ESnet

ENERGY SCIENCES NETWORK

The Energy Sciences Network: Overview, Update, Impact

ASCAC Meeting

American Geophysical Union

Washington, DC

March 24, 2015

Gregory Bell, Ph.D.

Director, Energy Sciences Network (ESnet)

Director, Scientific Networking Division

Lawrence Berkeley National Laboratory

Overview

Update

Impact

ESnet

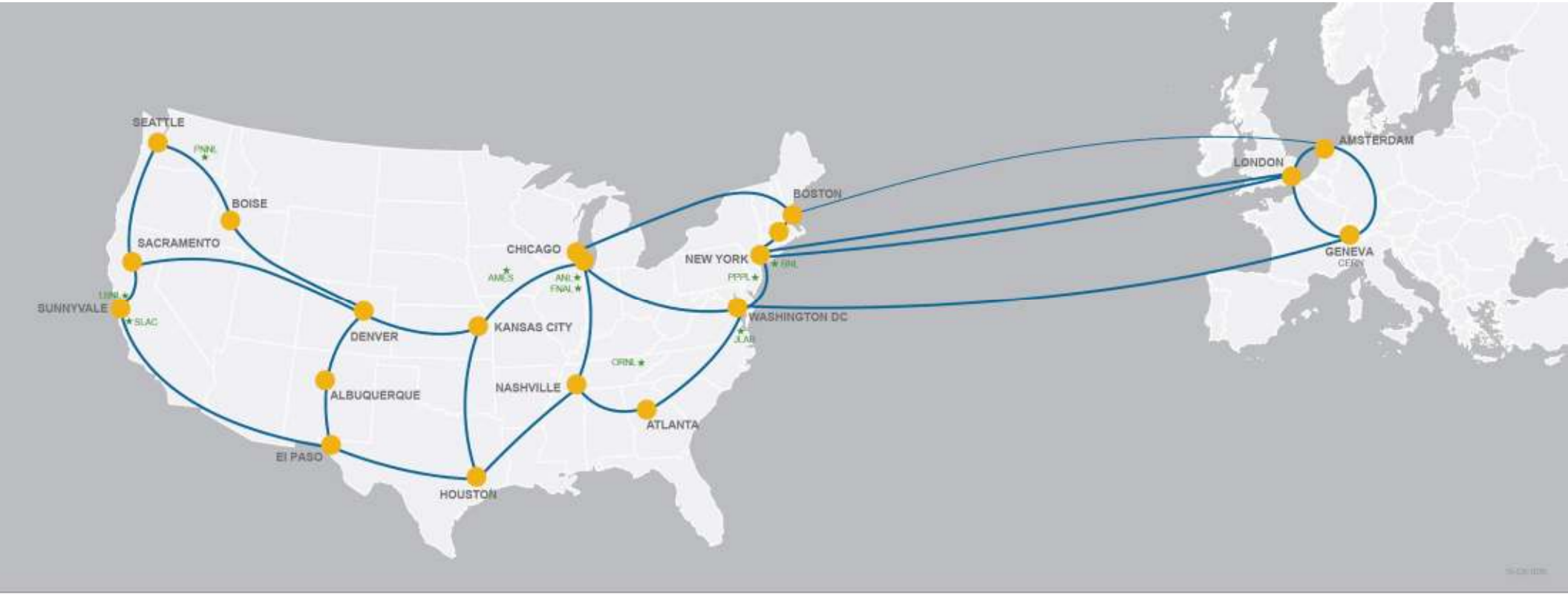
Overview

Update

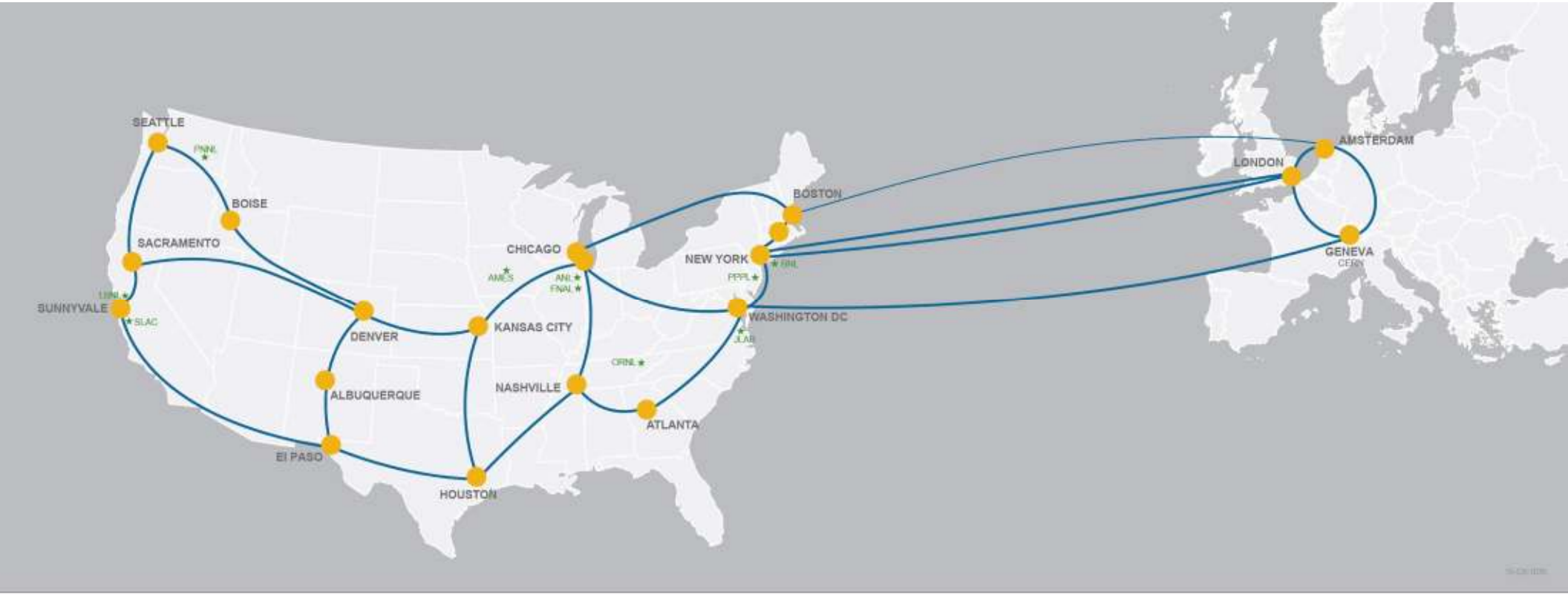
Impact

ESnet

This is not an ISP.

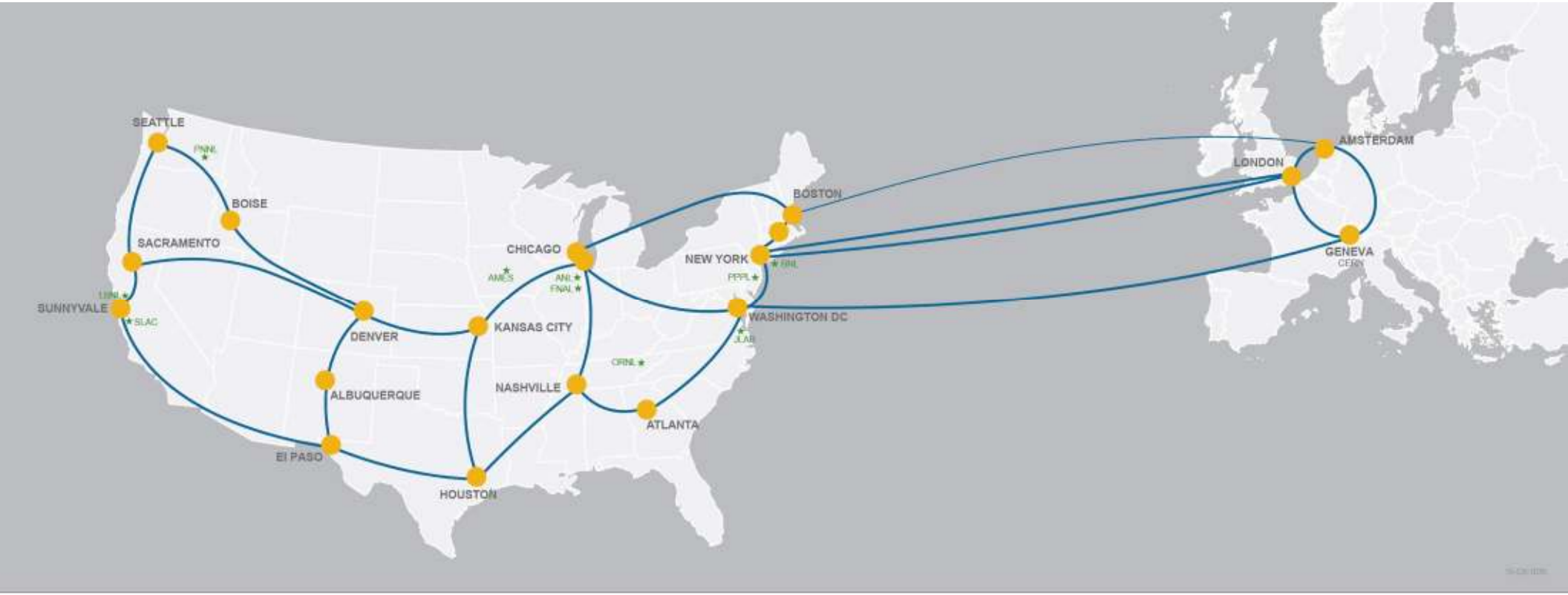


It's a DOE user facility engineered to overcome the constraints of geography.



We do this by offering unique capabilities, and optimizing the facility for data acquisition, data placement, data sharing, data mobility.

It's a DOE user facility engineered to overcome the constraints of geography.



ESnet's superficial resemblance to an international ISP is deceptive, and arguably a risk.

Our vision:

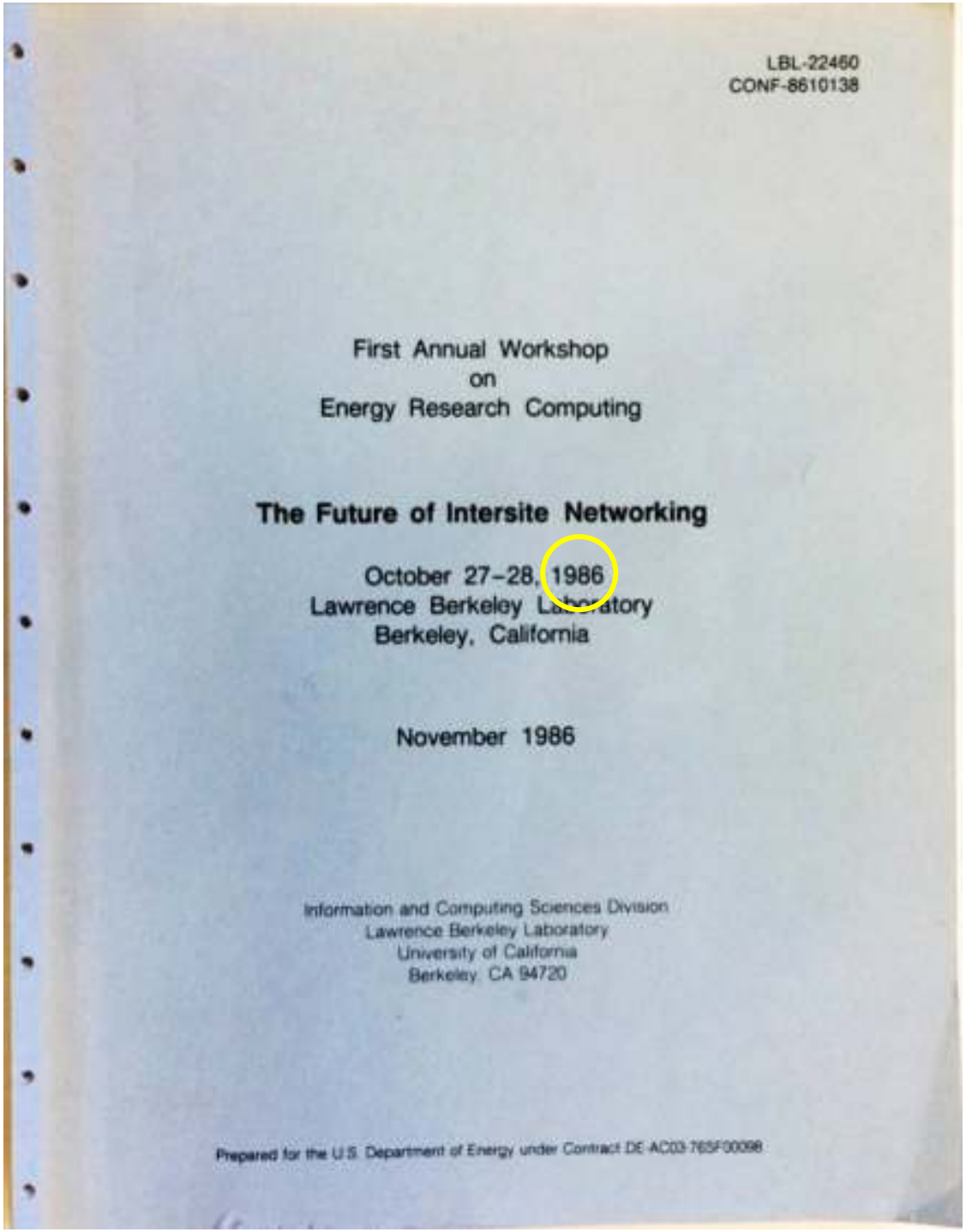
Scientific progress will be **completely unconstrained** by the physical location of instruments, people, computational resources, or data.

ESnet

This is not a *new* way of describing ESnet.

1. "What we can do on **LANs** today is indicative of what we wish to be able to do on **wide area networks**."

2. "Just as we expect a **computer** to perform as if we are the **only user**, we expect the **network** to give that same appearance."



LBL-22460
CONF-8610138

First Annual Workshop
on
Energy Research Computing

The Future of Intersite Networking

October 27–28, 1986
Lawrence Berkeley Laboratory
Berkeley, California

November 1986

Information and Computing Sciences Division
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

The basic facts (new or notable):

High-speed international networking facility, optimized for DOE science missions:

- connecting 50 labs, plants and facilities with >150 networks, universities, research partners globally
- 340Gbps transatlantic extension in production (Dec 2014)
- university connections to better serve LHC science
- $35M in FY15, 42FTE
- older than commercial Internet, growing ~twice as fast
- the DOE user facility that serves all others

$62M ARRA grant funded 100G upgrade in 2011:

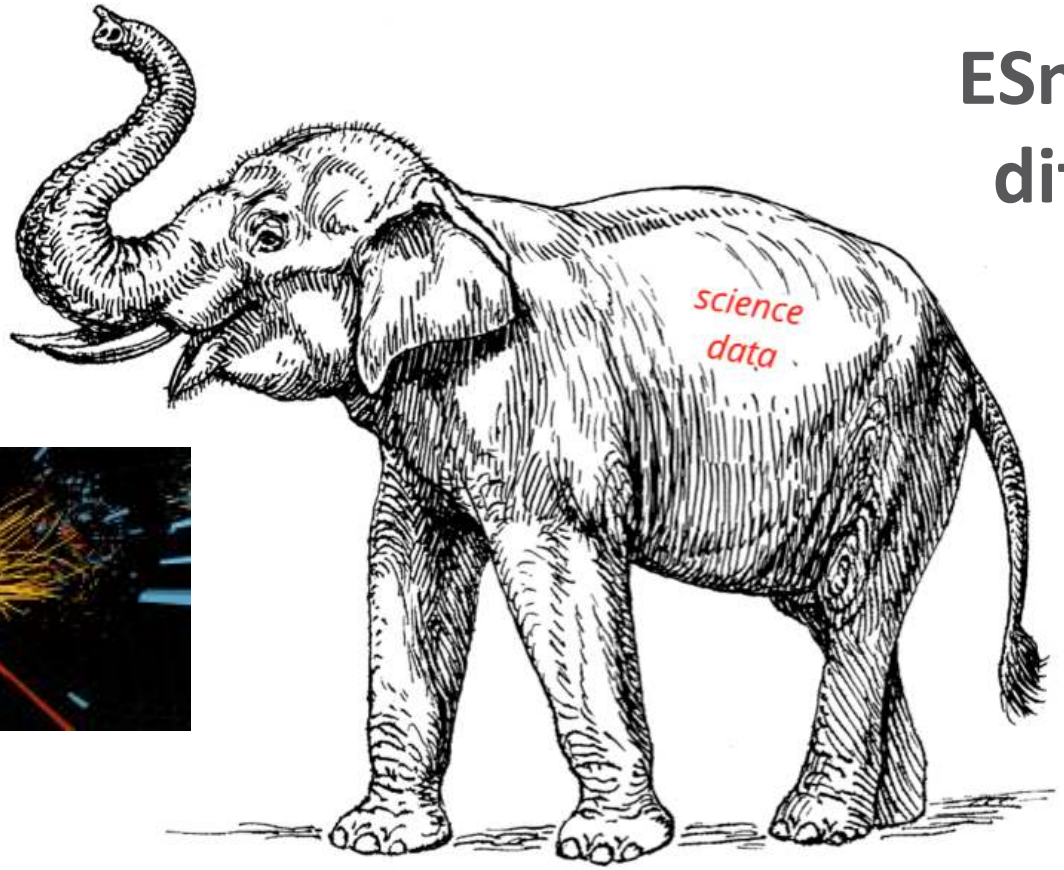
- fiber assets + access to spectrum, shared with Internet2
- new era of optical networking, abundant capacity
- world's first 100G network at continental scale

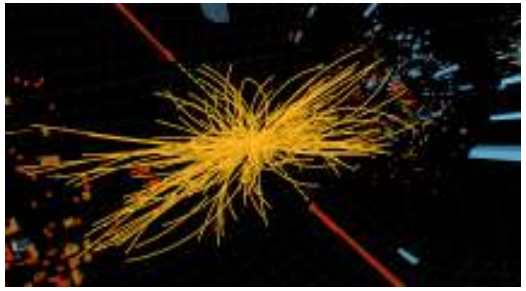
Culture of urgency:

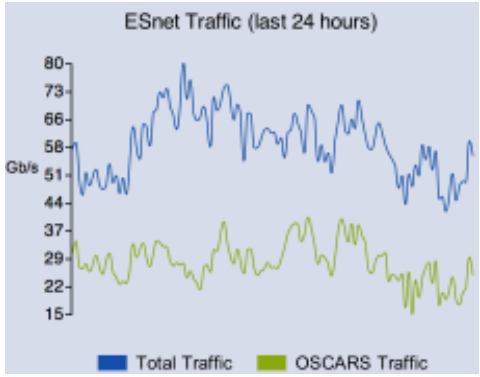
- several recent awards
- 80% engineers, highly motivated

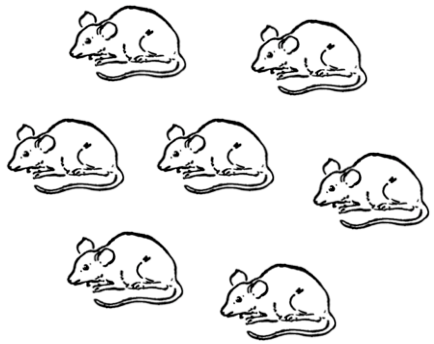
ESnet

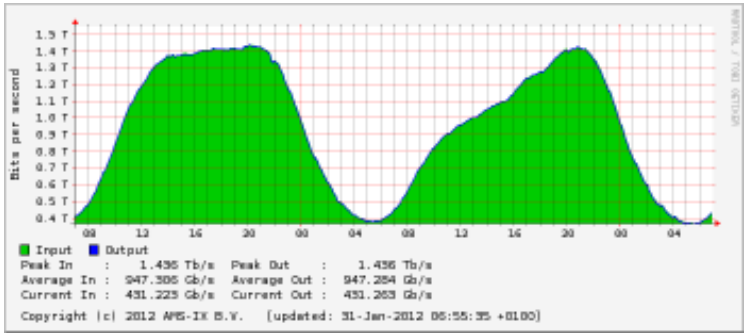
ESnet is designed for different goals than general Internet.



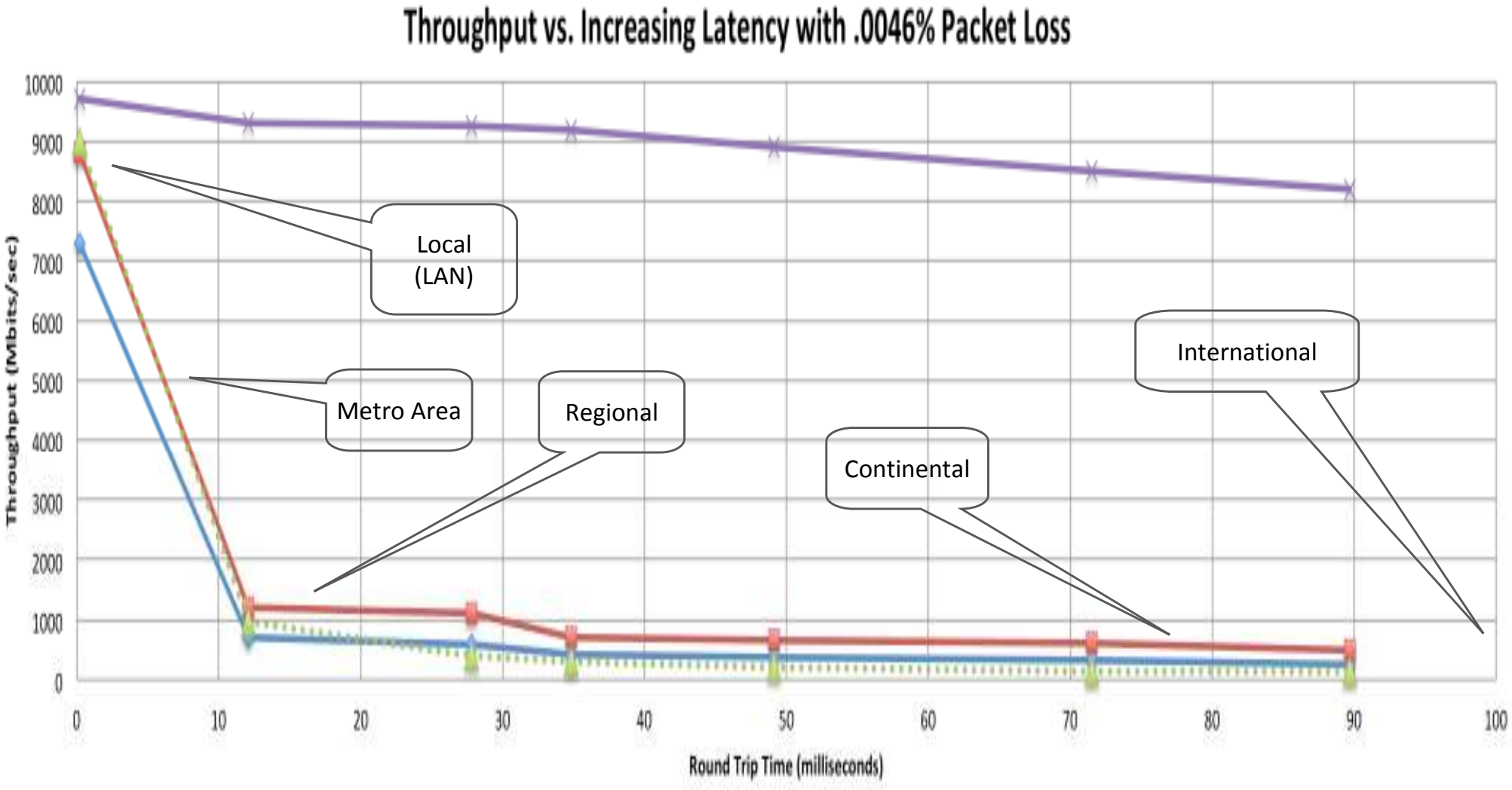




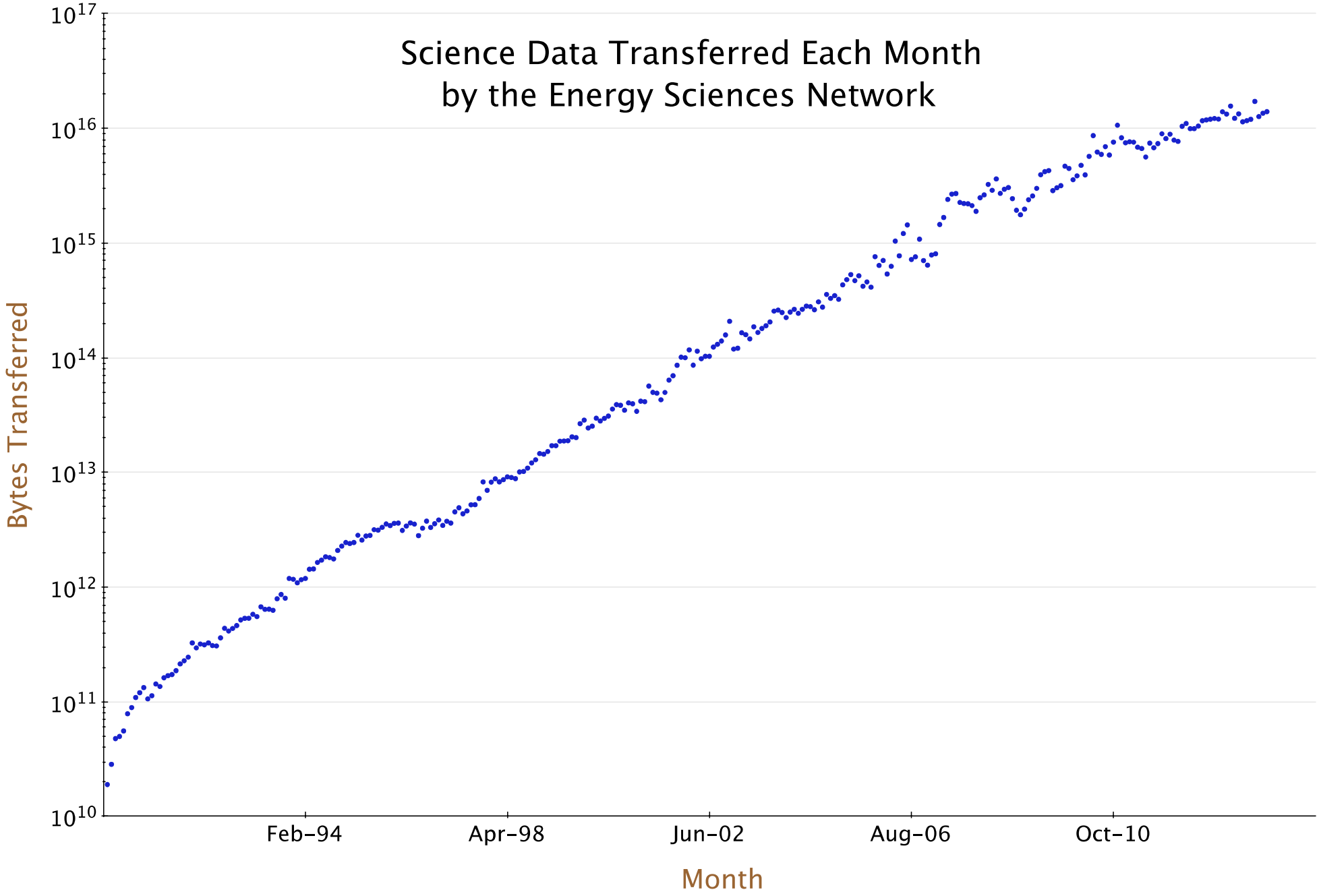


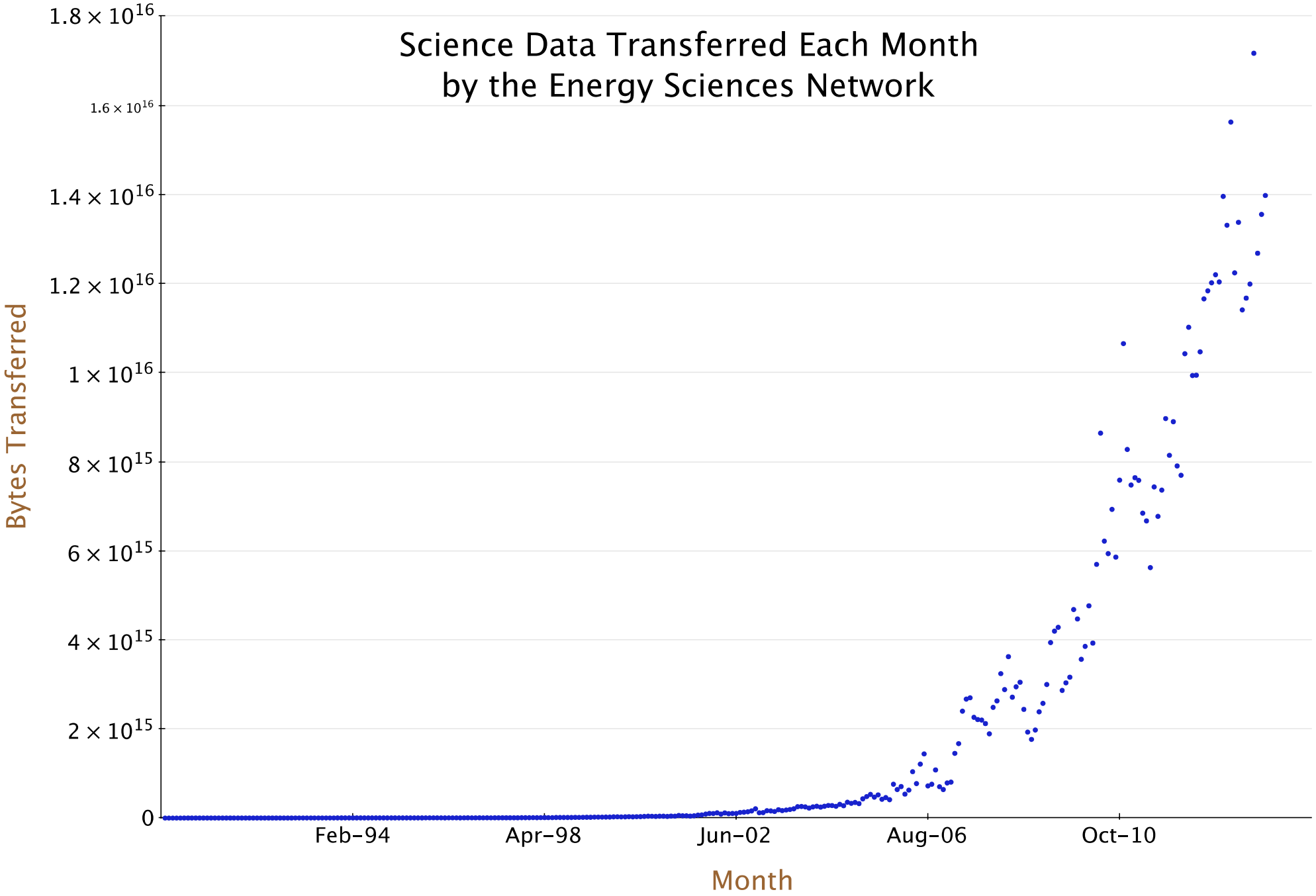


Elephant flows require almost *lossless* networks.

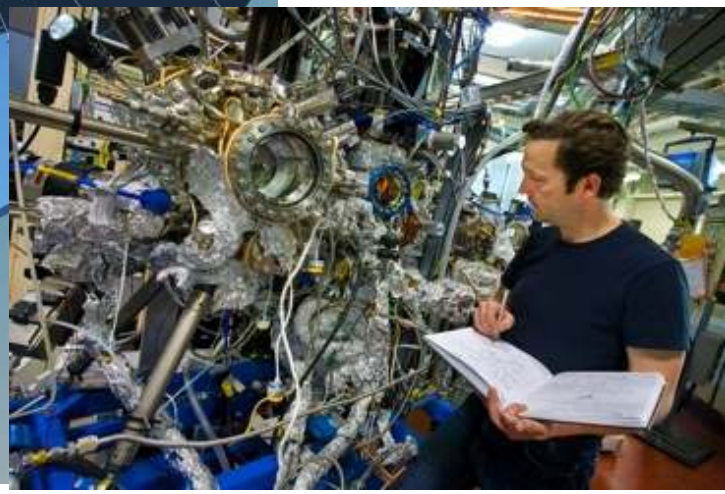


See Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, and Jason Zurawski. The Science DMZ: A Network Design Pattern for Data-Intensive Science. In *Proceedings of the IEEE/ACM Annual SuperComputing Conference (SC13)*, Denver CO, 2013.


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}
Feb-94
Apr-98
Jun-02
Aug-06
Oct-10
Month


Science Data Transferred Each Month
by the Energy Sciences Network
Bytes Transferred
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
Month

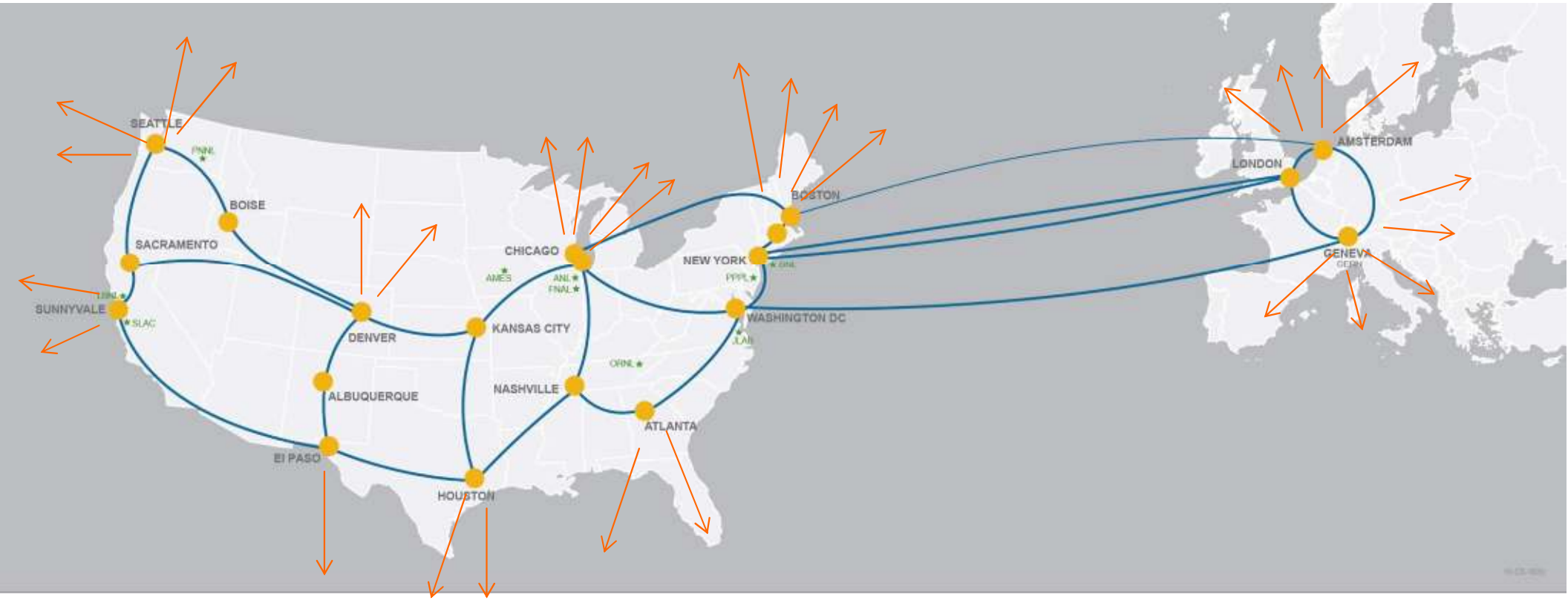

MinION



sensors and instruments.

ESnet

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



ESnet
ENERGY SCIENCES NETWORK

★ Department of Energy Office of Science National Labs

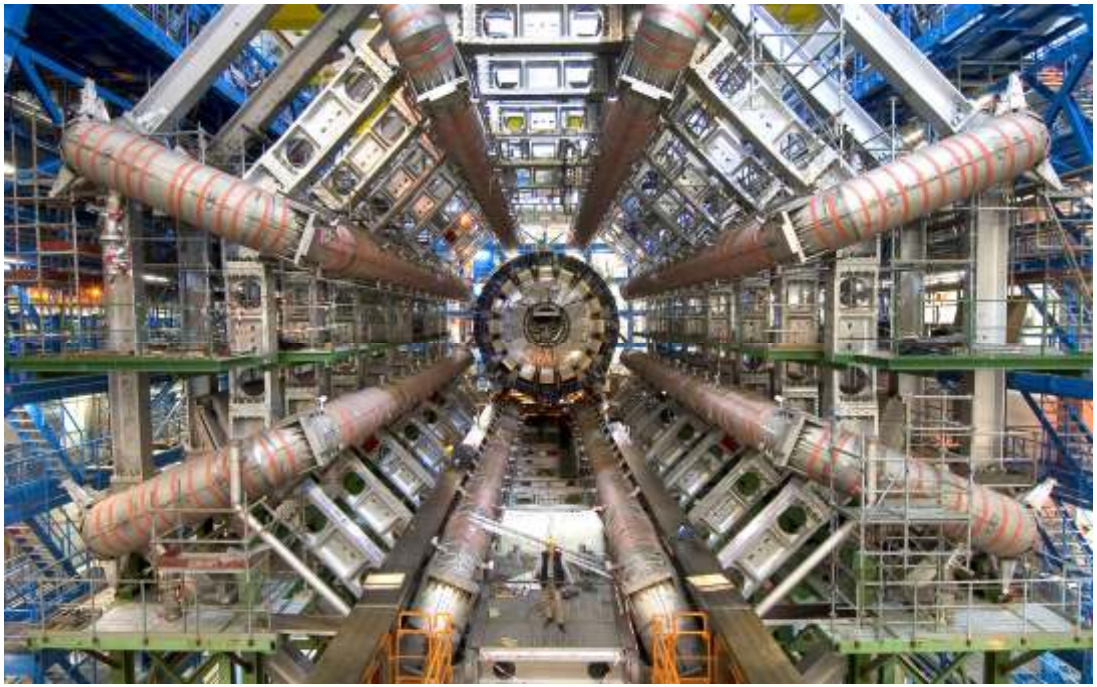
Ames	Ames Laboratory (Ames, IA)
ANL	Argonne National Laboratory (Argonne, IL)
BNL	Brookhaven National Laboratory (Upton, NY)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
LBNL	Lawrence Berkeley National Laboratory (Berkeley, 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)

In a nutshell:

- Data intensive science inevitably drives network intensity.
- DOE traffic continues to grow exponentially.
- Data ‘point sources’ are becoming more numerous, less expensive.
- DOE data flows typically include universities, global collaborations.
  - only 20% of ESnet data flows are DOE↔DOE

ESnet

Now, a few brief examples of ESnet's role in DOE workflows, starting with LHC science.



ESnet

Evolution of LHC data model:

In chronological order

1. Copy as much data as is feasible to analysis centers worldwide, with hierarchical distribution scheme ('Monarc' model, deterministic flows).
2. Relax the hierarchy, and rely on optimistic caching.
3. Use 'federated data stores' to fetch *portions* of relevant data sets from remote storage (anywhere), just before they're needed.

This evolution implies growing faith in networks, growing opportunity for Software Defined Networking.

ESnet

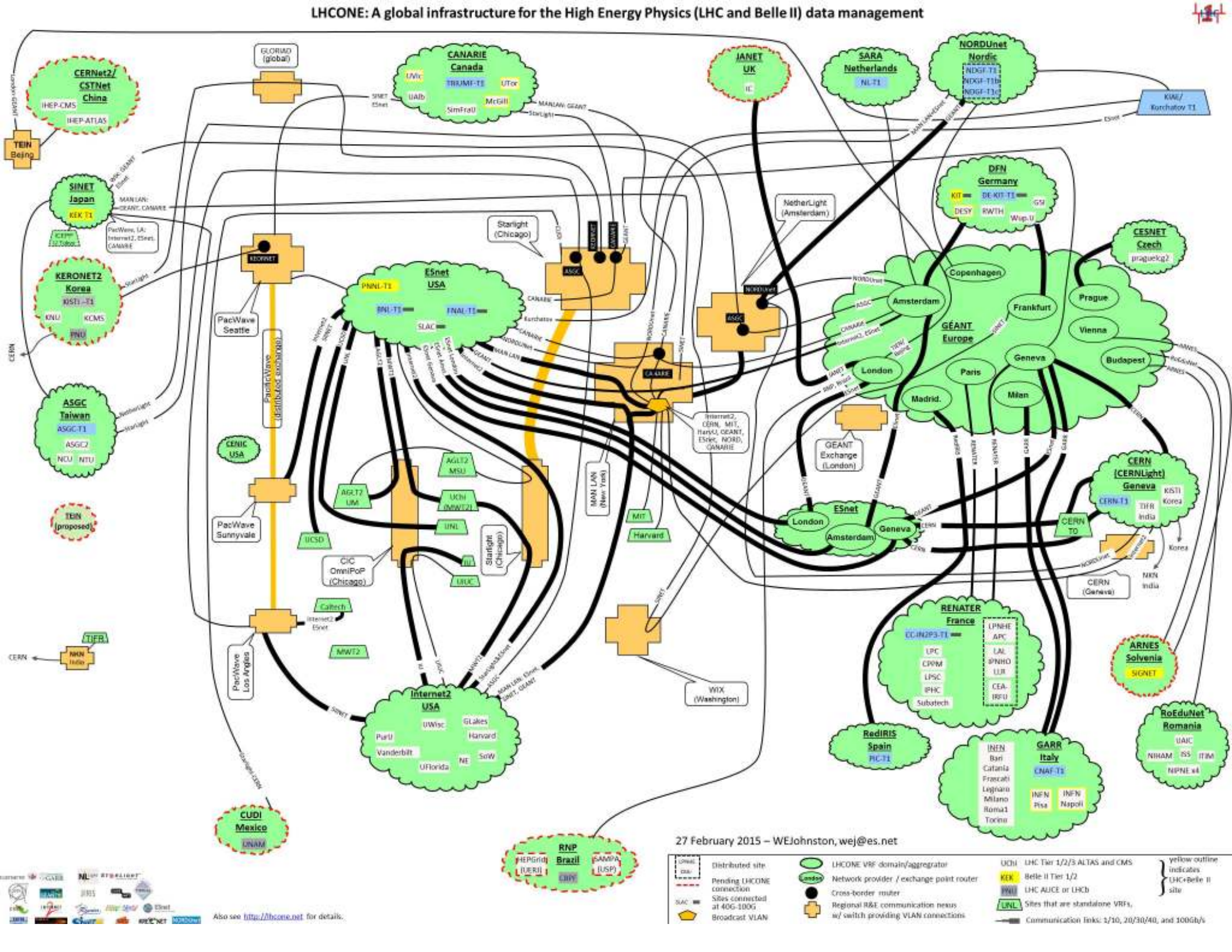
To support this evolving data model, research networks have built a vast global overlay.

LHC Open Network Environment (LHCONE):

- dedicated and isolated network overlay for LHC experiments
- gives consistent, high-performance access for LHC computing centers
  - extensive use of virtual circuits
- 30 networks (with ESnet as core participant), dozens of universities
- an international highway system optimized for LHC flows
  - custom global instrument, but also a *collaboration*

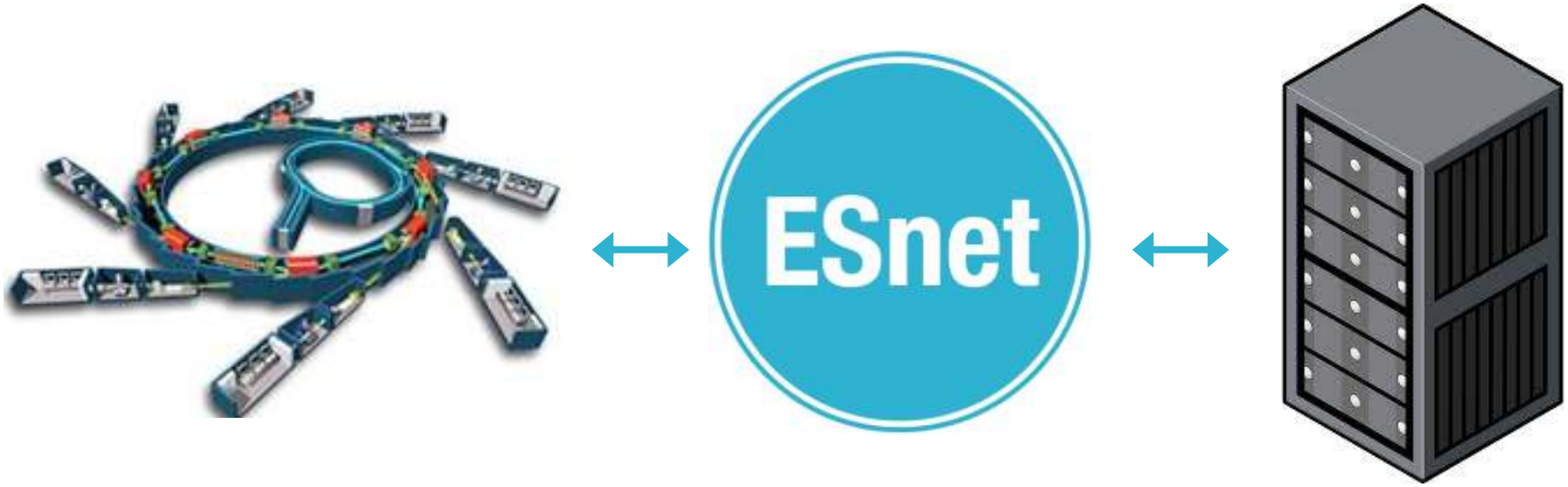
ESnet

LHCONE: A global infrastructure for the High Energy Physics (LHC and Belle II) data management



This architecture (instruments and computational resources coupled by networks) now spreading outside HEP: ‘super-facilities.’

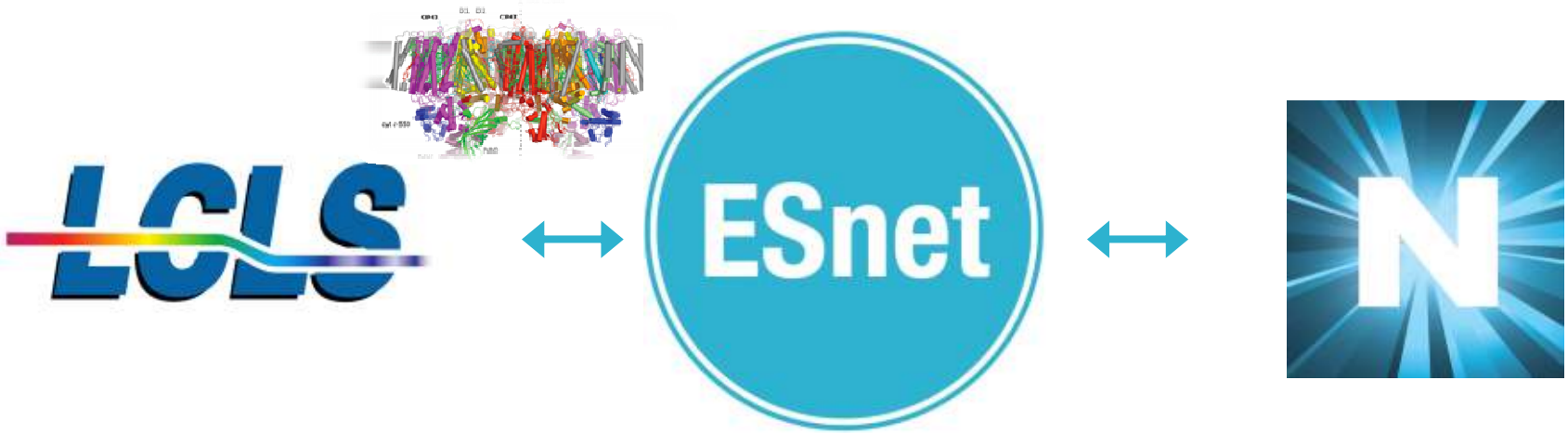
Experimental facilities are being transformed by new detectors, advanced mathematics, robotics, automation, advanced networks.

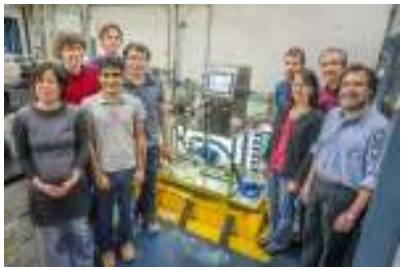


ESnet

Super-facility example #1:

Researchers from Berkeley Lab and SLAC conducted protein crystallography experiments at LCLS to investigate photoexcited states of PSII, with near-real-time computational analysis at NERSC.

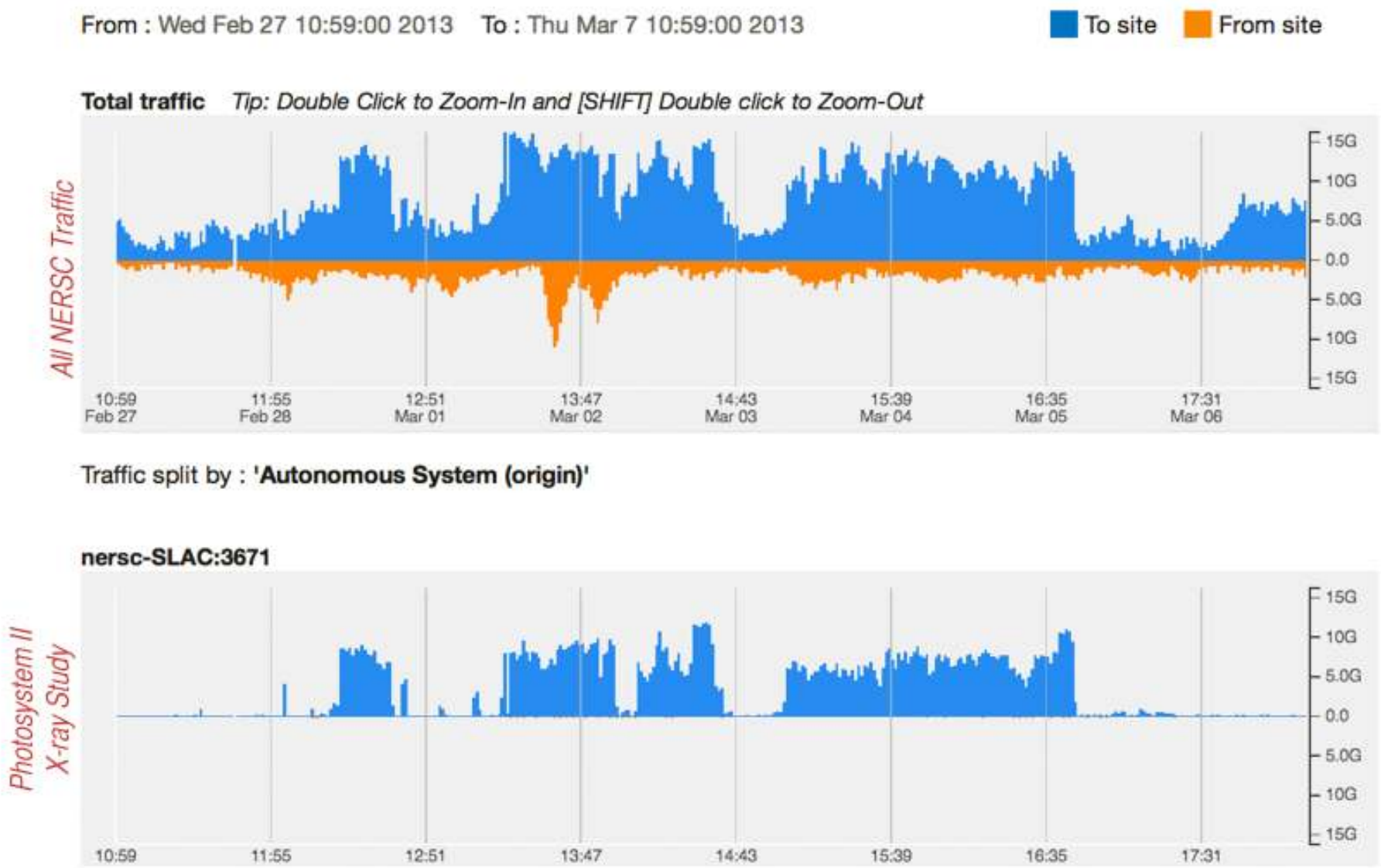




"Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy," *Nature Communications* 5, 4371 (9 July 2014)

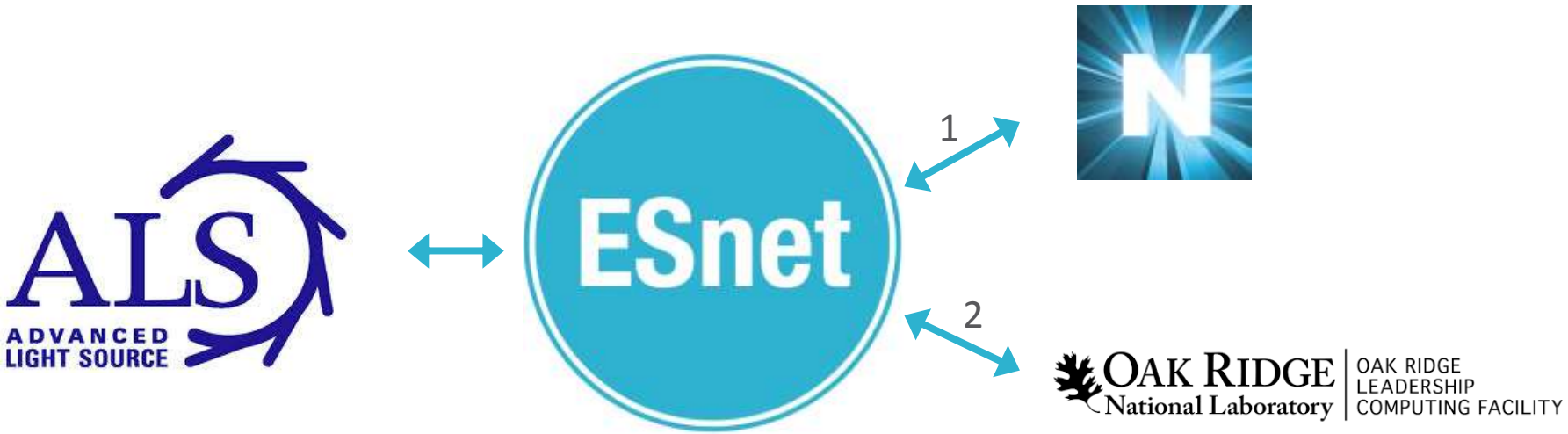
ESnet

Data flow from single LCLS detector *tripled* network utilization for major HPC center.



Super-facility example #2:

Real-time analysis of 'slot-die' technique for printing organic photovoltaics, using ALS + NERSC (SPOT Suite for reduction, remeshing, analysis) + OLCF (HipGISAXS running on Titan w/ 8000 GPUs).



http://www.es.net/news-and-publications/esnet-news/2015/esnet-paves-way-for-hpc-superfacility-real-time-beamline-experiments/

Results presented at March 2015 meeting of American Physical Society by Alex Hexemer.

Additional DOE contributions: **GLOBUS** (ANL), **CAMERA** (Berkeley Lab)

ESnet

Super-facility-on-demand demo at NSF GENI conference *tomorrow*.

SPADE instance - server at Argonne

ESnet, Internet2

ExoGENI SPADE VM @ Starlight, Chicago

ESnet

ExoGENI SPADE VM @ Oakland, California

Data from ALS beamline

- fictional - but realistic - workflow
- **dedicated systems for data transfer** and **network circuits** created *programmatically*
- future vision: application declares *intention* for super-facility, network responds
- "Science DMZ as a service"

Compute Cluster NERSC, LBL

ESnet

http://portal.nersc.gov/project/als/sc14/

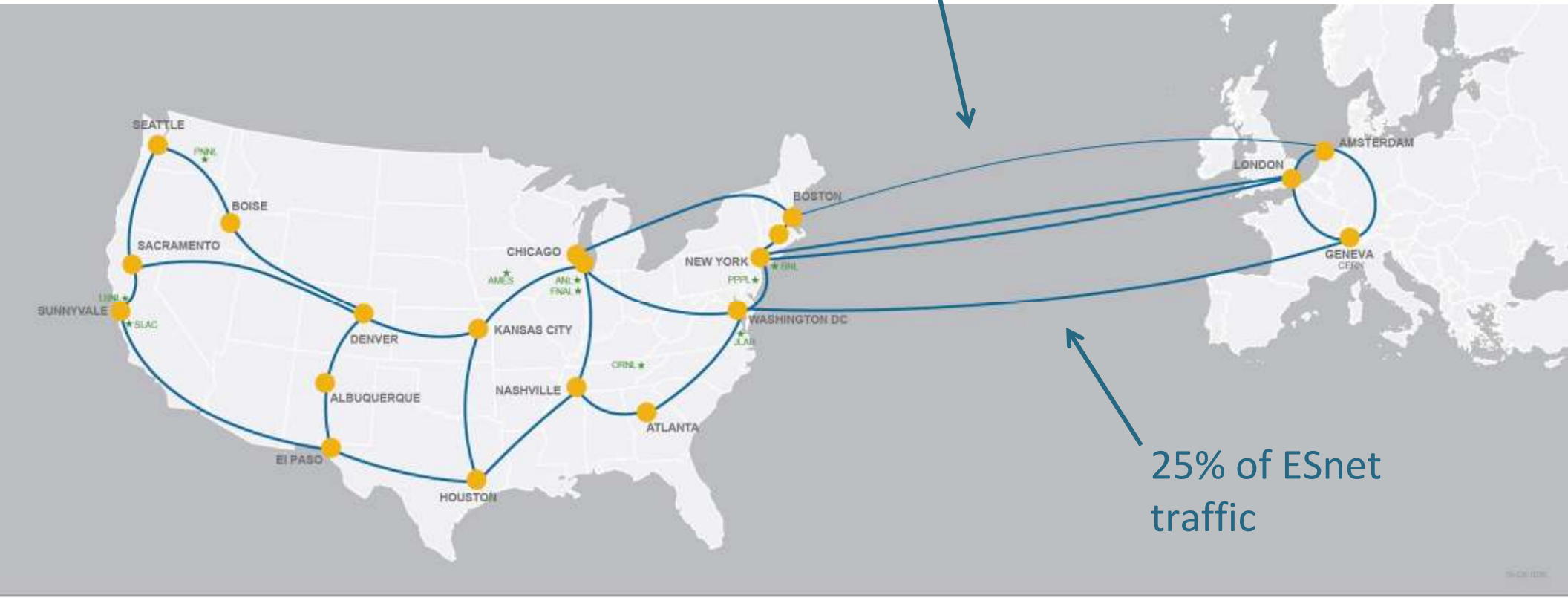
Overview

Update

Impact

ESnet

340Gbps extension. Why?



Immediate driver: more capacity for LHC Run 2. But the extension supports all DOE missions, reducing barriers to European collaborators / instruments.

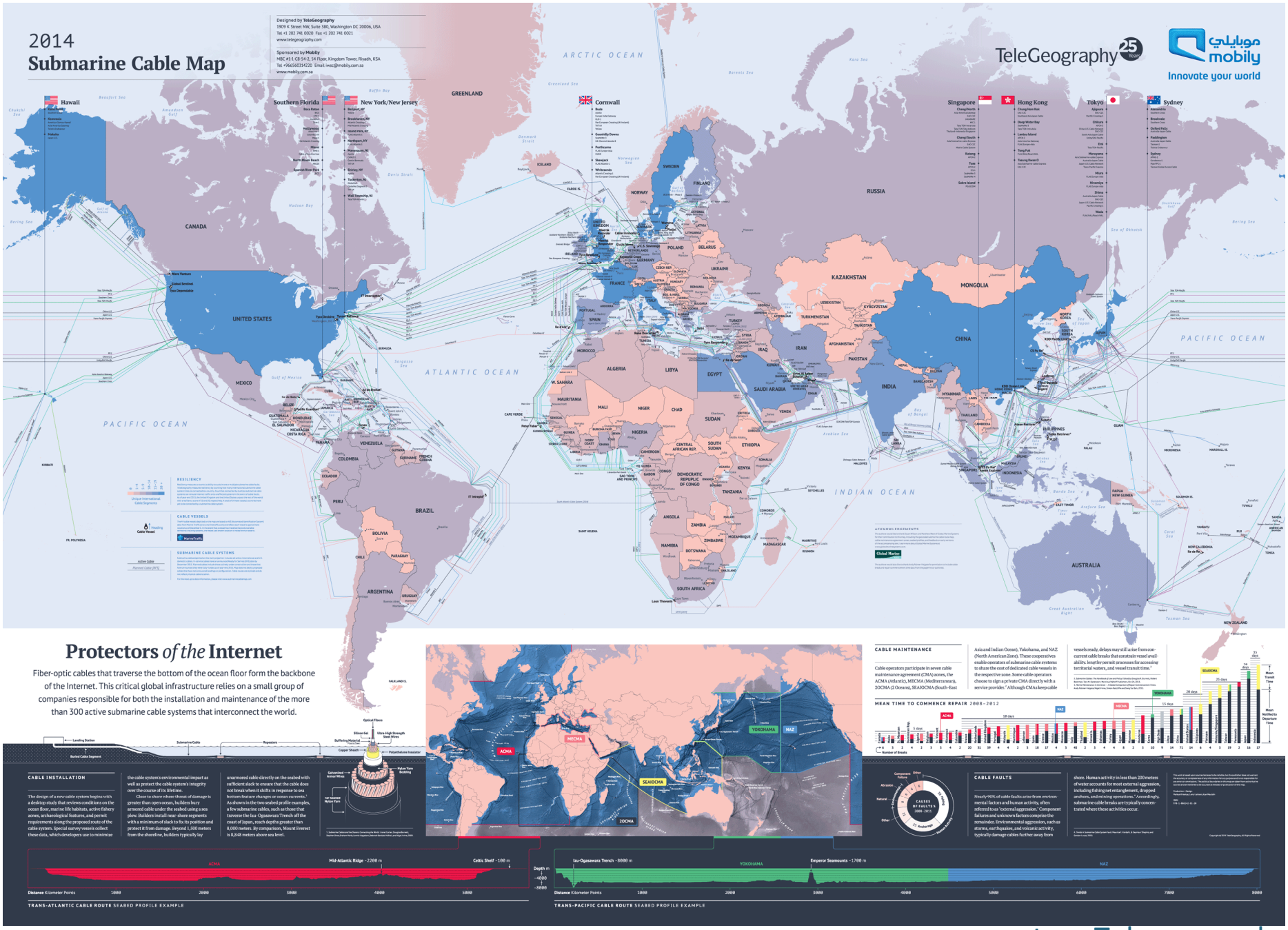
How? [short detour]

Internet's physical substrate, optical fiber…

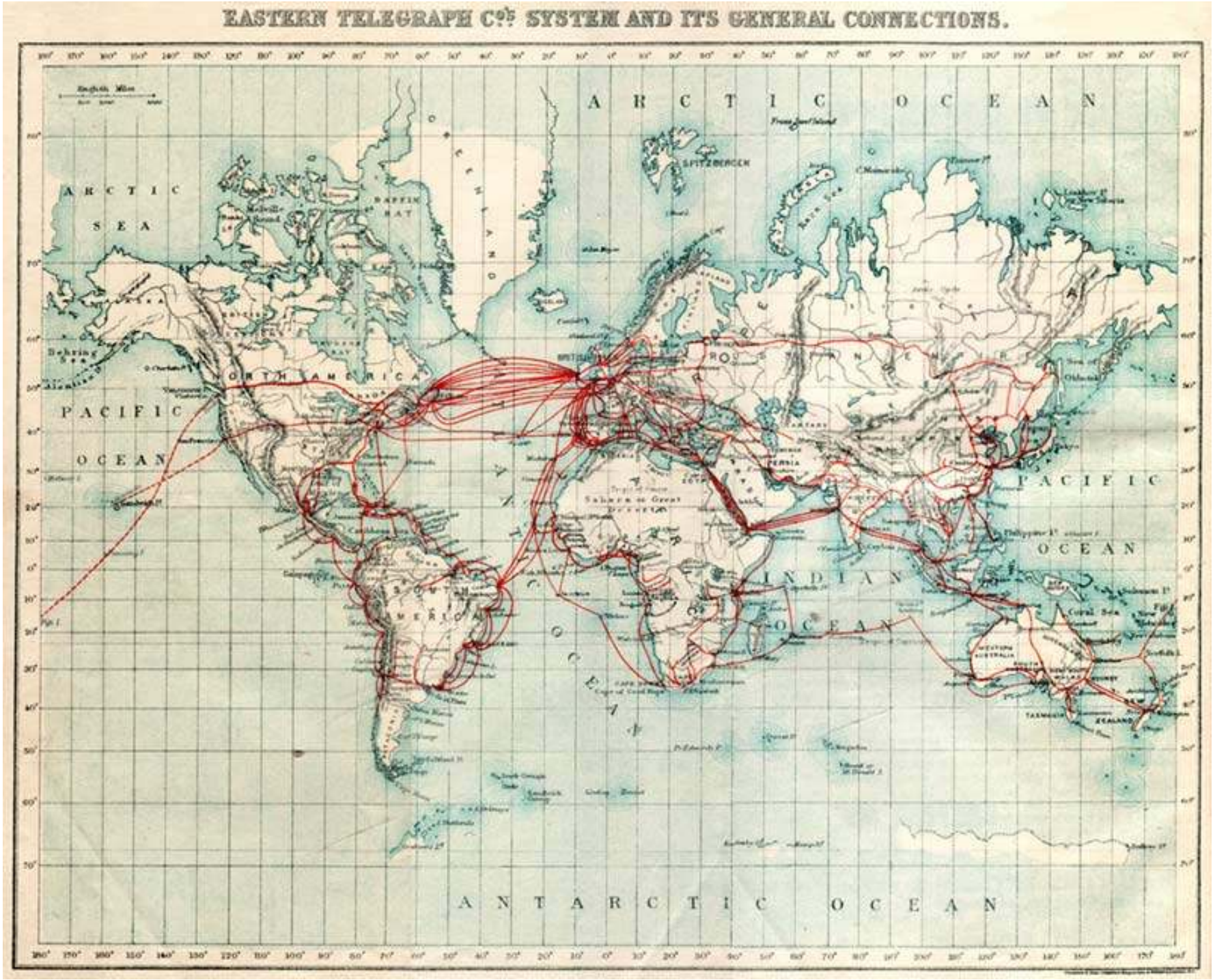
courtesy thefoa.org

courtesy http://shulihallak.com/

…also criss-crosses the ocean floors.



courtesy Telegeography



courtesy wikipedia

courtesy Tyco Reliance

courtesy http://shulihallak.com/


CAT
courtesy http://www.tstt.co.tt/


courtesy Rod Wilson, Ciena

courtesy Rod Wilson, Ciena

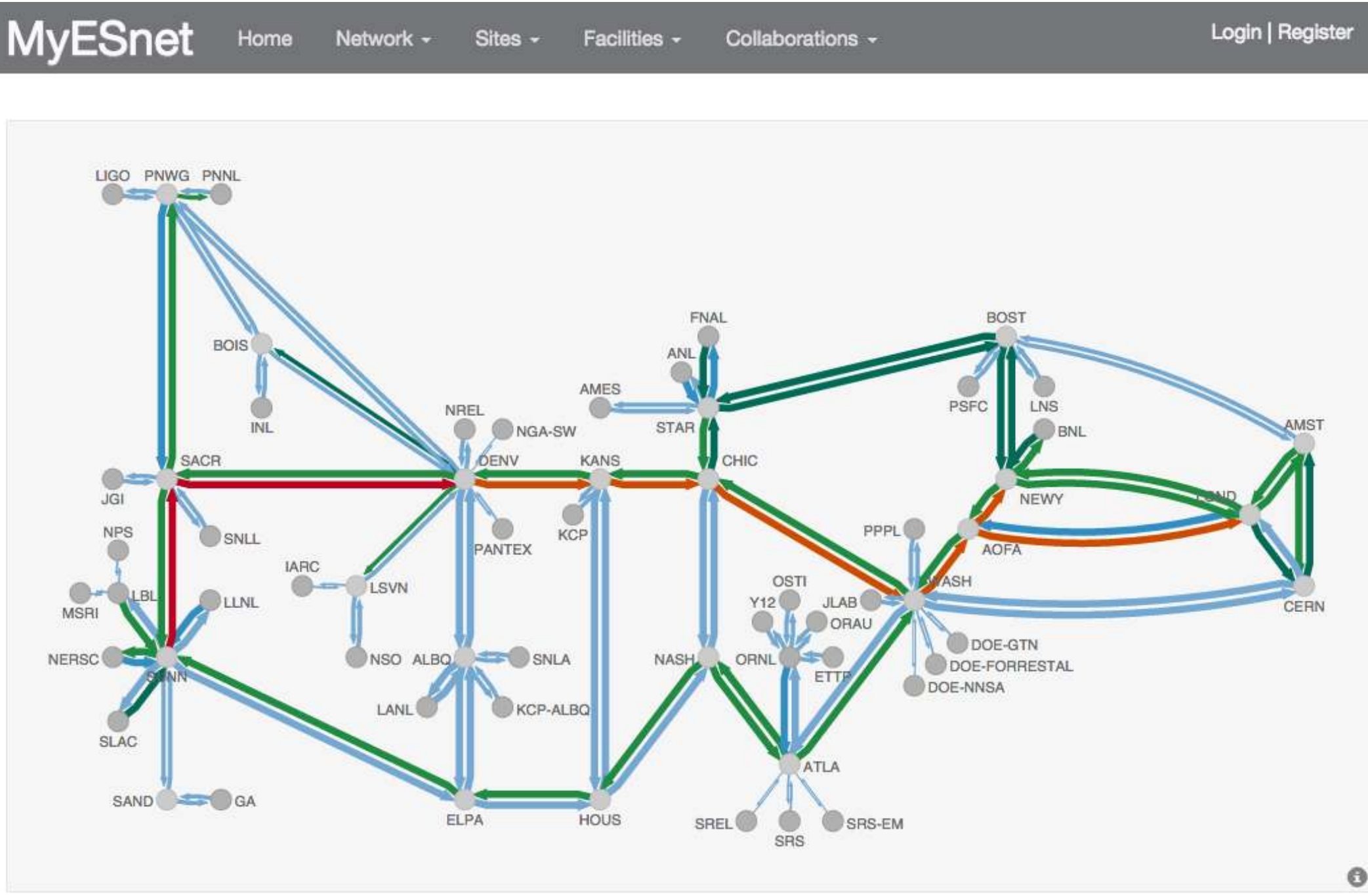
ESnet operations: focus on simplicity, automation, core mission.

Monitored hosts	433	394	-10%
Auto-patched hosts	111	187	44%
Auto-configured hosts	111	120	8%
Physical hosts	284	203	-40%
Virtual hosts	125	180	31%
Hypervisors	11	11	None
OS Versions	24	14	-71%

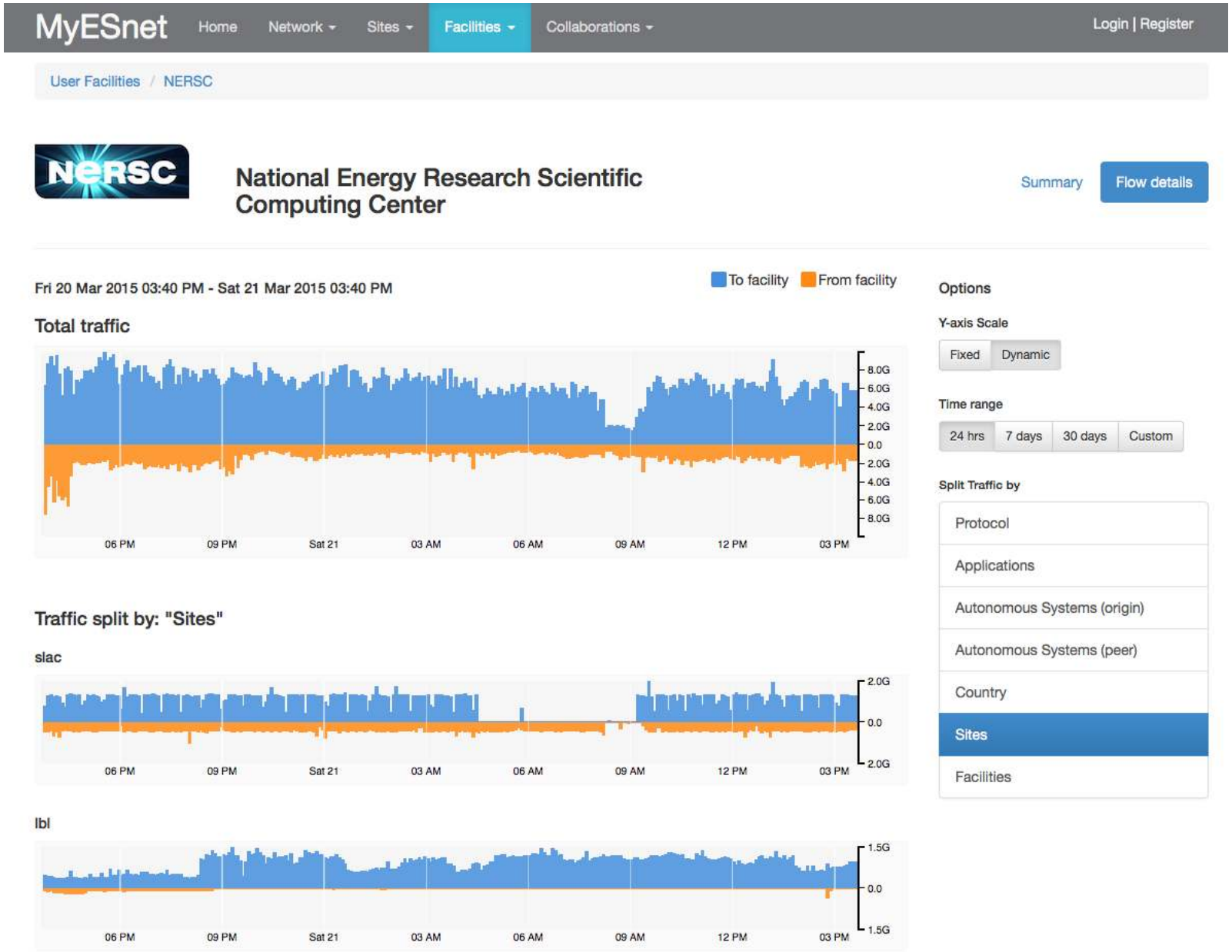
ESnet's video collaboration and X.509 services for DOE were highly distinctive at one time, but no longer. We have transitioned them to commercial providers, in the spirit of focusing on our core mission.

ESnet

Our portal (my.es.net) now greatly enhanced.



Our portal (my.es.net) now greatly enhanced.



Honors for ESnet5 deployment:

FierceGovernment chose the ESnet5 Deployment Team as a recipient of the annual Fierce 15 award, in recognition of "federal employees and teams who have done particularly innovative things."

Information Week named ESnet as one of the "top 15 innovators", among government entities at every layer: federal, state and local. It was the second time in four years ESnet received this award.

InformationWeek Government

ESnet

More recent honors and awards:

ESnet's OSCARS software was honored with a 2013 R&D100 Award, and more recently with a 2014 Secretary's Honor Award from DOE.

ESnet's Network Research Testbed received CENIC's 2015 *Innovations in Networking* award: "ESnet inspires us to do more for our communities, and to do better at what we do." (CENIC CEO Louis Fox)

Inder Monga, ESnet CTO and Division Deputy, named chair of Research Associate Council for Open Networking Foundation (most important membership org promoting SDN).

ESnet

Staffing remains lean.

- Brazil (422)
- UK (180)
- Czech Republic (142)
- Netherlands (138)
- Croatia (112)
- Internet2 (~100)
- Hungary (94)
- Australia (80)
- Norway (78)
- Switzerland (76)
- Greece (68)
- France (66)
- Italy (61)
- Germany (54)
- Ireland (52)
- Slovenia (51)
- Belgium (50)
- Portugal (47)
- ESnet (42)

Caveats: varying service and business models make comparisons difficult, but the large-scale pattern is instructive. (Headcount numbers interpolated from bar graph on page 80 of the most recent *GÉANT Association Compendium*, or described elsewhere in that report.)

ESnet

We are working to increase diversity in the field of network engineering.

- Sponsored two early-career women from DOE/SC labs to attend an important annual conference for network engineers (Internet2/ESnet Technology Exchange, Fall 2014).
- Co-organizing diversity track and panel at same conference.
- Participating in the steering committee of Internet2's gender diversity initiative for women in IT (across US university space).
- Grace Hopper conference, for career development as well as recruitment.
- In partnership with FRGP and others: proposal to NSF to fund SCinet participation for young female engineers.

ESnet

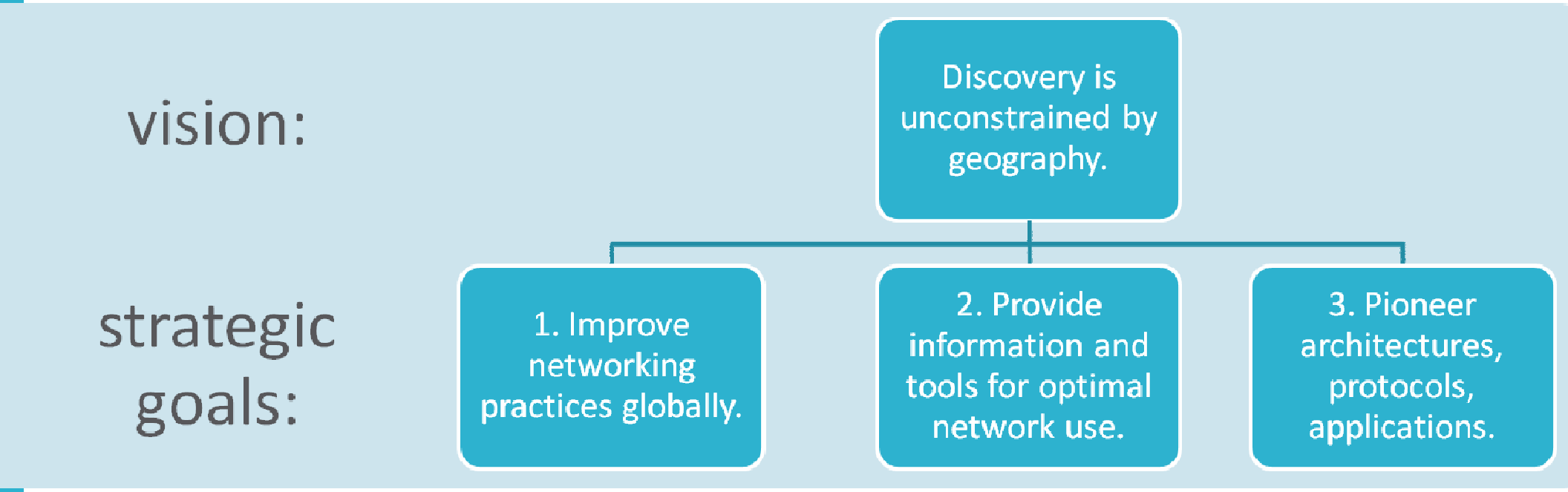
Overview

Update

Impact

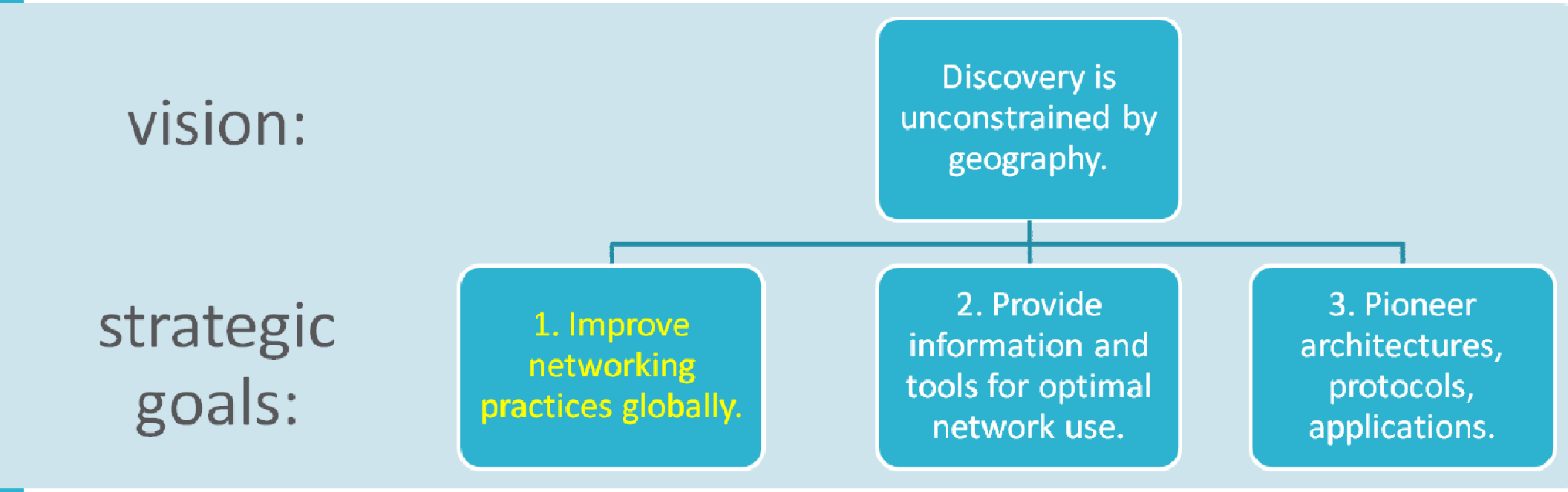
ESnet

Our vision and strategic goals guide impacts.



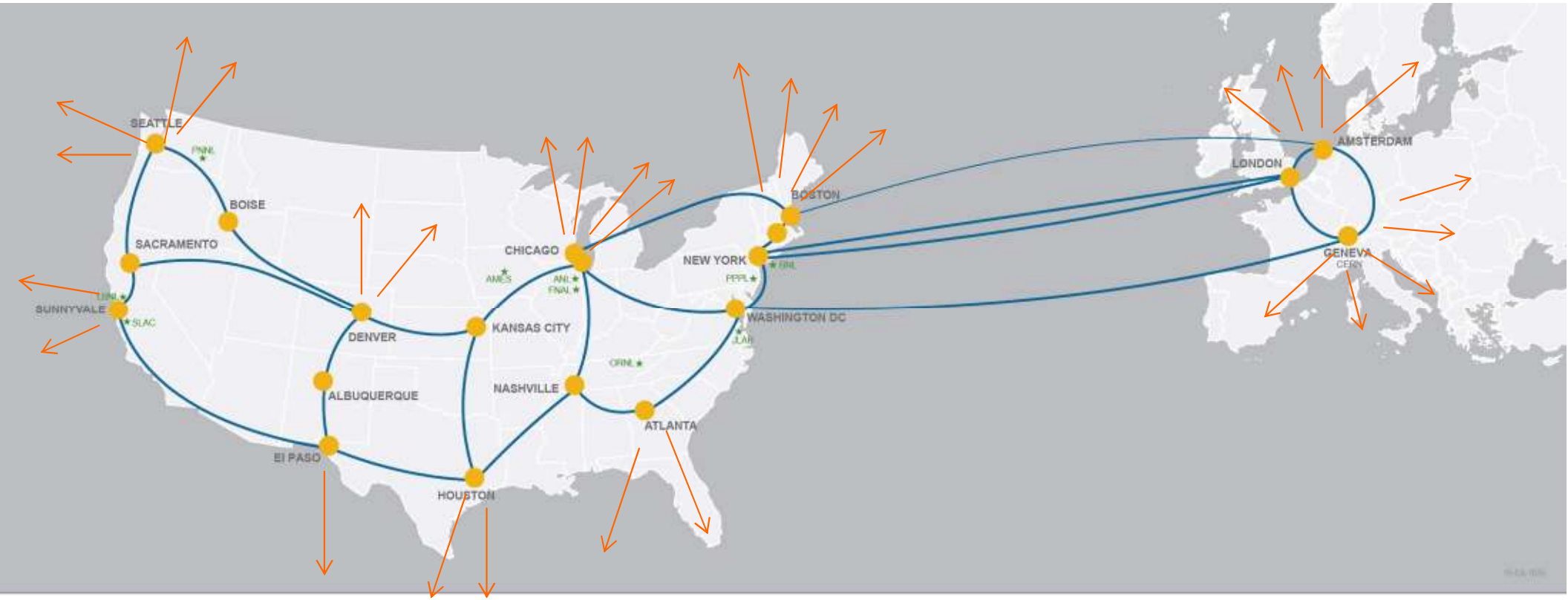
ESnet

Our vision and strategic goals guide impacts.



ESnet

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



ESnet's success is not sufficient, because networks share fate. SC invests nearly $1B/year in university research, and campus networks matter to DOE.

Introducing Science DMZ, a network design pattern for data-intensive science (origin: ESnet & NERSC).

Three components, all required

1. Friction-free network path:
   - highly-capable network devices (wire-speed, deep queues)
   - at or near site perimeter, with option for virtual circuit connections
   - security policies tailored for science
2. Dedicated, Data Transfer Nodes (DTNs):
   - hardware, operating system, libraries optimized for data transfer
   - appropriate tools such as Globus and GridFTP
3. Performance measurement / test nodes:
   - perfSONAR
   - testing, assurance, forensics

Much more information: http://fasterdata.es.net/science-dmz/

ESnet

Science DMZ now recognized as best practice.

NSF is investing $60M to promote adoption by US universities (among other CI goals). Fourth funding round underway.

>120 universities in the US have deployed this DOE architecture.

IN addition: USDA, NIH – with NASA, NOAA investigating.

Australian, Canadian universities following suit.

ESnet

What's next? Evolution of Science DMZ as a *regional* cyberinfrastructure platform.

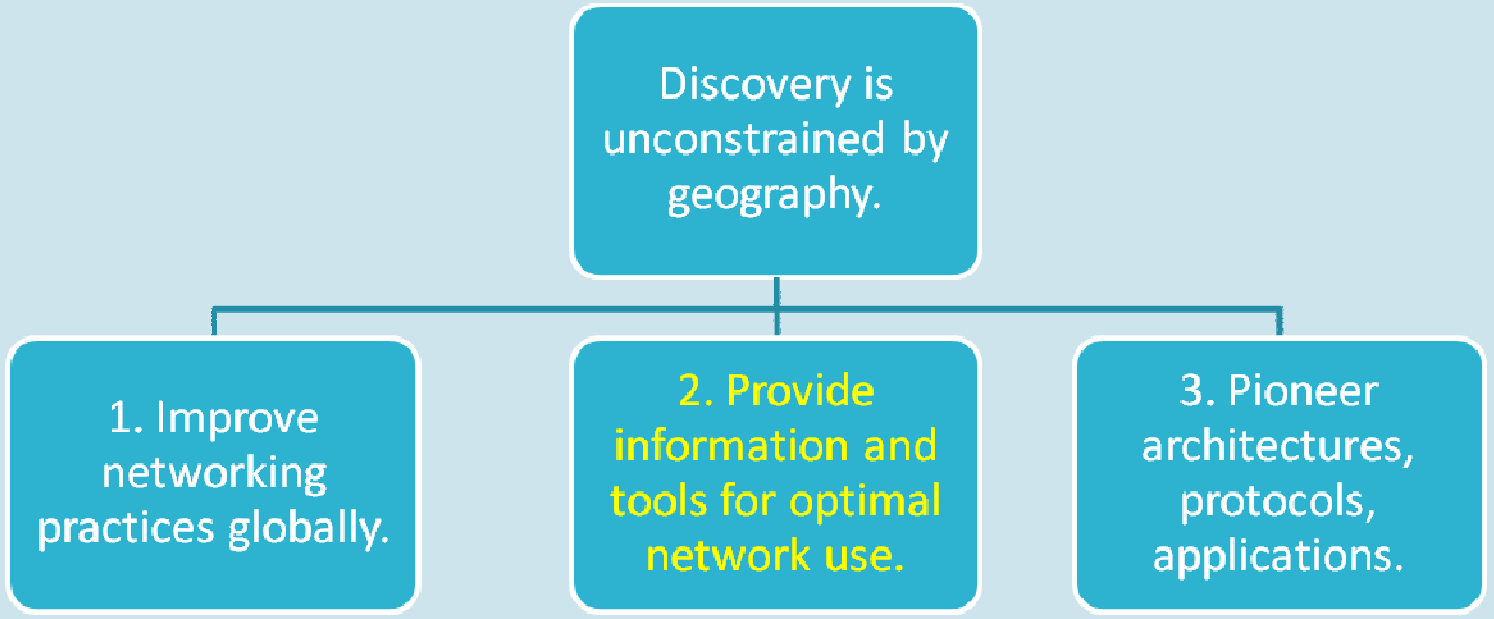
Pacific Research Platform initiative, lead by Larry Smarr (Calit2/UCSD)

- first large-scale effort to coordinate and integrate Science DMZs
- participation by **all major California R&E institutions, CENIC, ESnet**
- announced **March 6**: http://cenic.org/news/item/high-performance-big-science-pacific-research-platform-debuts-at-cenic

Back to impacts: ‘fasterdata’ knowledgebase.

vision:

strategic goals:



ESnet

Host Tuning

Background Information
Linux
Mac OSX
FreeBSD
MS Windows
Other OS
ESnet perfSONAR Tuning
NIC Tuning
Interrupt Binding
Virtual Machines
Packet Pacing
40G Tuning

Interrupt Binding

To fully maximize single stream performance (both TCP and UDP), you'll probably need to pay attention to which core its being used. To get the best performance you want the NIC interrupts going to 1 core, and the application IO thread to a nearby core, but not the same core.

For hosts with Intel Sandy/Ivy Bridge motherboards this is even more important. As you can see in the figure on the right, the PCI slot for the NIC is directly attached only one of the two processors. There is a large performance penalty if either the interrupts or the application is on the wrong processor because if that happens everything must cross the QPI bus.

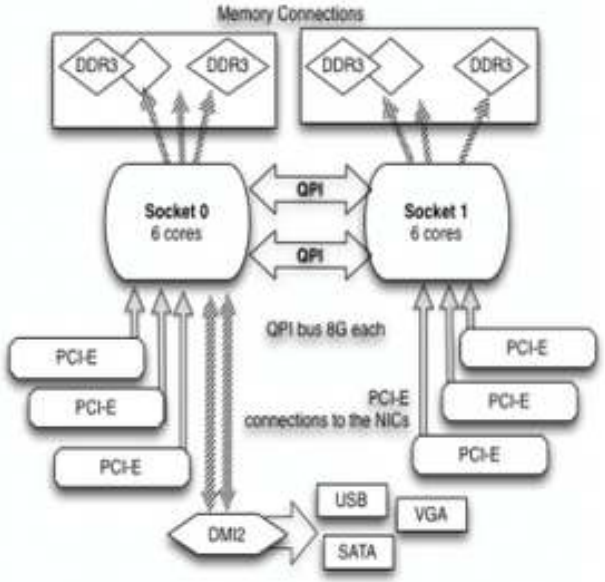
On a system with slow processors, or a 40G PIC gen-3 host, TCP and UDP performance increases of up to 2x have been observed by ensuring that the NIC driver interrupts and applications threads are handled by the right cores.

On Linux, you can use the sched_setaffinity() system call or the numuctl command line tool to bind a process to a core. For iperf3, you can use the "-A" flag, and for nuttcp you can use the "-xc" flag to do this.

To specify which core handles the NIC interrupts you need to **disable irqbalance**, and then bind the interrupts to a specfic core.

Some vendors provide scripts to do this IRQ binding at boot time.

Intel Sandy/Ivy Bridge



CI Plans

- UF Plan 2011-2015
- U of U Plan 2013
- UH Plan 2013
- AU Plan 2013
- UNH Plan 2008-2018
- UNC Plan
- TU Plan
- Great Plains Network Regional Plan
- KINBER Regional Plan
- Clemson University Regional Plan
- OARnet Regional Plan

Sample Campus & Regional Cyberinfrastructure Plans

The following are representative campus, and regional, cyberinfrastructure plans from facilities around the U.S. Each plan was submitted as part of a proposal to National Science Foundation's Campus Cyberinfrastructure programs (CC-IIE, CC-NIE, and CC-DNI). These materials are provided as examples for universities and institutions looking to develop their own campus CI plans, strategies and architectures to support research, education, and discovery.

University of Florida Cyberinfrastructure Plan

Executive summary As part of the strategy to make the University of Florida a member of the top 10 public universities in the United States, it is critical to build the right foundation for faculty and students to do their work in education and research. The University of Florida plans to build upon existing infrastructure and enhance it to reach the following goals and milestones: Infrastructure... READ MORE »

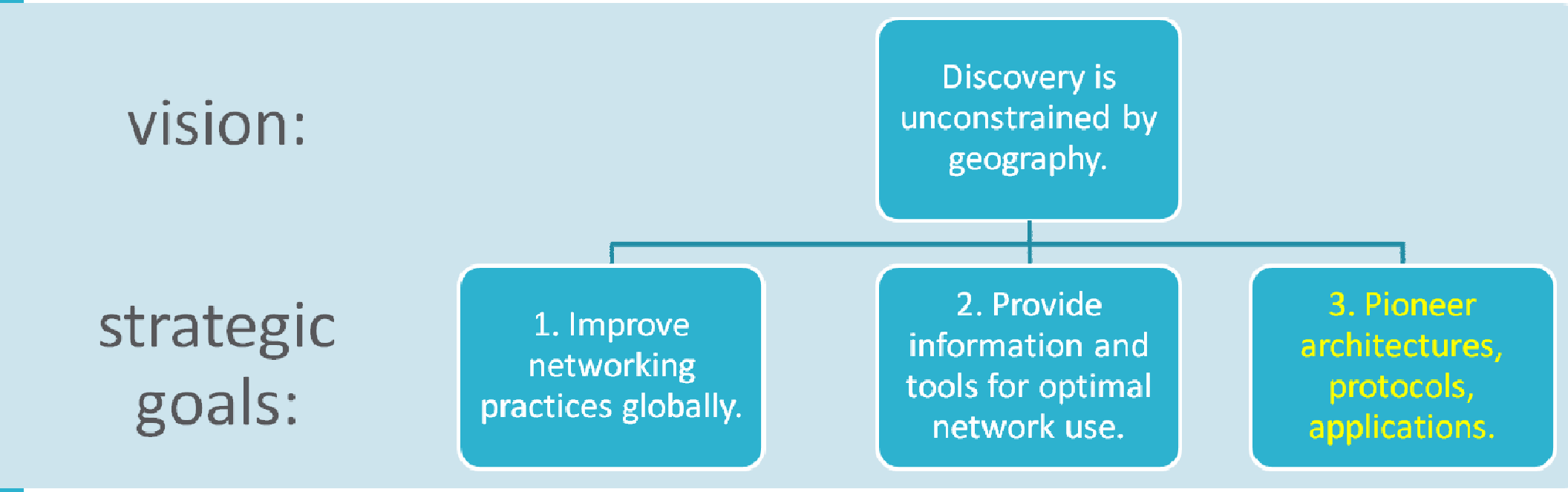
University of Utah Cyberinfrastructure Plan

Information Technology (IT) Governance Research Portfolio Document context The University of Utah Information Technology Research Portfolio, currently chaired by Prof. Thomas Cheatham, is a component of the newly implemented Information Technology (IT) governance structure of the University of Utah. The portfolio has replaced the earlier Campus Cyberinfrastructure (CI) Council that was... READ MORE »

University of Hawai'i Cyberinfrastructure Plan

Introduction The University of Hawai'i (UH) is already one of the nation's top research universities, with distinctive strengths in astronomy, earth and ocean sciences. In developing the new Hawai'i Innovation Initiative (HI2), which calls for bold expansion in these and other strategic areas such as agriculture and the health sciences, it became apparent that stronger capabilities... READ MORE »

Selected impacts in research and innovation.



ESnet

Ongoing impacts (universities, labs, networks):

- OSCARS for virtual circuits with service guarantees
  - adopted by **>40 networks & universities**
  - production software for almost 10 years
  - initially funded by ASCR NGNS, LBNL LDRD
  - Secretary's Honor Award, April 7

- perfSONAR for network measurement (diagnostics, assurance, forensics)
  - integral to ScienceDMZ
  - ESnet, Indiana, Internet2 and GÉANT major collaborators
  - **1200 deployments** worldwide

- Science DMZ
  - network design pattern for data intensive science
  - NSF support with **>120 campus deployments** in the US alone

ESnet

Ongoing impacts (industry):

- Frequently among the **first customers** for new technology (100G terrestrial, 100G trans-Atlantic, 400G terrestrial, fastest available NICs, etc).
- Deep collaboration with Infinera to demonstrate **first SDN for optical transport** (October 2012) resulted in new product, with major carrier announcement March 2015.
- Ongoing collaboration with different optical vendor will result in significant reduction in **product development cycle** (to be announced in May).
- Frequent engagement with relevant startups, for instance **Corsa Technology:**
  - first public customer for disruptive 'white box' networking gear
  - worked closely to develop packet-processing pipeline useful to DOE science missions
  - Corsa closed $16M series B funding this week

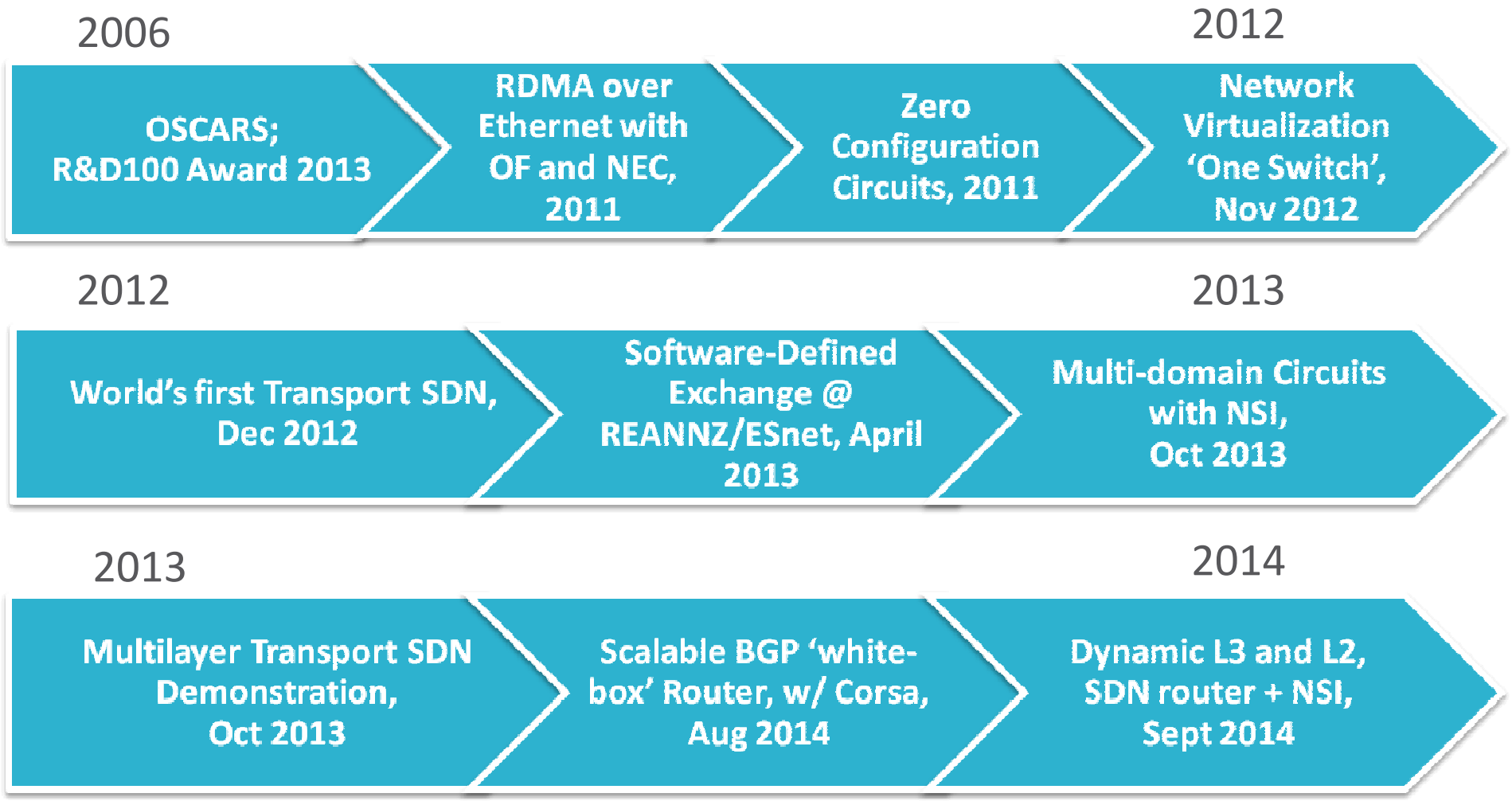
ESnet

Two sources of future impact:

- ESnet **testbed** – available to researchers and industry since 2011, now upgraded with low-cost 'white box' SDN equipment (Corsa Technology), on an international footprint. Intended uses:
  - –R&D for SDN, NFV, NDN, systems, protocols, security
  - –**ESnet6** prototypes
  - –federation with other testbeds
  - –platform for collaborations and demonstrations
- SDN innovation, including **ESnet Operating System**
  - –platform for science apps to express **intent**, simultaneously, across >1 network
  - –focus on requirements *not* being met by industry, open-source projects
  - –supported by Berkeley Lab LDRD
  - –multipoint VPN service demonstration in May

ESnet

ESnet SDN impacts, in a nutshell:



ESnet

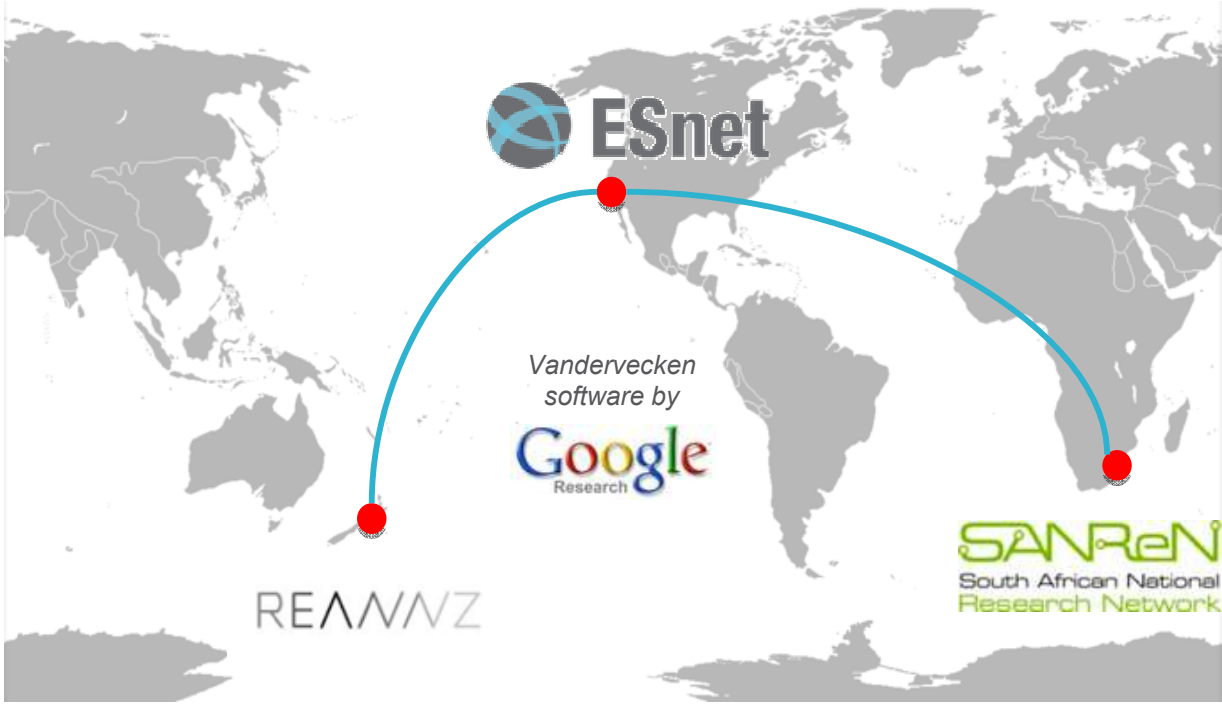
Spotlight on one project: global SDN BGP peering

Inter-operability with routing standards protocol (BGP) using SDN techniques.

Physically distinct control plane (using off the rack Unix server) and data plane (using OpenFlow switch) functions.

Demonstrated ~40% FIB compression (13,215 -> 7,577 routes)

Dynamic layer 2 setup (using OSCARS with NSI) for transport of layer 3 BGP protocol messages.



Goals

- Explore feasilbility of ‘white box’ designs for core Internet routing.
- Backward compatibility with standard protocols (especially BGP) without compromising performance, scalability, operability.
- Simplifly control and management through logical centralization.

ESnet

In conclusion, a reminder about our vision:

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

ESnet

ESnet is an instrument for discovery, and increasingly the *glue* for DOE super-facilities.

The new European extension supports LHC Run 2, plus all DOE missions.

ESnet innovation is impacting scientists, researchers, universities, industry – around the world.

ESnet

Thank you.

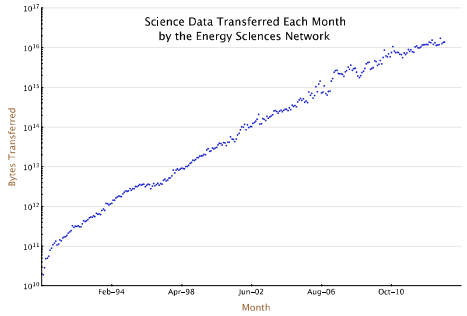
greg@es.net

ESnet

Additional Slides

Broad questions driving ESnet research and development activities:

1. How can we continue to scale up and handle **exponential traffic growth** with linear budgets?



1. Can we create useful abstractions to enable productive **interaction** between **science applications** and the network?

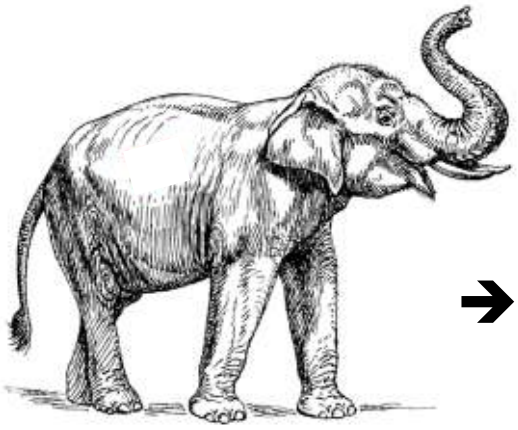
1. Can we transform ESnet into a **programmable platform** that can be **operationally supported**?

ESnet

Elephant Flows Place Great Demands on Networks

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

Result
99.9954% of water transferred, at “line rate.”



Network ‘pipe’ that drops packets at rate of .0046%.

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

essentially fixed

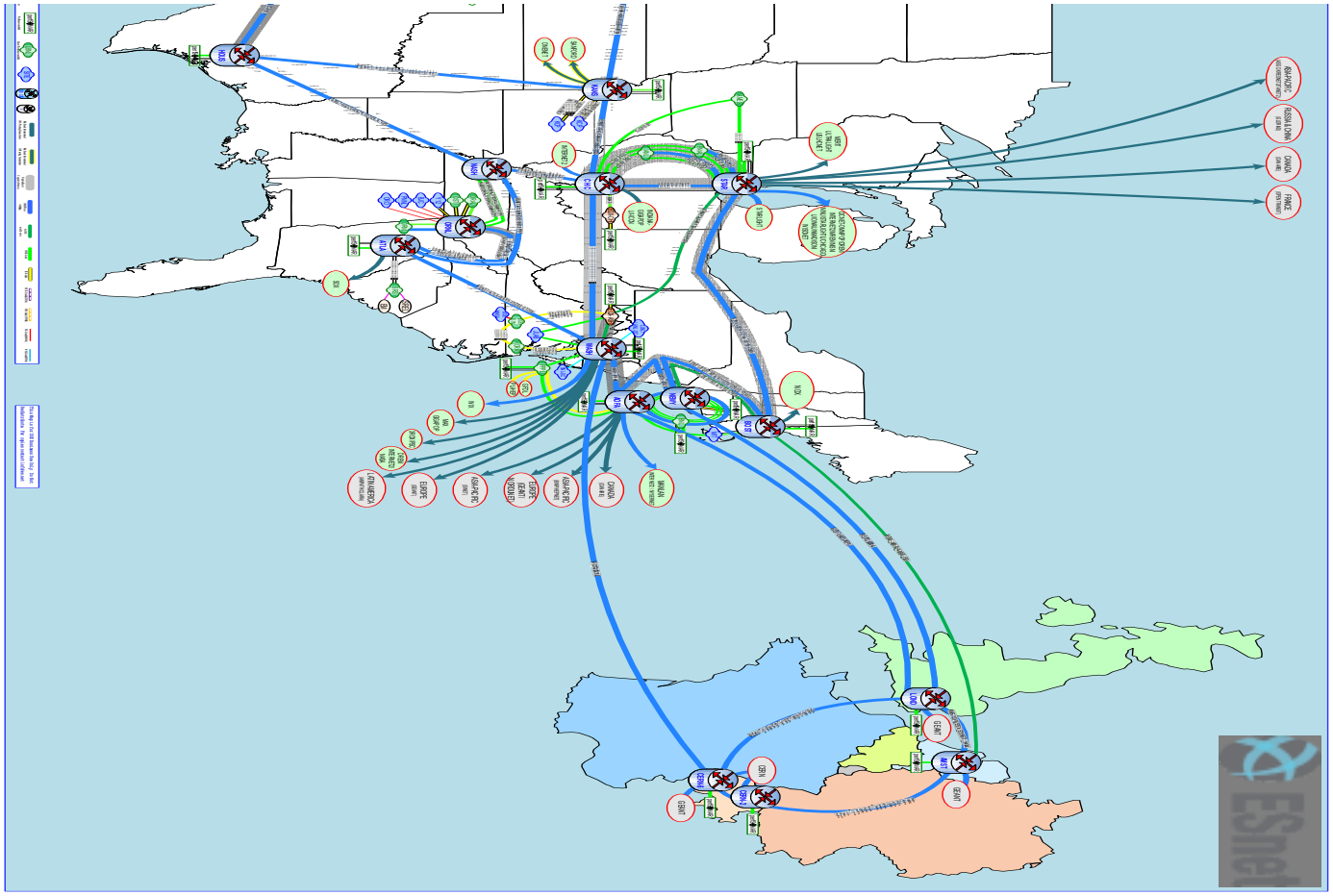
determined by speed of light

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

Through careful engineering, we can minimize packet loss.

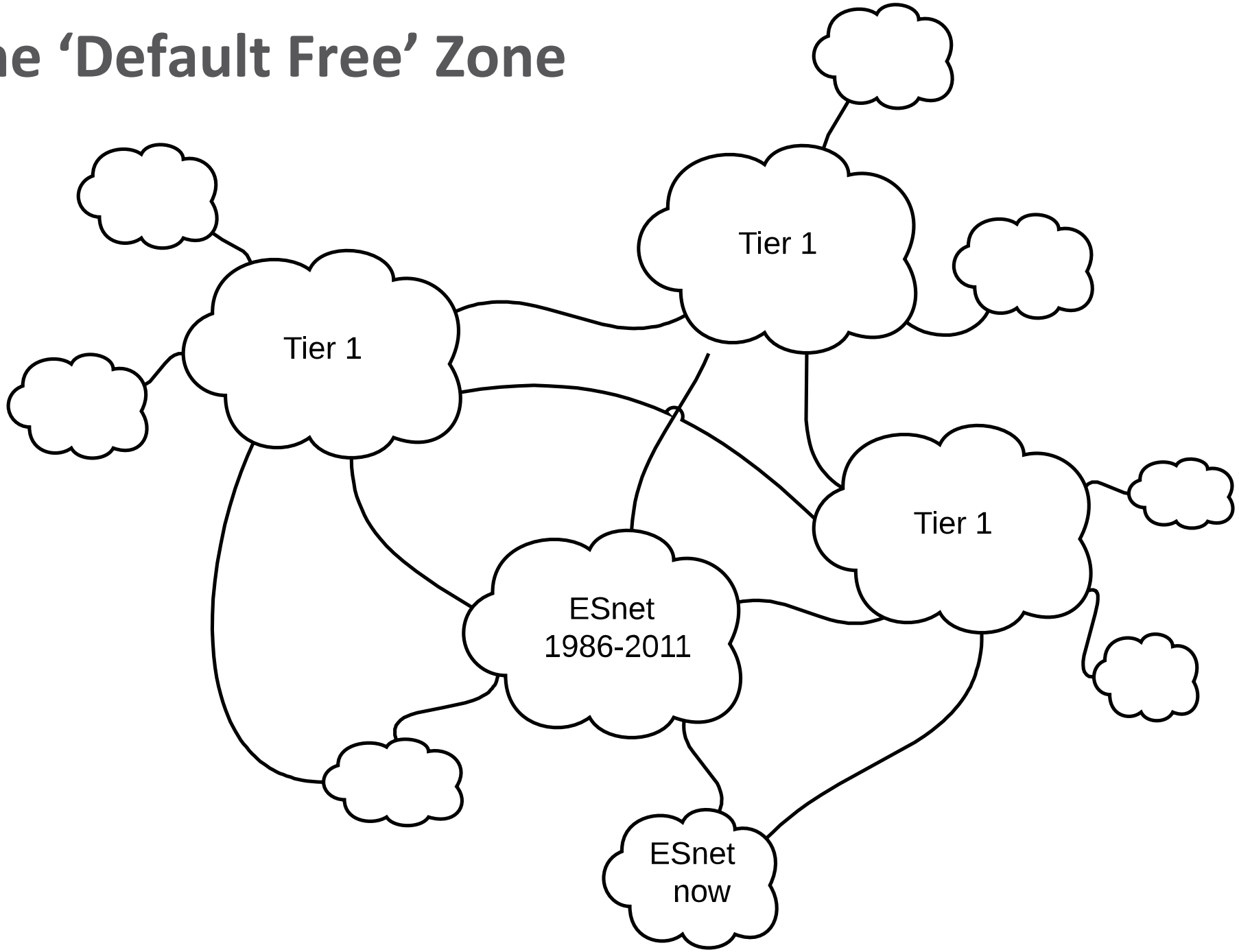
Assumptions: 10Gbps TCP flow, 80ms RTT.
See Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, and Jason Zurawski. The Science DMZ: A Network Design Pattern for Data-Intensive Science. In *Proceedings of the IEEE/ACM Annual SuperComputing Conference (SC13)*, Denver CO, 2013.

More detailed network diagram.

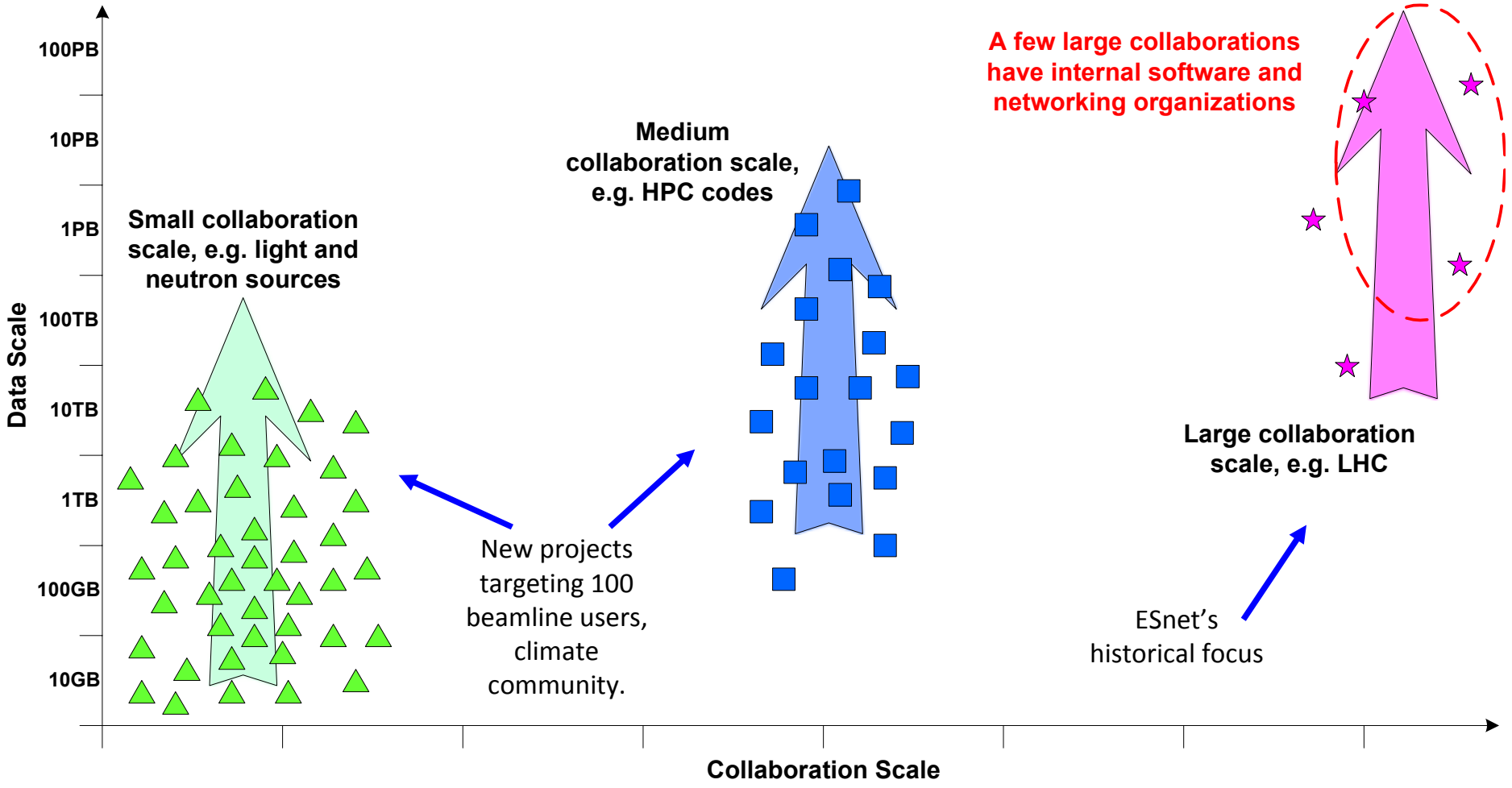


ESnet

The ‘Default Free’ Zone

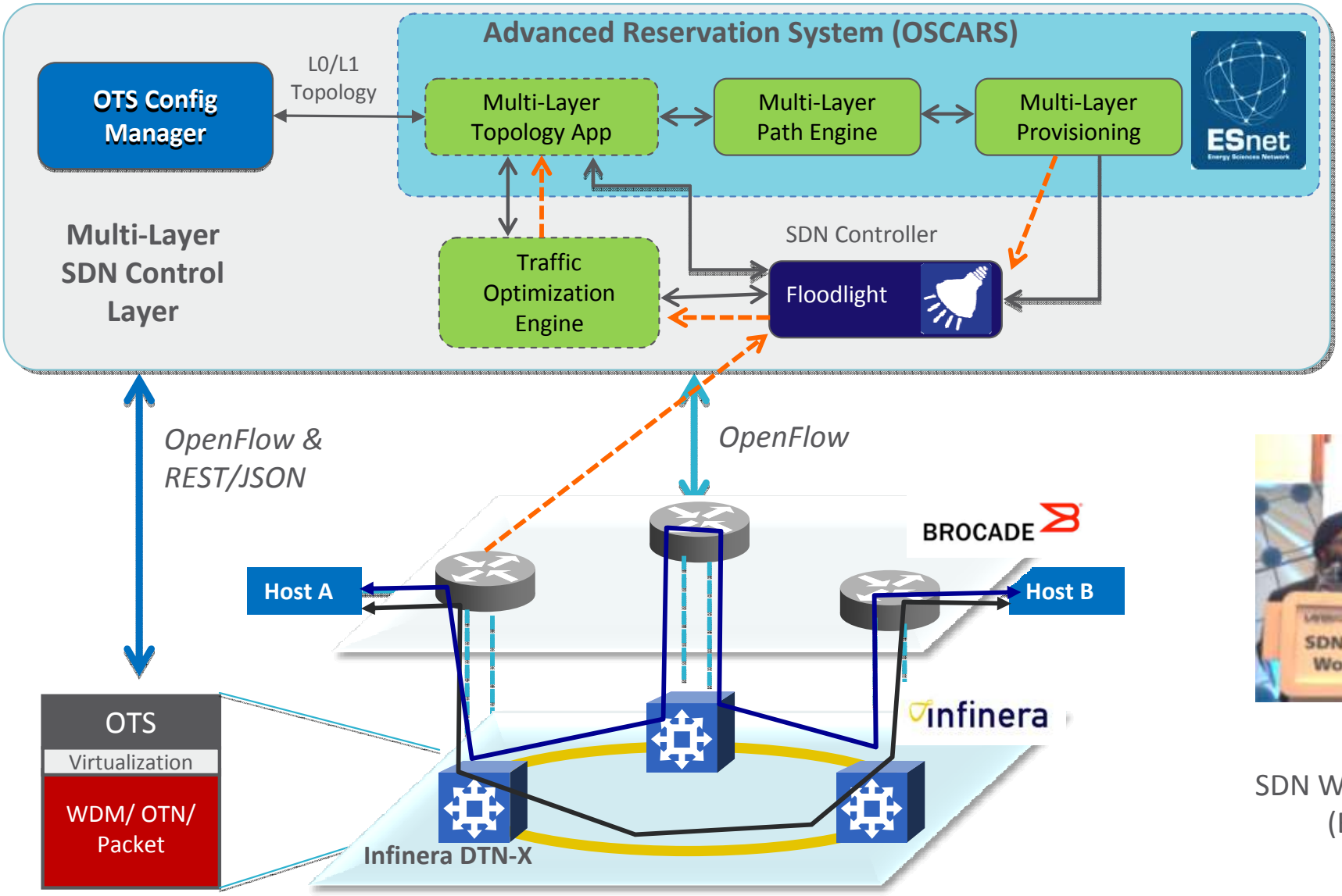


Segmenting the world of scientific collaboration.



ESnet

ESnet / Infinera / Brocade multi-layer packet-optical SDN demonstration (Oct 2013)



Presented at SDN World Congress (Bad Homburg, Germany)