:2811/dev/zero file:///tmp/test.out
    file:///dev/null
```

fasterdata.es.net

ESnet’s Network Performance Knowledge Base

Network Performance Knowledge Base: http://fasterdata.es.net/

Collection of instructions, theory, tutorials, and best practice documents

- Principal maintainers: Eli Dart and Brian Tierney, with contributions from ESnet engineering staff

Focused on the scientific community

- High-performance data transfer
- Commonly-deployed tools
- Incorporates lessons learned and best practices derived from ESnet staff experience solving customer problems
- Built on foundation of my TCP tuning site, evolving for over 10 years

Easy quick-reference Instructions for deploying common tools

- Many configuration instructions are cut-and-paste for ease of use and reduced errors
- Configuration guides for common network equipment (e.g. Cisco routers)

ESnet perfSONAR Service

Why perfSONAR?

PerfSONAR & the wizard gap

- ESnet is deeply involved in perfSONAR because typical users can't get what they need out of the network.

PerfSONAR allows us to:

- Find and fix problems that impact performance
- Help users understand network problems in their domain
- Demonstrate network capabilities & establish baselines
- Make sure problems stay fixed
- Quantify multi-domain network performance

ESnet Engineers are working 2-3 network performance problems in a typical week

- Problems are usually configuration/tuning related, or "soft failures"

ESnet perfSONAR Infrastructure

- Measurement Points: Bandwidth and Latency (>50 servers)
  - Every 10GE Hubs and 10GE site
  - 3 Equinix Locations (commercial peering locations)
  - Deploying 20+ more systems at low bandwidth sites in progress
- Full list of active services at:
  - http://stats1.es.net/perfSONAR/directorySearch.html
  - Instructions on using these services for network troubleshooting: http://fasterdata.es.net

These services have proven extremely useful to help debug a number of problems

ESnet perfSONAR Measurement Infrastructure Usage

Regular tests between ESnet systems

- Latency - A full mesh of continuous tests between 25 systems
- Bandwidth:
  - All Nodes run 4 tests a day to Chicago, Sunnyvale, Houston & Washington

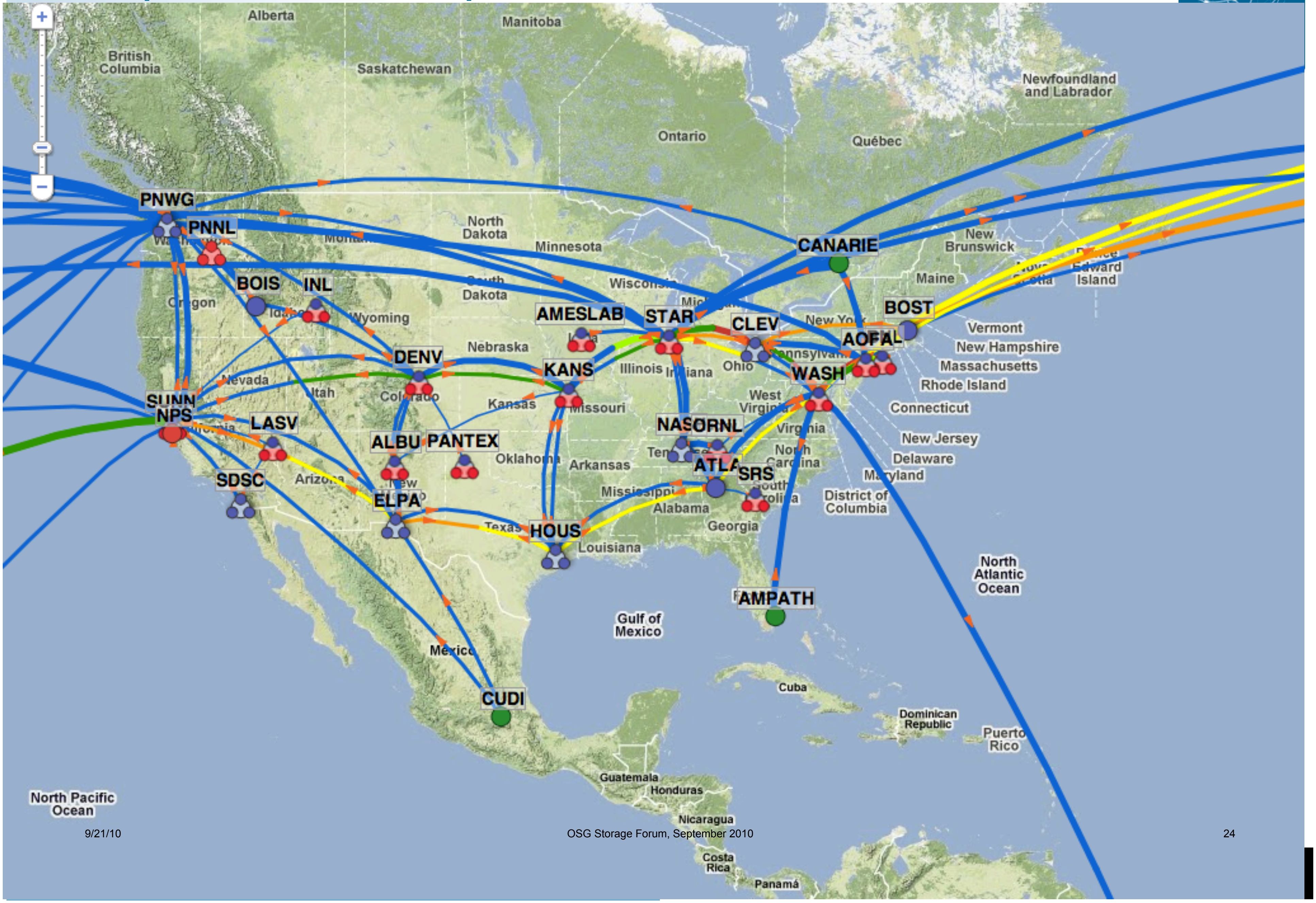
Other tests maintained by ESnet

- Regularly scheduled tests to Internet2, NLR & NOAA (JET)
- BNL, MIT, U-Penn, Indiana-U, LBL. SMU, Atlas Great Lakes T2

Some of the organizations running scheduled tests to ESnet

- PNL, ANL, LBL, USLHCnet, Hat Creek Radio Observatory, IHEP.AC.CN, lcg.ustc.edu.cn, UIUC, U-Utah, AARnet, U-Michigan, SMU, U-Penn, MIT

http://weathermap.es.net



Global PerfSONAR-PS Deployments

Based on "global lookup service" (gLS) registration, May 2010: currently deployed in over 100 locations

- ~ 100 bwctl and owamp servers
- ~ 60 active probe measurement archives
- ~ 25 SNMP measurement archives
- Countries include: USA, Australia, Hong Kong, Argentina, Brazil, Uruguay, Guatemala, Japan, China, Canada, Netherlands, Switzerland

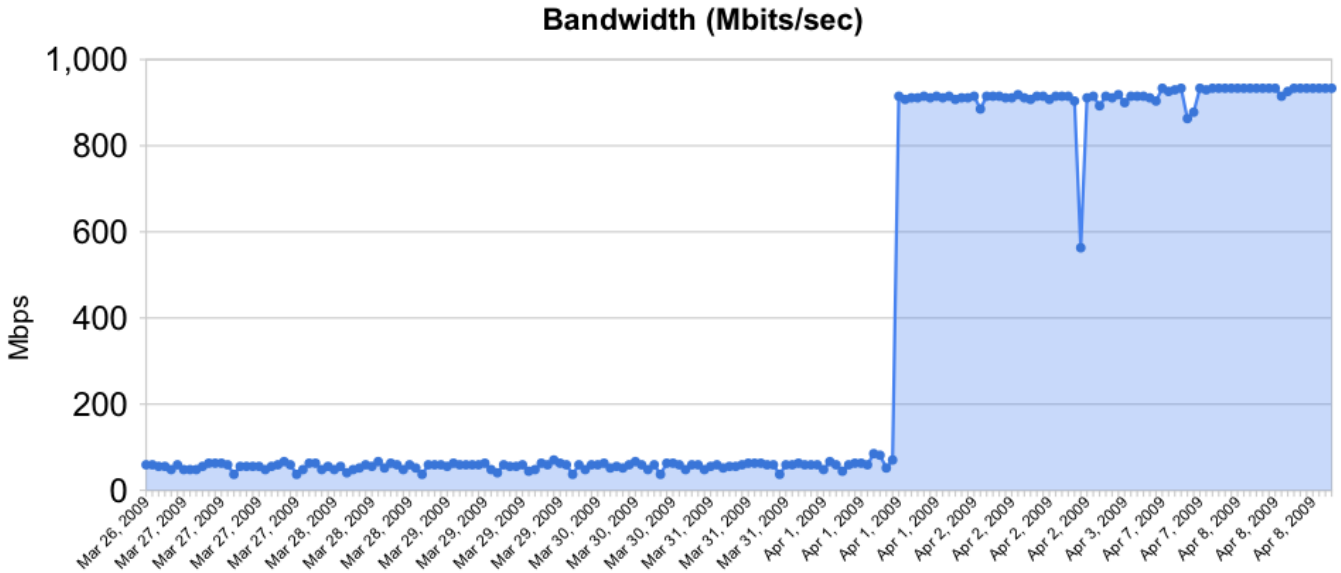
US Atlas Deployment

- Monitoring all "Tier 1 to Tier 2" connections

For current list of public services, see:

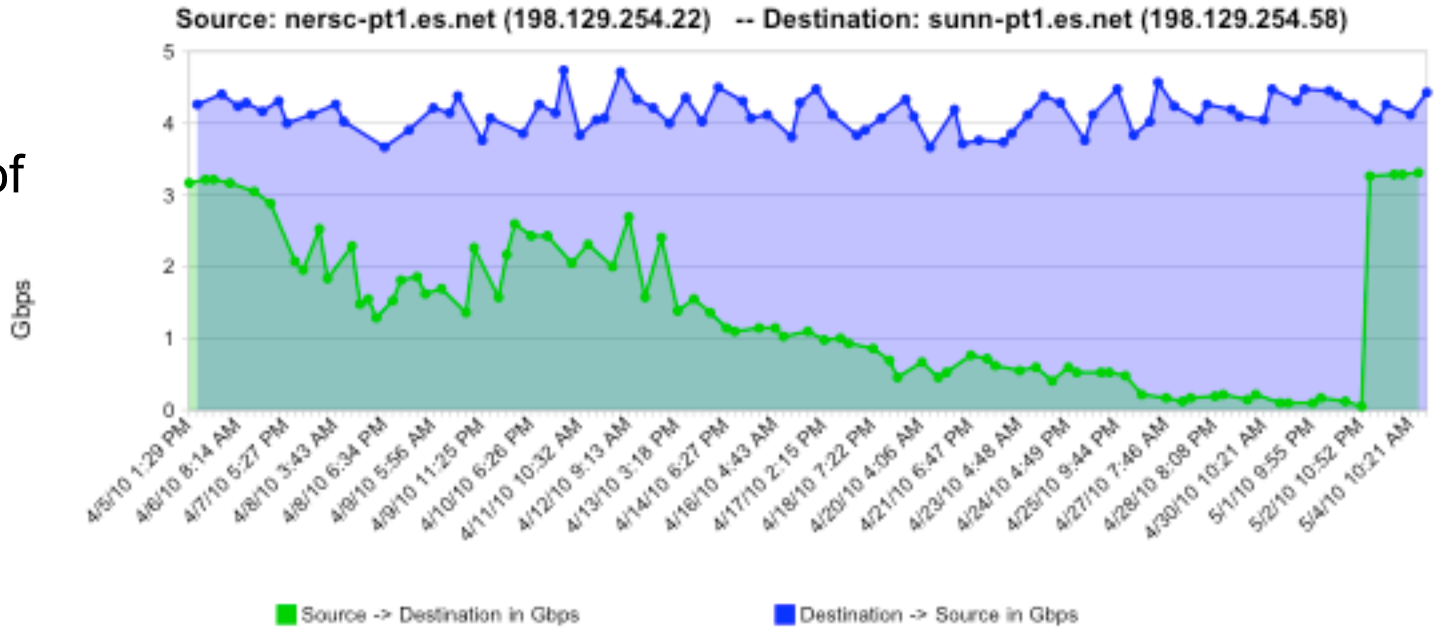
- http://stats1.es.net/perfSONAR/directorySearch.html

Sample Results: Finding/Fixing soft failures



Rebooted router with full route table

Gradual failure of optical line card



perfSONAR 3.2 release

New features

- RPM-based (Centos 5.5)
- New “Network Install” option alternative to LiveCD
- Many bug fixes / performance enhancements

Currently in beta (rc3 available for testing):

- http://packrat.internet2.edu/~aaron/pS-Performance_Toolkit-NetInstall-3.2rc3.iso
- http://packrat.internet2.edu/~aaron/pS-Performance_Toolkit-LiveCD-3.2rc3.iso

More Information

http://100gbs.lbl.gov/

http://sites.google.com/a/lbl.gov/ani-testbed/

http://fasterdata.es.net/

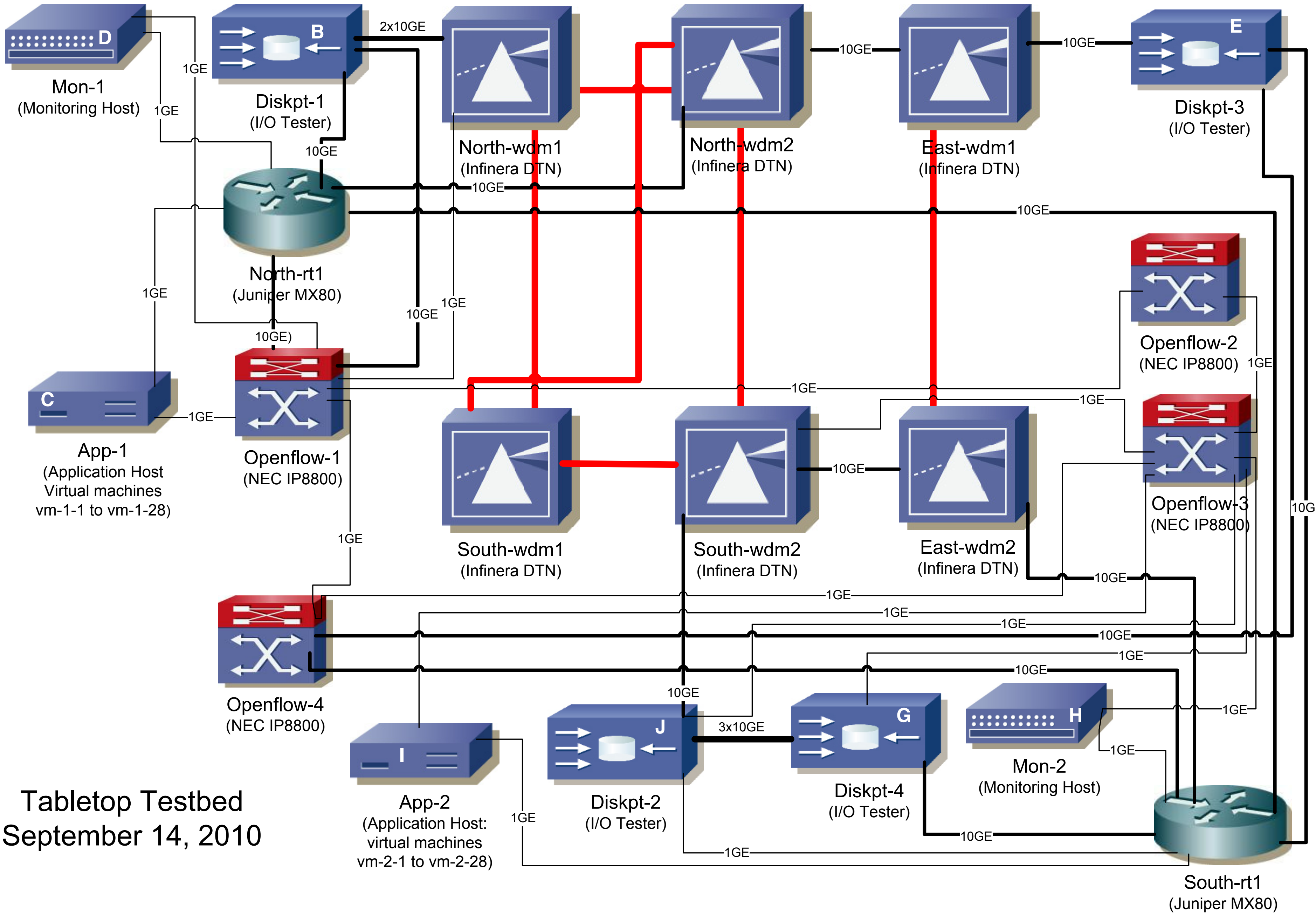
http://stats1.es.net/

http://www.perfsonar.net/

email: BLTierney@es.net

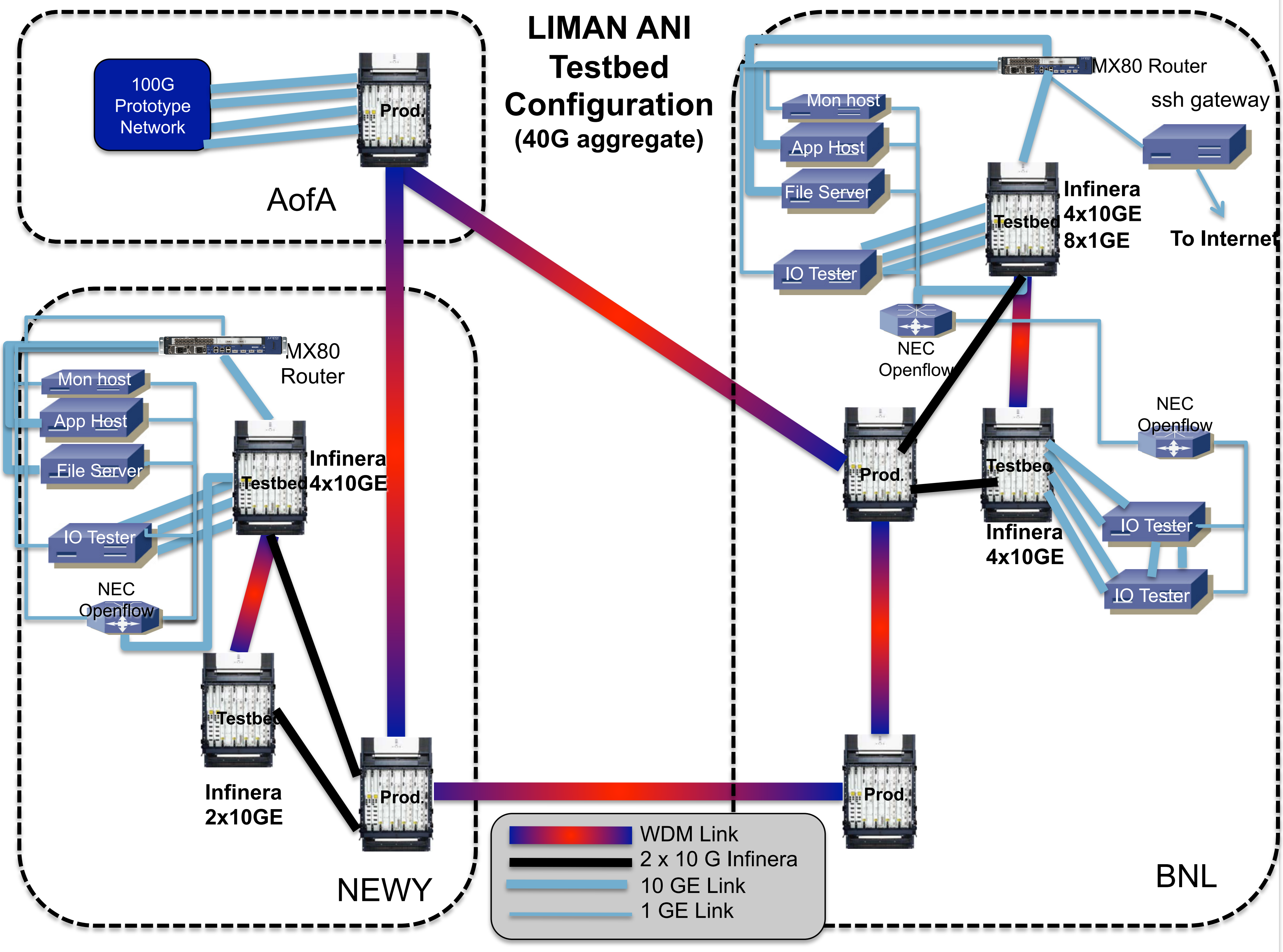
Extra Slides

Tabletop Testbed Configuration



Tabletop Testbed
September 14, 2010

LIMAN ANI Testbed Configuration (40G aggregate)



100G Prototype Network

Prod.

AofA

MX80 Router

ssh gateway

Mon host

App Host

File Server

IO Tester

Testbed

Infinera 4x10GE 8x1GE

To Internet

NEC Openflow

Prod.

Testbed

Infinera 4x10GE

NEC Openflow

IO Tester

IO Tester

Prod.

BNL

MX80 Router

Mon host

App Host

File Server

Testbed

Infinera 4x10GE

IO Tester

NEC Openflow

Testbed

Infinera 2x10GE

Prod.

NEWY

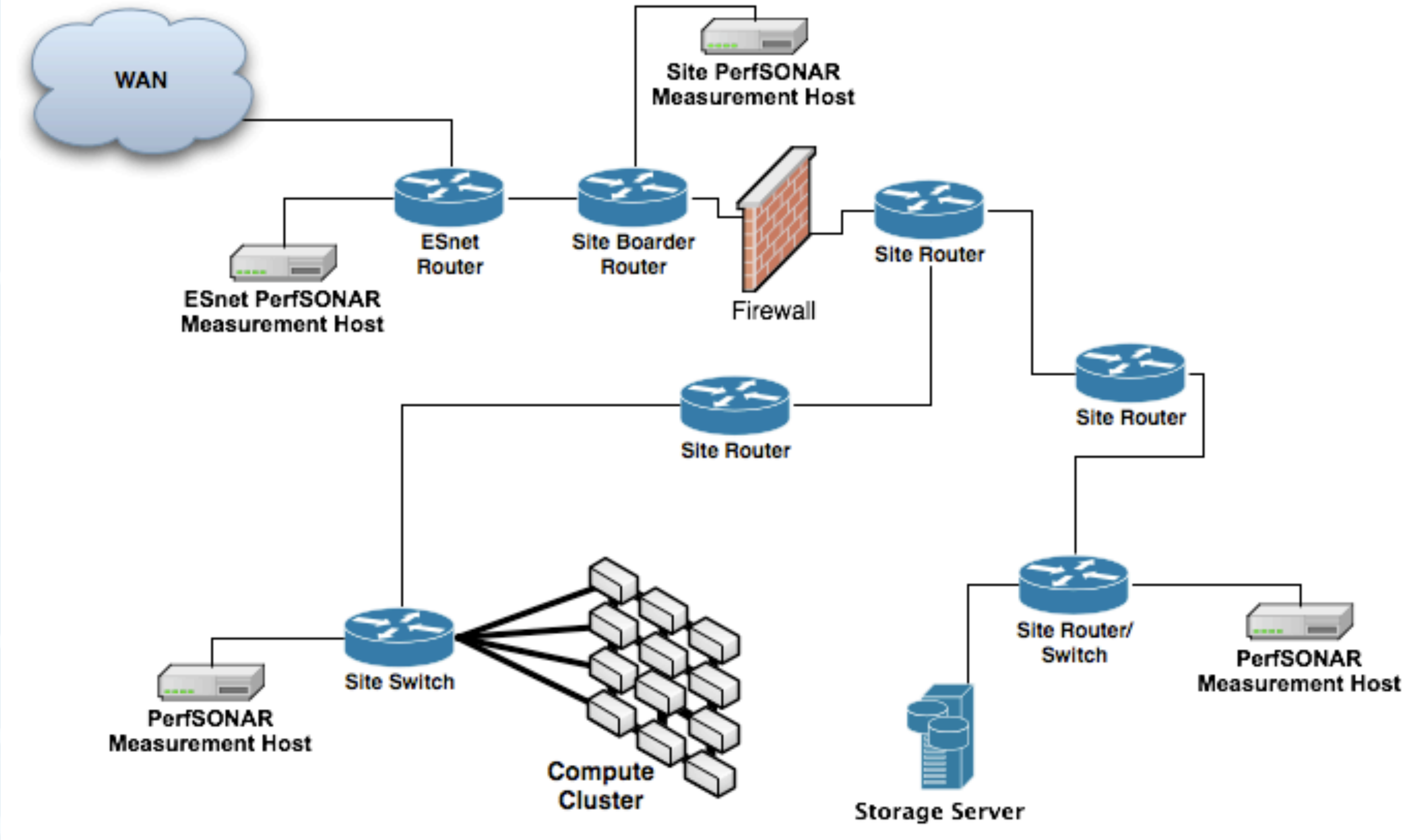
WDM Link

2 x 10 G Infinera

10 GE Link

1 GE Link

Sample PerfSONAR Site Deployment


WAN
Site PerfSONAR Measurement Host
ESnet Router
Site Boarder Router
Firewall
Site Router
ESnet PerfSONAR Measurement Host
Site Router
Site Router
Site Switch
Compute Cluster
Site Router/ Switch
PerfSONAR Measurement Host
PerfSONAR Measurement Host
Storage Server