ESnet
ENERGY SCIENCES NETWORK

Scientific Networking: Big Data in Motion

LBNL, June 18th, 2015

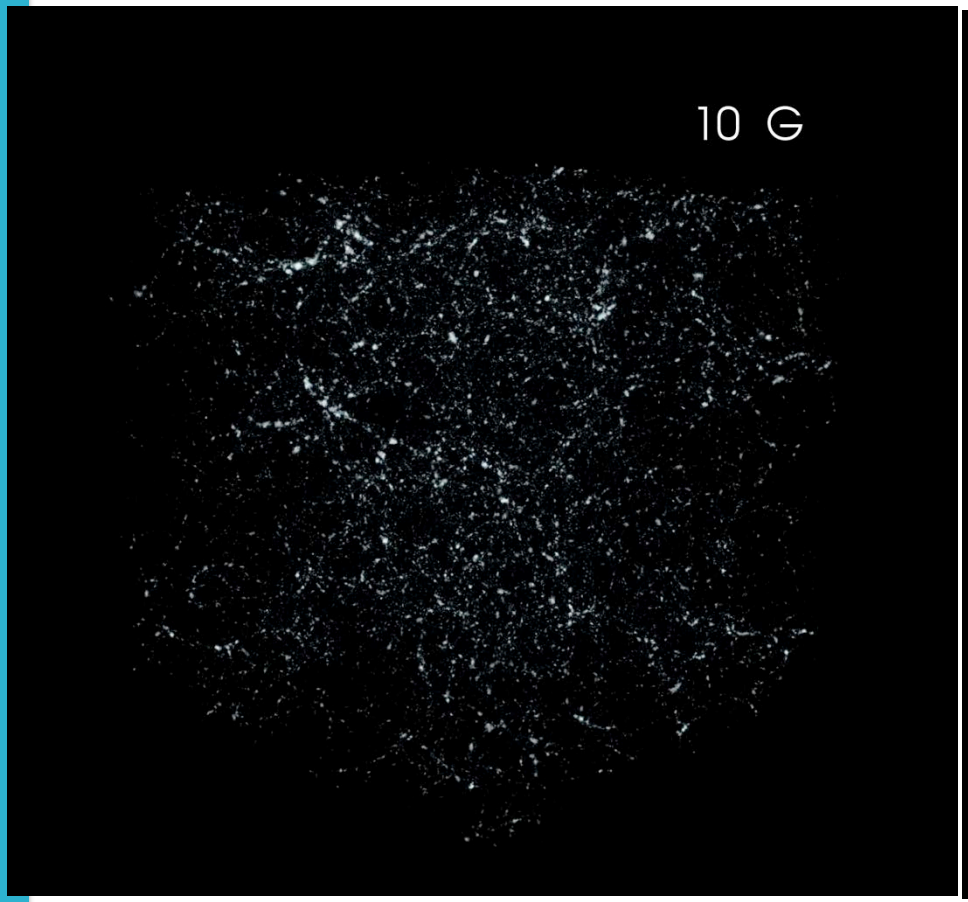
Eric Pouyoul
lomax@es.net

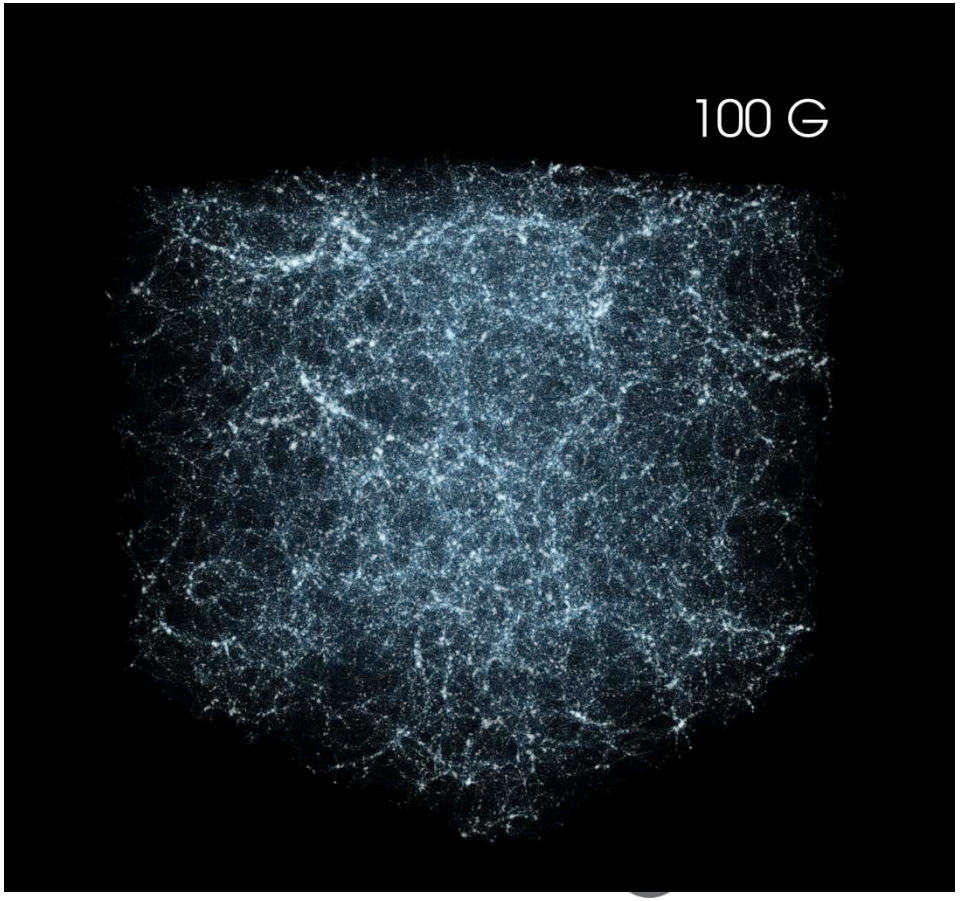
U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

- Why Science Networking ?
- What is ESnet ?
- Scientific Networking Issues
- ScienceDMZ
- Network Research

Data size does matter in sciences.

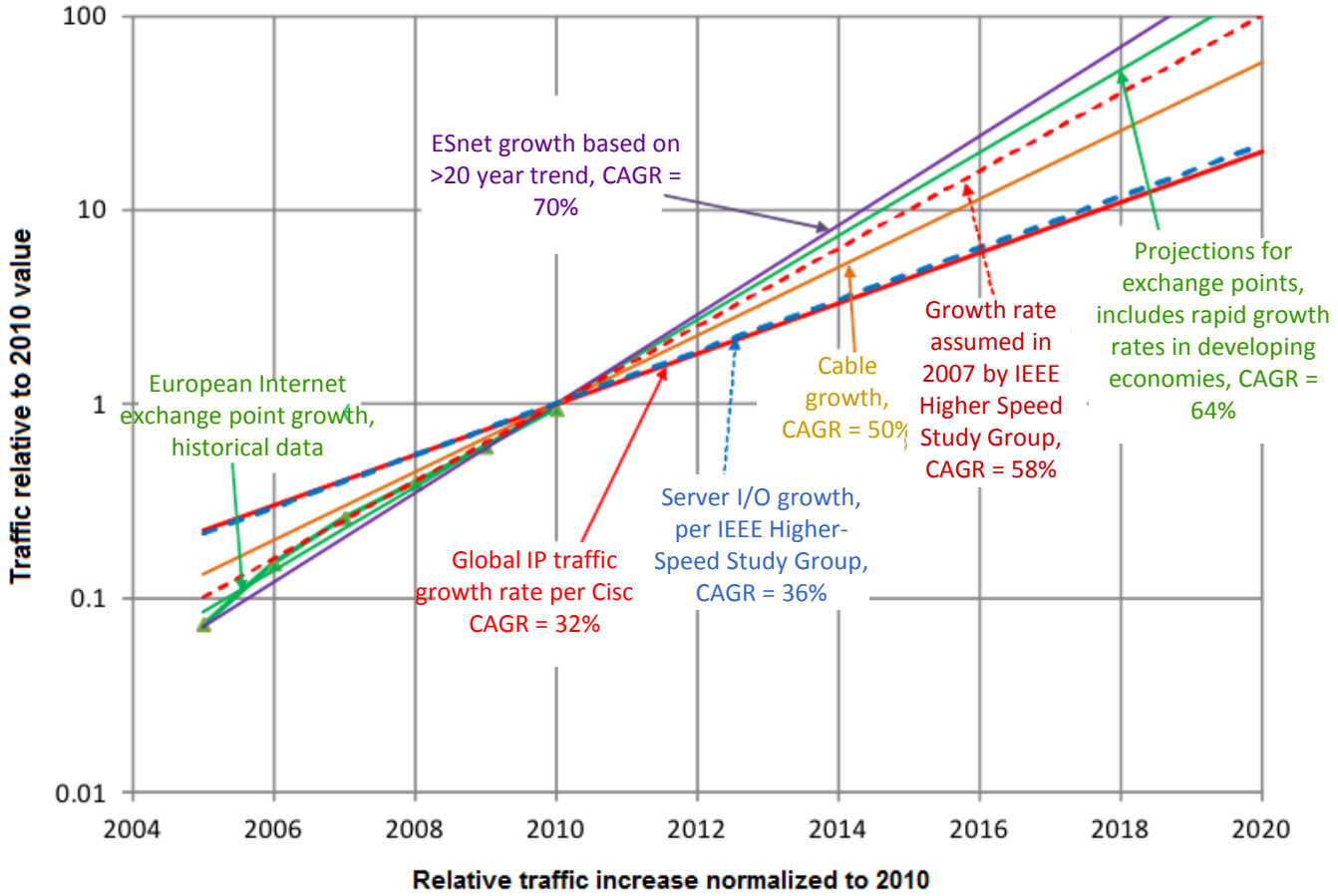




… so does it in everyday life.

ESnet

Growth in Science >> Global IP Traffic

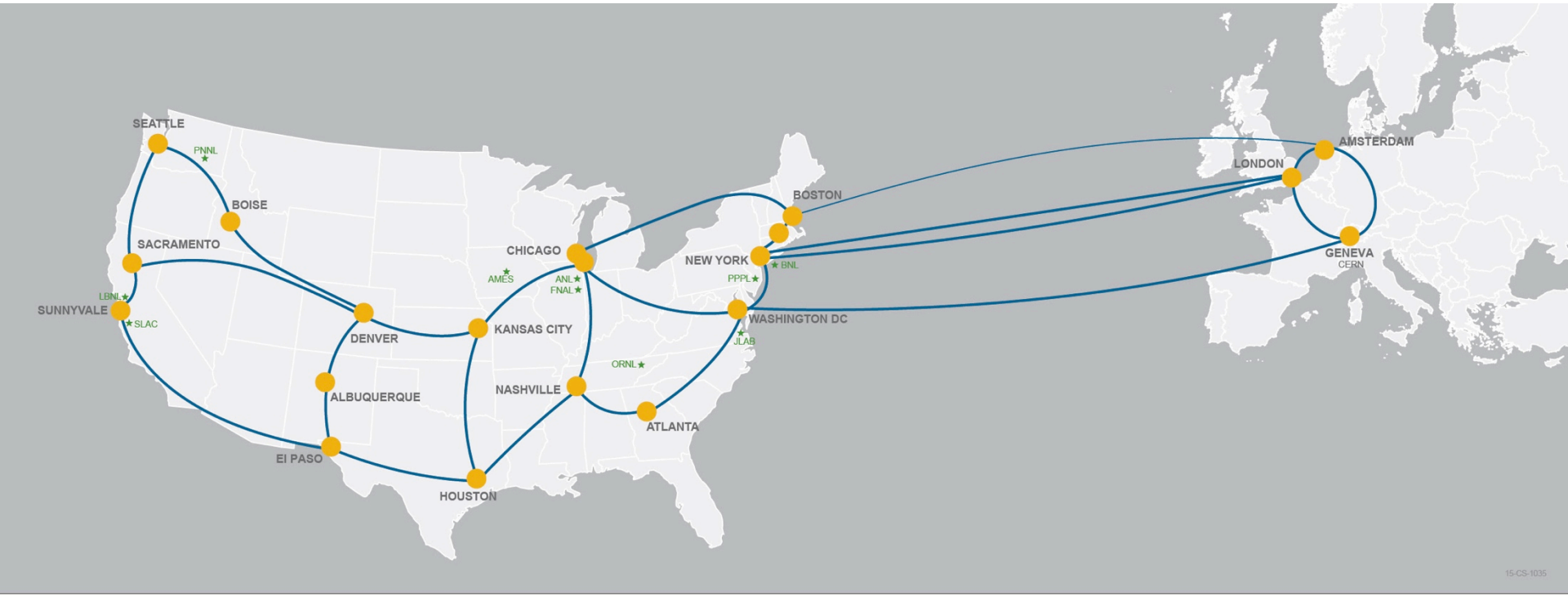


CAGR = compound annual growth rate.

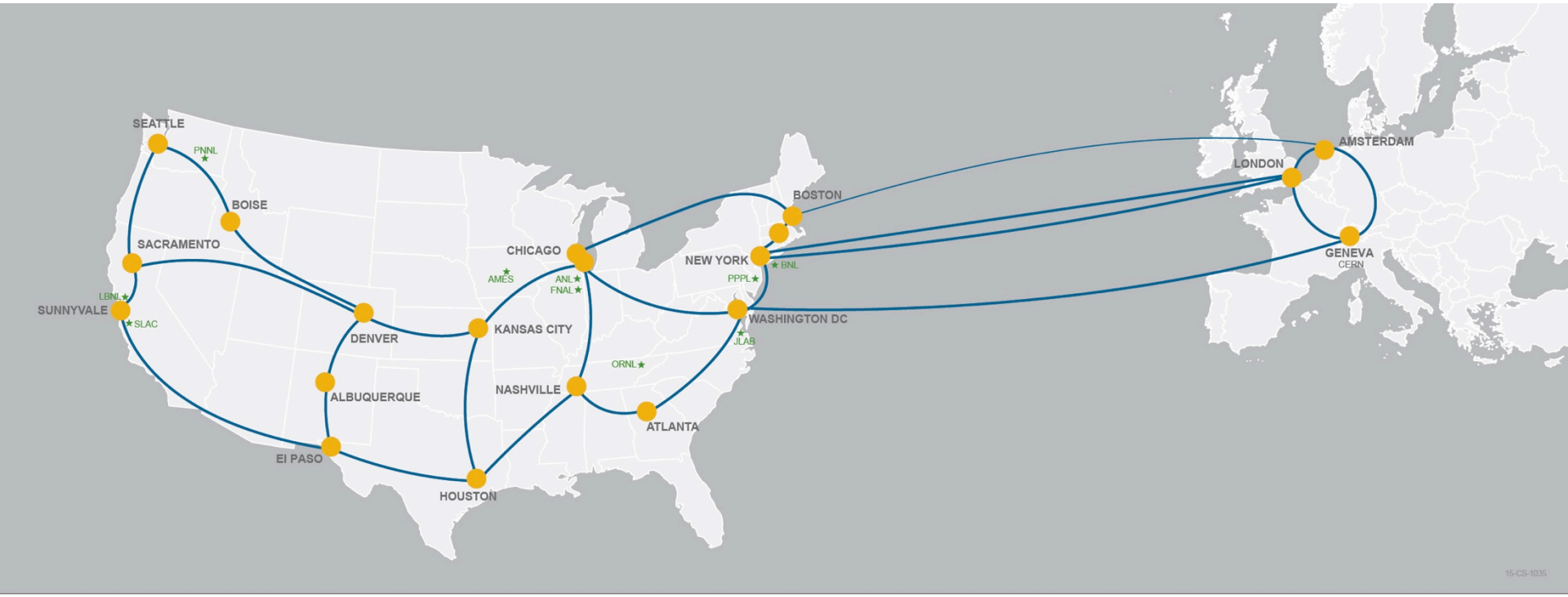
Relative traffic increase for various sectors normalized to 2010, **adapted from** IEEE 802.3™ Industry Connections Ethernet Bandwidth Assessment Ad Hoc Report.

ESnet

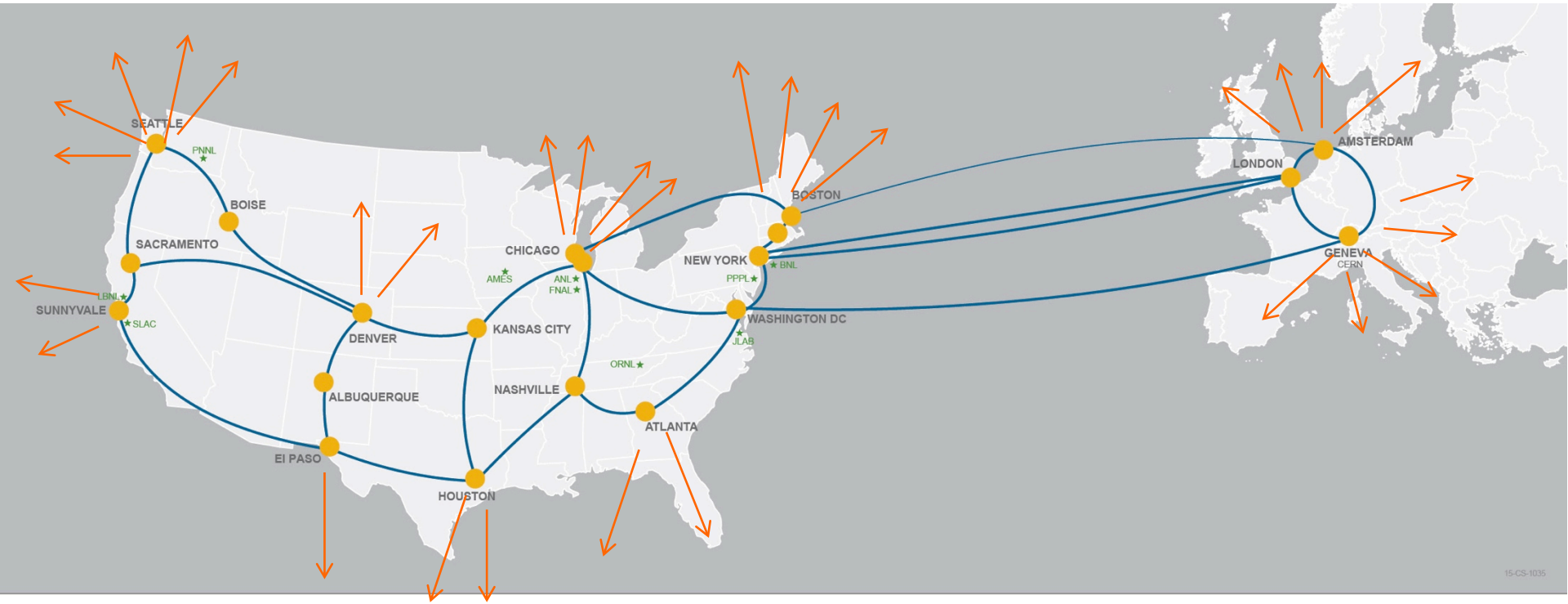
ESnet: Energy Science Network.



This is not an ISP.



80% of ESnet traffic originates or terminates outside the DOE complex.



ESnet
ENERGY SCIENCES NETWORK

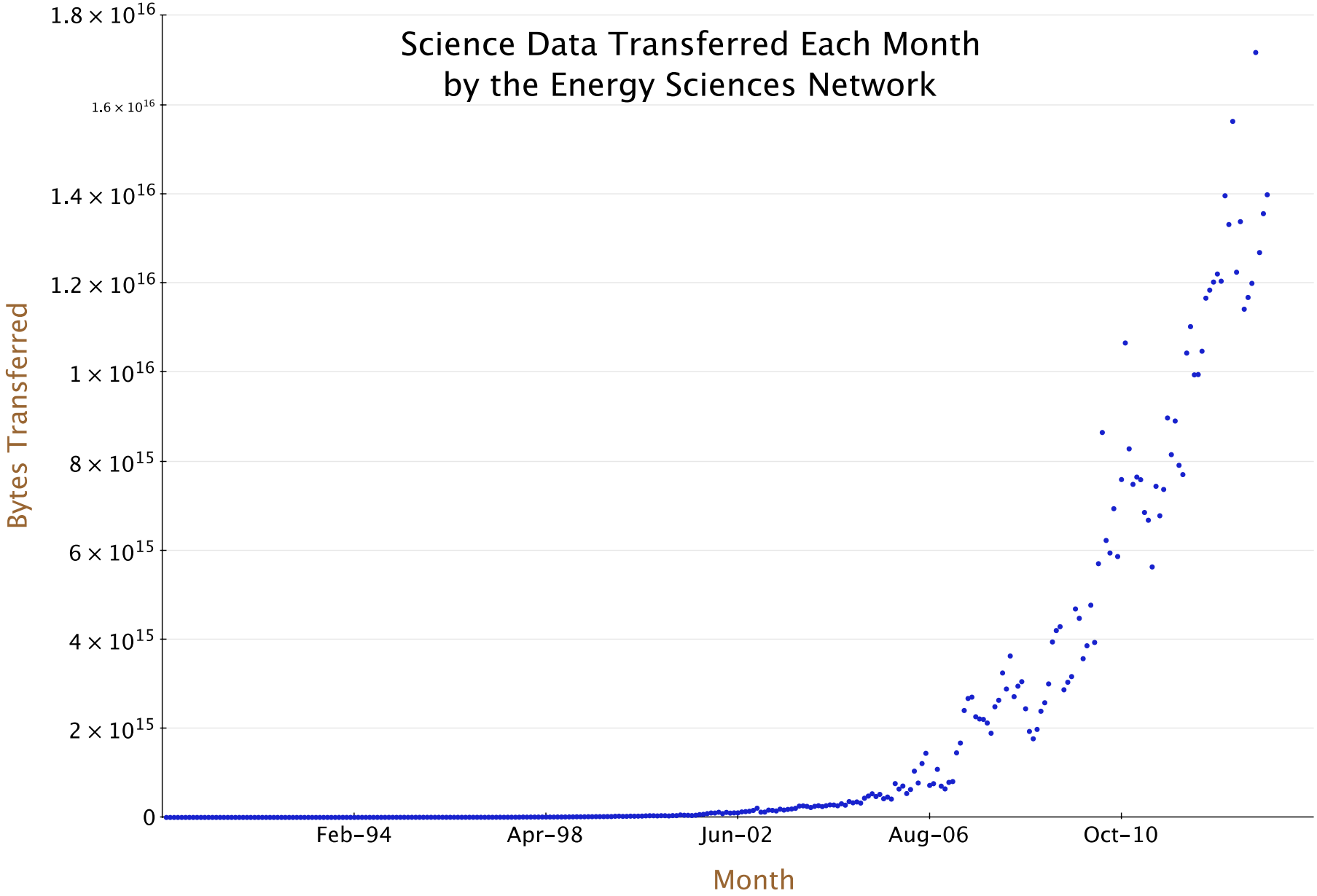
★ Department of Energy Office of Science National Labs

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	SLAC National Accelerator Laboratory (Menlo Park, CA)

Our vision:

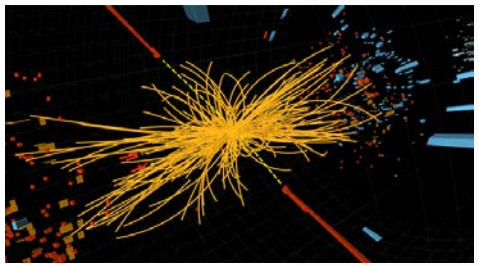
Scientific progress will be **completely unconstrained** by the physical location of instruments, people, computational resources, or data.

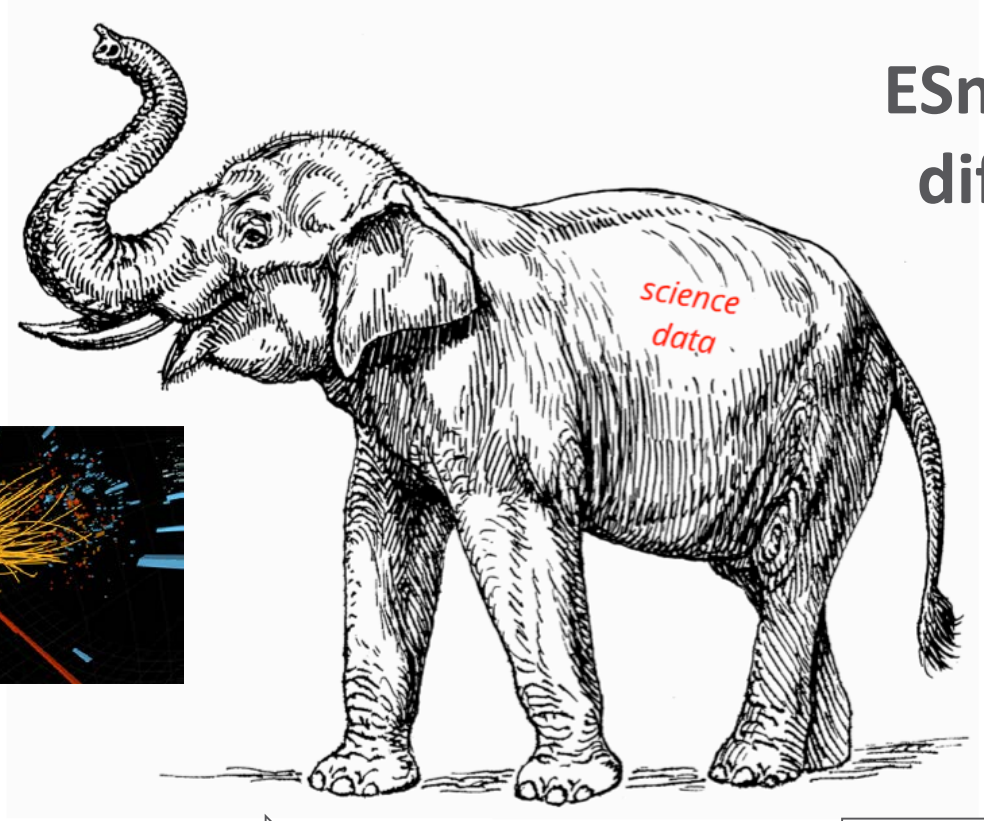
ESnet

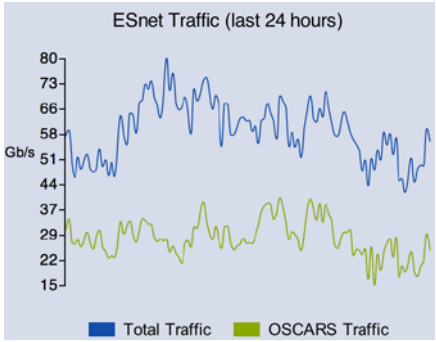

Science Data Transferred Each Month
by the Energy Sciences Network
Bytes Transferred
Month
1.8×10^{16}
1.6×10^{16}
1.4×10^{16}
1.2×10^{16}
1×10^{16}
8×10^{15}
6×10^{15}
4×10^{15}
2×10^{15}
0
Feb-94
Apr-98
Jun-02
Aug-06
Oct-10

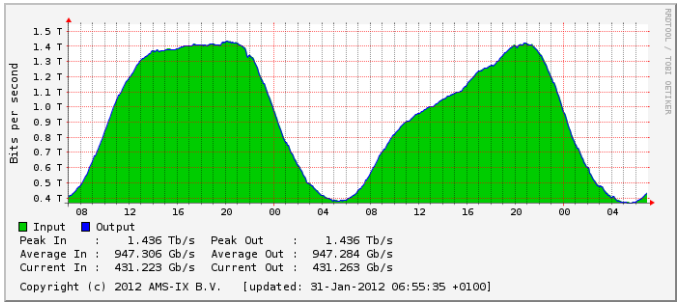
The Internet Can Not Transport Scientific Data

ESnet is designed for different goals than general Internet.

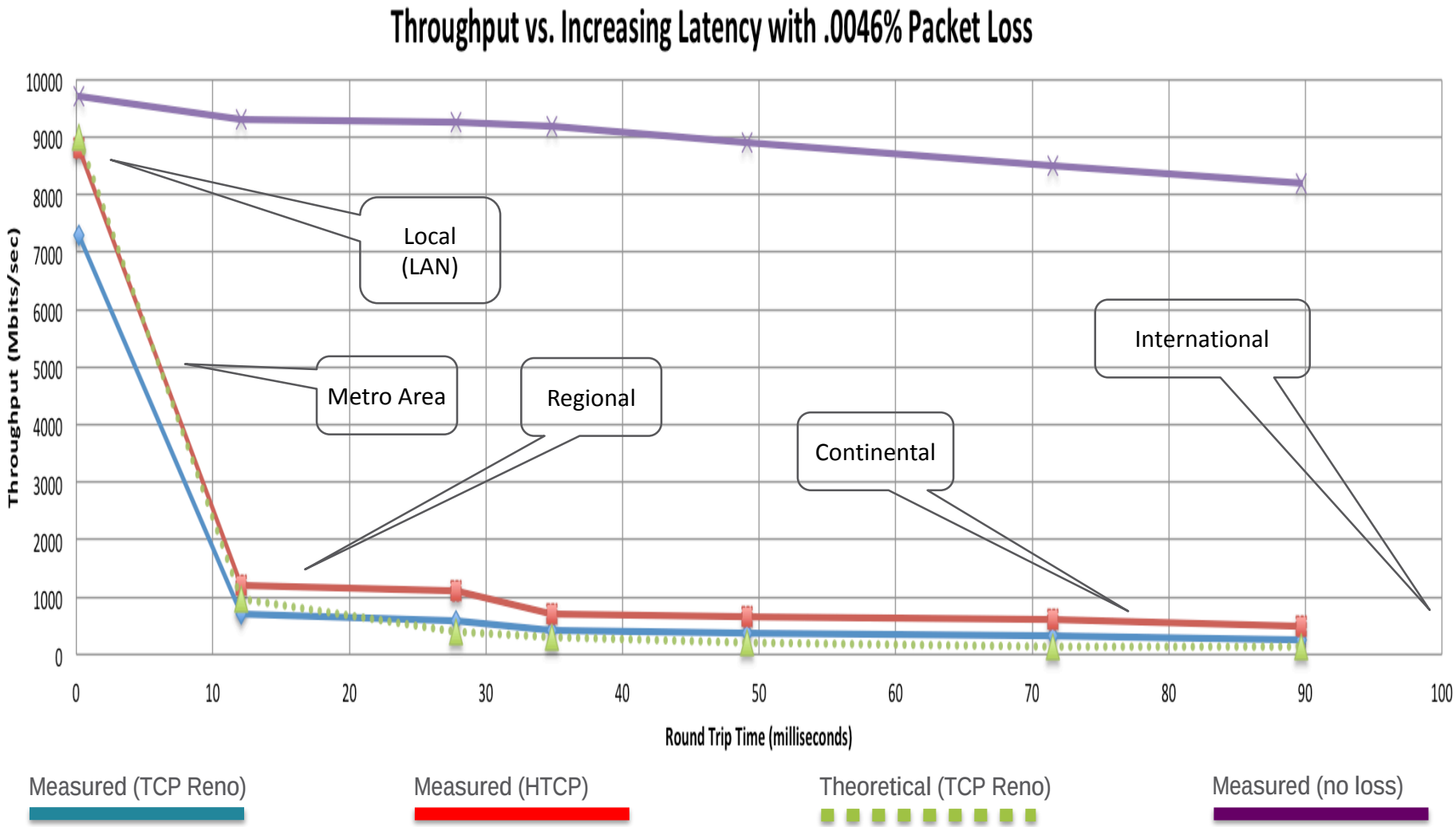








Elephant flows require almost *lossless* networks.

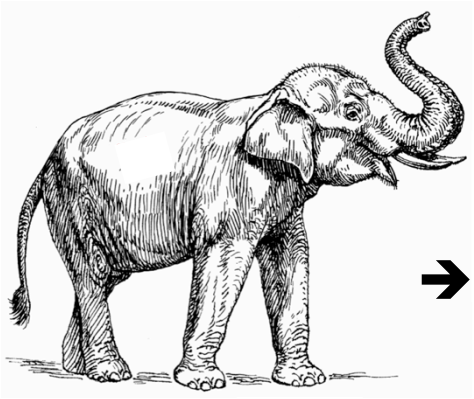


See Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, and Jason Zurawski. The Science DMZ: A Network Design Pattern for Data-Intensive Science. In *Proceedings of the IEEE/ACM Annual SuperComputing Conference (SC13)*, Denver CO, 2013.

Elephant Flows Place Great Demands on Networks

Physical pipe that leaks water at rate of .0046% by volume.

<u>Result</u>
99.9954% of water transferred, at "line rate."



Network 'pipe' that drops packets at rate of .0046%.

<u>Result</u>
100% of data transferred, *slowly*, at <<5% optimal speed.

essentially fixed

determined by speed of light

$$\frac{\text{maximum segment size}}{\text{round-trip time}} \times \frac{1}{\sqrt{\text{packet-loss rate}}}$$

Through careful engineering, we can minimize packet loss.

Assumptions: 10Gbps TCP flow, 80ms RTT.

See Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, and Jason Zurawski. The Science DMZ: A Network Design Pattern for Data-Intensive Science. In *Proceedings of the IEEE/ACM Annual SuperComputing Conference (SC13)*, Denver CO, 2013.

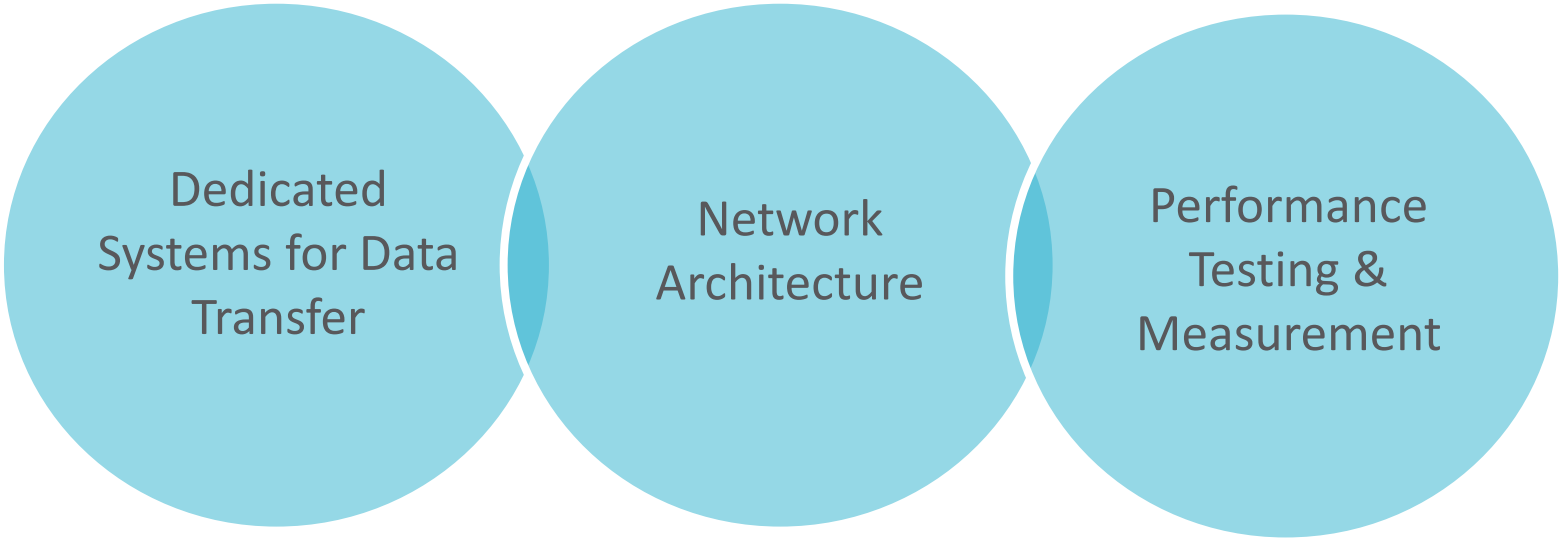
The Science DMZ

Raising Expectations: Time to Copy 1 Terabyte

- 10 Mbps network : 300 hrs (12.5 days)
- 100 Mbps network : 30 hrs
- 1 Gbps network : 3 hrs (are your disks fast enough?)
- 10 Gbps network : 20 minutes (requires RAID disk array)
- These figures assume some headroom left for other users
- Compare these speeds to:
  - USB portable disk
    - 20-25 hours to load 1 Terabyte

ESnet

The Data Transfer Trifecta: The "Science DMZ" Model



Data Transfer Node

- High performance
- Configured for data transfer
- Proper tools

Science DMZ

- Dedicated location for DTN
- Easy to deploy - no need to redesign the whole network

perfSONAR

- Enables fault isolation
- Verify correct operation
- Widely deployed in ESnet and other networks, as well as sites and facilities

ESnet

One motivation for Science DMZ model: Soft Network Failures

- Soft failures are where basic connectivity functions, but high performance is not possible.
- TCP was intentionally designed to hide all transmission errors from the user:
  - "As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users." (From RFC793, 1981)
- Some soft failures only affect high bandwidth long RTT flows.
- Hard failures are easy to detect & fix
  - soft failures can lie hidden for many months!
- One network problem can often mask others

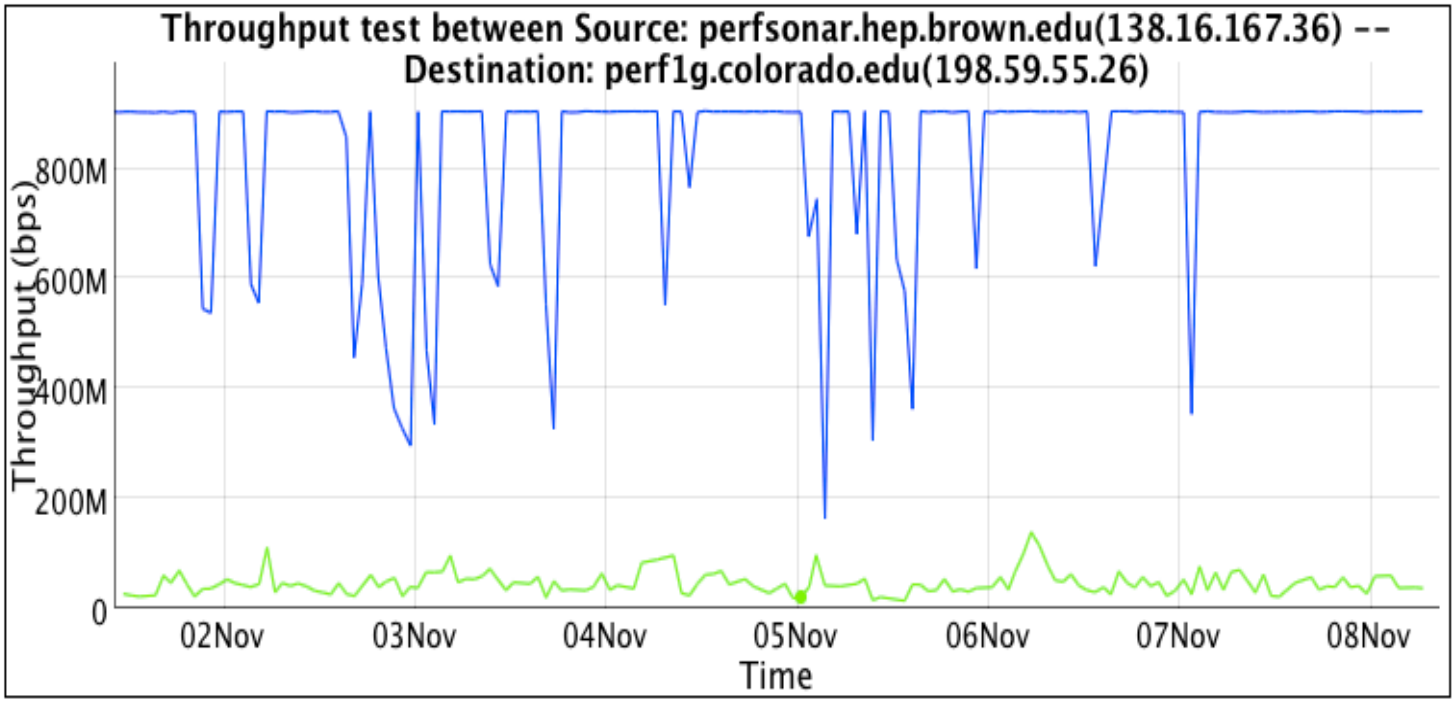
ESnet

Common Soft Failures

- Random Packet Loss
  - Bad/dirty fibers or connectors
  - Low light levels due to amps/interfaces failing
  - Duplex mismatch
- Small Router/Switch Buffers
  - Switches not able to handle the long packet trains prevalent in long RTT sessions and local cross traffic at the same time
- Un-intentional Rate Limiting
  - Processor-based switching on routers due to faults, acl's, or mis-configuration

ESnet

Another Common Problem: the Firewall



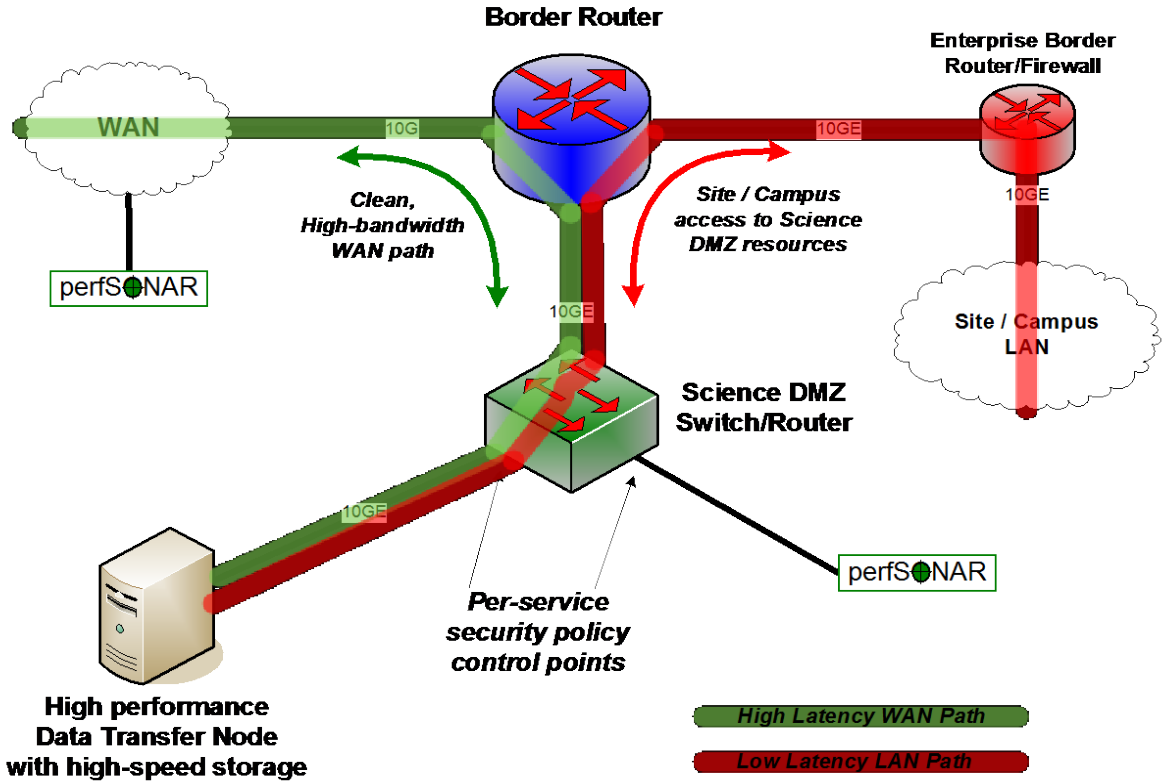
Graph Key

- Src-Dst throughput
- Dst-Src throughput

- Blue = outgoing, green = incoming
- This flow is between two 1G hosts on a 10G network

ESnet

Simple Science DMZ



perfSONAR

ESnet

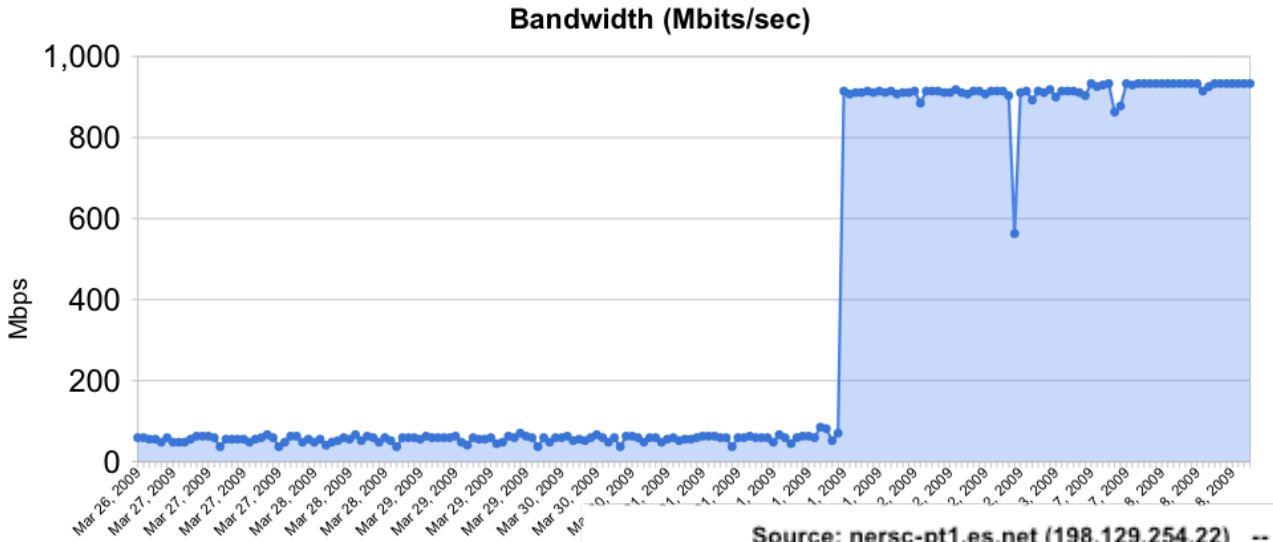
What is perfSONAR?

- perfSONAR is a set of tools to:
- Set network performance expectations
- Find network problems ("soft failures")
  - Isolate the cause of the problems
  - Make sure the problem stays fixed
- All in multi-domain environments
- These problems are all harder when multiple networks are involved

- perfSONAR is provides a standard way to publish active and passive monitoring data
  - This data is interesting to network researchers as well as network operators

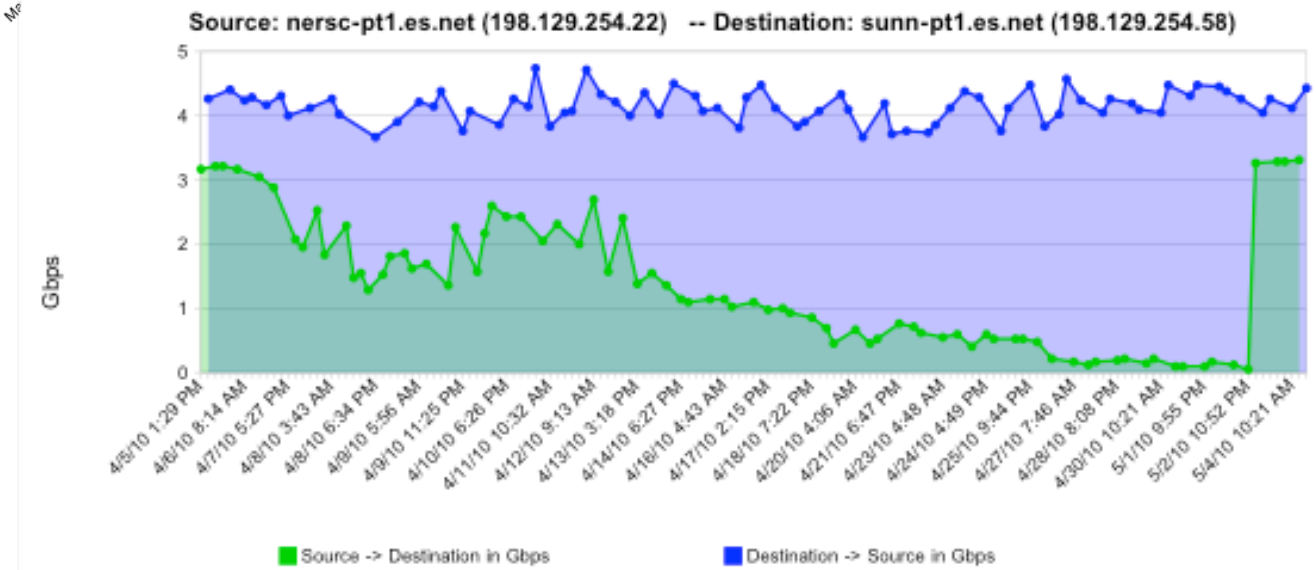
ESnet

perfSONAR Results: Sample Soft Failures as seen by perfSONAR



Rebooted router with full route table

Gradual failure of optical line card



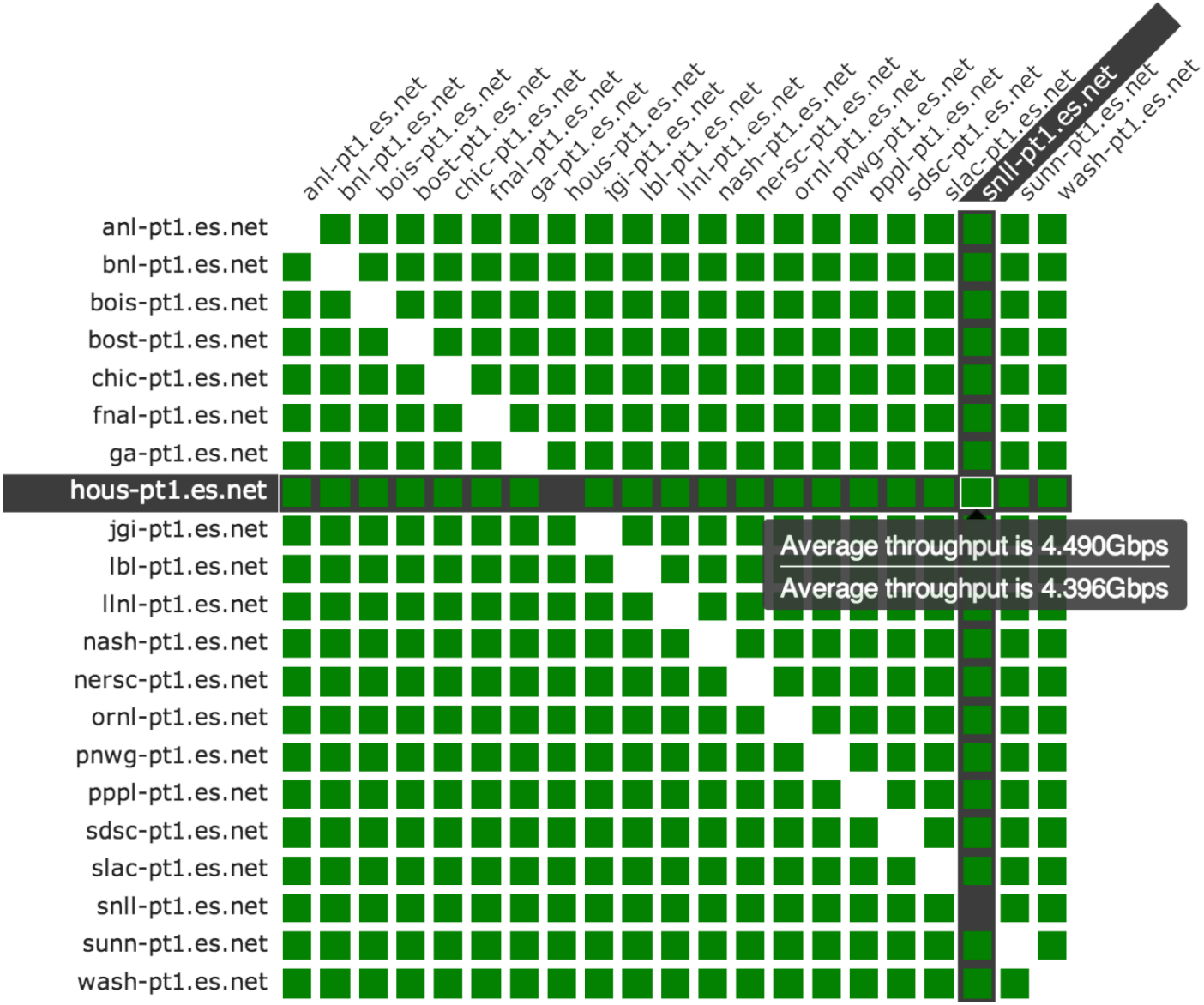
perfSONAR Toolkit Services

- PS-Toolkit includes these measurement tools:
  - BWCTL: network throughput
  - OWAMP: network loss, delay, and jitter
  - traceroute
- Test scheduler:
  - runs bwctl, traceroute, and owamp tests on a regular interval
- Measurement Archives (data publication)
  - SNMP MA – router interface Data
  - pSB MA -- results of bwctl, owamp, and traceroute tests
- Lookup Service: used to find services
- PS-Toolkit includes these web100-based Troubleshooting Tools
  - NDT (TCP analysis, duplex mismatch, etc.)
  - NPAD (TCP analysis, router queuing analysis, etc)

ESnet

World-Wide perfSONAR-PS Deployments:
535 bwctl nodes, 533 owamp nodes as of June '13

perfSONAR Dashboard: http://ps-dashboard.es.net



ESnet

Data Transfer Node and Data Transfer Tools

ESnet

The Data Transfer Node

- A DTN server is made of several subsystems. Each needs to perform optimally for the DTN workflow:
- **Storage:** capacity, performance, reliability, physical footprint
- **Networking:** protocol support, optimization, reliability
- **Motherboard:** I/O paths, PCIe subsystem, IPMI
- **Chassis:** adequate power supply, extra cooling

- DTNs are usually optimized for sequential reads/write of large files, and a moderate number of high bandwidth flows.
- The DTN is dedicated to data transfer, and not doing data analysis/manipulation

ESnet

DTN Tuning

http://fasterdata.es.net/science-dmz/DTN/tuning/

- Defaults are usually not appropriate for high throughput.

- What needs to be tuned:
- BIOS
- Firmware
- Device Drivers
- Networking
- File System
- Application

ESnet

Using the right tool is Critical!

- Sample Results: Berkeley, CA to Argonne, IL (near Chicago). RTT = 53 ms, network capacity = 10Gbps.

Tool	Throughput
• scp:	140 Mbps
• HPN patched scp:	1.2 Gbps
• ftp	1.4 Gbps
• GridFTP, 4 streams	6.6 Gbps (disk limited)

- Note that to get more than 1 Gbps (125 MB/s) disk to disk requires RAID.

ESnet

Science DMZ Summary

Consists of **three key components**, all required:

- "Friction free" network path
  - Highly capable network devices (wire-speed, deep queues)
  - Virtual circuit connectivity option
  - Security policy and enforcement specific to science workflows
  - Located at or near site perimeter if possible
- Dedicated, high-performance Data Transfer Nodes (DTNs)
  - Hardware, operating system, libraries all optimized for transfer
  - Includes optimized data transfer tools such as Globus Online and GridFTP
- Performance measurement/test node
  - perfSONAR

Details at http://fasterdata.es.net/science-dmz/

Networking Research

ESnet

Science applications are innovating…

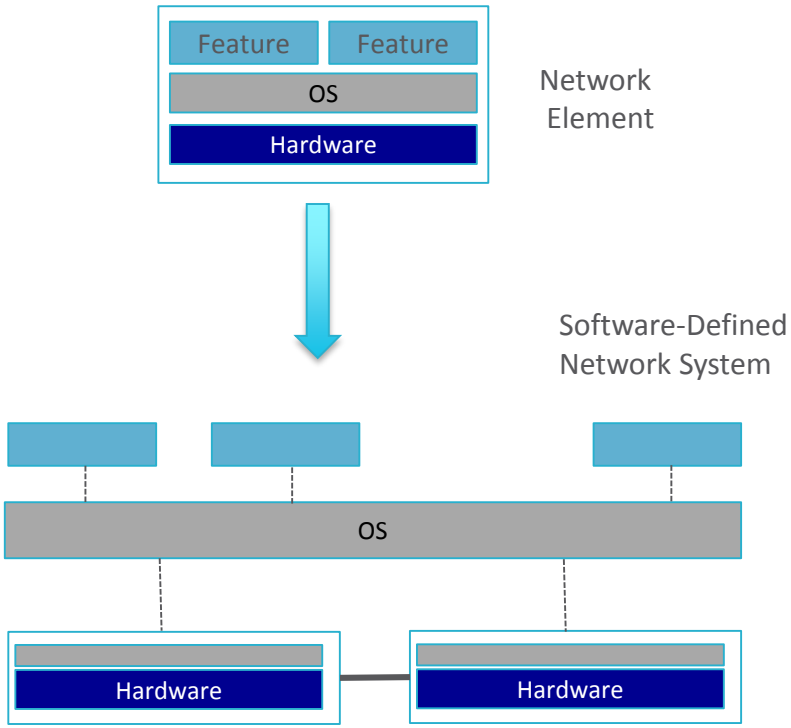
- …largely driven by Data-Intensity
- Use of complex workflows
- On-demand computing
- Distributed computing and storage
- Streaming from remote storage

- These new innovations increasingly **rely on the network** for success in meeting their objectives of higher productivity and better use of the resources

ESnet

Software Defined Networking (SDN) moves 'networking' out of the 'box'

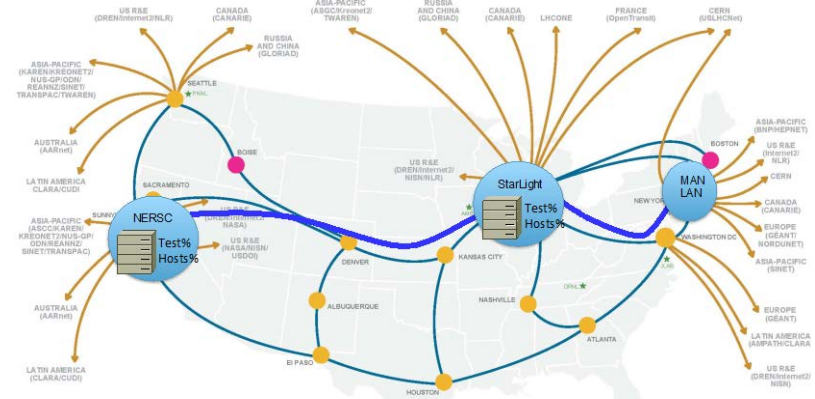
- SDN shifts the mechanisms to design and operate networks, from configuration to programming
- The network is not just seen as a composition of different elements
- User applications interact with the network operating system
- (Virtualized) network becomes a single programmable entity
- Various levels of abstraction, programmability can be applied depending on the application



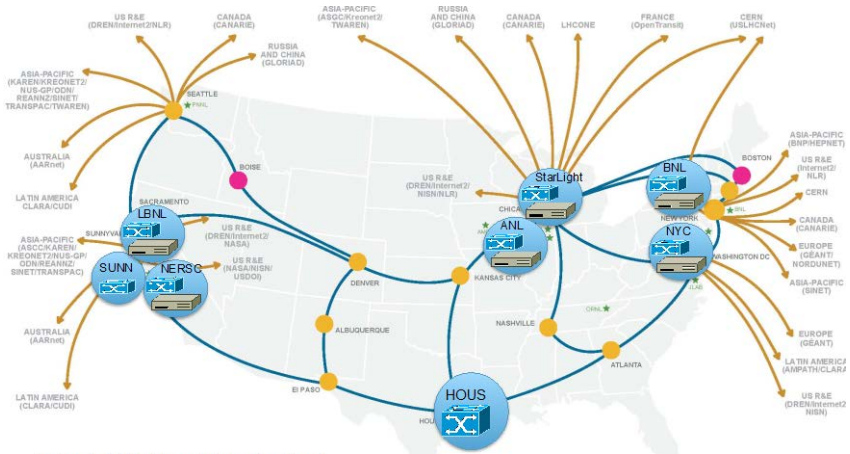
ESnet

ESnet Research Testbeds

- 100G Testbed
  - High-speed protocol research
  - Available since Jan 2012
  - Dedicated 100G wave from Oakland to Chicago to NYC
- OpenFlow Testbed
  - 10G Nationwide Footprint
- Dark Fiber Testbed
  - Continental-scale fiber footprint for disruptive research



ESnet 100G Testbed



ESnet 10G OpenFlow Testbed

Thank you!

lomax@es.net

http://fasterdata.es.net