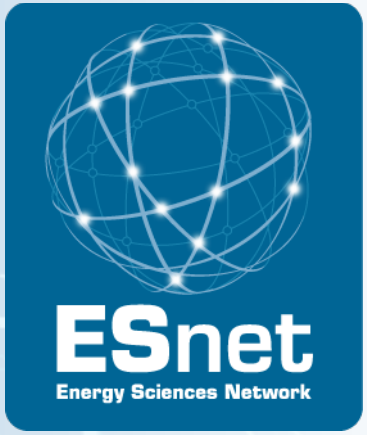


ESnet 100GTestbed: A Federated Model for Site Involvement

Brian Tierney

Berkeley National Lab / ESnet

ESCC, Jan, 2013

Leveraging the ESnet Testbed for Site Experiments

Example: Site A wants to run WAN benchmarks on a new DTN, but does not want to impact the production network.

- ESnet Testbed hosts (or another site), could be used as a data source/sink

Example: Fusion Scientist at ORNL needs to get data from a Tokamak in Korea as quickly as possible for analysis on OLCF, and want to experiment with RDMA protocols as an alternative to TCP

How to leverage the ESnet testbed to help with this testing?

Potential Benefits to Sites

Allows site researchers ability to do testing between their own systems at other testbed hosts.

- Ability to test/benchmark new DTNs, cloud resources, etc. in a controlled environment before putting them into production

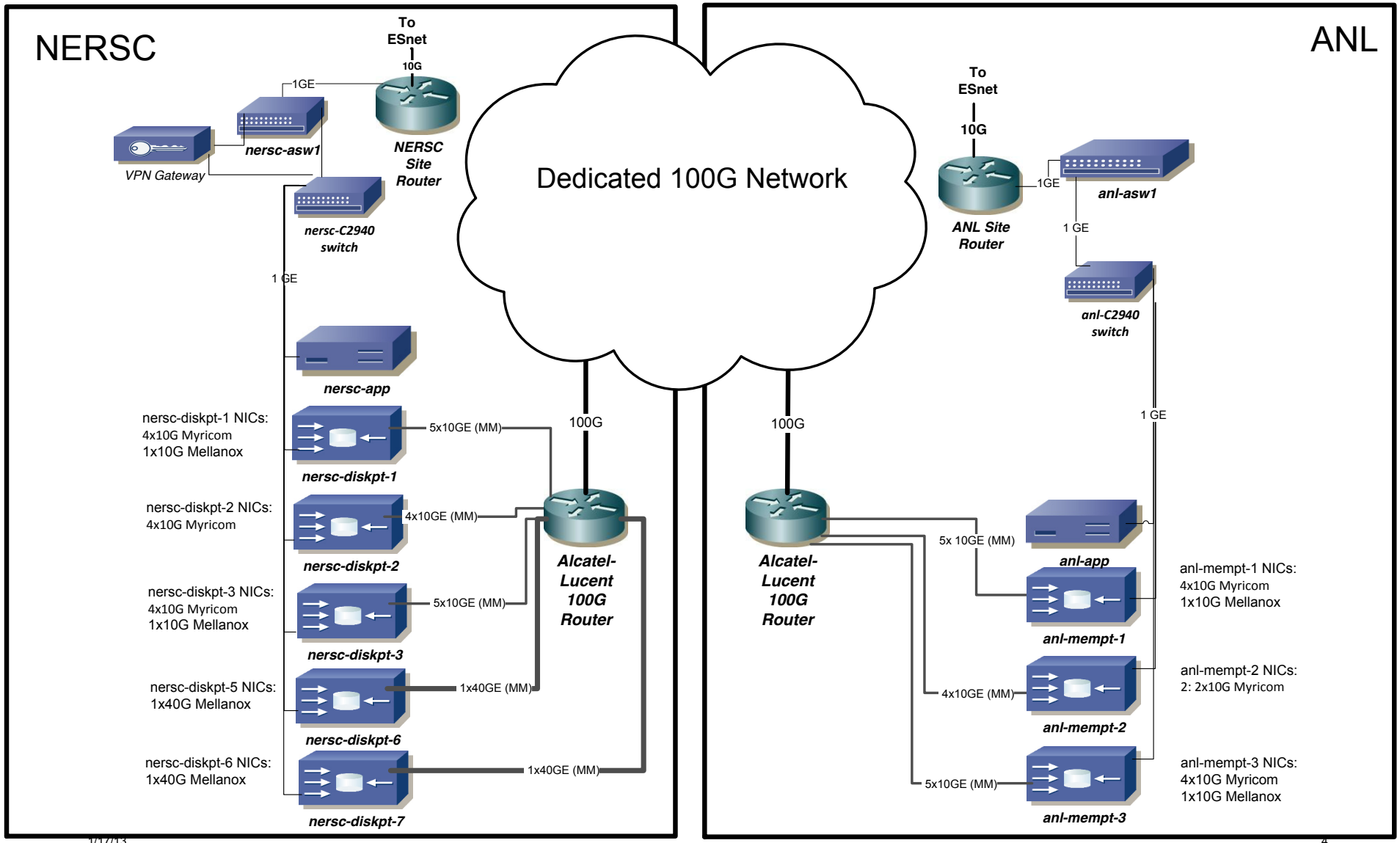
Allows site systems staff to experiment with next-gen hardware to help with procurement decisions.

Provides connection between site networking team and site scientists

Provides connection between sites and the network research community (useful recruiting tool)

Connection to Testbed strengthens research proposals from DOE sites

ESnet 100G Testbed



Unique Features of the ESnet 100G Testbed

A realistic national-scale environment for innovative network research that would be impossible for most researchers to create in their labs

- Many types of network research require realistic high-latency environments

Maximum flexibility: researchers get “super-user” to all hosts

- “bare metal” host access
- Each project get their own boot image with root access
  - ability to install custom OS on hosts

A controlled environment that supports reproducible results

- Complete control of every packet on the network

Sample Testbed Results

- Researchers at Indiana University found that the current upper limits of the traditional socket-based TCP and UDP protocols are 22 Gbps for Transmission Control Protocol (TCP) and 34.7 for User Datagram Protocol (UDP).
- Multiple projects were interested to see how remote direct memory access (RDMA)-based protocols would behave in a high-speed WAN environment. These experiments showed that with proper tuning, RDMA works quite well in the wide area, and can achieve 39.2 Gbps on a 40 Gbps link using only 1-2% of a CPU core.
- Using 2010-era hardware, only ten 10 GE NICs spread across three hosts are needed to completely saturate a 100 Gbps network.
- The Magellan project at Argonne National Laboratory (ANL) confirmed that using OpenStack, 10 virtual machines on 10 physical hosts can also saturate 100 Gbps.

Sample Results: RDMA over Converged Ethernet (RoCE)

Pronounced “Rocky”

Infiniband made to work over existing Ethernet networks

- Annex A16 of IB Architecture Specification

“Converged” Ethernet (i.e. Data Center Bridging) is not a requirement for RoCE

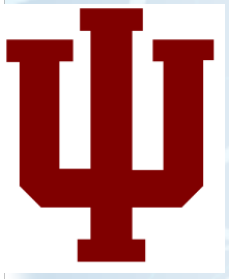
- Works over WANs

Requires certain guarantees about path characteristics

- Minimal packet loss / reordering
- Priority flow control (PFC)
- Hardware support in the NIC (Mellanox)

Recent Testbed Results: **Single** flow 40G

Tool	Protocol	Gbps	Send CPU	Recv CPU
netperf	TCP TCP-splice UDP	17.9 39.5 34.7	100% 34% 100%	87% 94% 95%
xfer_test	TCP TCP-splice RoCE	22 39.5 39.2	100% 43% 2%	91% 91% 1%
GridFTP	TCP UDT RoCE	13.3 3.6 13	100% 100% 100%	94% 100% 150%



Details in this paper:
http://www.es.net/assets/pubs_presos/eScience-networks.pdf

100G Testbed: Next Steps

Move ANL equipment to StarLight

- Adds ability to connect to other testbeds, including GENI

Add ExoGENI Racks

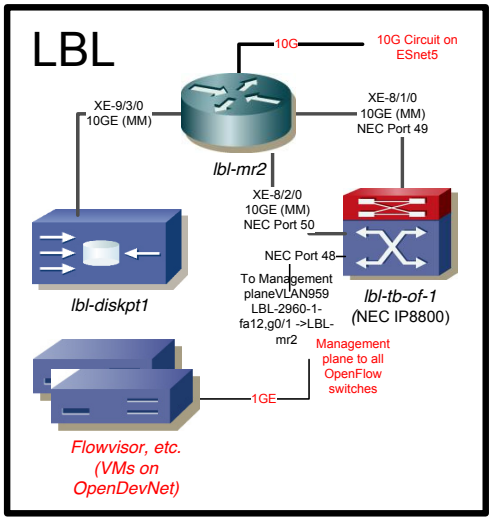
- At NERSC and Starlight, connected at 40G

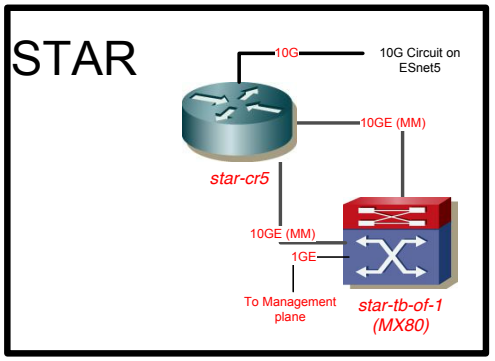
Connect to MANLAN and BNL

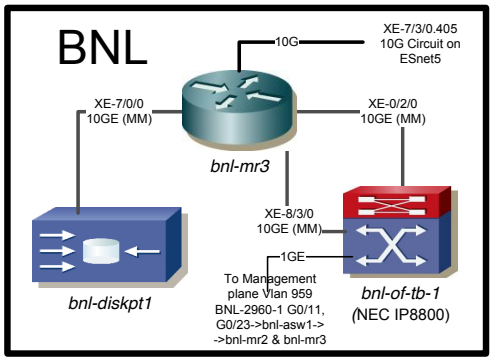
- Extend current testbed wave from Chicago to NYC, and connect to ESnet Long Island MAN to get to BNL

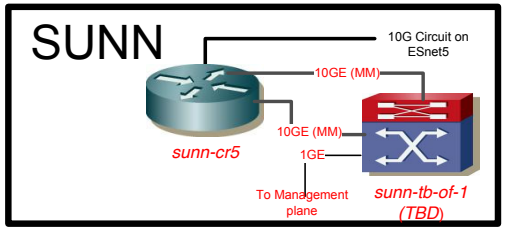
Also building an OpenFlow Testbed

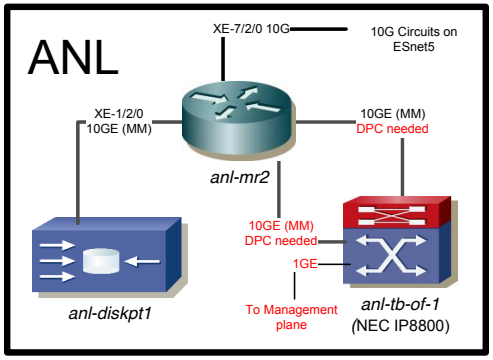
ESnet OpenFlow Testbed

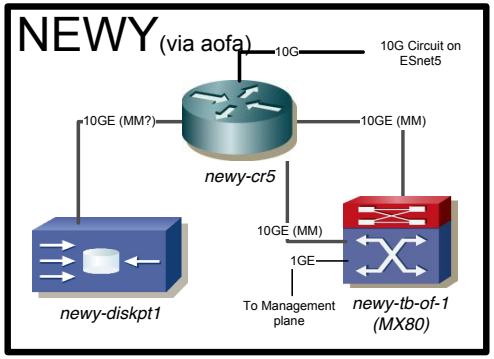


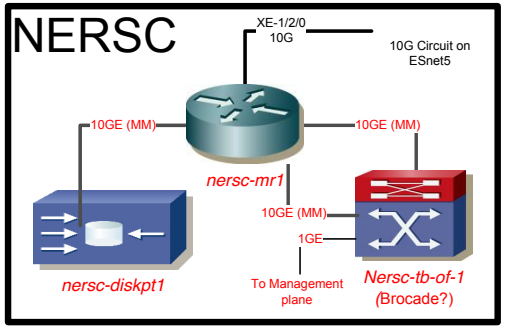


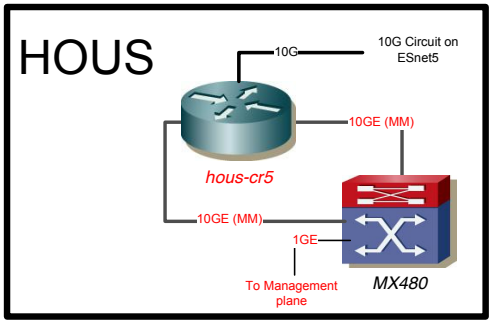


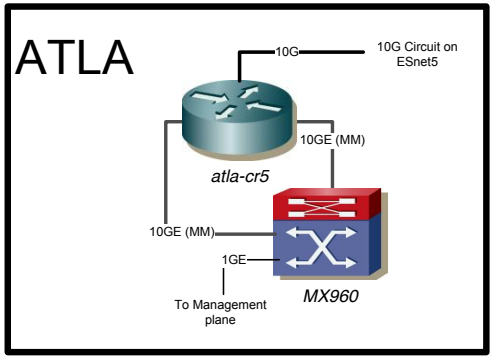












Phase 1: LBL, NERSC, BNL (Oct 2012)
Phase 2: ANL, STAR, NEWY (Feb 2013)
Phase 3: HOUS, ATLA, SUNN (March 2013)

Updated January 9, 2013

Expanding the Definition of Testbed

Any unused capacity on ESnet5 could become part of the testbed using OSCARS.

- Any site can request an OSCARS circuit to any other site for this testing
  - potentially to other networks too that support OSCARS
- Only need to reserve in the Testbed reservation calendar if using the NERSC to StarLight wave.
- Many network segments currently have 50-80G available for new OSCARS circuits.

Federation Model

Sites maintain full control of their resources

- Use existing site user account management process and access control process

Use layer-2 circuits to control what hosts can connect where

All circuits go via StarLight in Chicago

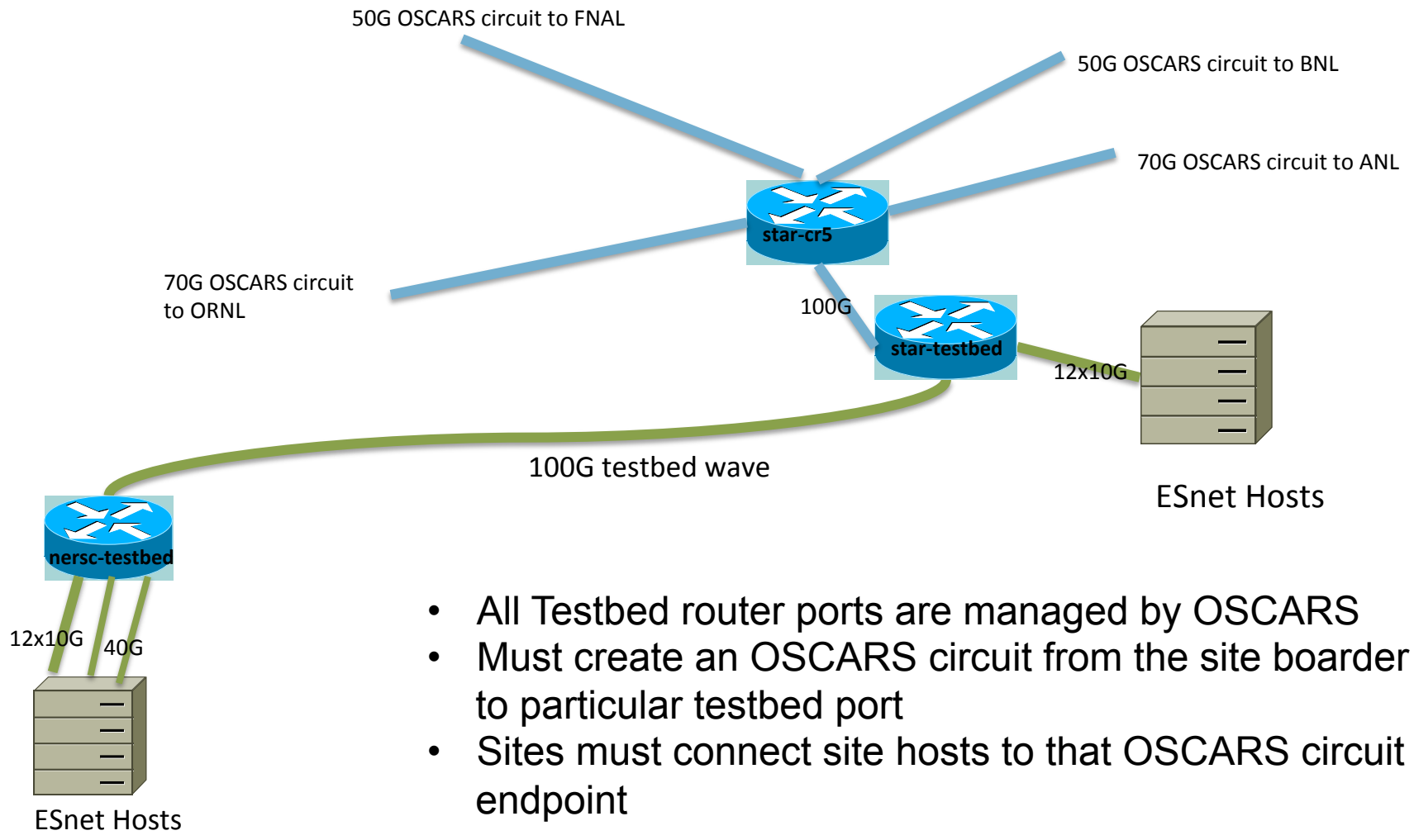
- This is the only location where the Testbed and ESnet5 are connected

All hosts on the testbed use RFC1918/private addresses to talk to each other

- Site hosts may also have public addresses as well
  - Sites are responsible to make sure end hosts do not route packets between the testbed network and the Internet.

ESnet will help facilitate remote end-points for tests

ESnet 100G Testbed with Sample Site connections



- All Testbed router ports are managed by OSCARS
- Must create an OSCARS circuit from the site boarder to particular testbed port
- Sites must connect site hosts to that OSCARS circuit endpoint

Testbed Security Model

Testbed is not routed to anywhere

External connectivity is via a VPN gateway host, connected to the internet at 1Gbps

- This is used to ssh into hosts, install software, etc.

Sites must ensure their hosts are not providing back-door routing

Site Resources Currently Committed

NERSC

- 8 DTN nodes are connected to testbed (2x10G each) and to NERSC production GPFS
- All NERSC resources could be connected

BNL

- 2 40G PCI gen3 hosts
- Juniper MX router with OpenFlow

FNAL

- 50Gbps of the 100G production network
- Nexus 7000 w/ 2-port 100GE module / 6-port 40GE module / 10GE copper module
- 6 nodes w/ 10GE Intel X540-AT2 (PCIe) / 8 cores / 16 GB RAM
- 2 nodes w/ 40GE Mellanox ConnectX®-2 (PCIe-3) / 8 cores w/ Nvidia M2070 GPU

Site Resources Under Discussion

ANL

- Connecting KBASE OpenStack nodes for OpenFlow experiments

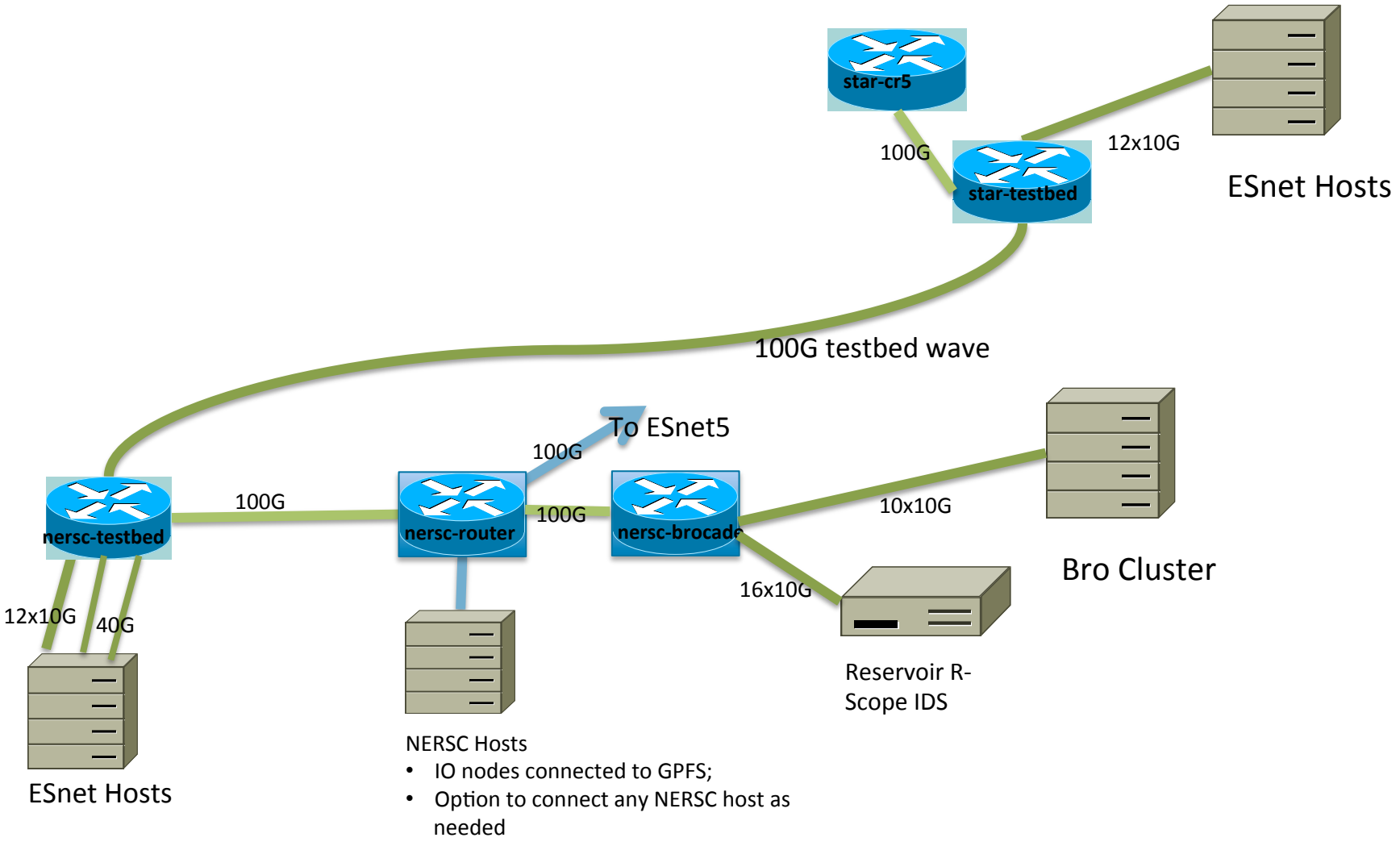
ORNL

- Resources may include
  - remote file system testing/connectivity, using Lustre
  - testing of expanded DTN capability prior to production
  - remote collaboration (K-Star as a potential example)

LLNL/LVOC

- Connecting some PCMDI DTNs

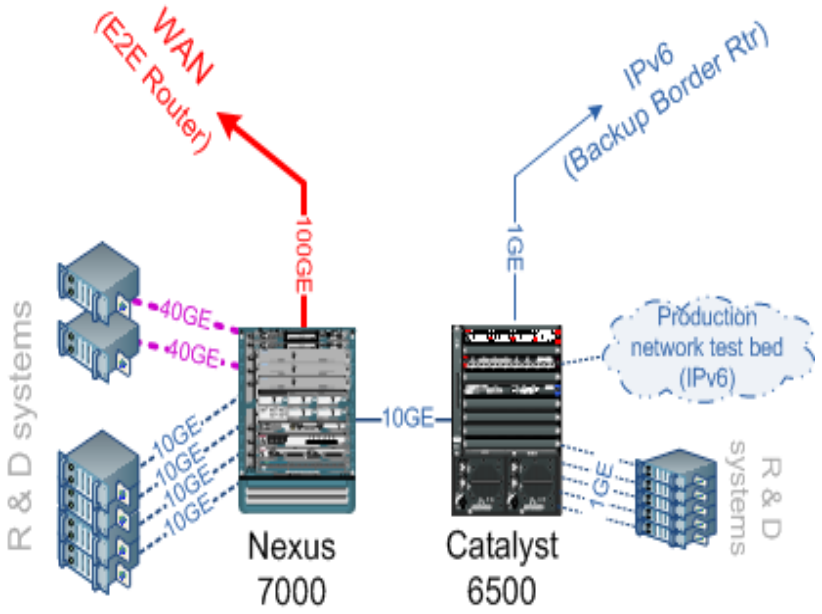
ESnet 100G Testbed with NERSC



FNAL Network R&D Test Environment

As previously discussed:

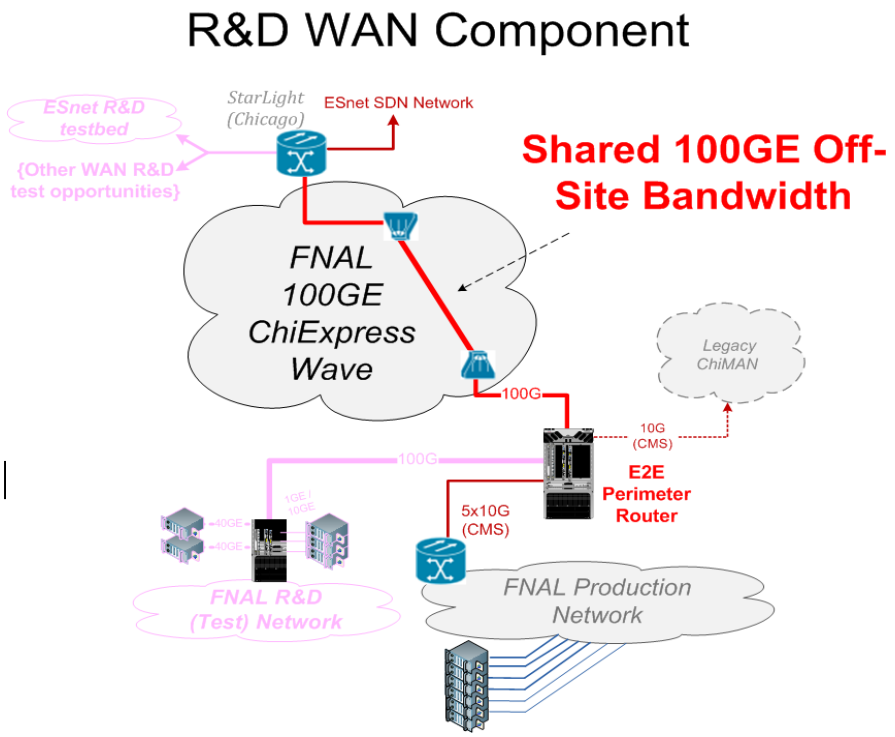
- 100GE-capable test environment, based on Nexus 7000
- 100GE connection to border
- Modest (~6) number of 10GE-connected systems
- Two PCI Gen3 systems w/ 40GE NICs
  - Nvidia GPUs as well…



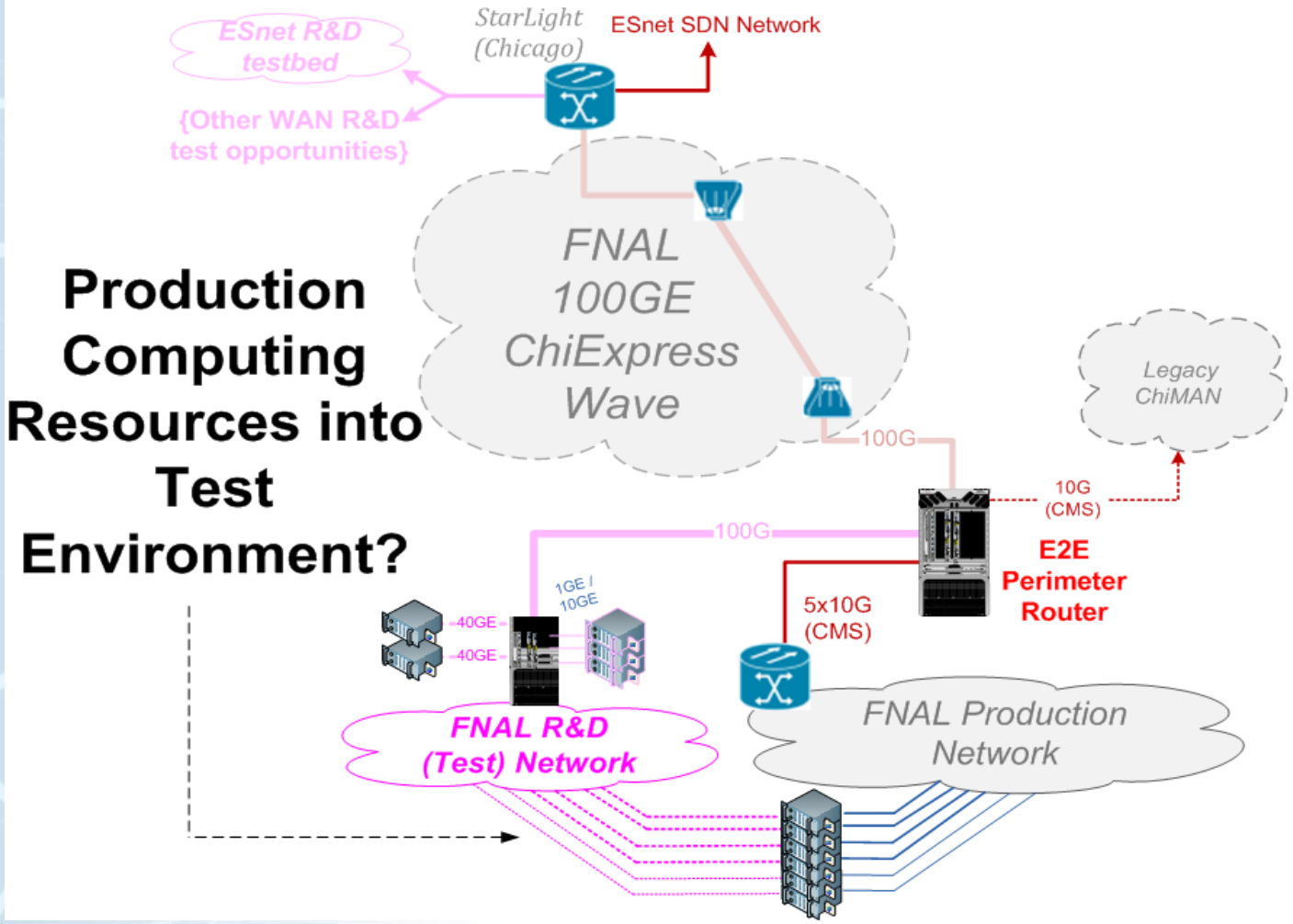
FNAL Attachment to ESnet 100GE Test Bed

Sharing 100GE MAN wave with CMS production traffic

- Currently expect to guarantee 50Gb/s for CMS
- With LHC shutdown until 2015, may provide more for R&D
- Also may want to use 100GE MAN wave for other collaborations
  - Outside of ESnet test bed…



Production Resources in Test Environment



Production Computing Resources into Test Environment?

FNAL “Federated” Test Bed Model Perspective

Desire direct participation in ESnet 100GE test bed

- Without exclusively dedicating resources to it

Security constraints:

- Need well-defined security requirements for us to participate
- Commitment to similar requirements by other participating sites

Addressing issues:

- RFC1918 addresses don’t fit well for our purposes

Acceptable Use Issues:

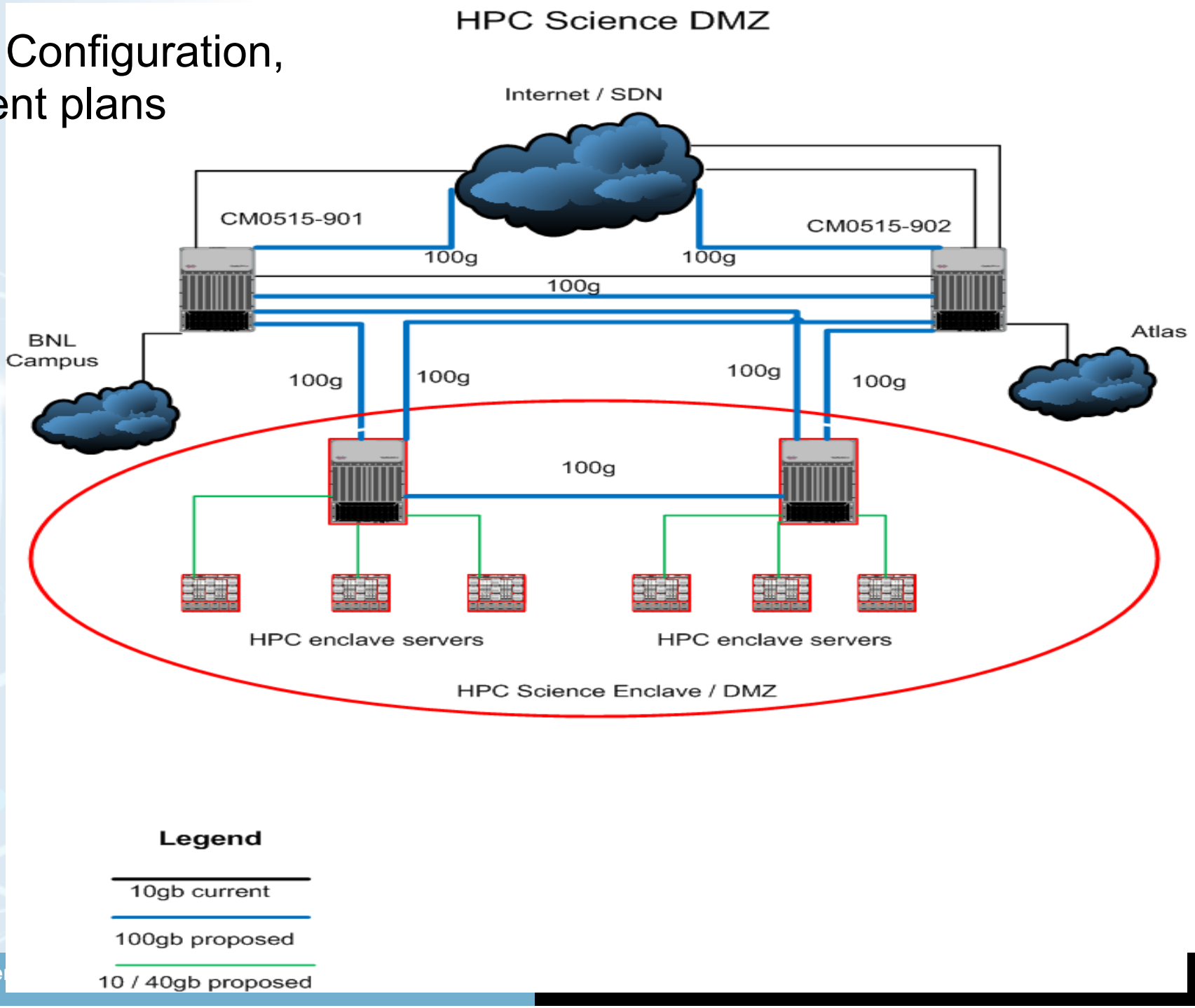
- Is site-to-site through test bed acceptable?

Well-defined technology requirements on us to participate in test bed

- Also what monitoring (PerfSONAR?) capabilities are required?

Local resource availability, reservation, & access under our control

BNL Configuration, current plans



Other Issues for discussion

Address Allocation

- How to manage 1918 address space?

Scheduling of site resources

- Depends on resource type
  - could use existing Testbed reservation calendaring system for 'reservable' components

Firewalls/IDS

- Layer 2 circuit to testbed would need to bypass security devices
- Same as the Science DMZ model

Monitoring

- Is it important for site to expose some level of monitoring data?
  - Perhaps ESnet's current SNMP data (graphite.es.net) is enough?

Need for multi-point circuits

- Adding this ability to OSCARS might simplify some of this

Next Steps

Sites should let me know what resources they have and might want to make available to other labs for test endpoints

ESnet will create a simple form for requesting circuits for this purpose.

More Information

http://www.es.net/testbed/

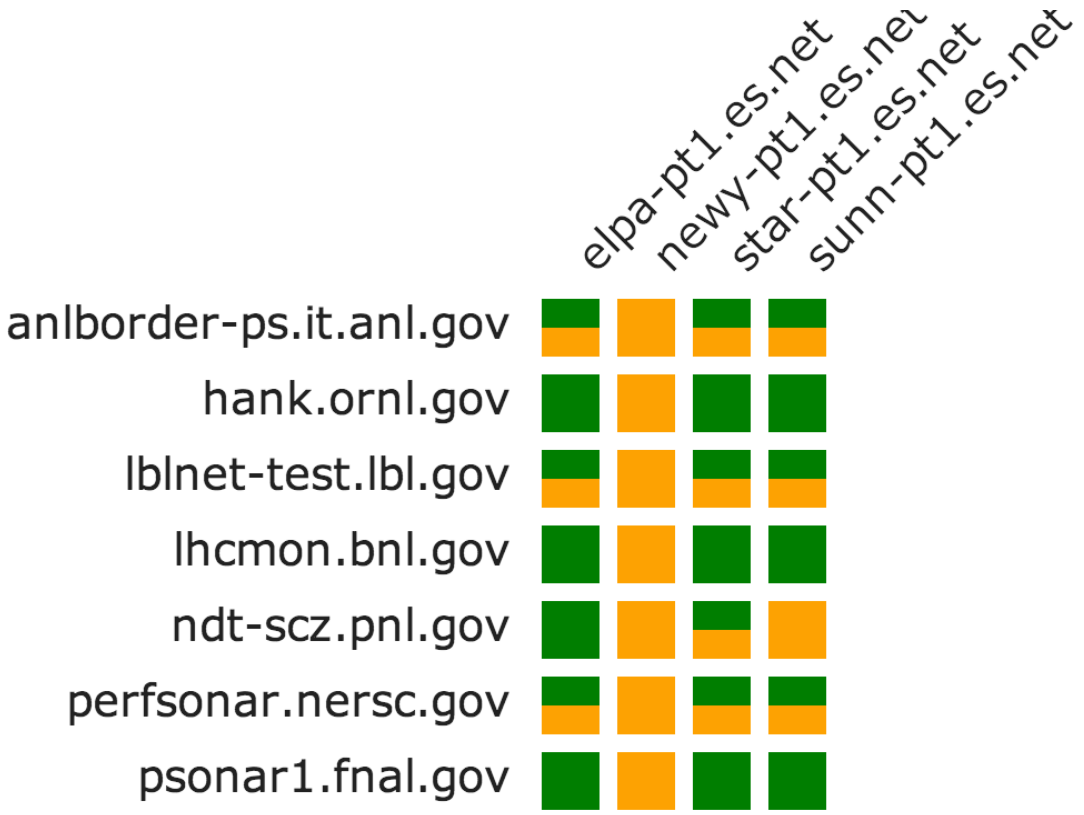
email: testbed-proposal@es.net, BLTierney@es.net

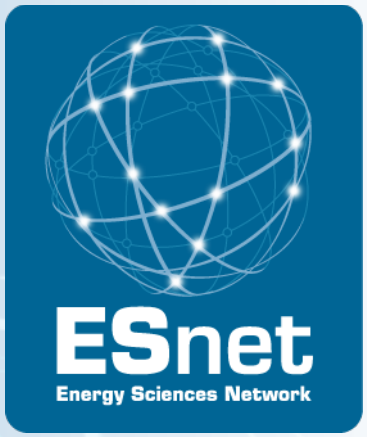
If time, demo this:

http://ps-dashboard.es.net/index.cgi?dashboard=3%3A%20ESnet%20to%20DOE%20Sites

ESnet to DOE Site Throughput Testing

- Throughput >= 500Mbps
- Throughput < 500Mbps
- Throughput < 100Mbps
- Unable to retrieve data
- Check has not yet run





Extra Slides

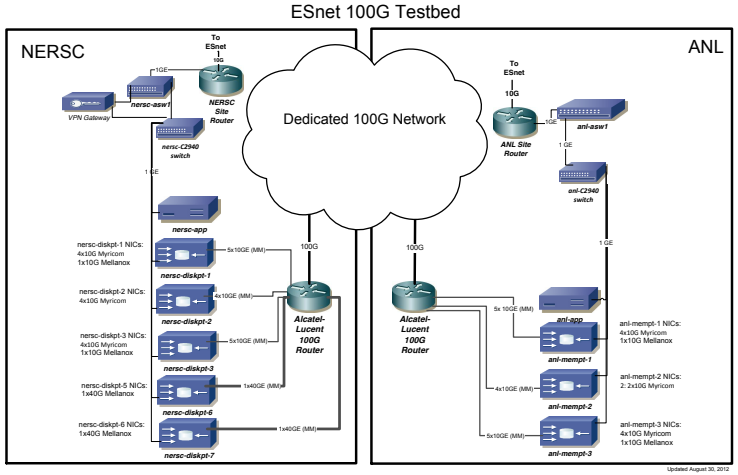
ESnet Research Testbeds

100G Testbed

- High-speed protocol research
- Routed network from Argonne to NERSC (dedicated 100G)
- See recent HPCwire

Dark Fiber Testbed

- Continental-scale fiber footprint for disruptive research



Testbed Access

Proposal process to gain access described at:

http://www.es.net/RandD/100g-testbed/proposal-process/

Testbed is available to anyone:

- DOE researchers
- Other government agencies
- Industry

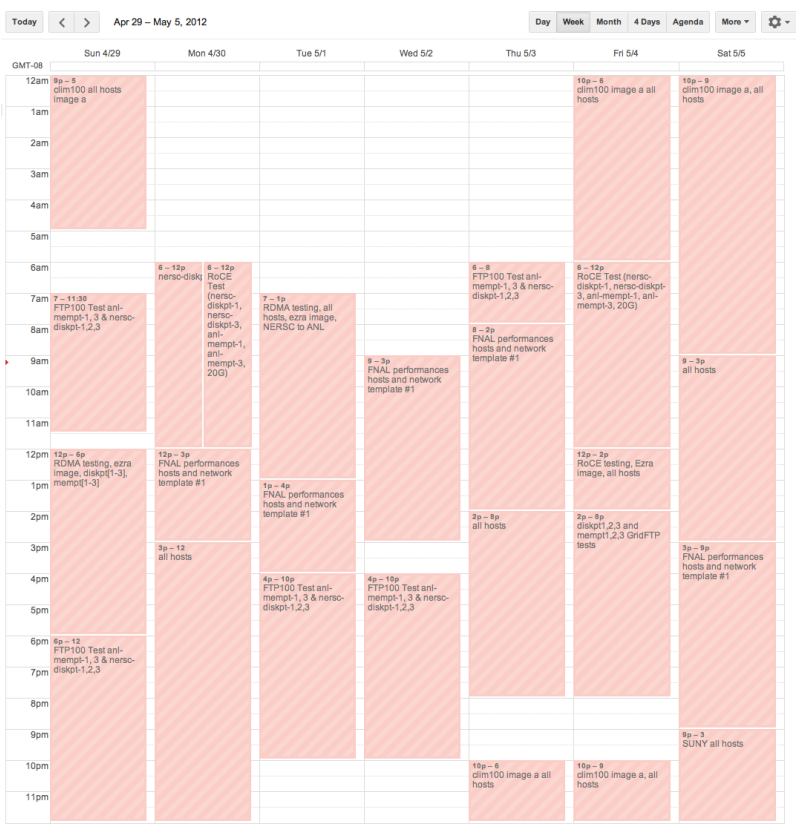
Must submit a short proposal to the testbed review committee

- Committee is made up of members from the R&E community and industry

Goal is to accept roughly five proposals every 6 month review cycle

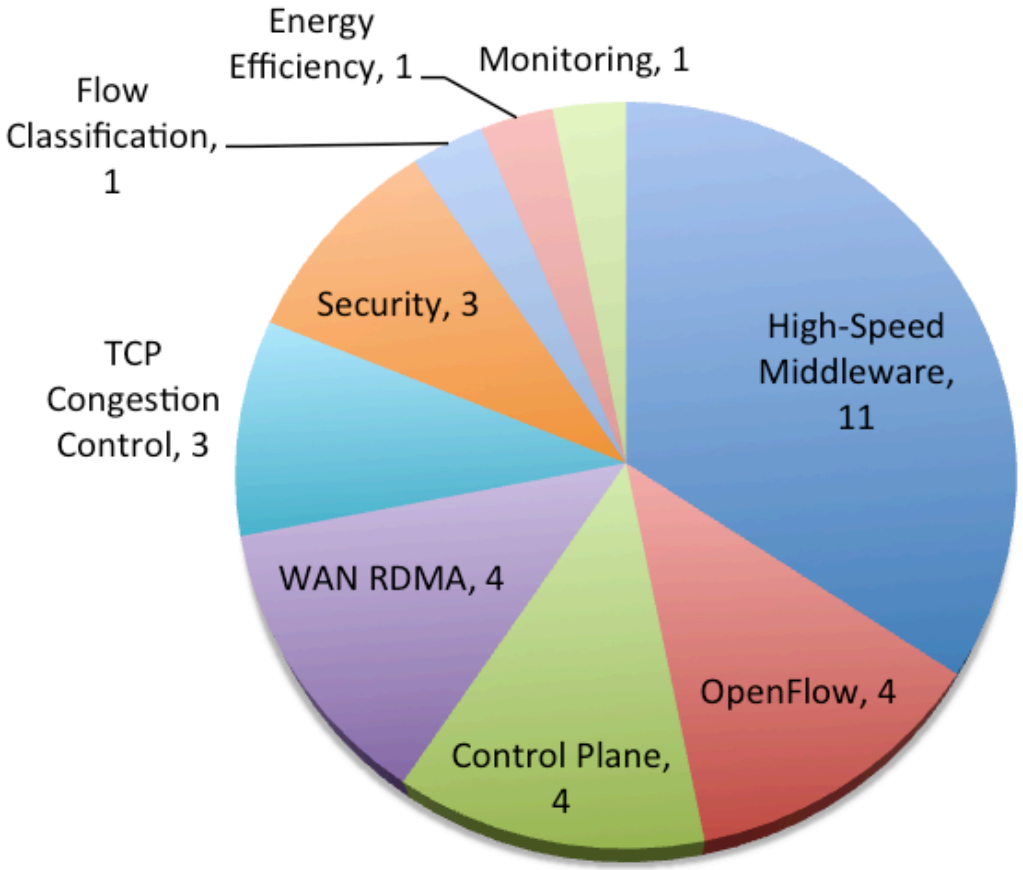
- Next round of proposals is due April 1, 2013

100G Testbed: Significant Demand

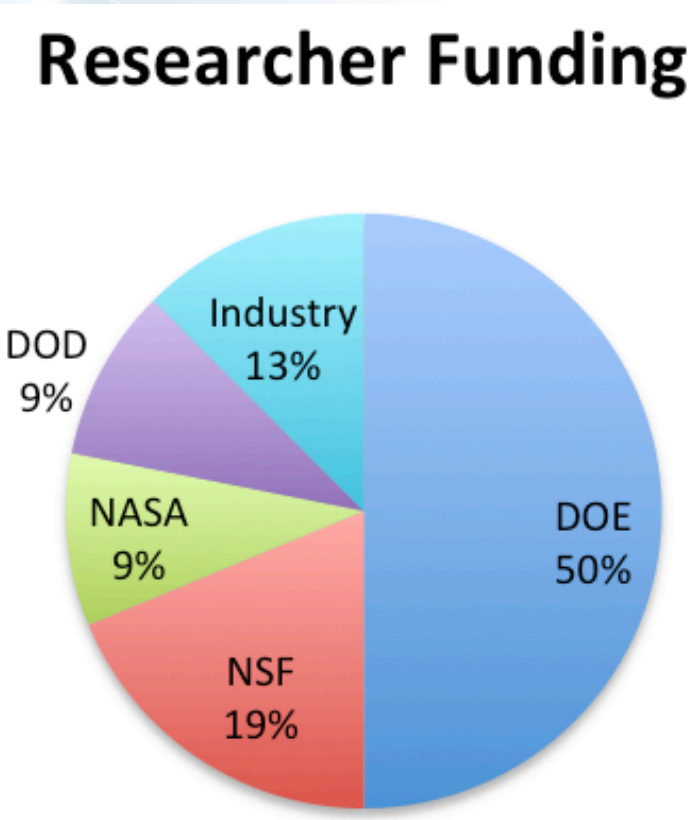


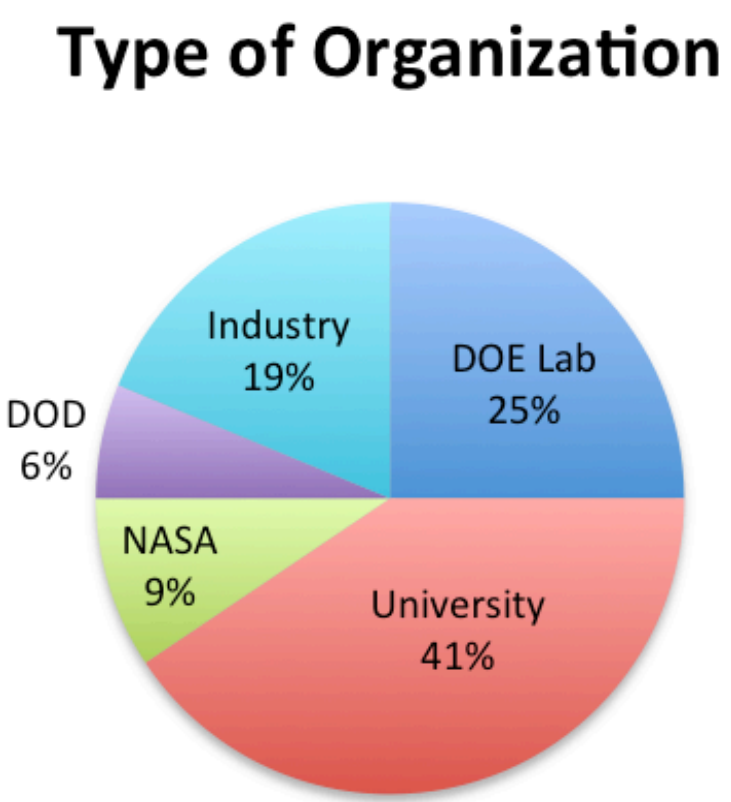
Community impact

- 32 projects accepted
- Strong diversity in research topics
- Research on the testbed has generated
  - eight peer-reviewed publications
  - one poster
  - three news releases
  - several papers under submission

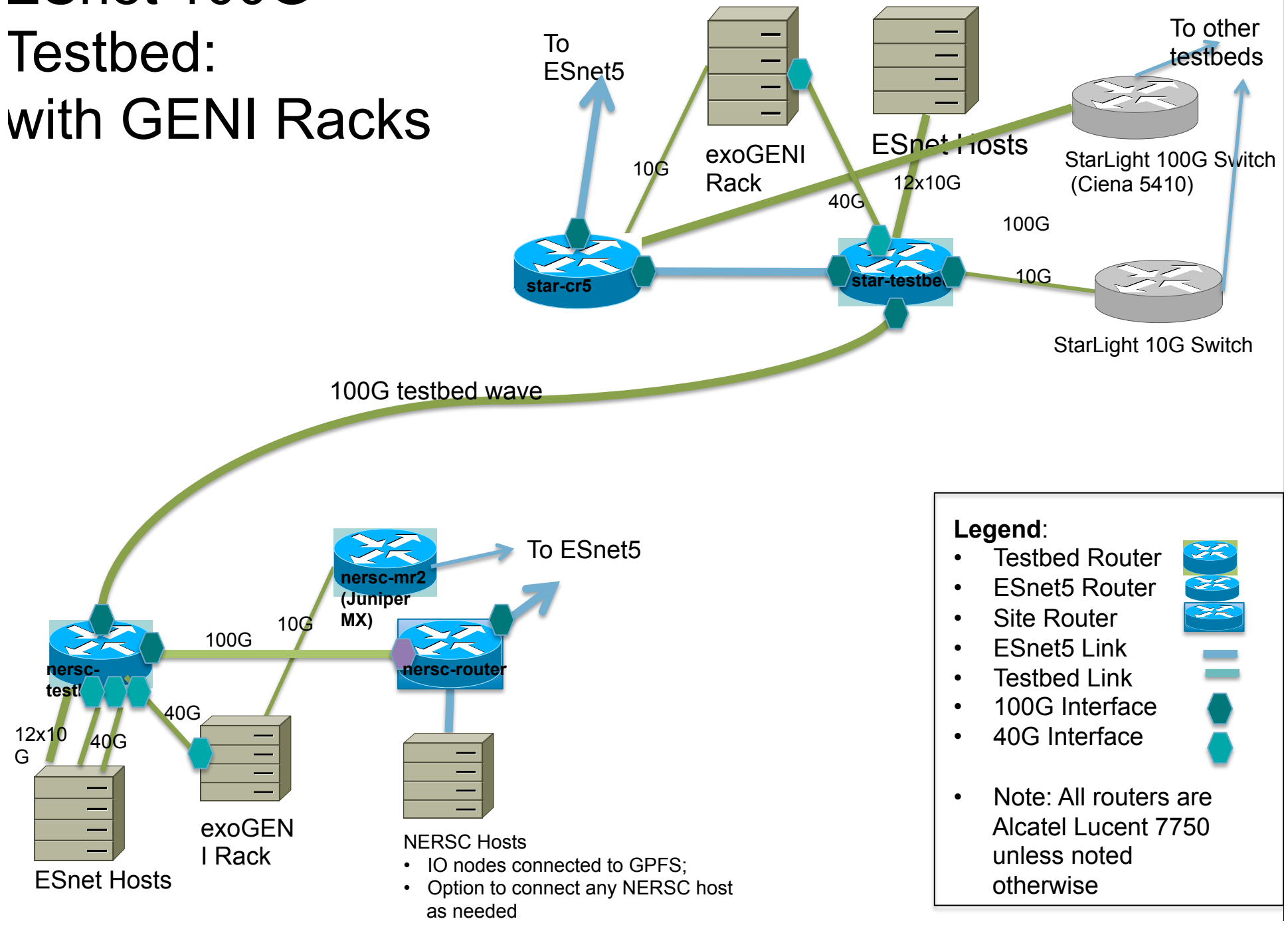


Accepted Testbed Projects





ESnet 100G Testbed: with GENI Racks



Industry Use of the Testbed

- Alcatel-Lucent used the testbed in May 2012 to verify the performance of its new 7950 XRS core router.
- Bay Microsystems used the testbed to verify that its 40 Gbps IBEx InfiniBand extension platform worked well over very long distances.
- Infinera used the testbed to demonstrate the telecommunication industry's first successful use of a prototype software-defined networking (SDN) open transport switch (OTS).
- Acadia Optronics used the testbed to test ITS 40 Gbps and 100 Gbps host NICs, and to debug the Linux device driver for its hardware.
- Orange Silicon Valley is using the testbed to test a 100G SSD-based video server
- Reservoir Labs is using the testbed to test their 100G IDS product under development

Results Since January 2012: 8 Publications Accepted

Yufei Ren, Tan Li, Dantong Yu, Shudong Jin, et. al, *Protocols for Wide-Area Data-intensive Applications: Design and Performance Issues*, Proceedings of IEEE Supercomputing 2012, November 12, 2012.

Zhengyang Liu, Malathi Veeraraghavan, Zhenzhen Yan, Chris Tracy, et.al, *On Using Virtual Circuits for GridFTP Transfers*, Proceedings of IEEE Supercomputing 2012, November 12, 2012.

Brian Tierney, Ezra Kissel, Martin Swany, and Eric Pouyoul, *Efficient Data Transfer Protocols for Big Data*, Proceedings of the 8th International Conference on eScience, IEEE, October 9, 2012.

Zhenzhen Yan, Chris Tracy, and Malathi Veeraraghavan, *A Hybrid Network Traffic Engineering System*, Proceedings of the IEEE 13th Conference on High Performance Switching and Routing (HPSR). June 2012, Belgrade, Serbia.

Mehmet Balman, Eric Pouyoul, Yushu Yao, E. Wes Bethel, Burlen Loring, Prabhat, John Shalf, Alex Sim, and Brian L. Tierney, *Experiences with 100Gbps Network Applications*, The Fifth International Workshop on Data Intensive Distributed Computing (DIDC 2012), June 2012.

Yufei Ren, Tan Li, Dantong Yu, Shudong Jin, and Thomas Robertazzi, *Middleware Support for RDMA-based Data Transfer,* Cloud Computing High-Performance Grid and Cloud Computing Workshop, May 2012.

G. Garzolglio, D. Dykstra, P. Mhashilkar, and H. Kim, *Identifying Gaps in Grid Middleware on Fast Networks with the Advanced Network Initiative*, International Conference on Computing in High Energy and Nuclear Physics (CHEP 2012), May 2012.

H. Pi, I. Sfiligoi, F. Wüerthwein, and D. Bradley, *Data Transfer Test with 100 Gpbs Network for Open Science Grid (LHC) Application*, International Conference on Computing in High Energy and Nuclear Physics (CHEP 2012), May 2012.

See: http://www.es.net/RandD/100g-testbed/results/

Our (Separate) Dark Fiber Testbed

Purpose of Dark Fiber Testbed

Ultimately, this is our ‘roll pair’. In the medium term, a testbed for innovative research and development projects.

For validation of new network architectures & technologies.

- imagine a future architecture, and try it out
- potentially disruptive, incompatible with existing optical systems
- requiring dedicated fiber

Possible examples:

- dynamic optical switching, packet-optical architectures
- higher-speed networking (superchannels, >100G waves)
- clean-slate energy designs

Dark Fiber Projects in the Pipeline (all awaiting funding)

quantum key encryption

- proposal to DARPA
- JPL/Boeing

reduce network energy > 90%

- ARPA-E project proposal
- Bell Labs, HP Labs, Texas Instruments, UCSD, UoA, Columbia

NIST precision time keeping

packet-optical integration and >100Gbps

- mainly vendors
- still in discussions