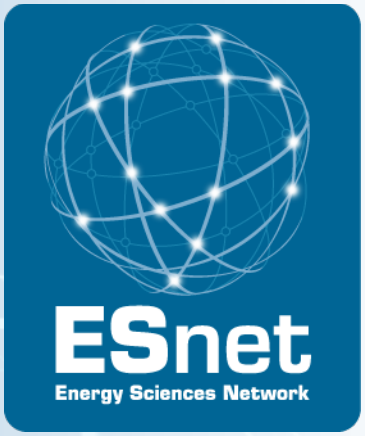


ESnet Support for WAN Data Movement

Eli Dart, Network Engineer

ESnet Science Engagement Group

Joint Facilities User Forum on Data Intensive Computing

Oakland, CA

June 16, 2014

Outline

ESnet overview

Support for data movement – specific engineering concerns

- Effect of packet loss
- Science DMZ model

Data movement – rates and volumes

Current network – ESnet5

Network enhancements

What are others doing?

ESnet by the Numbers

High-speed national network, optimized for DOE science missions:

- high performance networking facility for DOE/SC
- connecting 40 labs, plants and facilities with >100 networks
- $32.6M in FY14, 42FTE
- older than commercial Internet, growing twice as fast

$62M ARRA grant for 100G upgrade:

- transition to new era of optical networking
- fiber assets + access to spectrum shared with Internet2
- world's first 100G network at continental scale
- first connections were to DOE supercomputer centers

Culture of urgency:

- 4 awards in past 3 years
- R&D100 in FY13
- "5 out of 5" for customer satisfaction in last review

ESnet Is A Science Network First

We specifically engineer the network to support data-intensive science

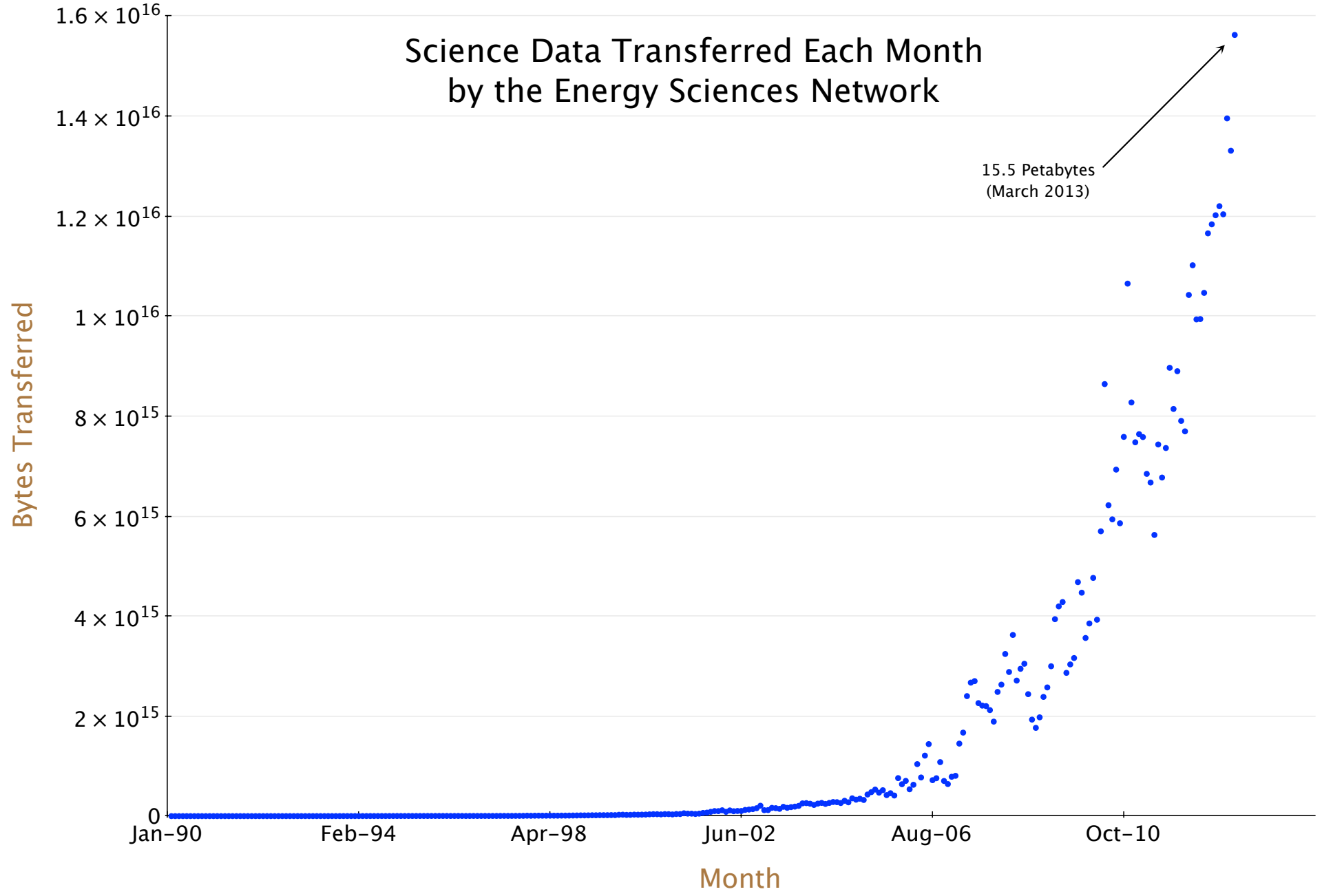
High-speed, high-capability connections to sites, facilities, and other networks

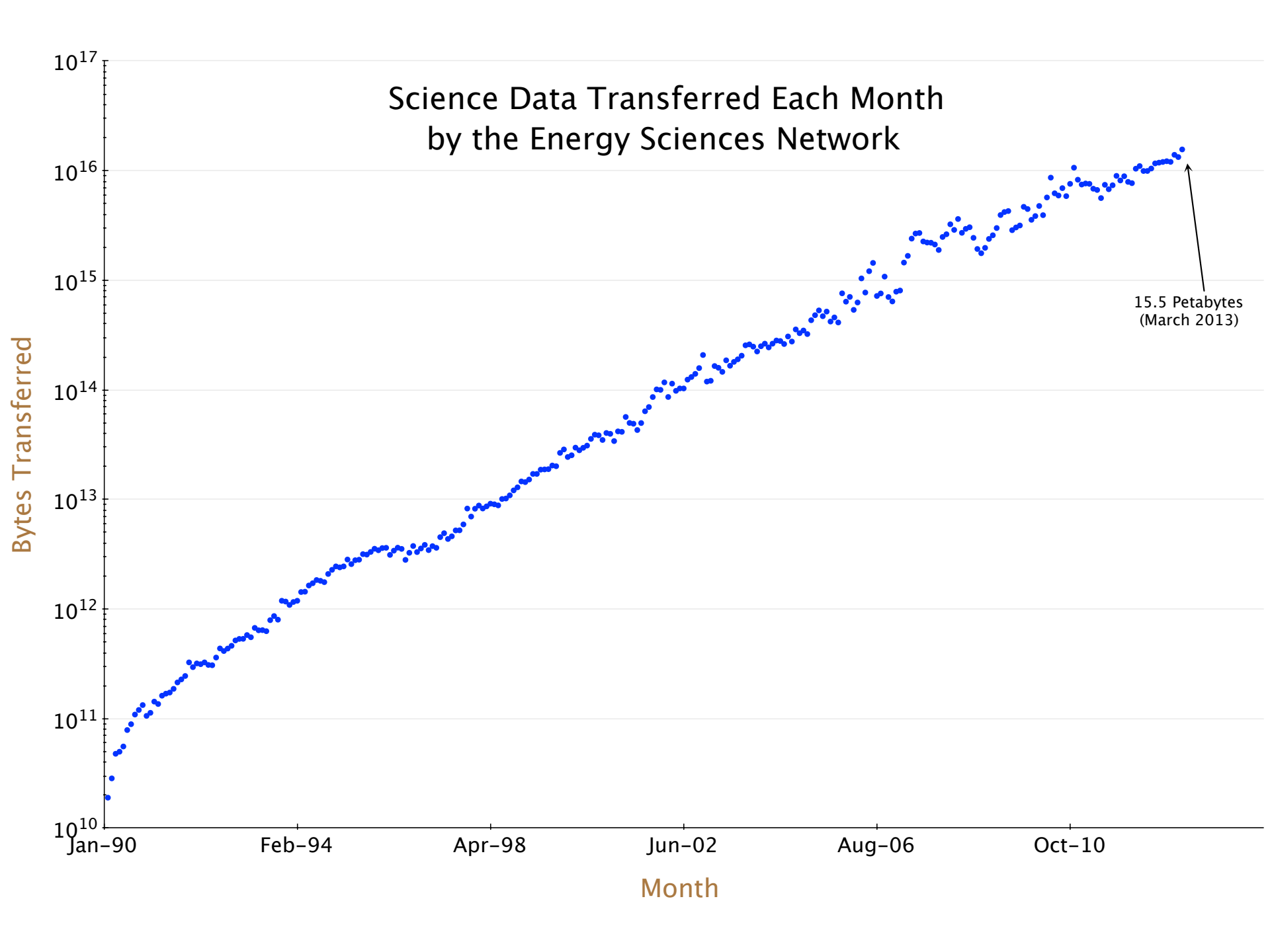
- 100G to ANL, LBNL, FNAL, NERSC, ORNL
- 100G to BNL, LANL, LLNL, SLAC coming soon
- 100G connections to major peering exchanges

Test and measurement instrumentation to ensure performance

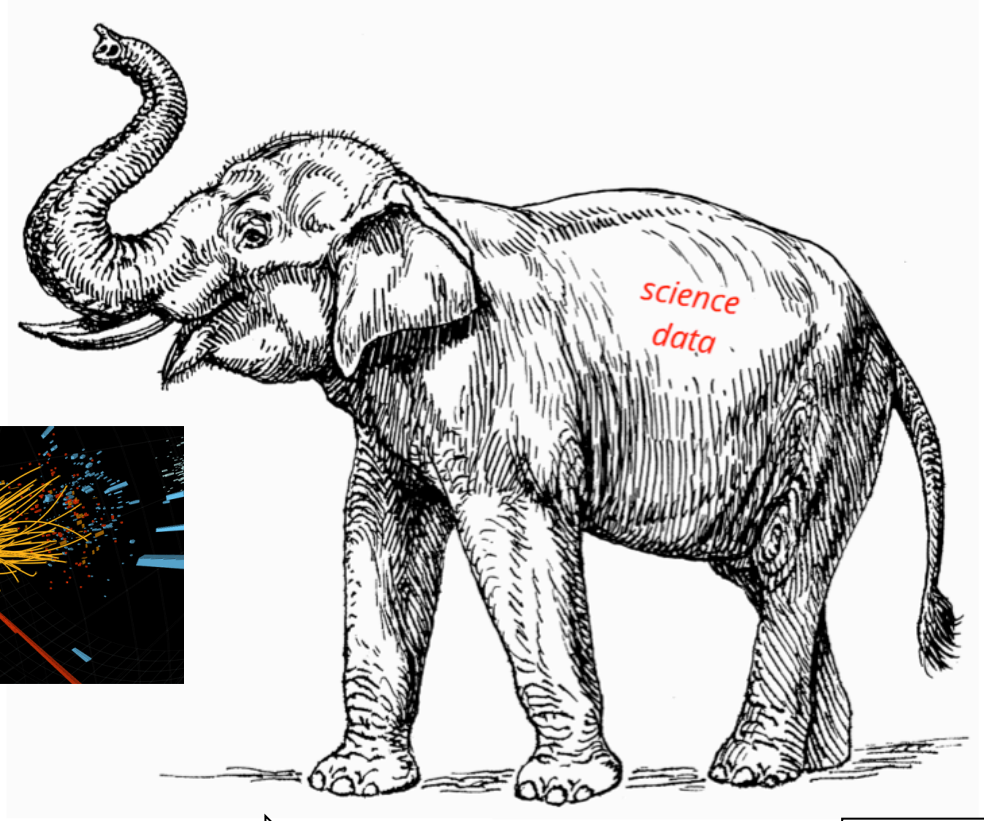
Engagement with science experiments, facilities, and sites to help get the maximum benefit from the infrastructure

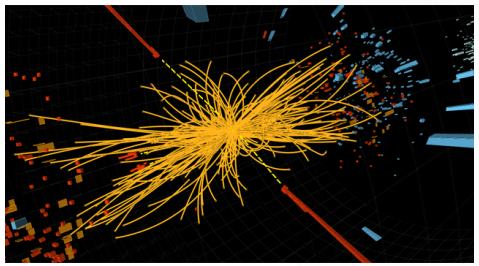
ESnet's capabilities are derived from the requirements of the science

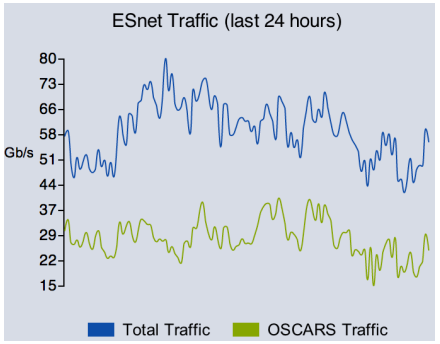

Science Data Transferred Each Month
by the Energy Sciences Network
15.5 Petabytes
(March 2013)
Bytes Transferred
Month
1.6 × 10^16
1.4 × 10^16
1.2 × 10^16
1 × 10^16
8 × 10^15
6 × 10^15
4 × 10^15
2 × 10^15
0
Jan-90
Feb-94
Apr-98
Jun-02
Aug-06
Oct-10

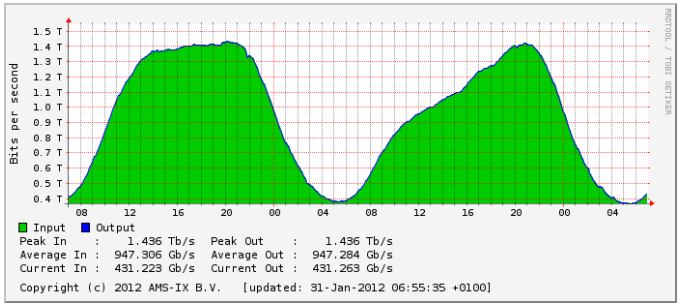

Science Data Transferred Each Month
by the Energy Sciences Network
Bytes Transferred
10^{17}
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
15.5 Petabytes
(March 2013)
Jan-90
Feb-94
Apr-98
Jun-02
Aug-06
Oct-10
Month

ESnet is not the Commercial Internet





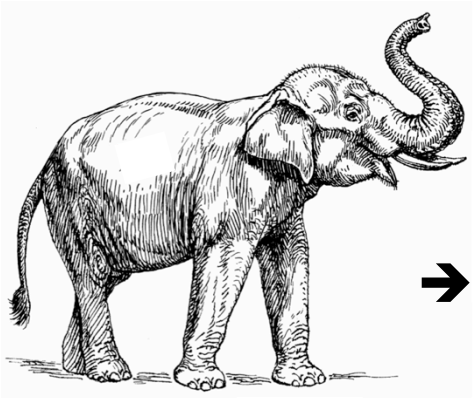




Elephant Flows Place Great Demands on Networks

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

Result
99.9954% of water transferred.



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

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

essentially fixed

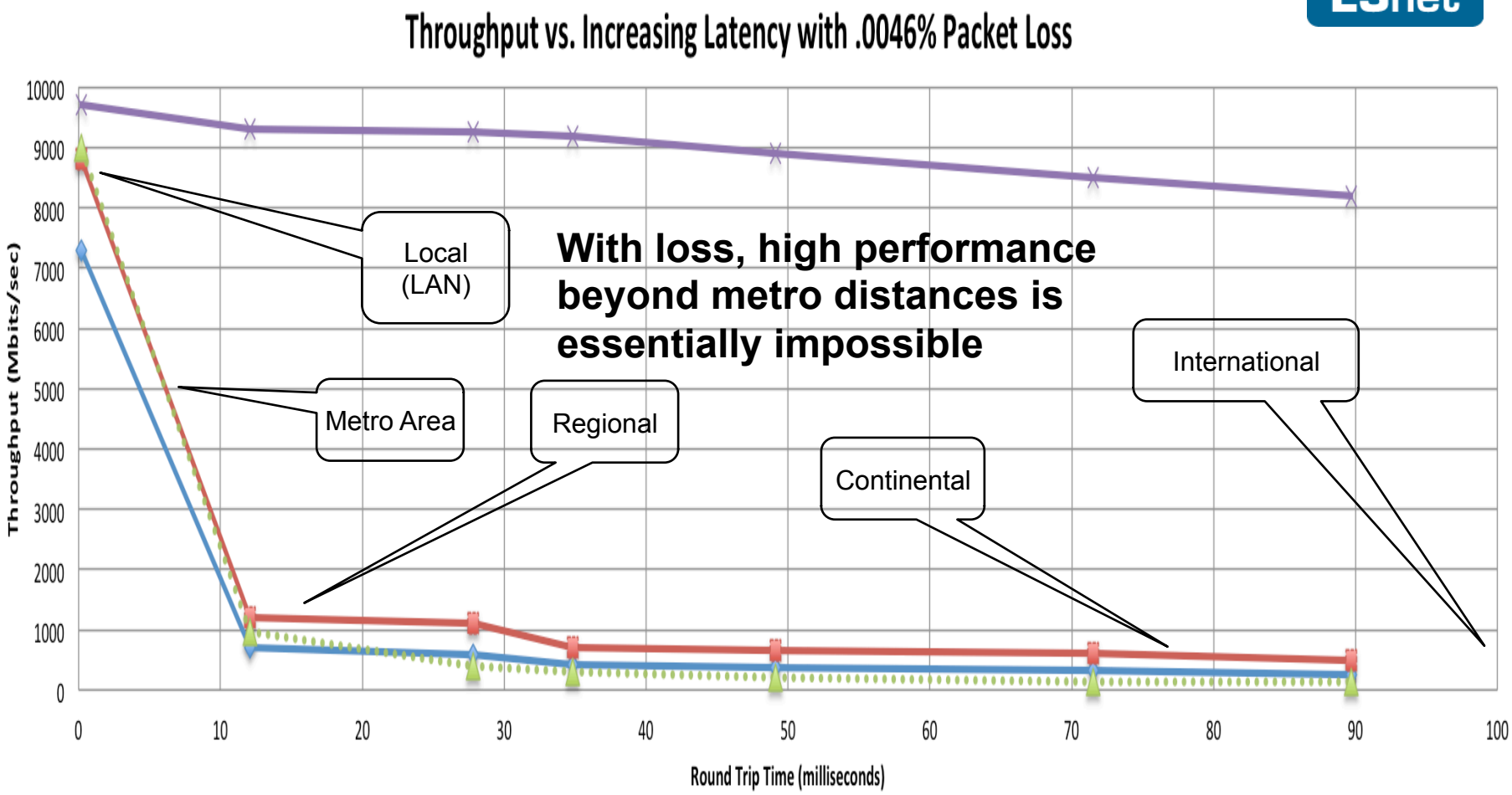
determined by speed of light

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

With proper 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.

A small amount of packet loss makes a huge difference in TCP performance



Working With TCP In Practice

Far easier to support TCP than to fix TCP

- People have been trying to fix TCP for years – limited success
- Like it or not we're stuck with TCP in the general case

Pragmatically speaking, we must accommodate TCP

- Sufficient bandwidth to avoid congestion
- Zero packet loss
- Verifiable infrastructure
  - Networks are complex
  - Must be able to locate problems quickly
  - Small footprint is a huge win – small number of devices so that problem isolation is tractable

HPC facilities must also do this – "the other end" must also do this

Putting A Solution Together

Effective support for TCP-based data transfer

- Design for correct, consistent, high-performance operation
- Design for ease of troubleshooting

Easy adoption is critical

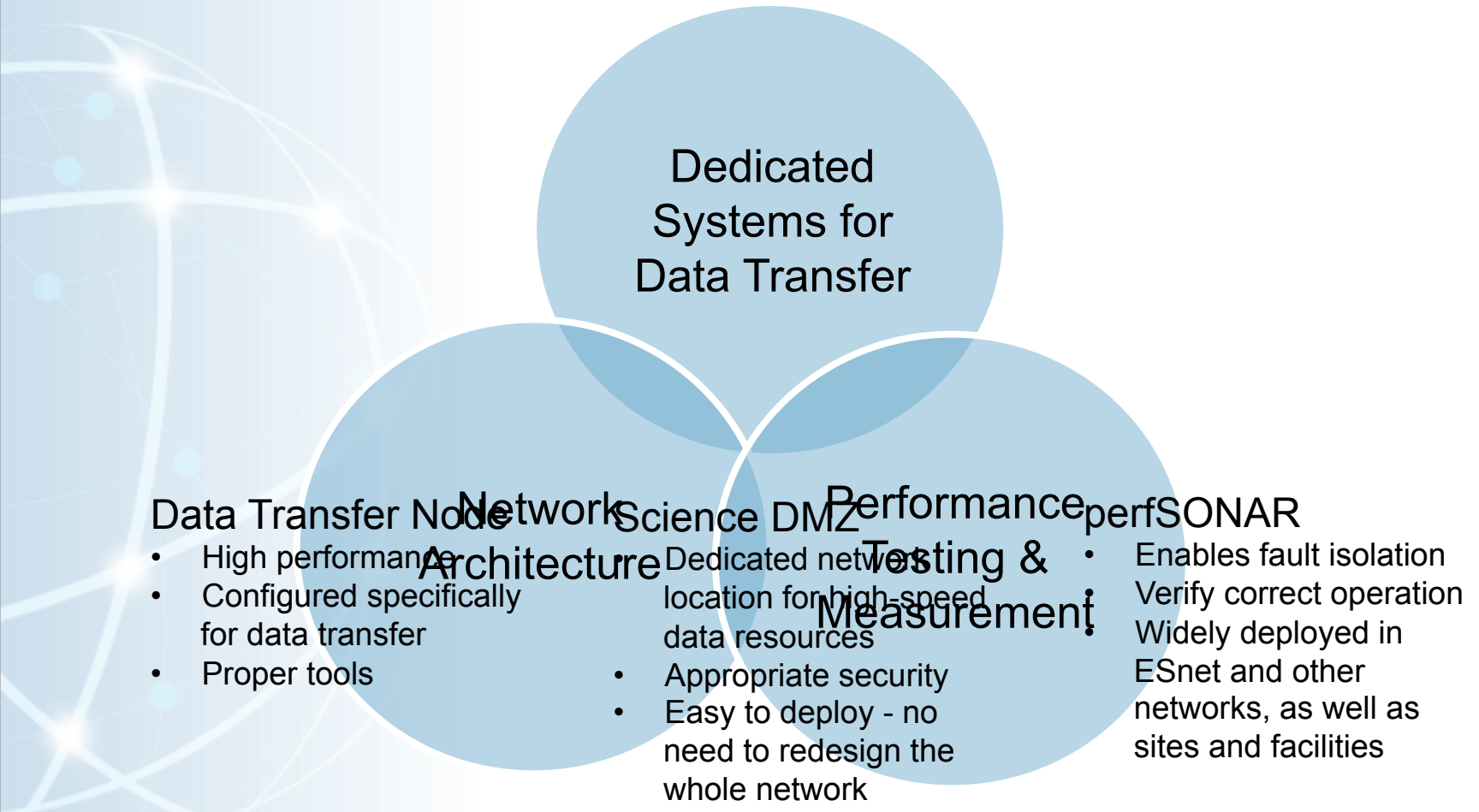
- Large laboratories and universities have extensive IT deployments
- Drastic change is prohibitively difficult

Cybersecurity – defensible without compromising performance

Borrow ideas from traditional network security

- Traditional DMZ – separate enclave at network perimeter ("Demilitarized Zone")
  - Specific location for external-facing services
  - Clean separation from internal network
- Do the same thing for science – ***Science DMZ***

The Science DMZ Design Pattern



Data Transfer Node
- High performance
- Configured specifically for data transfer
- Proper tools

Science DMZ
- Dedicated network location for high-speed data resources
- Appropriate security
- 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

Science DMZ Implications

Many Science DMZs have been or are being deployed

- DOE laboratories and facilities
- Supercomputer centers (e.g. XSEDE)
- Many tens of universities
  - NSF CC-NIE and CC*IIE programs (over $40M and still counting)
  - LHC infrastructure
- International
  - Australia (~$50M program)
  - Europe
  - Brazil

This is how many of your users will move data to/from your systems

- Data transfer nodes
- Globus
- High-speed connections from Science DMZ to research networks

Sample Data Transfer Rates

Data set size	1 Minute	5 Minutes	20 Minutes	1 Hour
	Time to transfer			
10PB	1,333.33 Tbps	266.67 Tbps	66.67 Tbps	22.22 Tbps
1PB	133.33 Tbps	26.67 Tbps	6.67 Tbps	2.22 Tbps
100TB	13.33 Tbps	2.67 Tbps	666.67 Gbps	222.22 Gbps
10TB	1.33 Tbps	266.67 Gbps	66.67 Gbps	22.22 Gbps
1TB	133.33 Gbps	26.67 Gbps	6.67 Gbps	2.22 Gbps
100GB	13.33 Gbps	2.67 Gbps	666.67 Mbps	222.22 Mbps
10GB	1.33 Gbps	266.67 Mbps	66.67 Mbps	22.22 Mbps
1GB	133.33 Mbps	26.67 Mbps	6.67 Mbps	2.22 Mbps
100MB	13.33 Mbps	2.67 Mbps	0.67 Mbps	0.22 Mbps

This table available at:

http://fasterdata.es.net/fasterdata-home/requirements-and-expectations/

User Requirements

What have users asked for?

- 2Gbps, 5Gbps, 10Gbps ***per experiment***
  - Light sources
  - Microscopy
- 1PB/week data replication – cosmological simulations
- Petascale data replication – climate science

To from a capacity perspective, the network part is done

- 1PB/week = 14Gbps in round numbers
- Light source workflows are bursty – lots of them fit in 100G

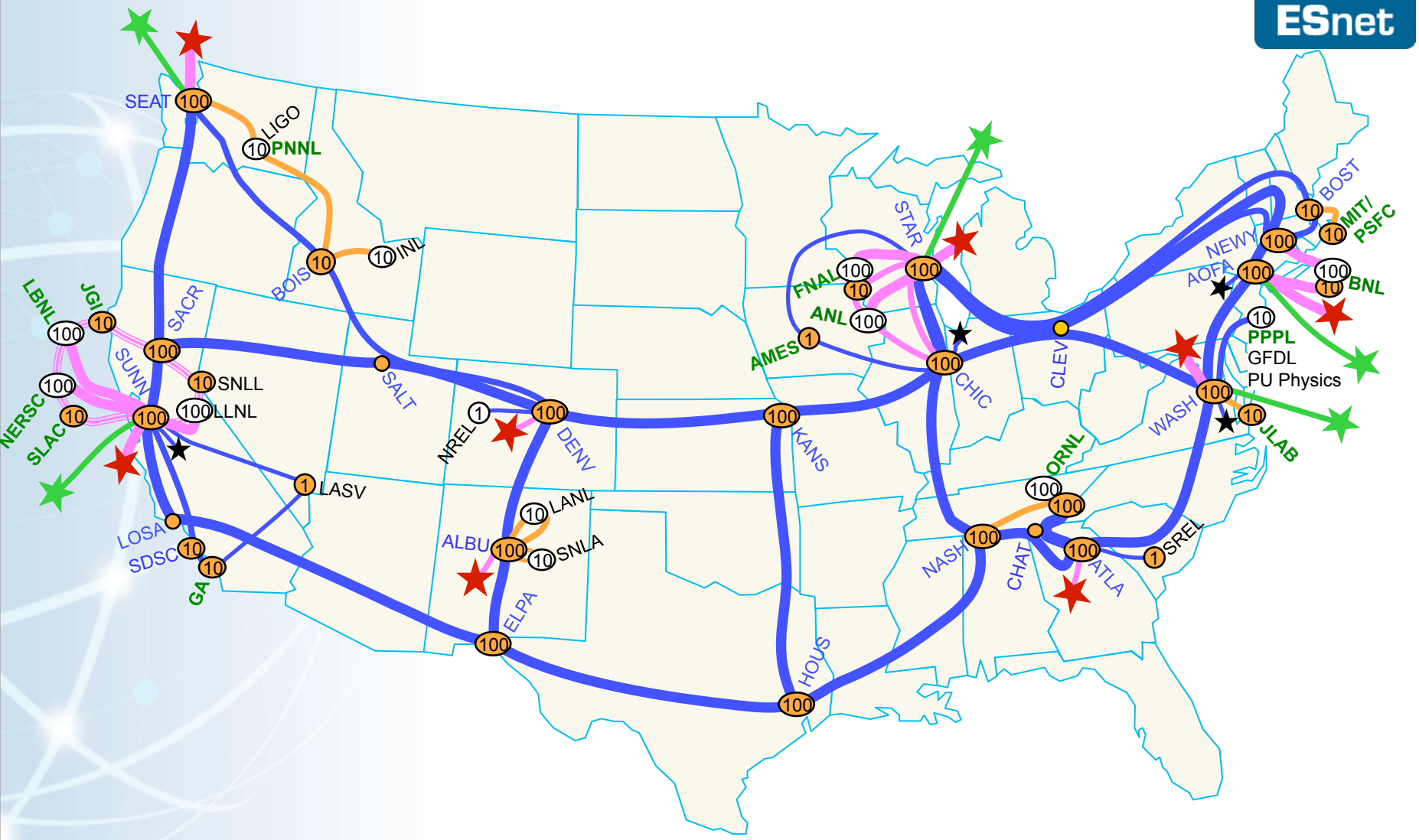
Effective use of the network is now gated on two things

- End system resources, particularly storage (but tools too)
- Squeaky-clean network path (zero packet loss)

Current ESnet5 Capabilities

- 100G connections to DOE/ASCR computing facilities (ANL, NERSC, ORNL)
- 100G connections to additional labs (BNL, FNAL, LBNL, LLNL)
- 100G connections to more sites soon (LANL, SLAC)
- 100G connections to major peering exchanges
  - MANLAN (New York)
  - OmniPOP (Chicago)
  - Starlight (Chicago)
  - WIX (Washington, DC)
  - Pacific Wave coming soon (US West Coast)
- 100G connections to major science networks (CENIC, Internet2)
- Performance Assurance using perfSONAR (we watch for performance problems and we fix them)
- Multiple 10G Data Transfer Nodes for end-to-end testing of disk to disk transfers
- Current 100G core runs at 5% to 20% load, with bursts to 40%

ESnet5 Map


SEAT
LIGO
PNNL
BOIS
INL
LBNL
JGI
SACR
SUNN
SNLL
LLNL
NERSC
SLAC
SALT
LASV
LOSA
SDSC
GA
NREL
DENV
LANL
ALBU
SNLA
ELPA
KANS
HOUS
FNAL
ANL
AMES
STAR
CHIC
CLEV
NEWY
AOFA
BOST
MIT/
PSFC
BNL
PPPL
GFDL
PU Physics
WASH
JLAB
ORNL
NASH
CHAT
ATLA
SREL

EEX – Future Expansion of the ESnet network into Europe

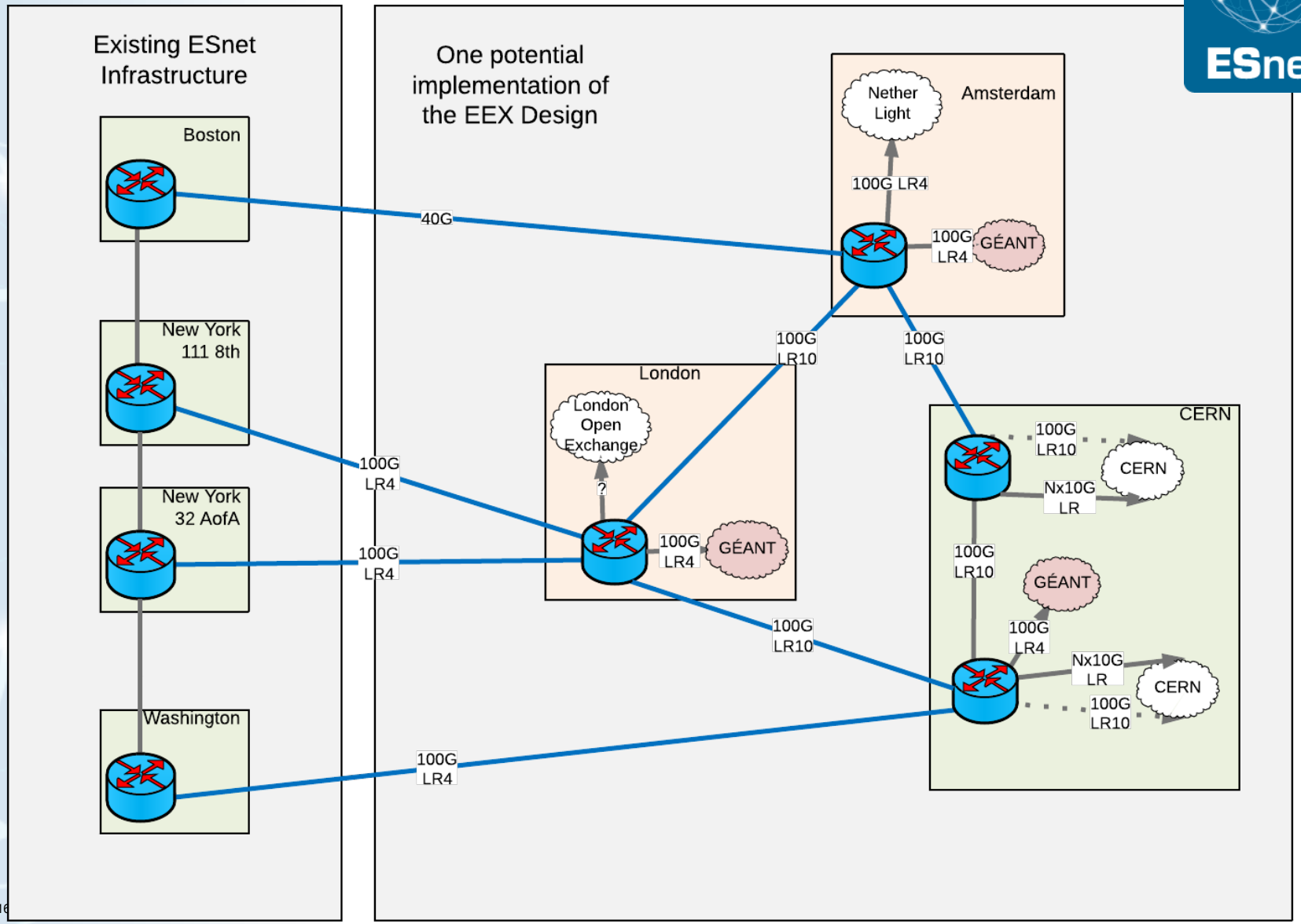
Drivers: Improve the quality and quantity of transit to Europe for:

- All currently supported ESnet programs
- All US LHC science

Proposed plan: extend ESnet5 geography

- Hubs at London, Amsterdam and 2 at CERN
- Links
  - 100G Ring in Europe
  - 100G T/A - WASH-CERN, AOFA-LOND, NEWY-LOND
  - 40G T/A - BOST-AMS
- Peering connections to GEANT, Netherlight & London Open

One Possible EEX Design



Implications of EEX for ESnet constituents

To first order, the ESnet 100G core would be extended into Europe

Significant upgrade to existing capability

- Current capacity is showing its limitations
  - 3x10GE from New York to Amsterdam
  - 3x10GE from Washington to Frankfurt
- EEX plans include diverse 100GE with 40G failover

Significant improvement in operational model

- Current connectivity is shared between many players
  - Different players at each end
  - Very difficult to reach resolution on problems (months!)
- EEX would be run by ESnet as part of the ESnet Facility

Future Upgrades

ESnet is already planning for core bandwidth upgrades

- Anticipation of LHC Run 2 in early 2015
- Increased demand from HPC facilities, light sources, etc
- Remember the loss plot – we have to stay ahead

Upgrades will occur where they are needed

- SF Bay Area
- Chicago / New York / Washington (major LHC footprint)

If you know of significant needs, please tell us ASAP

- If you are planning to support something, ESnet should be planning too
- If we don't know about it, we can't help you
- Some procurements take months

It is ESnet's intent to stay ahead of demand – we don't like being the bottleneck

Other Players

NSF is currently soliciting proposals for IRNC (international connectivity)

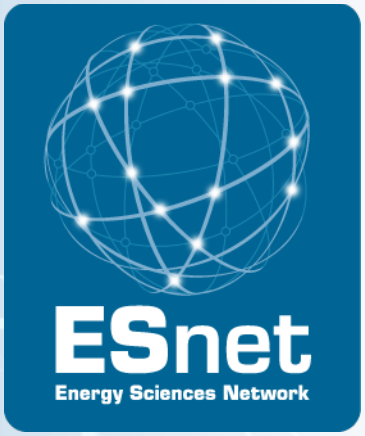
- This funds much of the current international science connectivity for the US
- Expect increased bandwidth to Asia, Latin America in the coming years

GEANT (Pan-European network) has 100G core now

- Many European NRENs (national networks) are 100G also
  - We get to them via GEANT
  - France
  - Germany
  - Netherlands
  - UK
- EEX will provide robust connectivity

Many, many science DMZs – folks at “the other end” are really upping their game

All of this comes to a HPC facility at its DTNs



Questions?

Thanks!

Eli Dart - dart@es.net

http://www.es.net/

http://fasterdata.es.net/