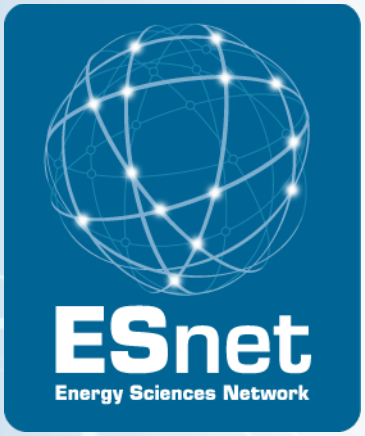
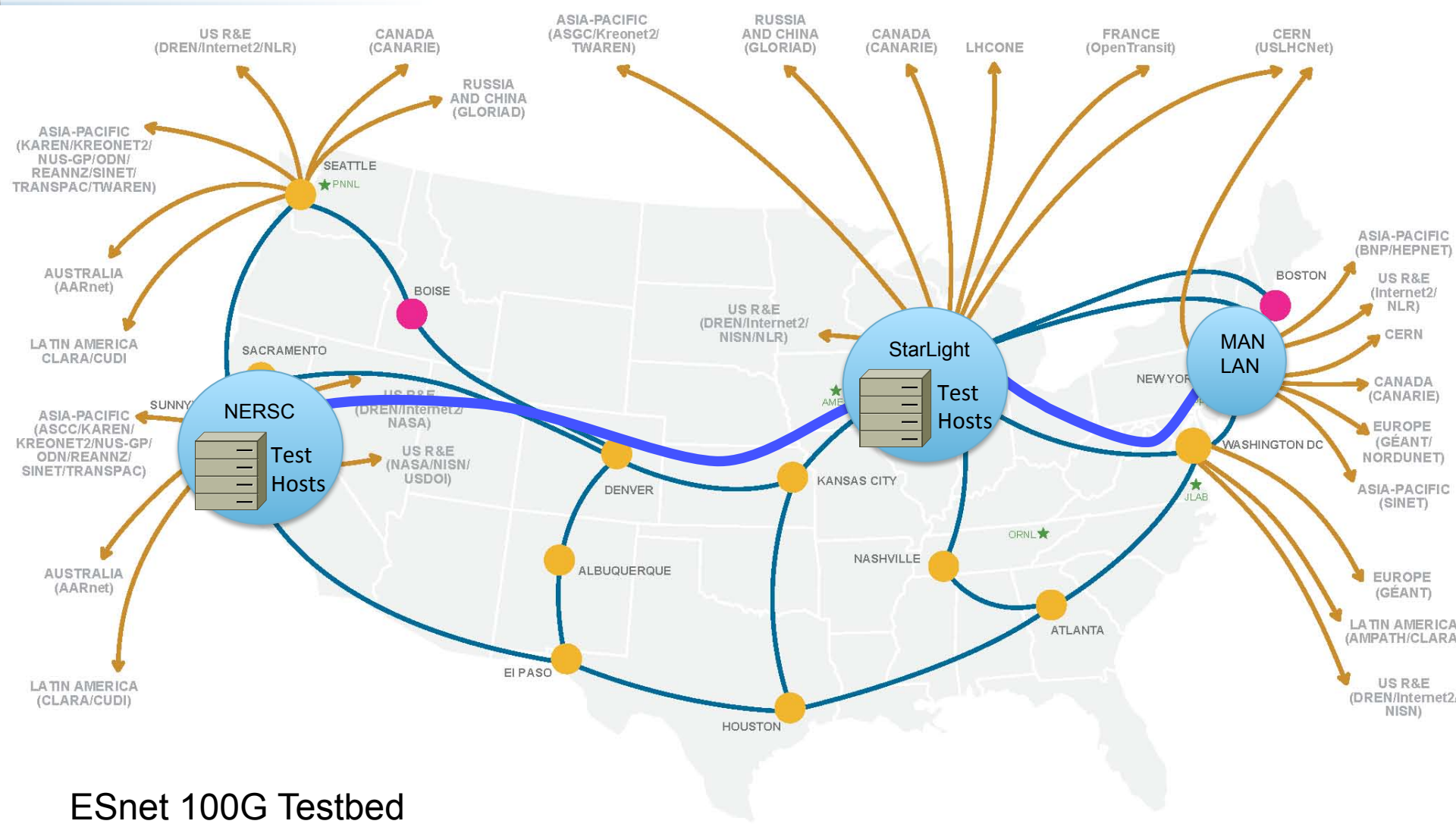


ESnet's 100G Network Testbed

Brian Tierney, Eric Pouyoul

Berkeley National Lab / ESnet

November 17, 2013



ESnet 100G Testbed

ESnet
Energy Sciences Network

100G IP Hubs

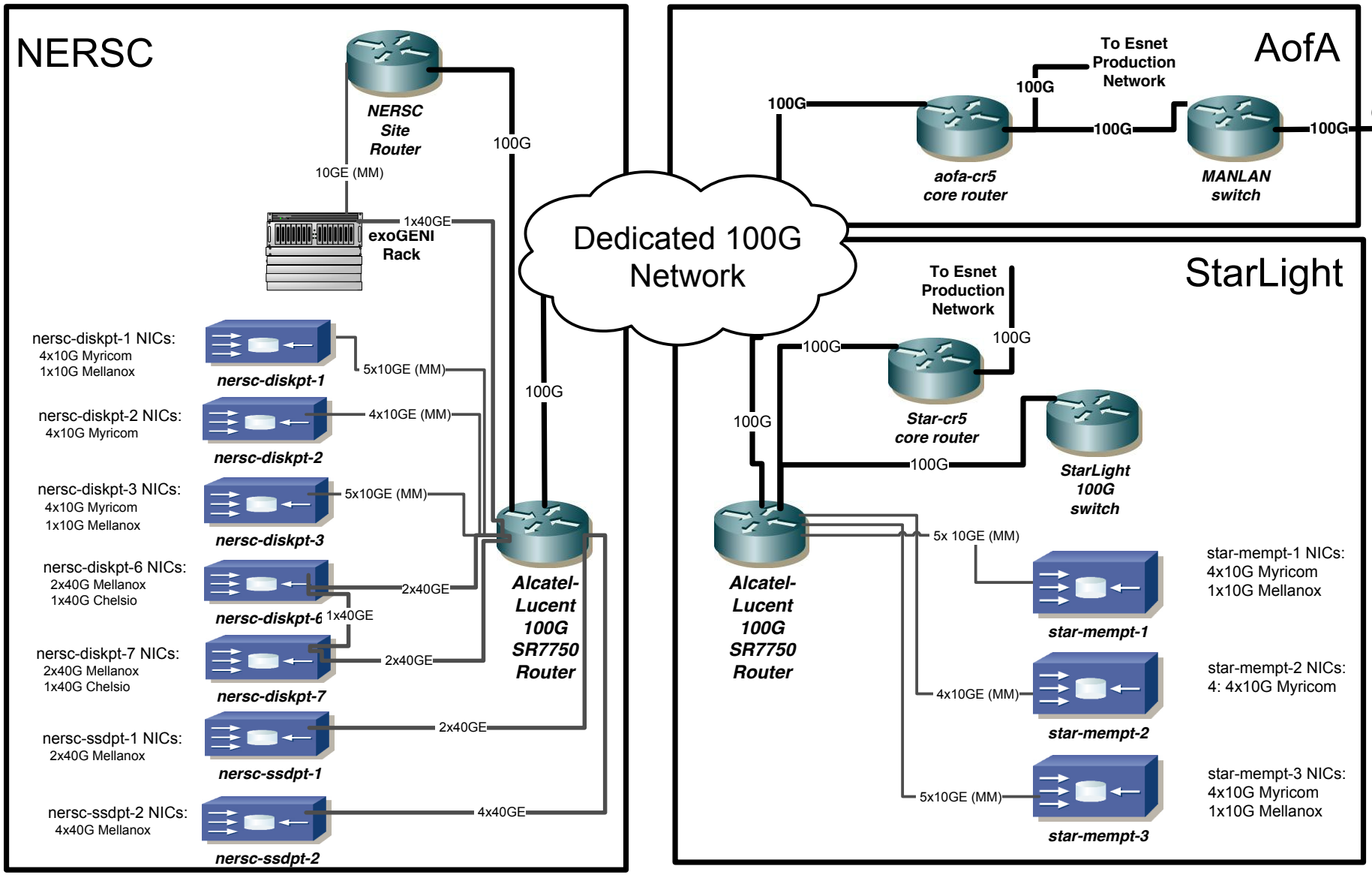
4x10G IP Hub

Major R&E and International peering connections

★ Office of Science National Labs

Ames	Ames Laboratory (Ames, IA)
ANL	Argonne National Laboratory (Argonne, IL)
BNL	Brookhaven National Laboratory (Upton, NY)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
PNNL	Pacific Northwest National Laboratory (Richland, WA)
PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
SLAC	Stanford Linear Accelerator Center (Menlo Park, CA)

ESnet 100G Testbed



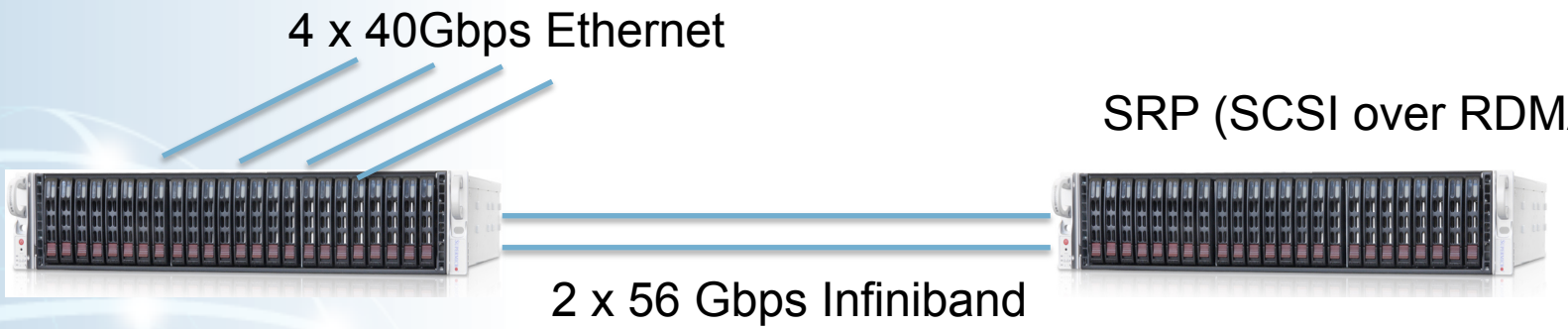
100G Testbed Capabilities

This testbed is designed to support research in high-performance data transfer protocols and tools.

Capabilities:

- "bare metal" access to very high performance hosts
  - Up to 100Gbps memory to memory, and 70 Gbps disk to disk
- each project gets their own disk image, which root access
  - Can experiment with custom kernels, custom network protocols , etc.

New “SSD” test host



- 2 x Sandy Bridge 2.9 Ghz (2 x 6 cores)
- 128 GB RAM
- 2 x Dual Port 40G Ethernet (4 x 40G)
- 1 x Dual Port Infiniband HCA
- 24 x SSD (250GB)
- 2 x HDD system drives
- CentOS 6.4

- 2 x Sandy Bridge 2.9 Ghz
- 64 GB RAM
- 1 x Dual Port Infiniband HCA
- 24 x HDD (250GB)
- 2 x HDD system drives
- ESOS (SRP-Target OS)

Testbed Access

Proposal process to gain access described at:

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

Testbed is available to anyone:

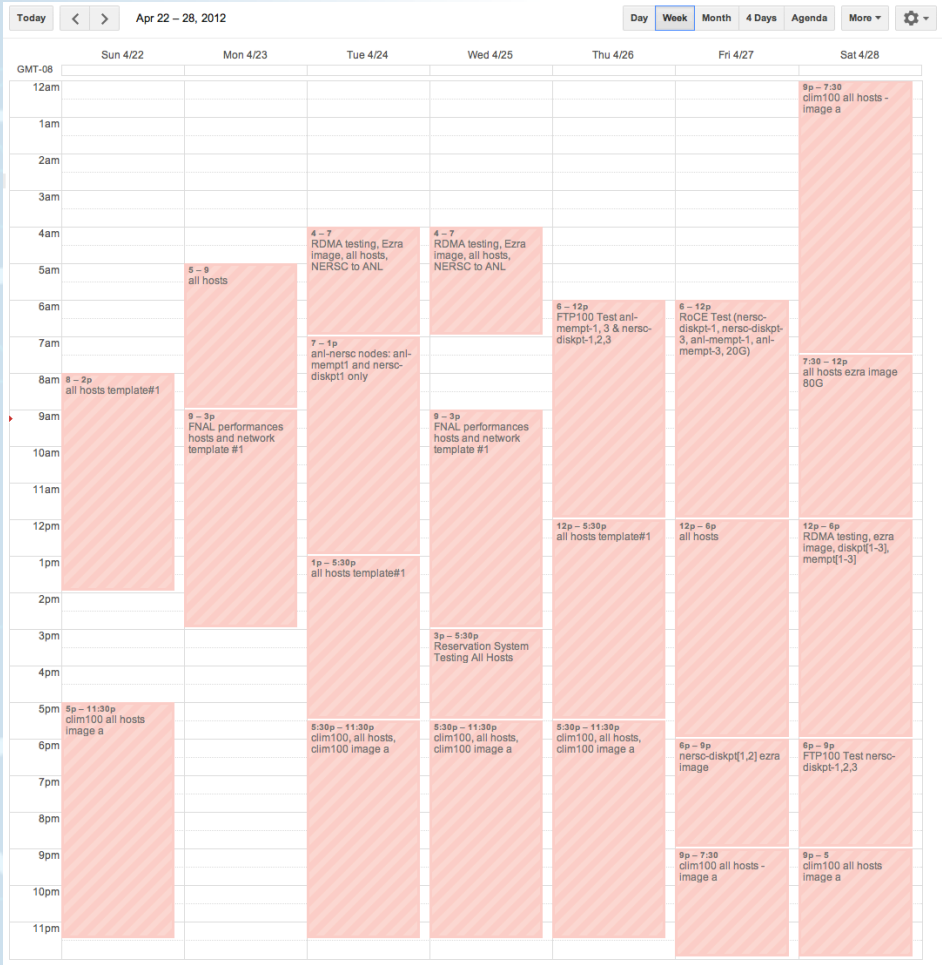
- DOE researchers
- Other government agencies
- Industry

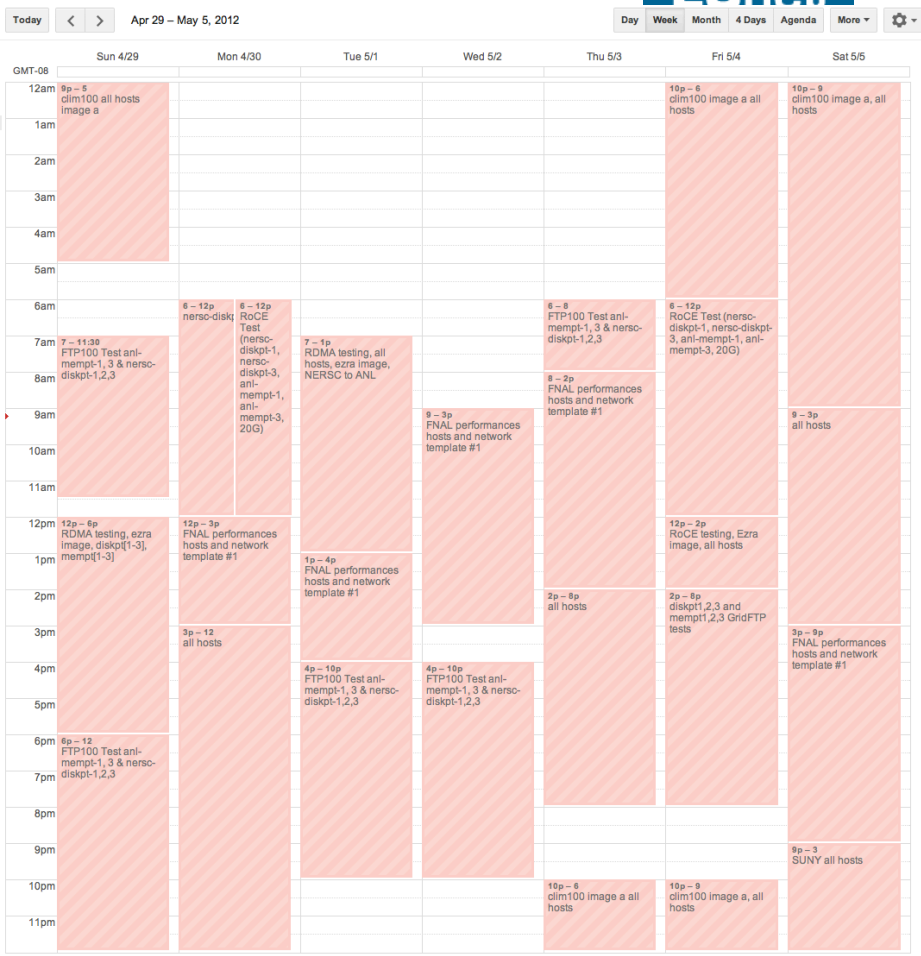
Must submit a short proposal to ESnet (2 pages)

Review Criteria:

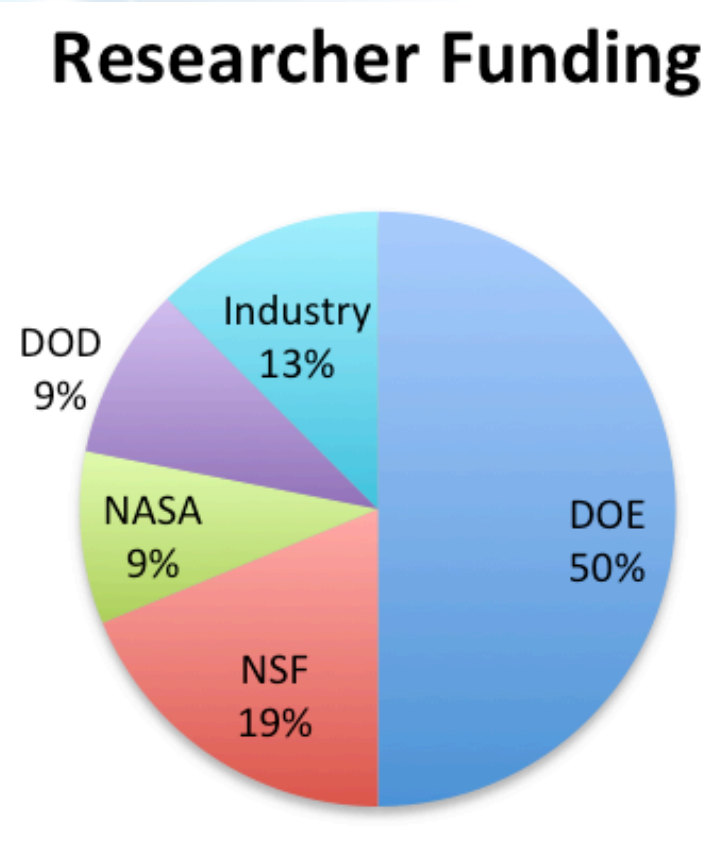
- Project “readiness”
- Could the experiment easily be done elsewhere?

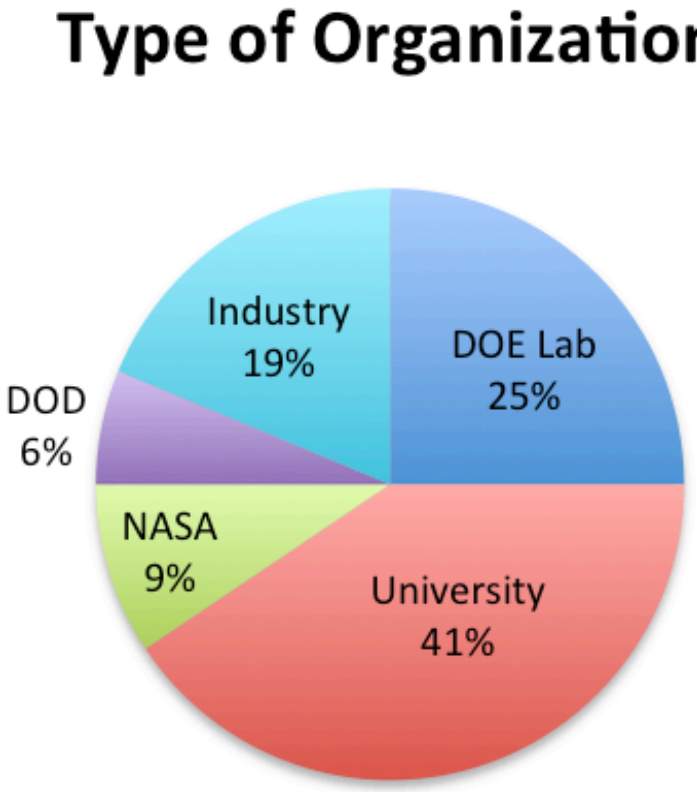
100G Testbed: Significant Demand





Accepted Testbed Projects





Publications based on Testbed Results

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

100G Testbed became available in January, 2012.

The testbed has already provided results for 20 accepted papers!

- 2012: 8 publications
- 2013: 11 publications
- 2014: 1 already

Specific Conferences:

- SC12: 1 paper
- SC13: 1 paper
- NDM 2013: 4 Papers

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

“Federated” Testbed

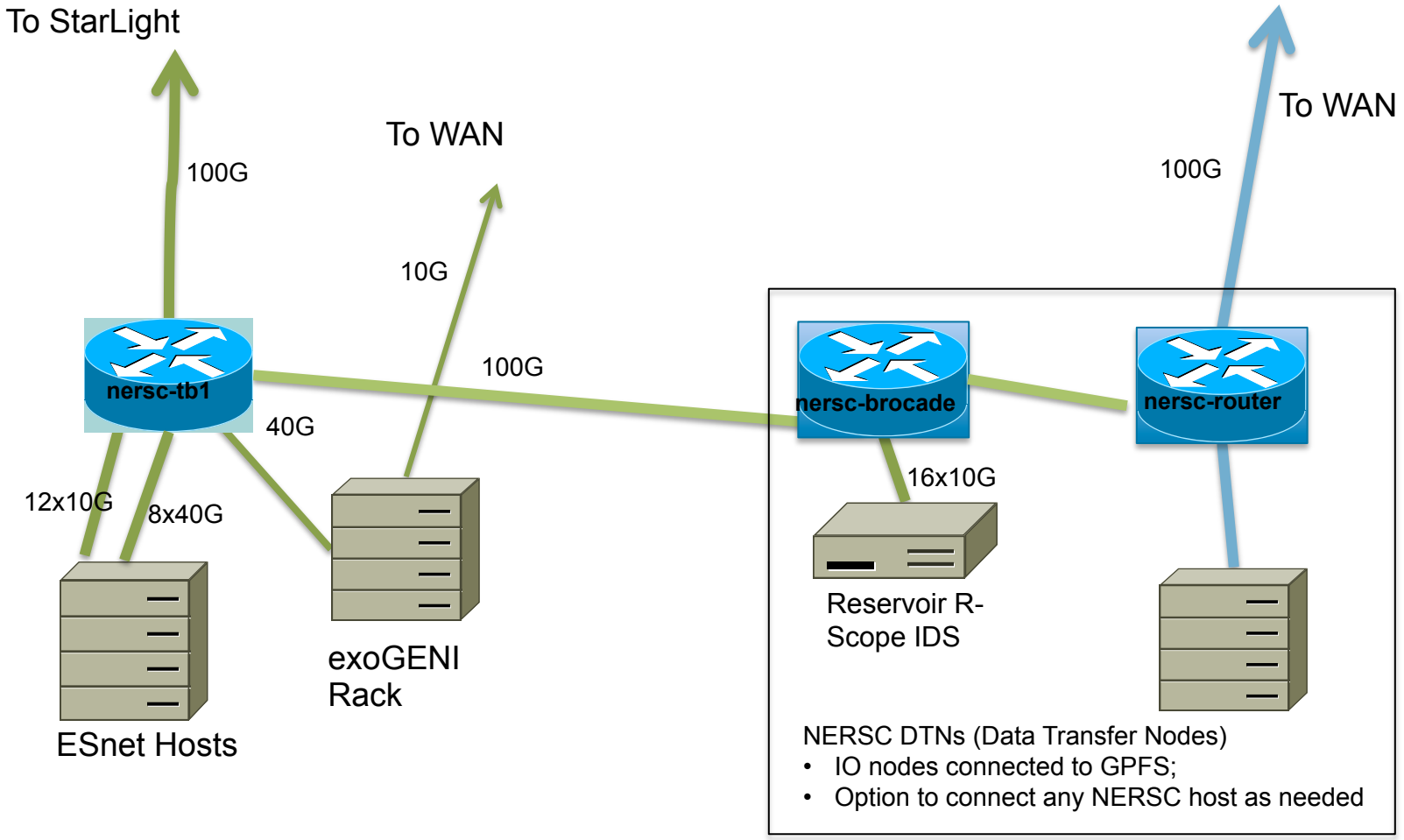
Using Layer-2 circuits, external hosts can be connected to the ESnet testbed

- ESnet has 50-80G of spare capacity on much of it’s footprint at this time available for testing, as does Internet2

So far we have connected the following resources to the ESnet testbed for testing

- FNAL: 2x40G hosts
- BNL: 3x40G hosts
- NERSC: 100G connection to NERSC Security router (see next slide)
- University of Chicago: 4x10G to “Kenwood” and “Goldberg” clusters
- NASA Goddard: 1 host with 4x40G
- Navel Research Lab: 2x10G hosts

ESnet 100G Testbed: NERSC Connections for 100G IDS testing



ExoGENI Rack (https://wiki.exogeni.net/)

ESnet

ExoGENI Testbed

geni Exploring Networks of the Future

XO

- 14 GPO-funded racks
  - Partnership between RENCI, Duke and IBM
  - IBM x3650 M4 servers (X-series 2U)
    - 1x146GB 10K SAS hard drive +1x500GB secondary drive
    - 48G RAM 1333Mhz
    - Dual-socket 8-core CPU
    - Dual 1Gbps adapter (management network)
    - 10G dual-port Chelseo adapter (dataplane)
  - BNT 8264 10G/40G OpenFlow switch
  - DS3512 6TB sliverable storage
    - iSCSI interface for head node image storage as well as experimenter slivering
- Each rack is a small networked cloud
  - OpenStack-based
  - EC2 node sizes (m1.small, m1.large etc)
- http://www.exogeni.net

DUKE COMPUTER SCIENCE

IBM

Lessons Learned

Tuning for 40G is not just 4x Tuning for 10G

Some of the conventional wisdom for 10G Networking is not true at 40Gbps

e.g.: Parallel streams more likely to hurt than help

UDP needs to be tuned even more than TCP

“Sandy Bridge” Architectures require extra tuning as well

Lots of details at http://fasterdata.es.net/science-dmz/DTN/tuning/

Sample results: TCP Single vs Parallel Streams


```
1 stream: iperf3 -c 192.168.102.9
[ ID] Interval           Transfer     Bandwidth       Retransmits
[  4]   0.00-1.00   sec  3.19 GBytes  27.4 Gbits/sec    0
[  4]   1.00-2.00   sec  3.35 GBytes  28.8 Gbits/sec    0
[  4]   2.00-3.00   sec  3.35 GBytes  28.8 Gbits/sec    0
[  4]   3.00-4.00   sec  3.35 GBytes  28.8 Gbits/sec    0
[  4]   4.00-5.00   sec  3.35 GBytes  28.8 Gbits/sec    0

2 streams: iperf3 -c 192.168.102.9 -P2
[ ID] Interval           Transfer     Bandwidth       Retransmits
[  4]   0.00-1.00   sec  1.37 GBytes  11.8 Gbits/sec    7
[  6]   0.00-1.00   sec  1.38 GBytes  11.8 Gbits/sec   11
[SUM]   0.00-1.00   sec  2.75 GBytes  23.6 Gbits/sec   18
……
- - - - - - - - - - - - - - - - - - - - - - - - -
[  4]   8.00-9.00   sec  1.43 GBytes  12.3 Gbits/sec    8
[  6]   8.00-9.00   sec  1.42 GBytes  12.2 Gbits/sec    7
[SUM]   8.00-9.00   sec  2.85 GBytes  24.5 Gbits/sec   15
- - - - - - - - - - - - - - - - - - - - - - - - -
[  4]   9.00-10.00  sec  1.43 GBytes  12.3 Gbits/sec    4
[  6]   9.00-10.00  sec  1.43 GBytes  12.3 Gbits/sec    6
[SUM]   9.00-10.00  sec  2.86 GBytes  24.6 Gbits/sec   10
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bandwidth       Retransmits
[  4]   0.00-10.00  sec  13.8 GBytes  11.9 Gbits/sec   78          sender
[  4]   0.00-10.00  sec  13.8 GBytes  11.9 Gbits/sec               receiver
[  6]   0.00-10.00  sec  13.8 GBytes  11.9 Gbits/sec   95          sender
[  6]   0.00-10.00  sec  13.8 GBytes  11.9 Gbits/sec               receiver
[SUM]   0.00-10.00  sec  27.6 GBytes  23.7 Gbits/sec  173          sender
[SUM]   0.00-10.00  sec  27.6 GBytes  23.7 Gbits/sec               receiver
```

iperf3: https://code.google.com/p/iperf/

Sample results: TCP On Intel “Sandy Bridge” Motherboards

30% Improvement using the right core!

```
nuttcp -i 192.168.2.32
 2435.5625 MB /   1.00 sec = 20429.9371 Mbps     0 retrans
 2445.1875 MB /   1.00 sec = 20511.4323 Mbps     0 retrans
 2443.8750 MB /   1.00 sec = 20501.2424 Mbps     0 retrans
 2447.4375 MB /   1.00 sec = 20531.1276 Mbps     0 retrans
 2449.1250 MB /   1.00 sec = 20544.7085 Mbps     0 retrans

nuttcp -i1 -xc 2/2 192.168.2.32
 3634.8750 MB /   1.00 sec = 30491.2671 Mbps     0 retrans
 3723.8125 MB /   1.00 sec = 31237.6346 Mbps     0 retrans
 3724.7500 MB /   1.00 sec = 31245.5301 Mbps     0 retrans
 3721.7500 MB /   1.00 sec = 31219.8335 Mbps     0 retrans
 3723.7500 MB /   1.00 sec = 31237.6413 Mbps     0 retrans
```

nuttcp: http://lcp.nrl.navy.mil/nuttcp/beta/nuttcp-7.2.1.c

Sample results: TCP On Intel “Sandy Bridge” Motherboards: Fast host to Slower Host


```
nuttcp -i1 192.168.2.31
  410.7500 MB /   1.00 sec = 3445.5139 Mbps     0 retrans
  339.5625 MB /   1.00 sec = 2848.4966 Mbps     0 retrans
  354.5625 MB /   1.00 sec = 2974.2888 Mbps   350 retrans
  326.3125 MB /   1.00 sec = 2737.3022 Mbps     0 retrans
  377.7500 MB /   1.00 sec = 3168.8220 Mbps   179 retrans

nuttcp -i1 192.168.2.31
 2091.0625 MB /   1.00 sec = 17540.8230 Mbps     0 retrans
 2106.7500 MB /   1.00 sec = 17672.0814 Mbps     0 retrans
 2103.6250 MB /   1.00 sec = 17647.0326 Mbps     0 retrans
 2086.7500 MB /   1.00 sec = 17504.7702 Mbps     0 retrans
```

http://fasterdata.es.net/host-tuning/interrupt-binding/

Sample results: UDP Tuning

Defaults:

```
nuttcp -i1 -u -R10G -T4 10.26.202.10
 1125.4844 MB /   1.00 sec = 9441.1434 Mbps     0 / 144062 ~drop/pkt  0.00 ~%loss
 1125.7031 MB /   1.00 sec = 9443.1295 Mbps     0 / 144090 ~drop/pkt  0.00 ~%loss
 1125.7031 MB /   1.00 sec = 9443.0634 Mbps     0 / 144090 ~drop/pkt  0.00 ~%loss
 1125.5000 MB /   1.00 sec = 9441.3689 Mbps     0 / 144064 ~drop/pkt  0.00 ~%loss
```

Bigger Packets:

```
nuttcp -i1 -u -R10G -T4 -l8972 10.26.202.10
 1135.5752 MB /   1.00 sec = 9525.7906 Mbps     0 / 132717 ~drop/pkt  0.00 ~%loss
 1134.8051 MB /   1.00 sec = 9519.4546 Mbps     0 / 132627 ~drop/pkt  0.00 ~%loss
 1133.8297 MB /   1.00 sec = 9511.2531 Mbps     0 / 132513 ~drop/pkt  0.00 ~%loss
 1133.6672 MB /   1.00 sec = 9509.8989 Mbps     0 / 132494 ~drop/pkt  0.00 ~%loss
```

Bigger window:

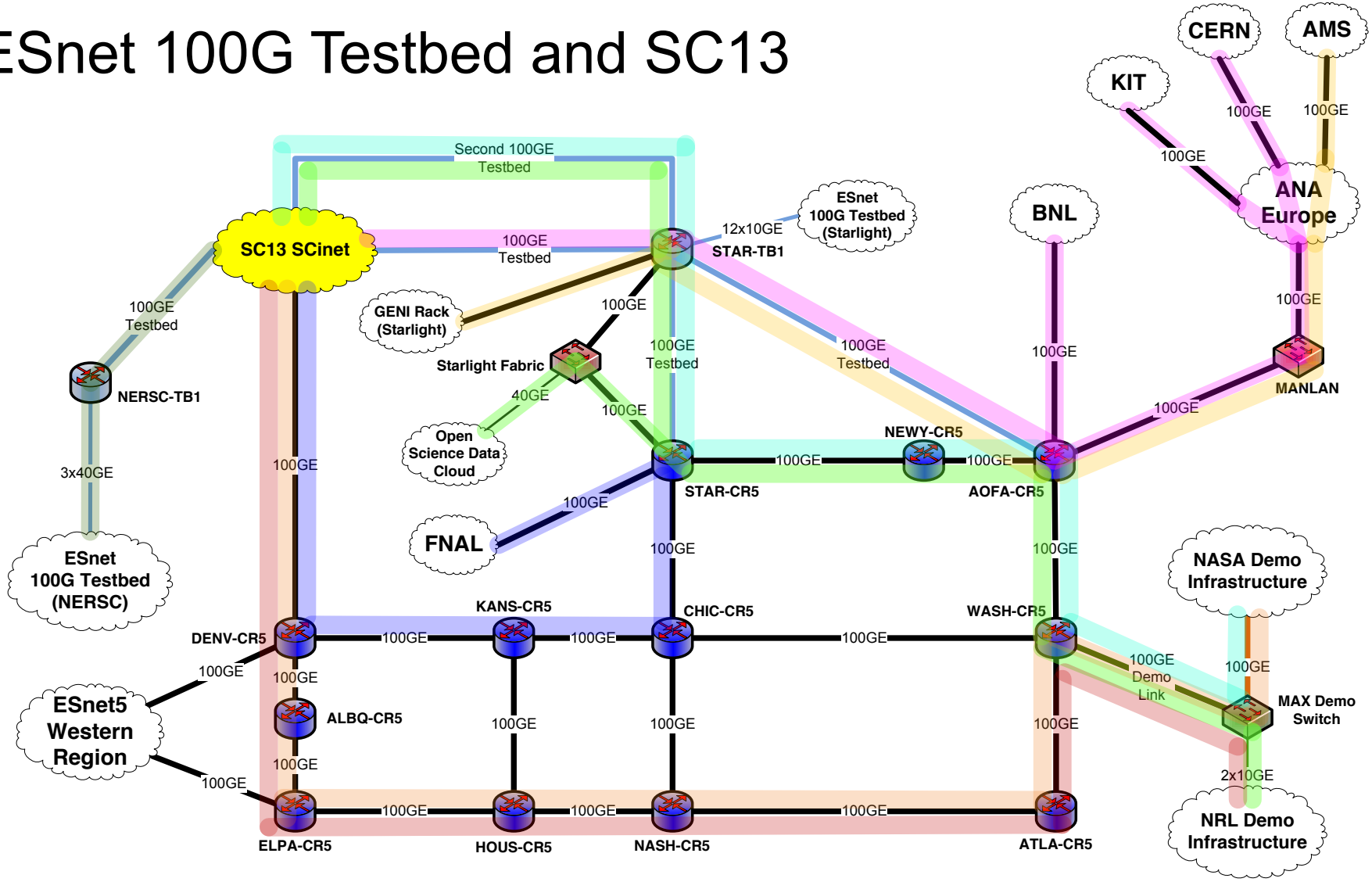
```
nuttcp -i1 -u -R10G -T4 -l8972 -w4m 10.26.202.10
 1182.1475 MB /   1.00 sec = 9916.4432 Mbps     0 / 138160 ~drop/pkt  0.00 ~%loss
 1181.6513 MB /   1.00 sec = 9912.4488 Mbps     0 / 138102 ~drop/pkt  0.00 ~%loss
 1181.6513 MB /   1.00 sec = 9912.3893 Mbps     0 / 138102 ~drop/pkt  0.00 ~%loss
 1181.6855 MB /   1.00 sec = 9912.7260 Mbps     0 / 138106 ~drop/pkt  0.00 ~%loss
```

nuttcp: http://lcp.nrl.navy.mil/nuttcp/beta/nuttcp-7.2.1.c

Single flow 40G Results

Tool	Protocol	Gbps	Send CPU	Recv CPU
netperf	TCP TCP-sendfile 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%

ESnet 100G Testbed and SC13



- NRL demo Northern path (20G)
- NRL demo Southern path (20G)
- NASA demo – production path (50G)
- NASA demo – testbed path (50G)
- OpenFlow/SDN demo – ANA path (100G)
- Caltech demo – ANA path (100G)
- Caltech demo – FNAL path (60G)
- Caltech demo – NERSC TB path (100G)

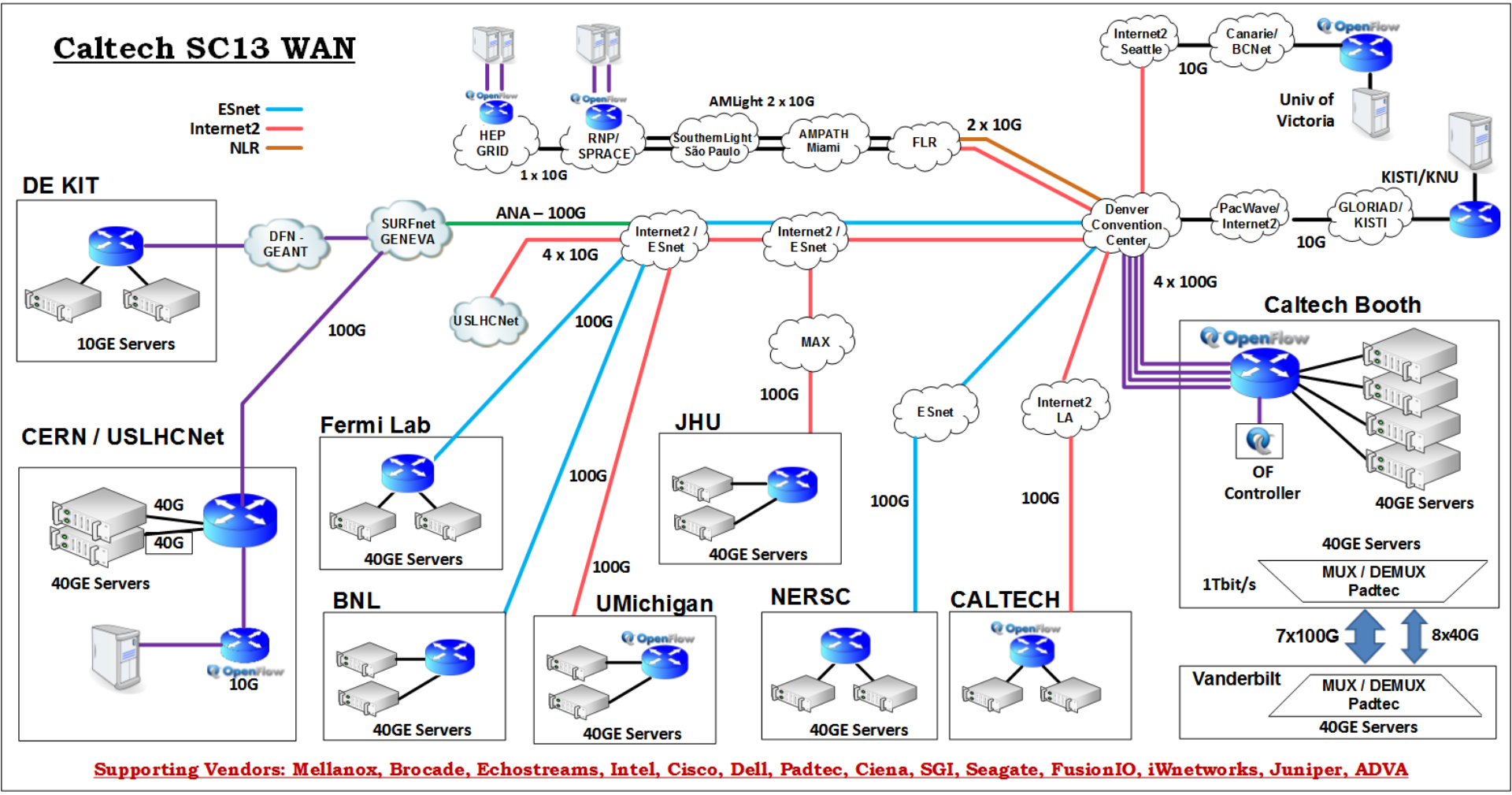
ESnet
Energy Sciences Network

SC13 demos – ESnet5 map

Eli Dart, ESnet 11/14/2013

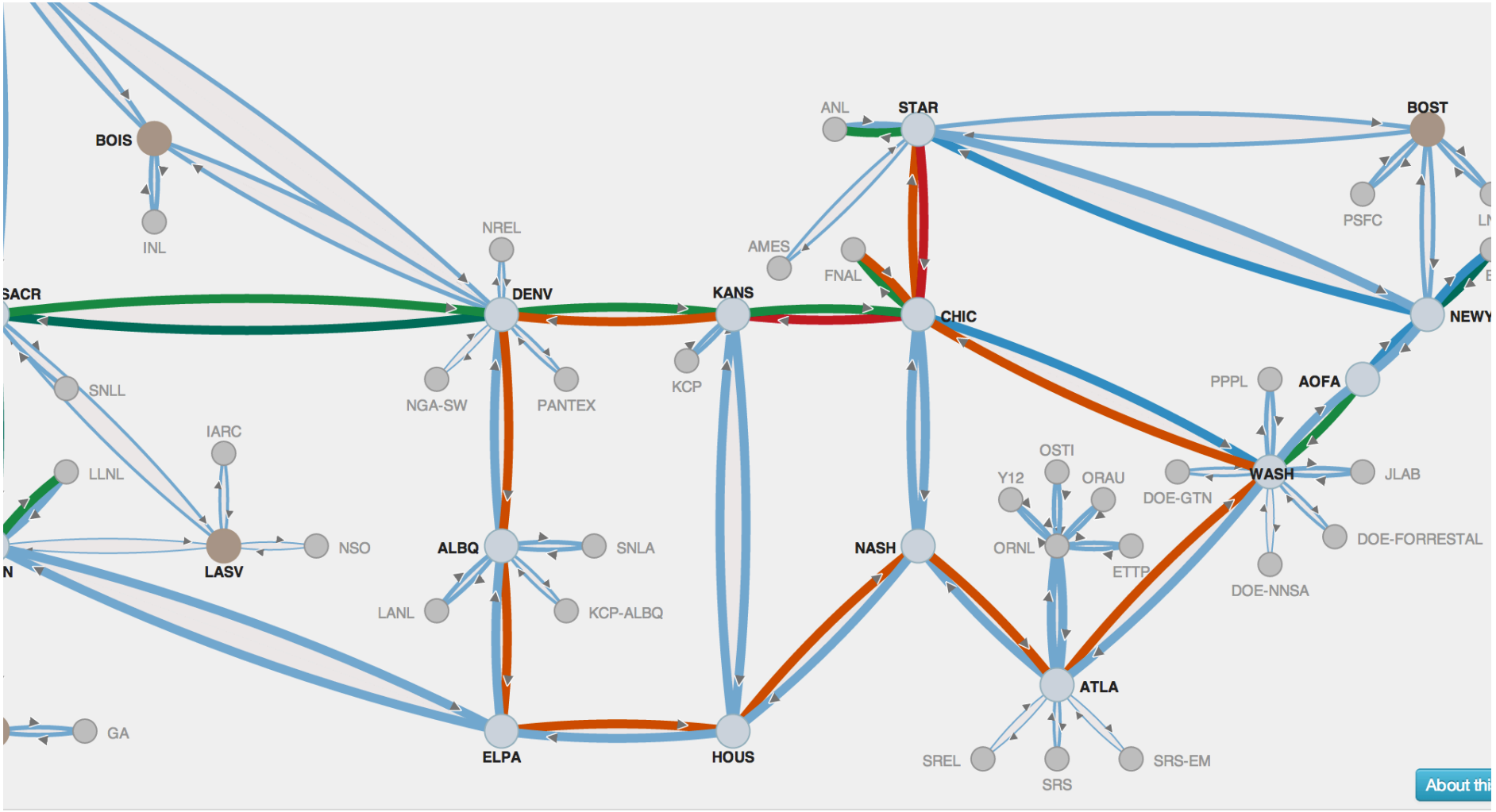
FILENAME SC13-DEMOS-V24.VSD

ESnet 100G Testbed and SC13: 400Gbps to the Caltech Booth



Supporting Vendors: Mellanox, Brocade, Echostreams, Intel, Cisco, Dell, Padtec, Ciena, SGI, Seagate, FusionIO, iWnetworks, Juniper, ADVA

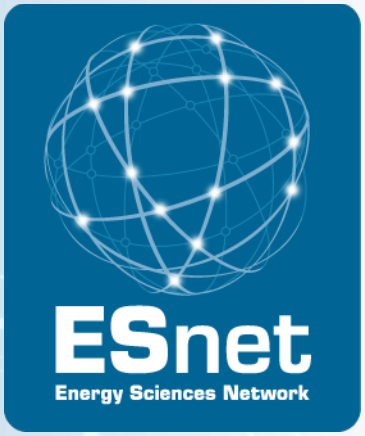
Loop Test From NASA last week: my.es.net



More Information

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

email: BLTierney@es.net



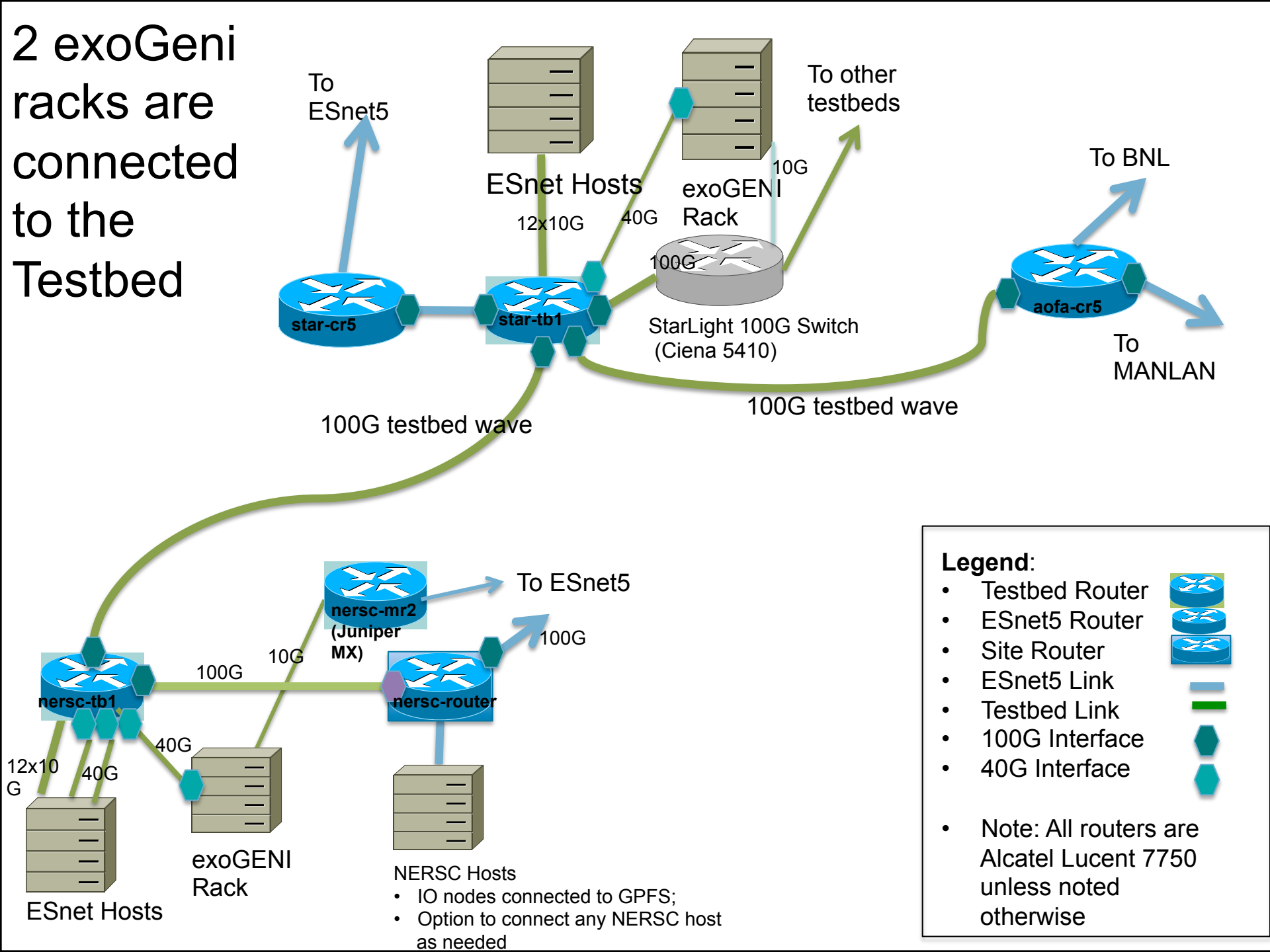
Extra Slides

New SSD Host Performance

Test	incoming	outgoing	Both at once
Network only (memory to memory test, nuttcp)	75 Gbps	75 Gbps	N Gbps
Disk to Network test (GridFTP)	14 Gbps	75 Gbps	N Gbps

Note: This is using 2 40G interfaces, connecting to 2 hosts with 1 40G interface

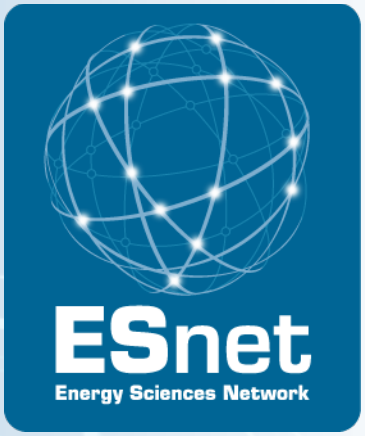
2 exoGeni racks are connected to the Testbed



Legend:

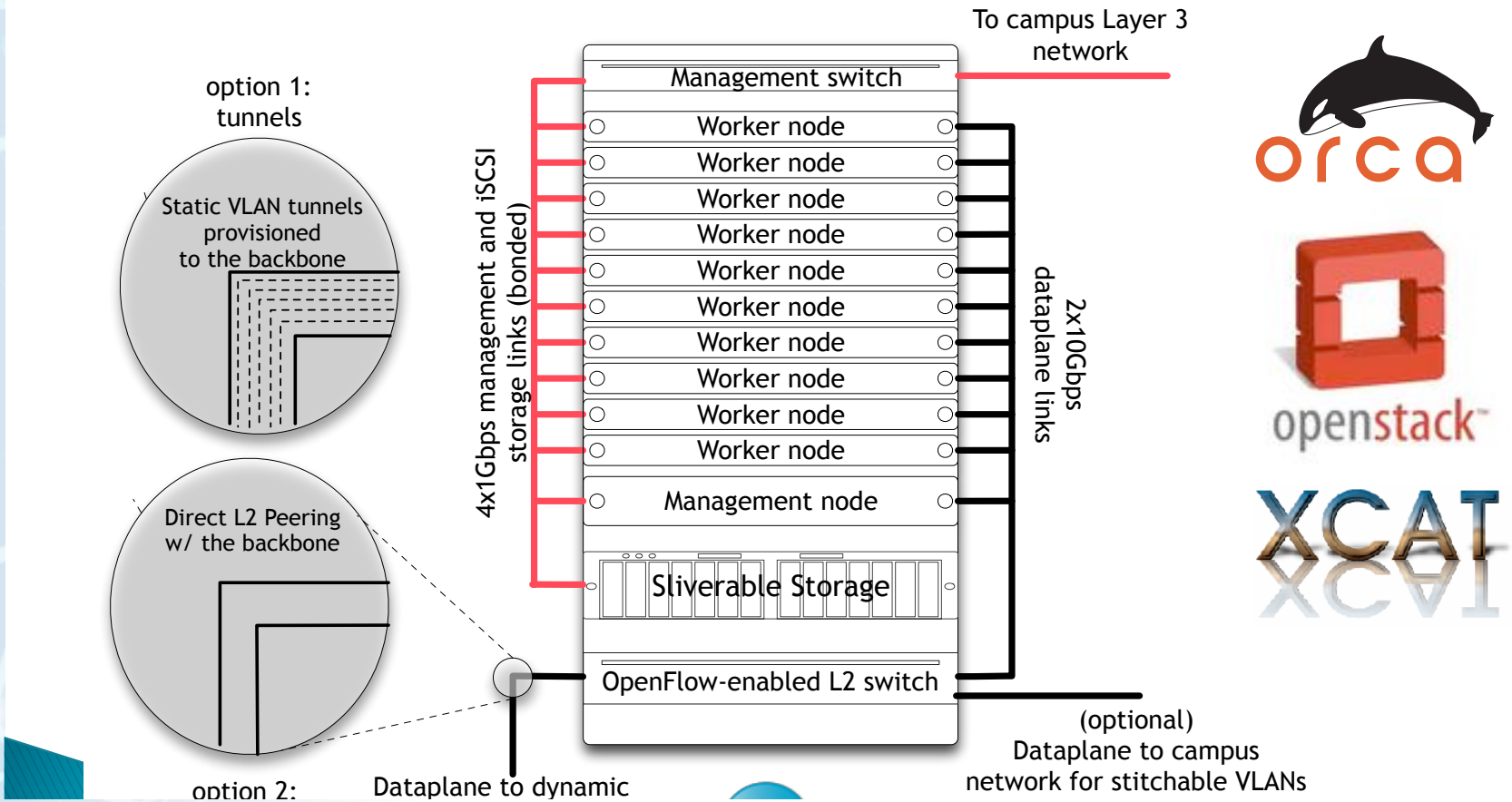
- Testbed Router
- ESnet5 Router
- Site Router
- ESnet5 Link
- Testbed Link
- 100G Interface
- 40G Interface

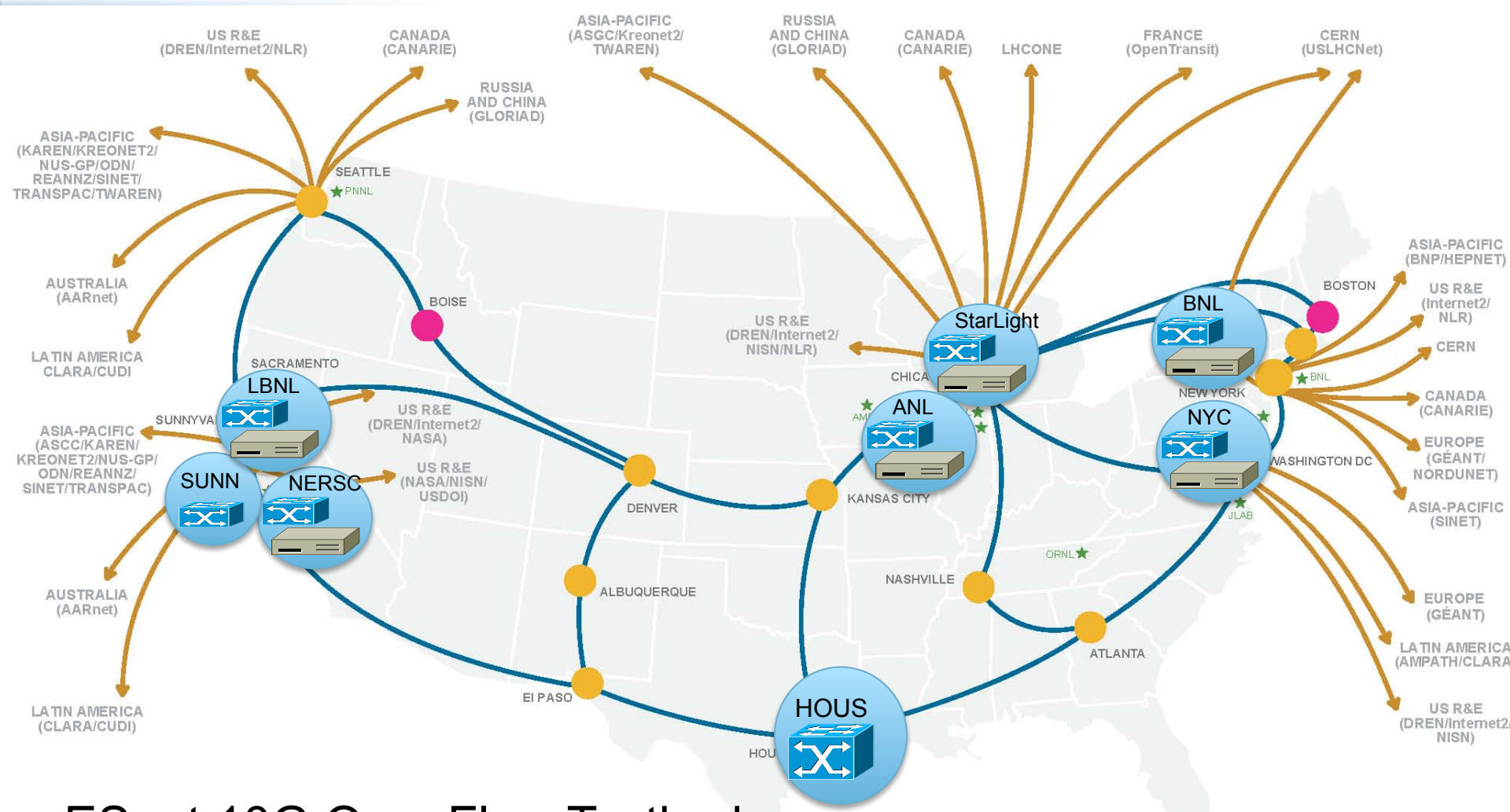
- Note: All routers are Alcatel Lucent 7750 unless noted otherwise



OpenFlow Testbed

ExoGeNI Rack Details





ESnet 10G OpenFlow Testbed

ESnet
Energy Sciences Network

100G IP Hubs

4x10G IP Hub

Major R&E and International peering connections

★ Office of Science National Labs

Ames	Ames Laboratory (Ames, IA)
ANL	Argonne National Laboratory (Argonne, IL)
BNL	Brookhaven National Laboratory (Upton, NY)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
PNNL	Pacific Northwest National Laboratory (Richland, WA)
PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
SLAC	Stanford Linear Accelerator Center (Menlo Park, CA)

OpenFlow Testbed

- Uses 10G circuits on 100G backbone
- 8 OpenFlow Switches
  - 6 of which have 10G hosts directly connected
  - Multi-Vendor
    - NEC, Juniper, Brocade, IBM, pica8, noviflow
- Available to ESnet collaborators

OpenFlow Testbed Capabilities

Researchers can:

- Experiment with multiple types of controllers and hardware
- Experiment with multiple paths
- Connect to other testbeds

Capabilities

- Sliceable with FlowVisor
- Support for internal and external open flow controllers
  - i.e. running within ESnet, or accessed from the internet)
- Data plane provided by ESnet OSCARS, provides QoS
  - Nationwide footprint
- Support for topology virtualization
- Integration of other ESnet services (perfSONAR, SNMP collector, Topology service, NSI)

Sample Use of the OpenFlow Testbed

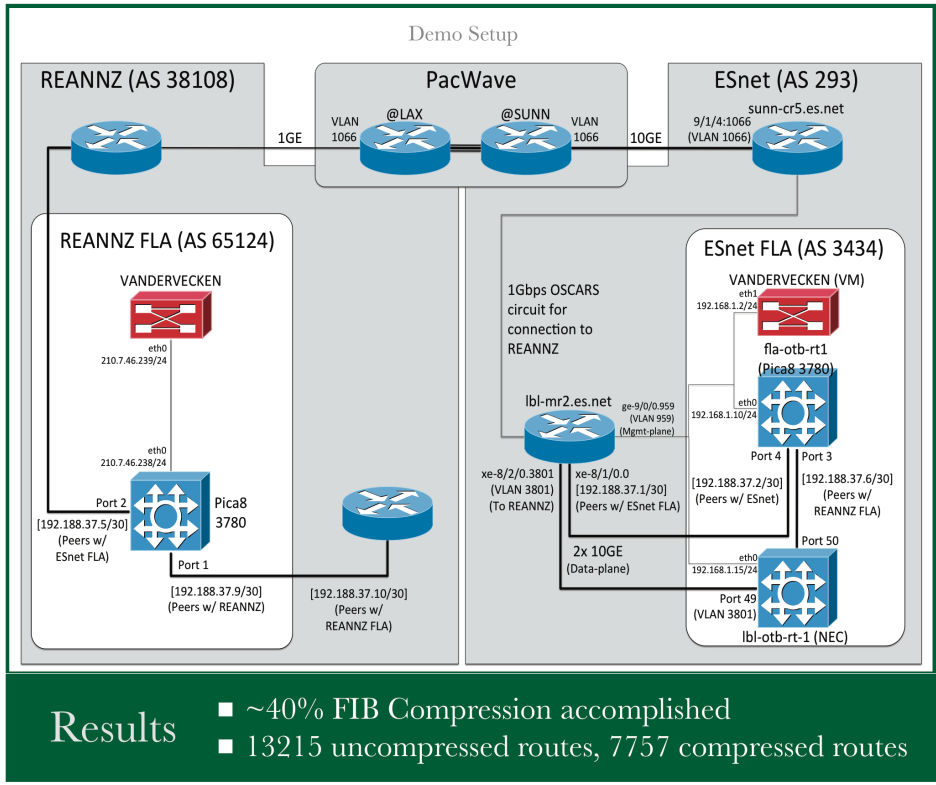
- Demonstration at Open Networking Summit (April)

Front-Line Assembly

DEMO

First international BGP peering using SDN in production between two national-scale network providers

- Innovative FIB compression enables using commodity OpenFlow switches for peering
- Leverages community open-source packages. RouteFlow and Quagga



Results

- ~40% FIB Compression accomplished
- 13215 uncompressed routes, 7757 compressed routes

Insights

- SDN networks can interface with existing Internet
- New techniques need to be developed to scale controller-based networking

Demonstration Team:
Google Network Research – Josh Bailey, Scott Whyte
REANNZ – Dylan Hall, Sam Russell, James Wix, Steve Cotter
ESnet – Inder Monga, Chin Guok, Eric Pouyoul, Brian Tierney

Acknowledgements - Joe Stringer

OpenFlow Testbed Experiment: A Virtual Switch Implementation:

Create Virtual switch:

- Specify edge OF ports
- Specify backplane topology and bandwidth
- Policy constraints like flowspace
- Store the switch into a topology service