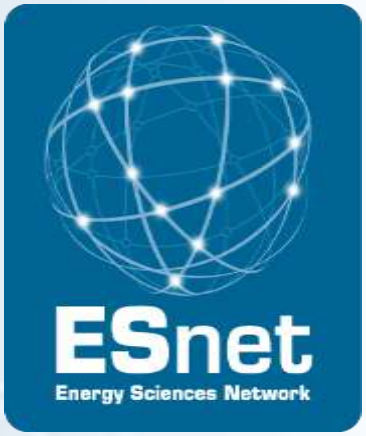


Network as Instrument: Faster Data for Ontario

Gregory Bell, Ph.D.

Director, Energy Sciences Network

Director, Scientific Networking Division

Lawrence Berkeley National Laboratory

2013 ORION THINK Conference

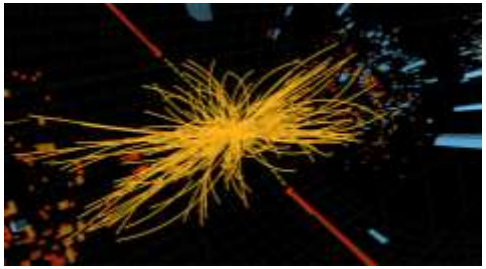
Toronto

April 25, 2013

Outline

Energy Sciences Network

Network as Discovery Instrument



Faster Data

Outline

Energy Sciences Network

Network as Discovery Instrument

Faster Data

26 Years as a Mission Network

Mission of DOE Office of Science:
Deliver knowledge and tools for transforming our understanding of the universe.

Mission of Energy Sciences Network:
Accelerate research and discovery for DOE Office of Science.

$5B/year for the US National Lab Complex, which includes:

- world's largest collection of scientific user facilities (32)
- supercomputers, accelerators, xray / neutron sources, electron microscopes, sequencers, fusion facilities, Energy Sciences Network
- >100 Nobel Prizes

The Energy Sciences Network (ESnet)

A national network, optimized for science:

- connecting 40 labs & facilities with >100 networks
- optimized for massive science data flows
- offering capabilities not available commercially
- $34.5M in FY12 (40 staff)

$62M ARRA grant funding:

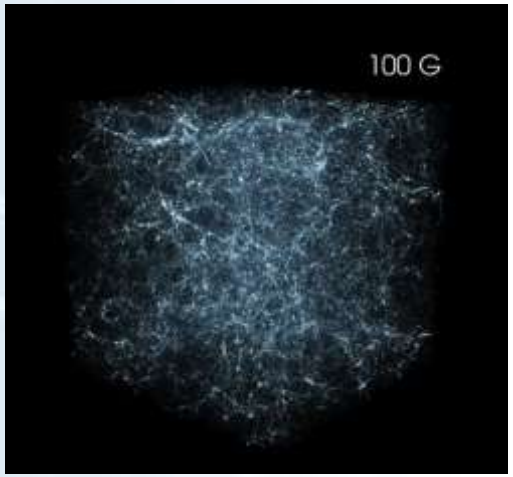
- world's first 100G network at continental scale
- optical fiber assets, access to spectrum
- easier (and cheaper) growth

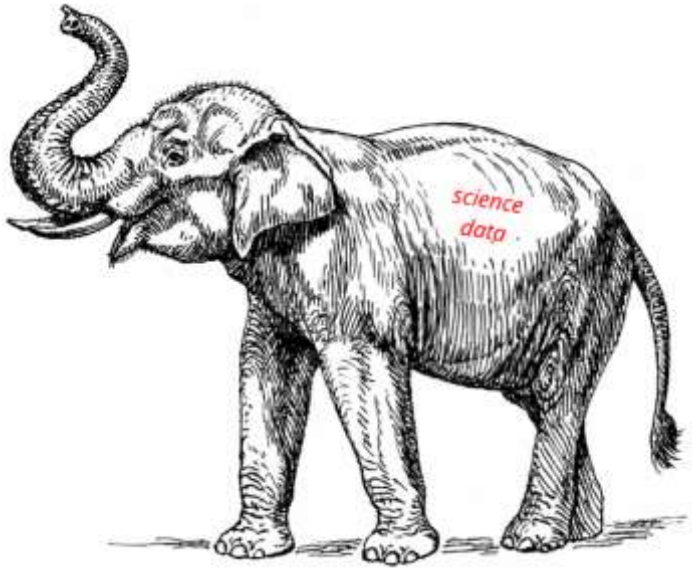
On the web:

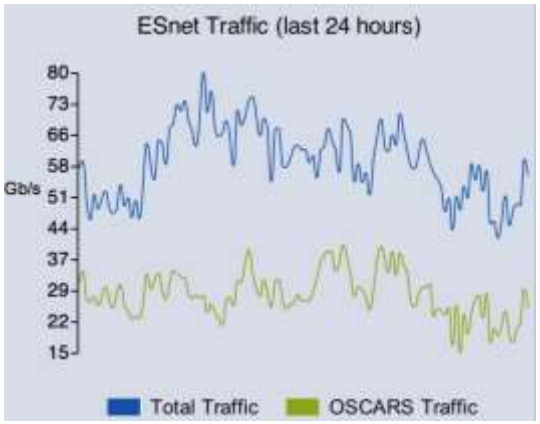
- www.es.net
- fasterdata.es.net
- my.es.net

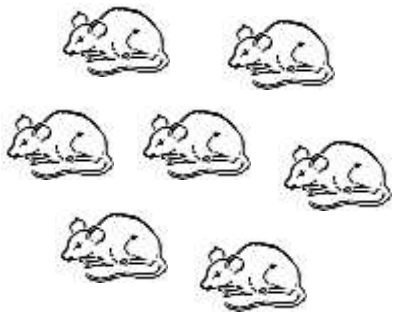
Our vision: scientific discovery unconstrained by geography.

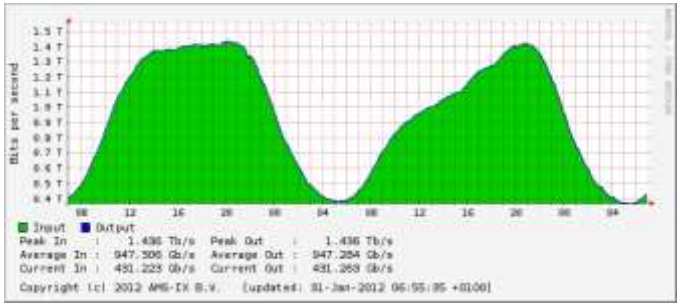
Engineered for the Elephants





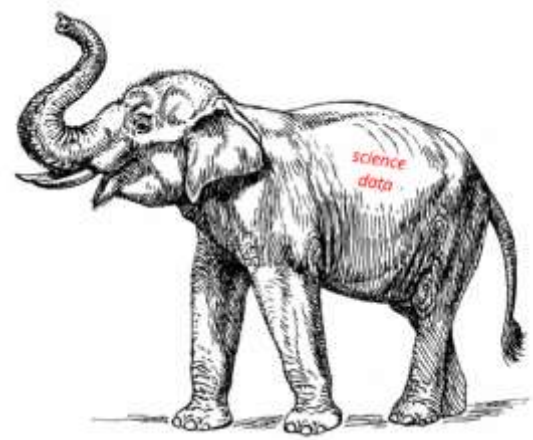




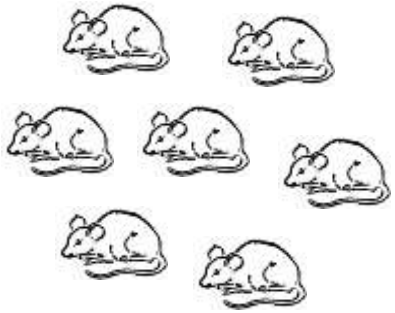


Sensitive Elephants, Robust Mice

Effect of 0.0046% packet loss (1 out of 22000 packets) on data transfer rates for *elephant* and *mouse* flows:



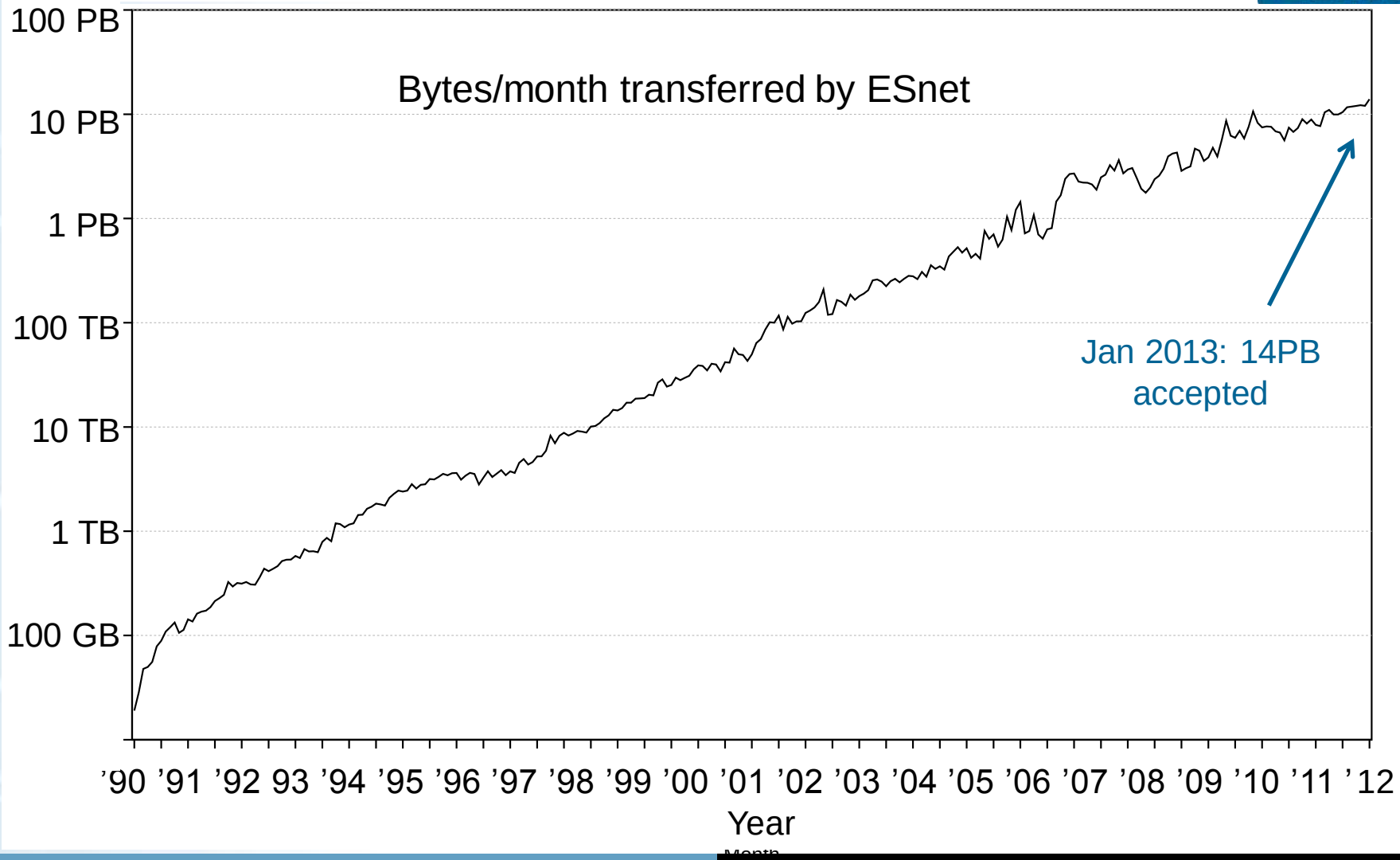
80x reduction in data transfer rate at DOE-relevant distances (ANL to NERSC) and speeds (10Gpbs).



Negligible.

(As measured recently by ESnet research scientist Brian Tierney.)

Unrelenting Growth in Traffic for >20 Years



Among the Drivers for Extreme Data: *Compounded* Moore's Law Growth Rates

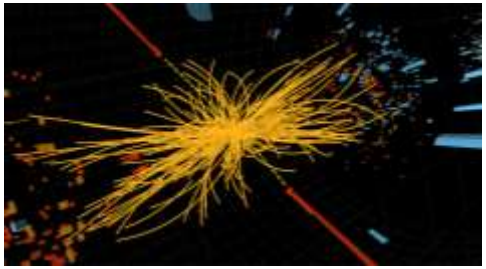
Characteristics of one new detector at Berkeley Lab's Advanced Light Source:

- a 'video microscope' for materials science
- megapixel detector resolution x KHz data rates ➔compounded growth rate
  - 10^6 increase in data output compared to previous generation ($4x10^8$ integers per second)
  - 8 Gbps flow to NERSC for real-time analysis!
  - detector installations this year: 4
- how many beamlines at the ALS? detectors per beamline?
- how many light sources like the ALS in the world?

Outline

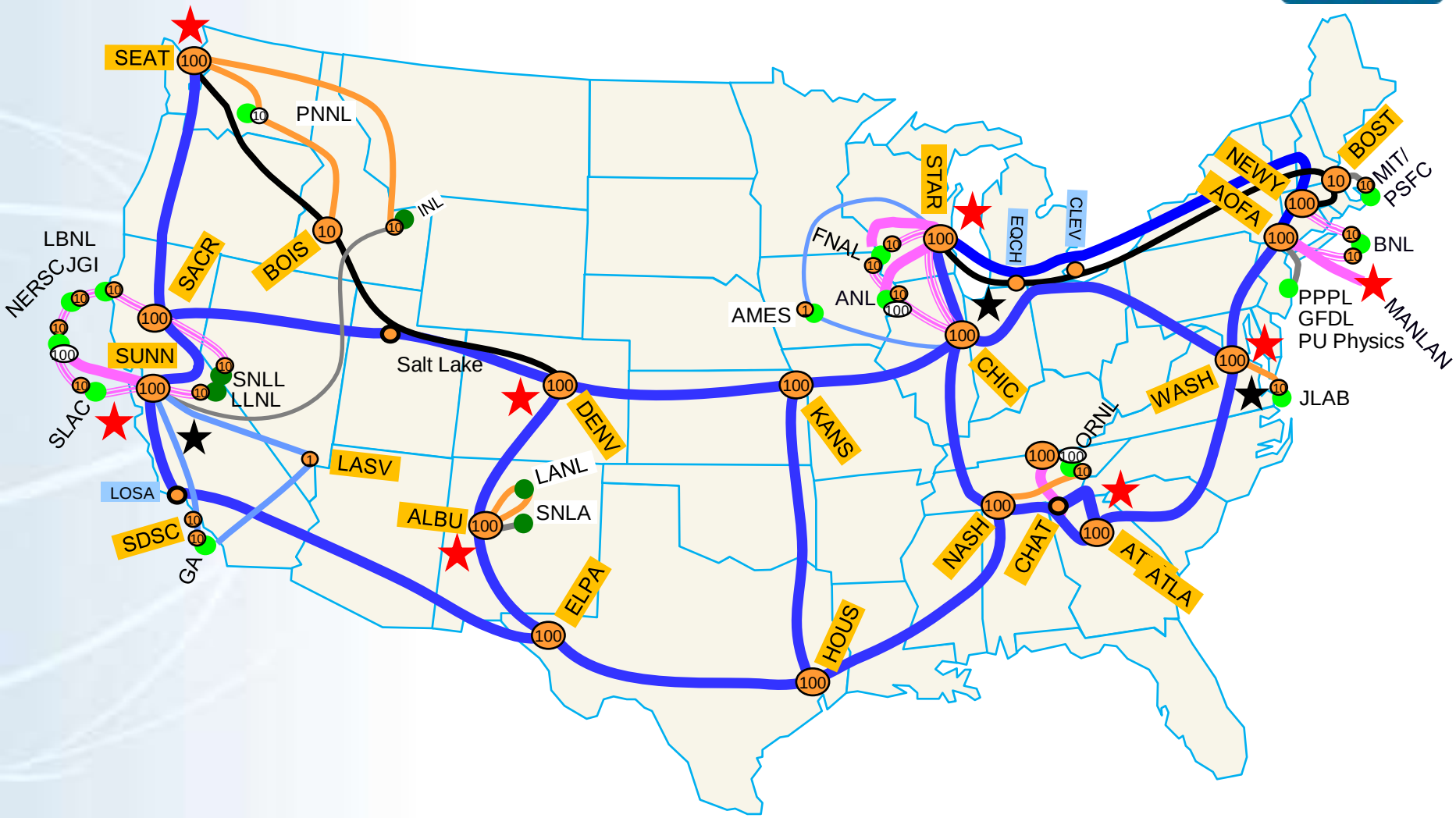
Energy Sciences Network

Network as Discovery Instrument



Faster Data

Our Past: Network as Infrastructure



Three Historical Inflection Points for Global Research Networks

1. Abundant capacity (88 λ x 100Gbps)

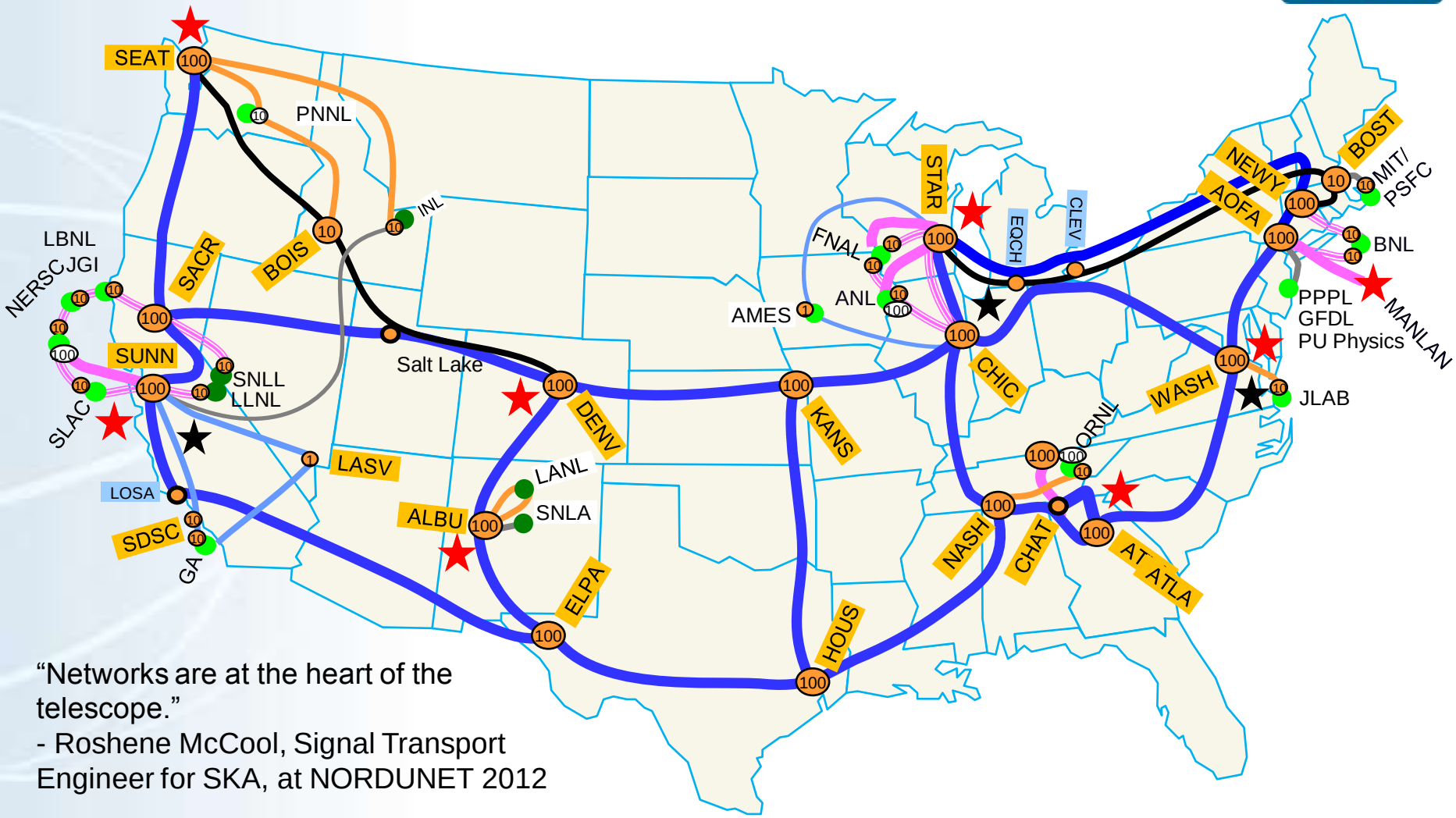
2. Programmability

SDN Network NSI Services Software Networking Interface OpenFlow Defined Programmable Networks

3. Campus architectures newly optimized for data mobility.

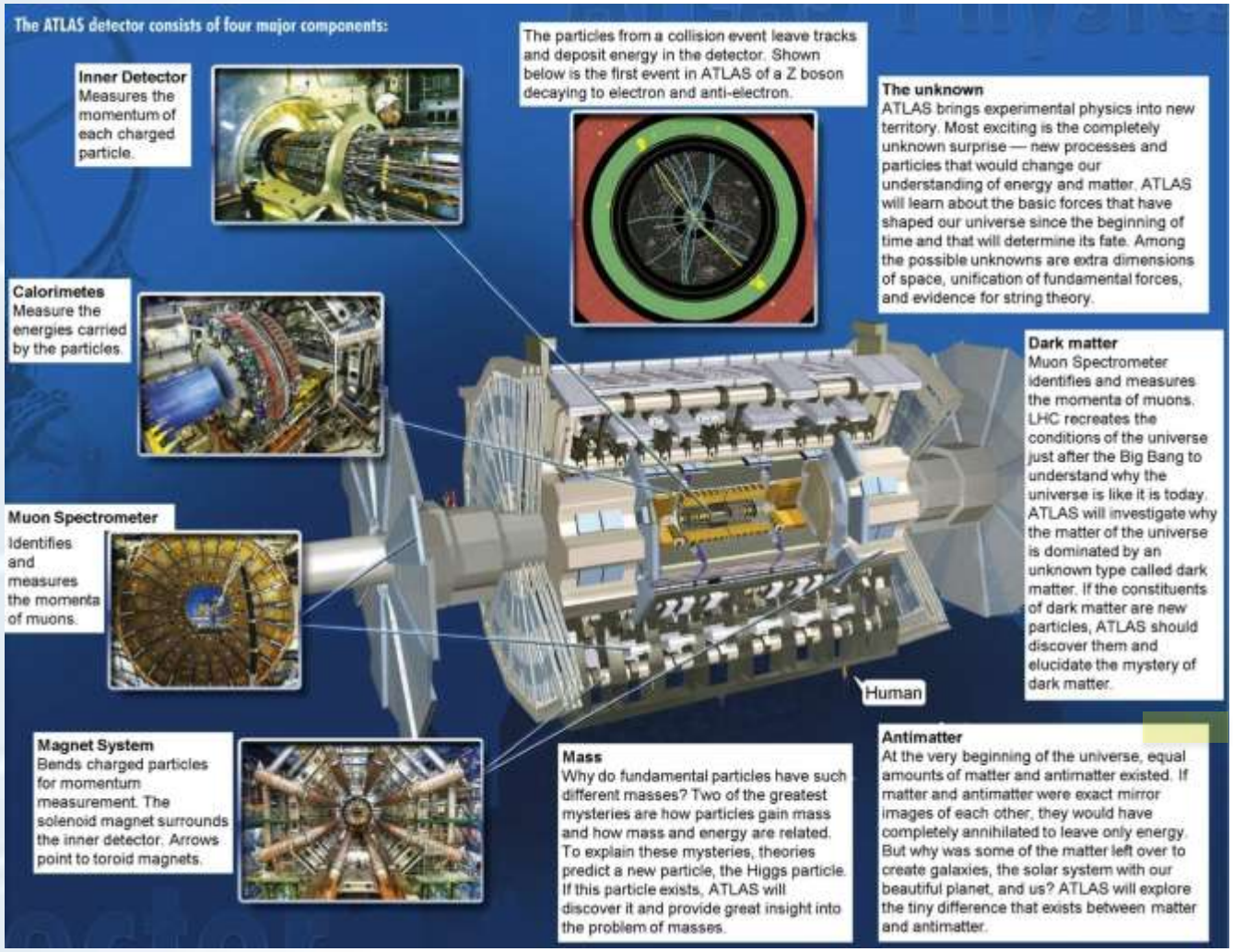
Science DMZ +
NSF grants.

Our Future: Network as *Instrument*



"Networks are at the heart of the telescope."
- Roshene McCool, Signal Transport Engineer for SKA, at NORDUNET 2012

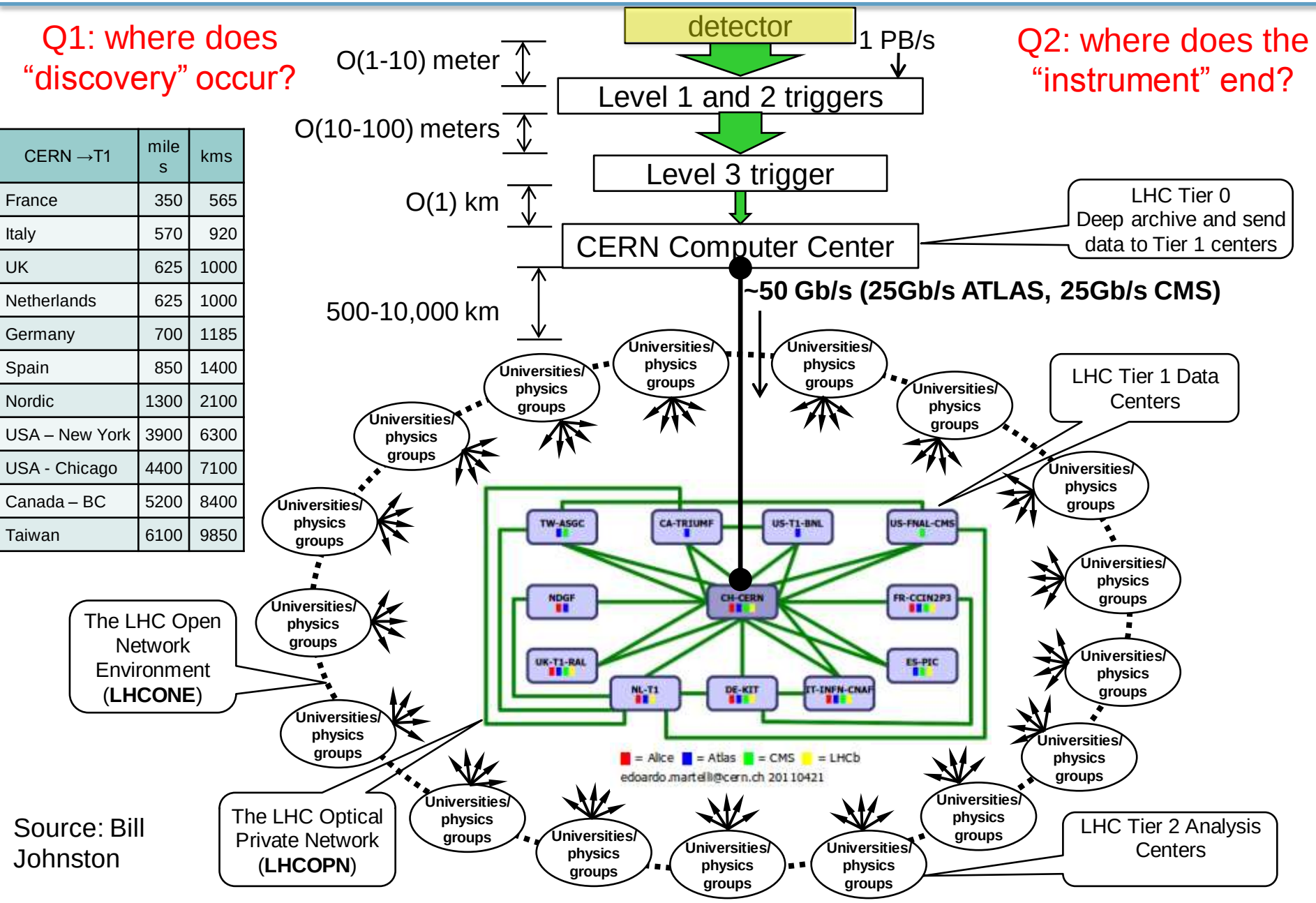
Wait! How can a network amount to an ‘instrument’ for discovery?



Network-Centric View of Large Hadron Collider (@CERN)

Q1: where does "discovery" occur?

CERN →T1	miles	kms
France	350	565
Italy	570	920
UK	625	1000
Netherlands	625	1000
Germany	700	1185
Spain	850	1400
Nordic	1300	2100
USA – New York	3900	6300
USA - Chicago	4400	7100
Canada – BC	5200	8400
Taiwan	6100	9850



Q2: where does the "instrument" end?

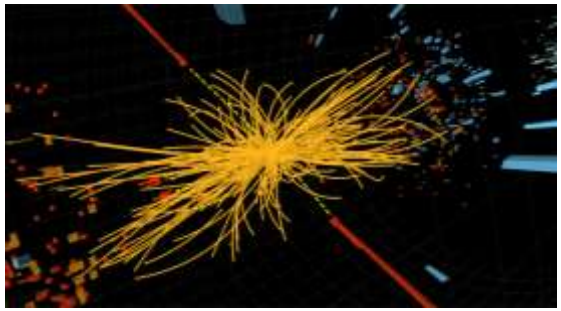
Source: Bill Johnston

Evolution of LHC Data Model

In chronological order:

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

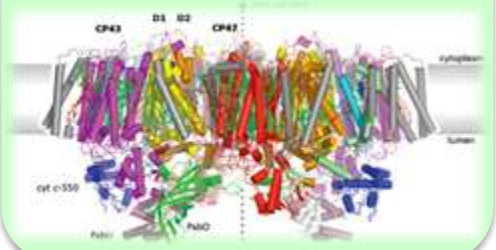
Increasing faith in global science networks.



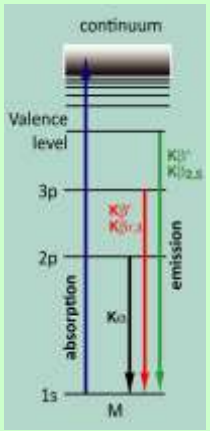
Facilities and instruments now **coupled** by research networks.

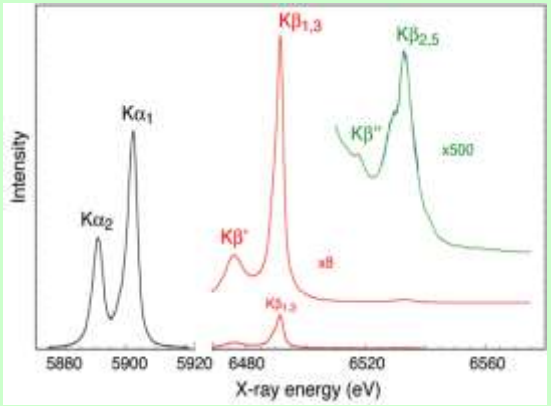
- Recent beam time on free-electron laser (LCLS) at SLAC to take ‘snapshots’ of catalytic reaction in Photosystem II (Nick Sauter et al).
- Data transported to a nearby HPC resource (NERSC) for real-time computational analysis.
- This one experiment *tripled* NERSC’s network utilization.

X-ray diffraction
(structure)

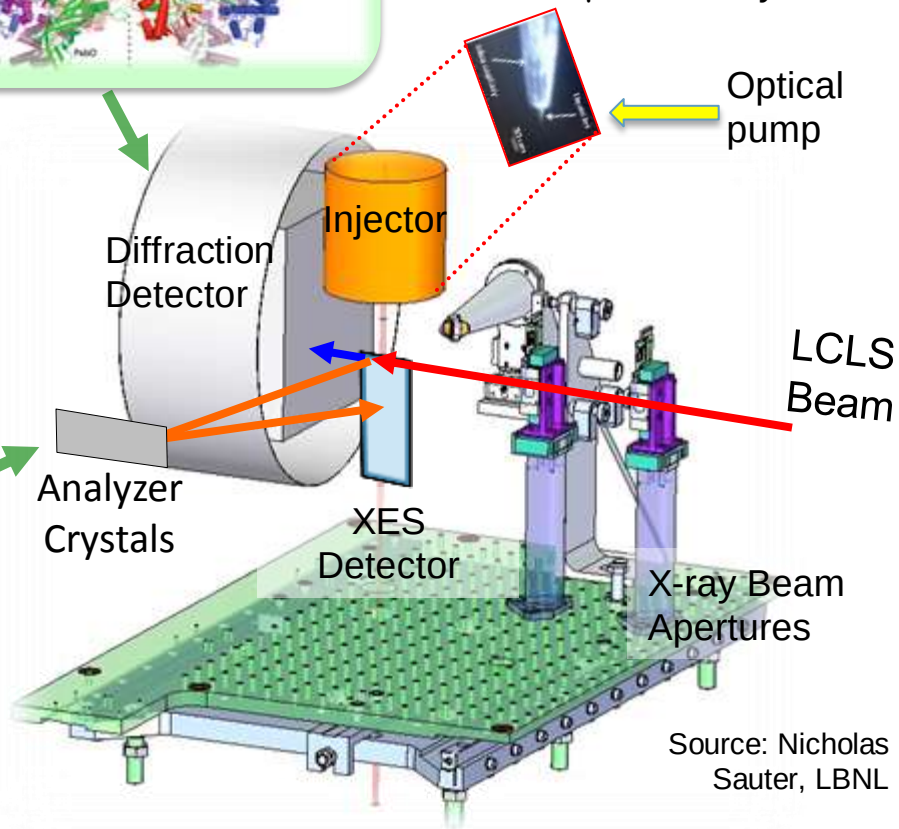


X-ray emission spectroscopy
(Chemistry at the catalytic site)





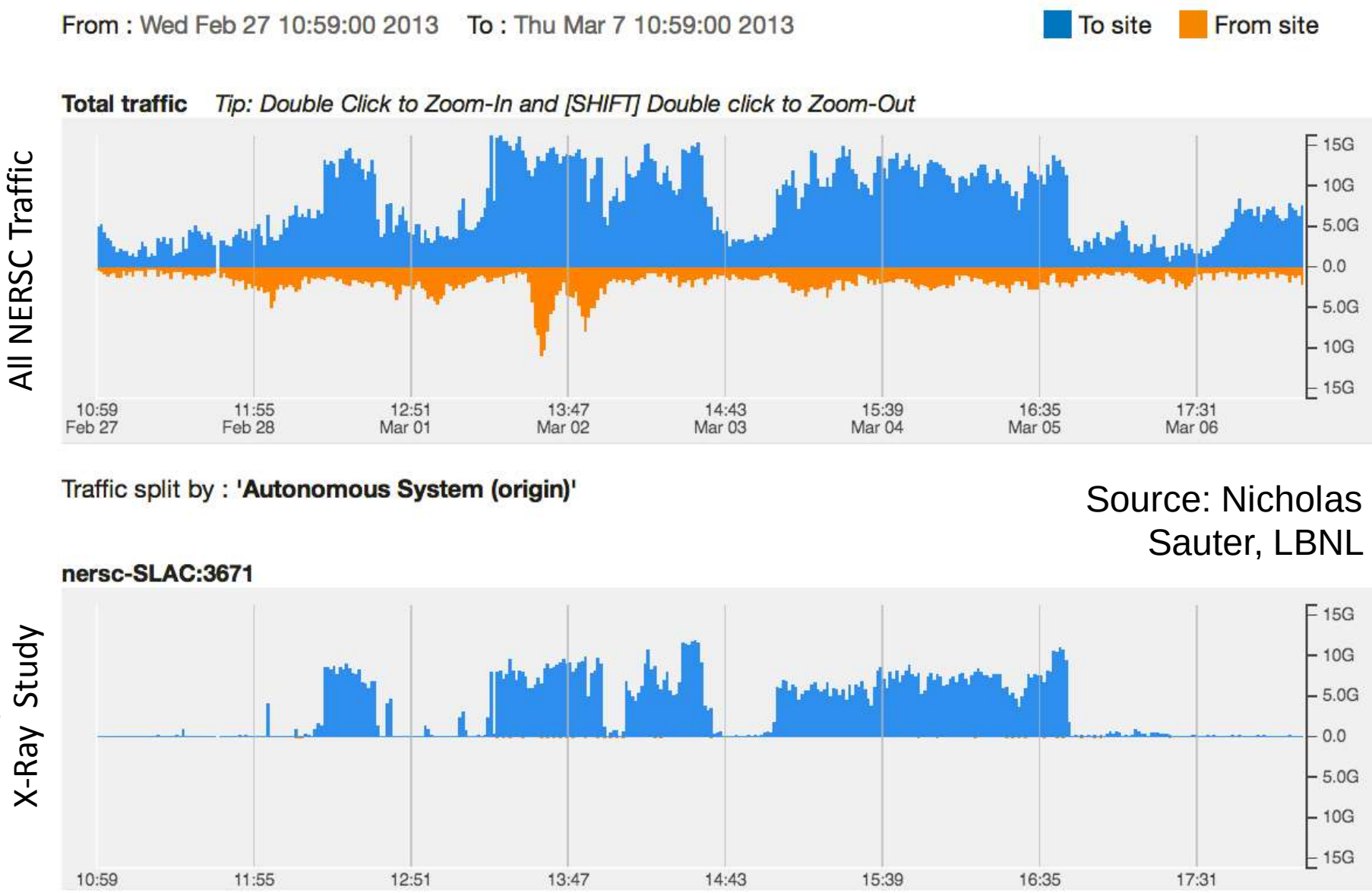
- charge density/spin state
- ligand environment



Source: Nicholas Sauter, LBNL

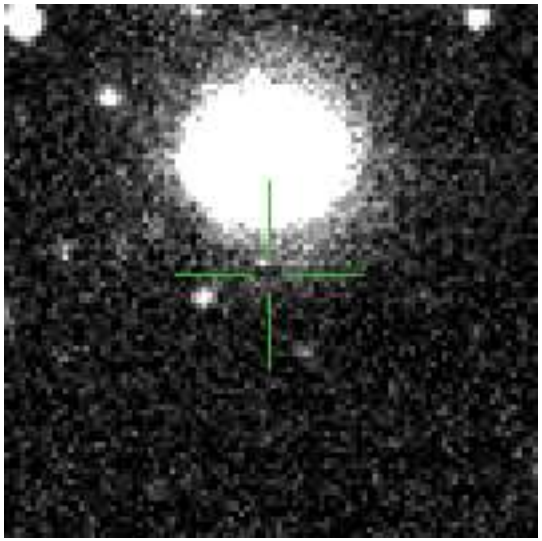
Kern et al (2012) PNAS 109: 9721
Sierra et al (2012) Acta Cryst D68: 1584
Mori et al (2012) PNAS 109: 19103

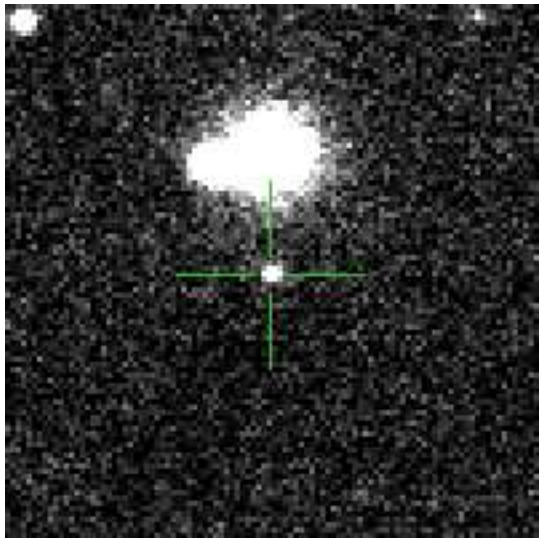
SLAC / LCLS ➔NERSC (http://my.es.net)

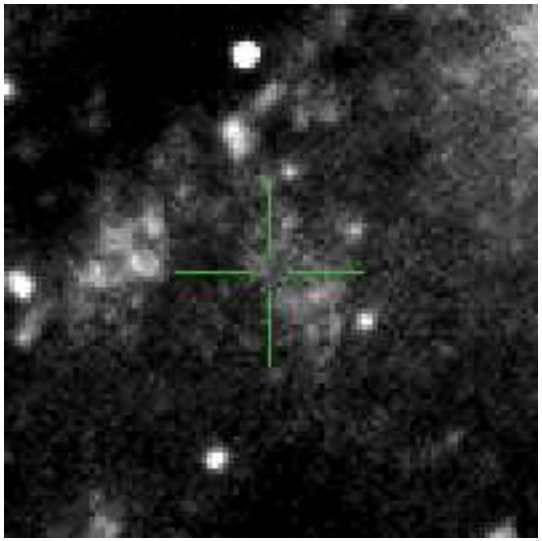


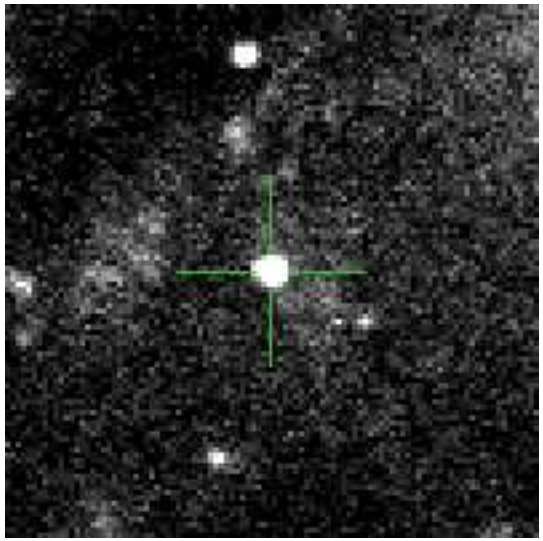
Source: Nicholas Sauter, LBNL

Coupled facilities lead to rapid discovery of earliest observed supernova (August 2011):









Caught at magnitude ~17.4, 100,000 times fainter than the eye can see.

20% rise between first 2 detections separated by 1 hour.

Peak was 1000x brighter.

Source: Peter Nugent
LBNL / UCB

Role of network in discovery of most-observed supernova in history

Palomar Transient Factory (PTF) survey data, automatically collected by robotic instrument.

Apparent mechanism: "Internet just works."

Actual mechanism: dedicated, science-optimized networks linking large-scale instruments and computational resources.

Berkeley Lab / NERSC

perfSONAR

[Science DMZ]

perfSONAR

California High-Performance Wireless Research and Education Network

Energy Sciences Network

perfSONAR

155 Mbps point-to-point microwave systems.

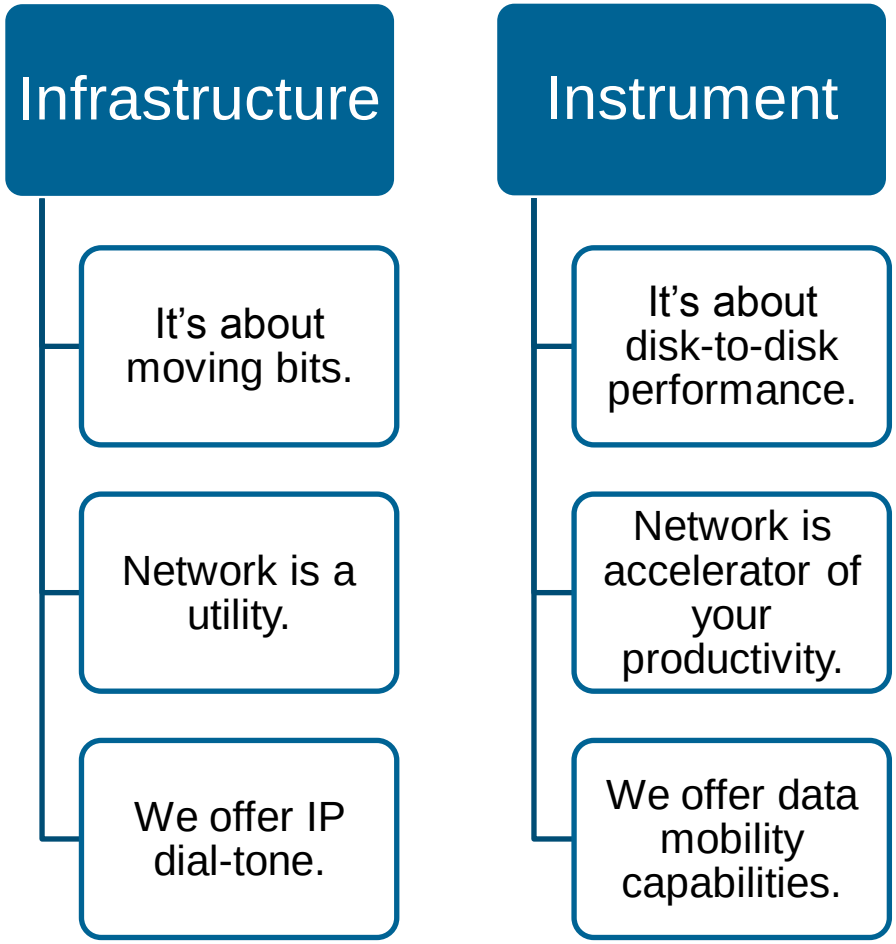
perfSONAR

[Science DMZ]

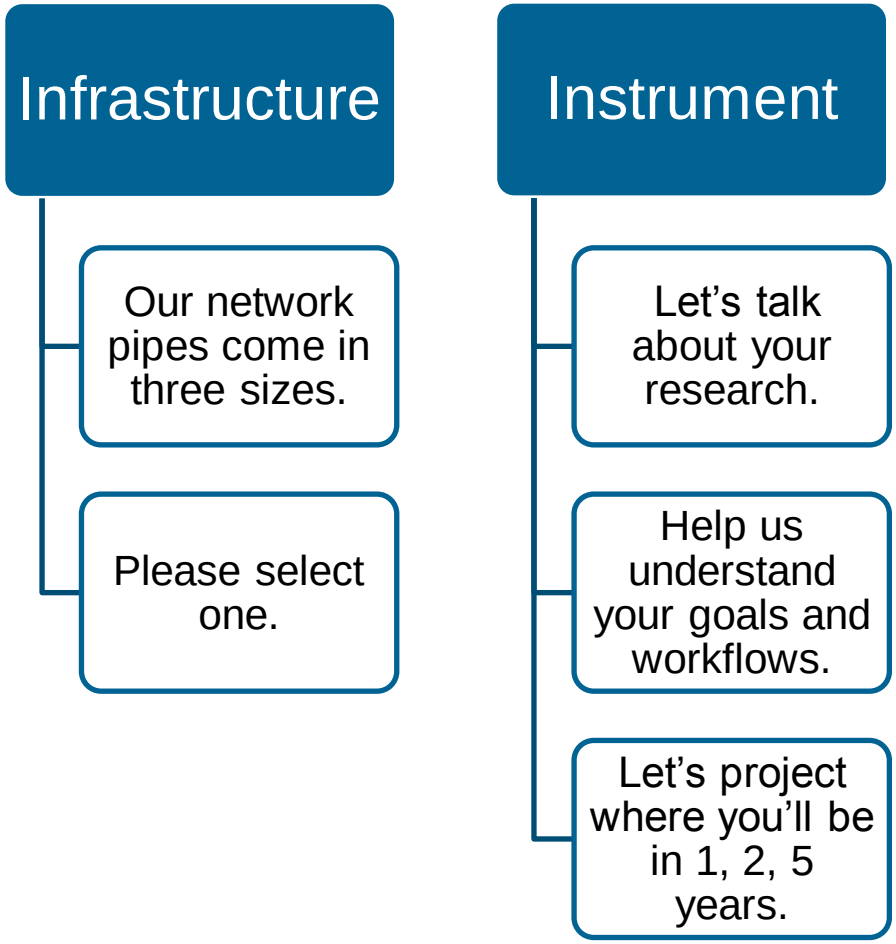
perfSONAR

UCSD and San Diego Supercomputer Center

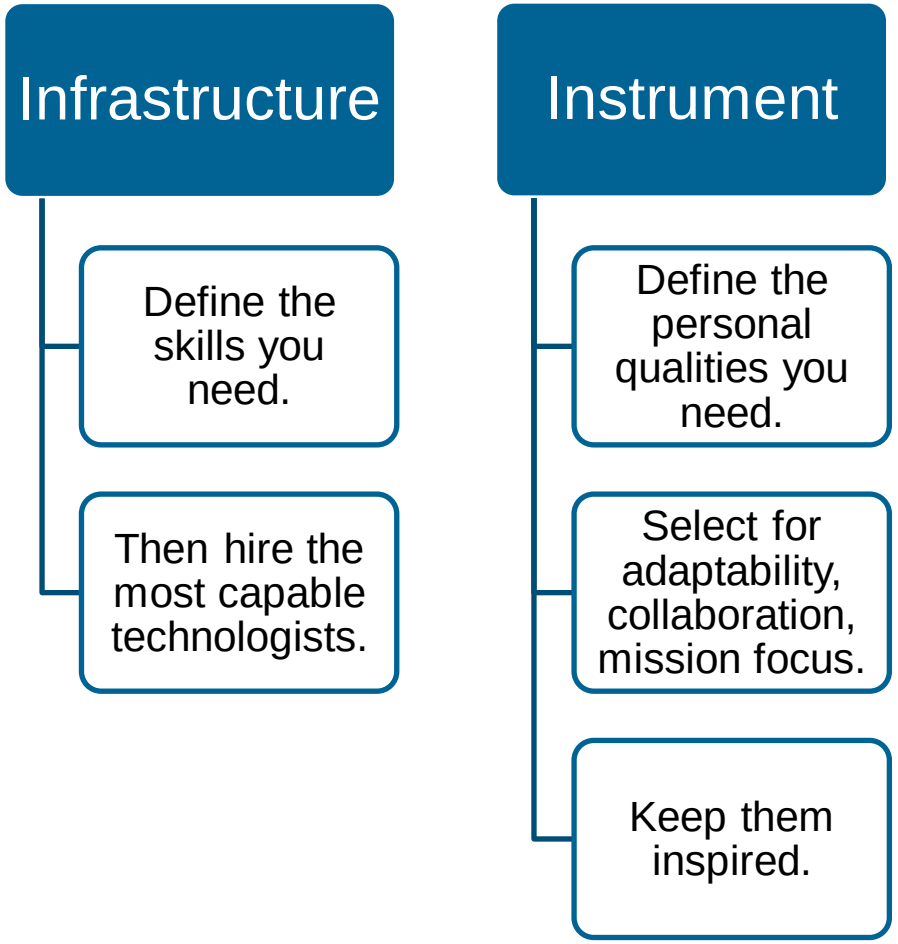
Two Perspectives on Data Mobility



Two Perspectives on Requirements



Two Perspectives on Hiring



Outline

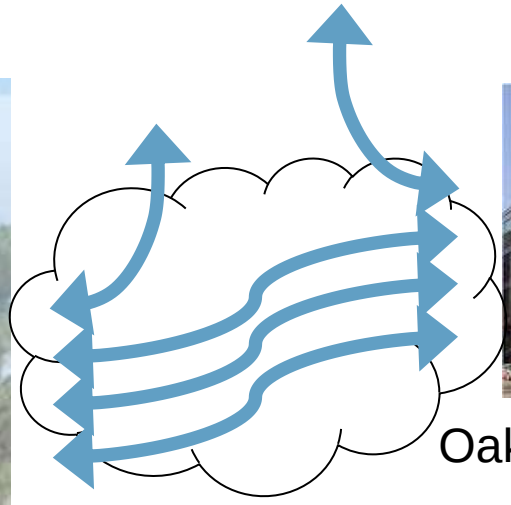
Energy Sciences Network

Network as Discovery Instrument

Faster Data

Research Networks and Campuses Now Making Leap to 100G Networking

Computational Research and Theory Building, Berkeley Lab



ESnet

Oakland Scientific Facility

- 2 x 100G for file system links
- 1 x 100G for inter-site traffic
- 2 x 100G for WAN traffic

Can you move 1TB in 20 minutes? If not, please raise your expectations.

File size	1 Minute	5 Minutes	20 Minutes	1 Hour
100TB	13,333.3 Gbps	2,666.7 Gbps	666.7 Gbps	222.2 Gbps
10TB	1,333.3 Gbps	266.7 Gbps	66.7 Gbps	22.2 Gbps
1TB	133.3 Gbps	26.7 Gbps	6.7 Gbps	2.2 Gbps
100GB	13.3 Gbps	2.7 Gbps	666.7 Mbps	222.2 Mbps
10GB	1.3 Gbps	266.7 Mbps	66.7 Mbps	22.2 Mbps
1GB	133.3 Mbps	26.7 Mbps	6.7 Mbps	2.2 Mbps
100MB	13.3 Mbps	2.7 Mbps	666.7 Kbps	222.2 Kbps
10MB	1.3 Mbps	266.7 Kbps	66.7 Kbps	22.2 Kbps
1MB	133.3 Kbps	26.7 Kbps	6.7 Kbps	2.2 Kbps

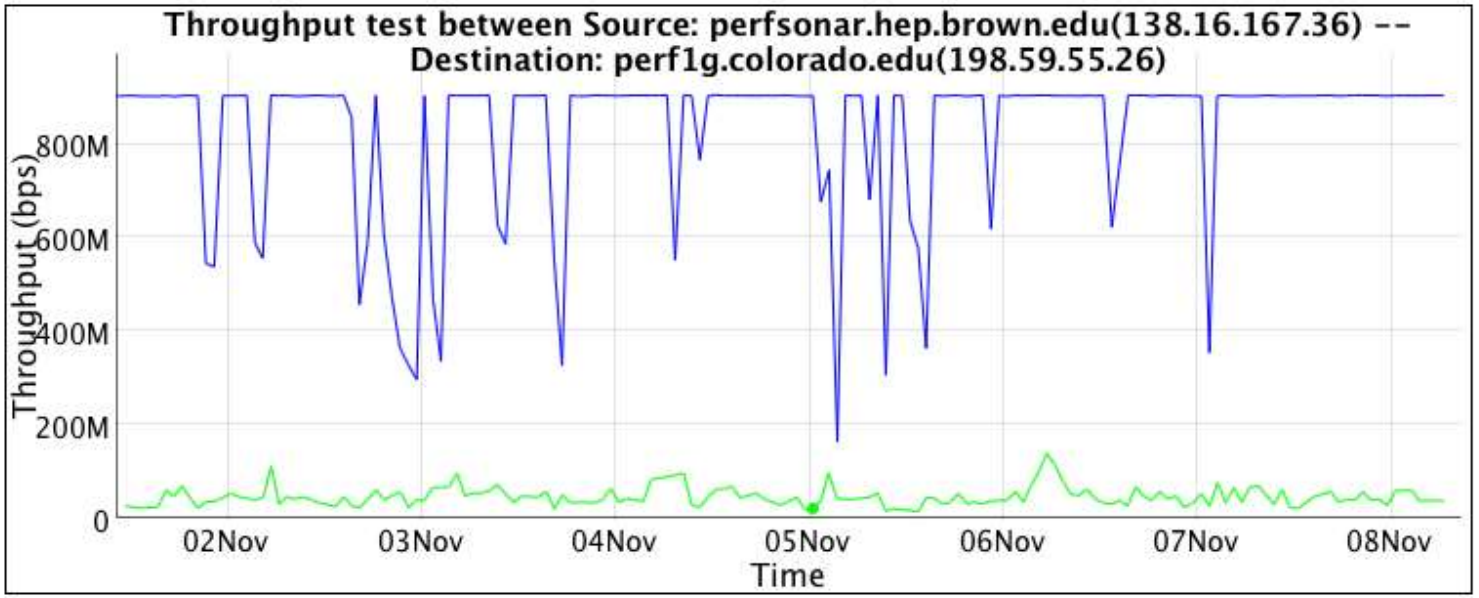
Time to transfer

File size	8 Hours	24 Hours	7 Days	30 Days
10PB	2,777.8 Gbps	925.9 Gbps	132.3 Gbps	30.9 Gbps
1PB	277.8 Gbps	92.6 Gbps	13.2 Gbps	3.1 Gbps
100TB	27.8 Gbps	9.3 Gbps	1.3 Gbps	308.6 Mbps
10TB	2.8 Gbps	925.9 Mbps	132.3 Mbps	30.9 Mbps
1TB	277.8 Mbps	92.6 Mbps	13.2 Mbps	3.1 Mbps
100GB	27.8 Mbps	9.3 Mbps	1.3 Mbps	308.6 Kbps
10GB	2.8 Mbps	925.9 Kbps	132.3 Kbps	30.9 Kbps
1GB	277.8 Kbps	92.6 Kbps	13.2 Kbps	3.1 Kbps
100MB	27.8 Kbps	9.3 Kbps	1.3 Kbps	0.3 Kbps

Time to transfer

Network throughput required to move Y bytes in X time (http://fasterdata.es.net).

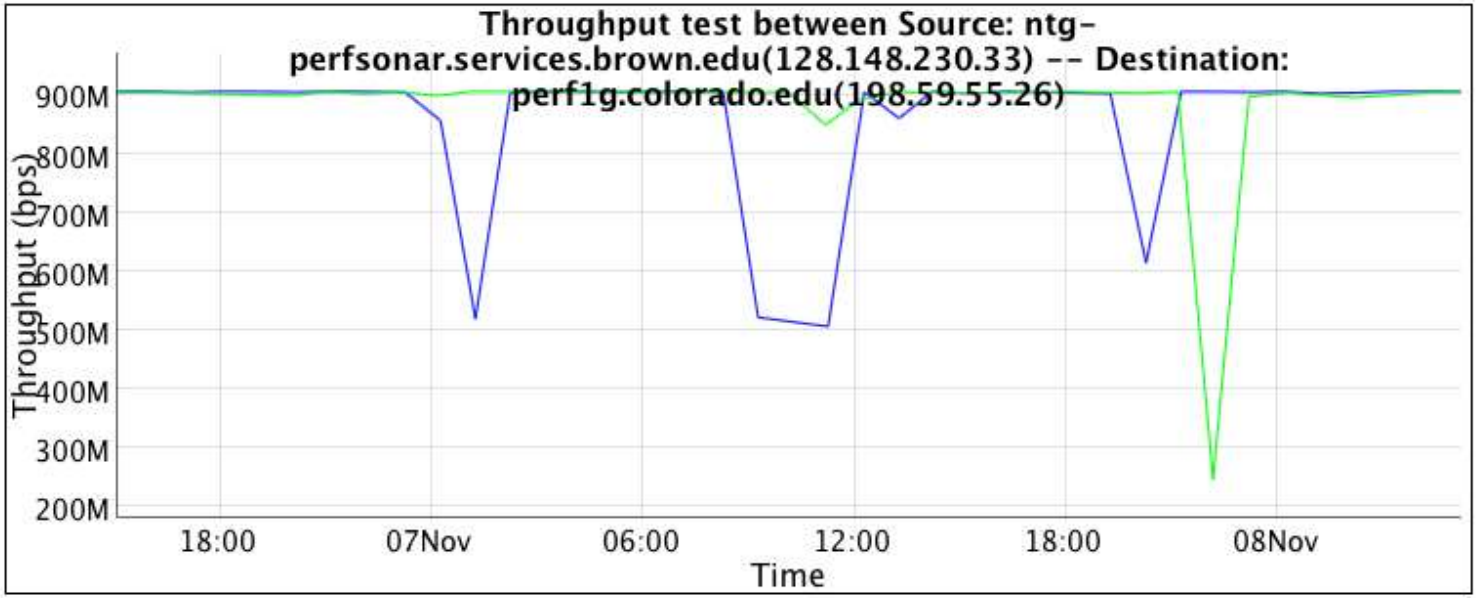
Firewalls Harmful to Large Data Flows (and Unnecessary)



Graph Key

Src-Dst throughput

Dst-Src throughput



Graph Key

Src-Dst throughput

Dst-Src throughput

Source: Jason Zurawski, Internet 2

The Solution: Science DMZ Design Pattern

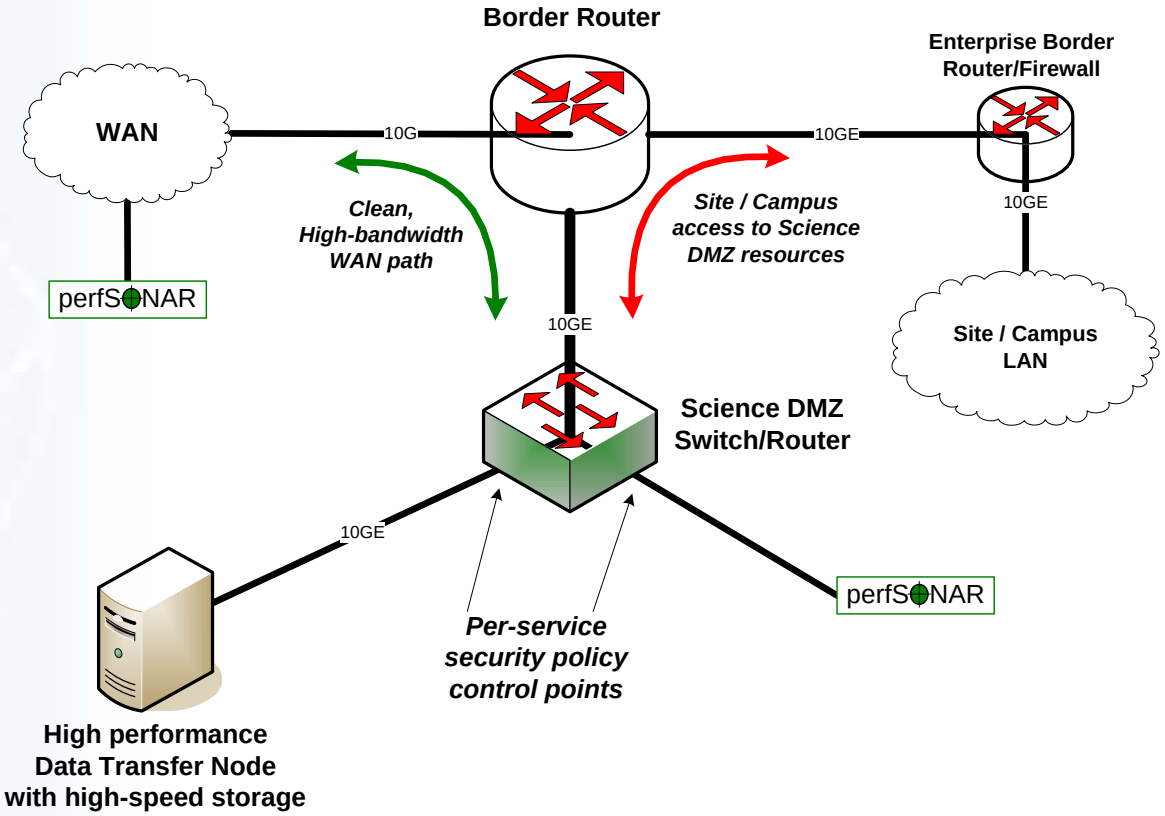
Adopted by HEP (CMS/ATLAS), Climate (ESG), supercomputer centers, dozens of US universities and national labs.

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

- Friction-free network path
  - A few highly-capable network devices (wire-speed, deep queues)
  - Option for virtual circuit connectivity
  - Security policies tailored for science workflows
  - Located at or near site perimeter
- Dedicated, high-performance Data Transfer Nodes (DTNs)
  - Hardware, operating system, libraries all optimized for transfer
  - Appropriate data transfer tools such as Globus Online and GridFTP
- Performance measurement system
  - perfSONAR
  - forensics and testing

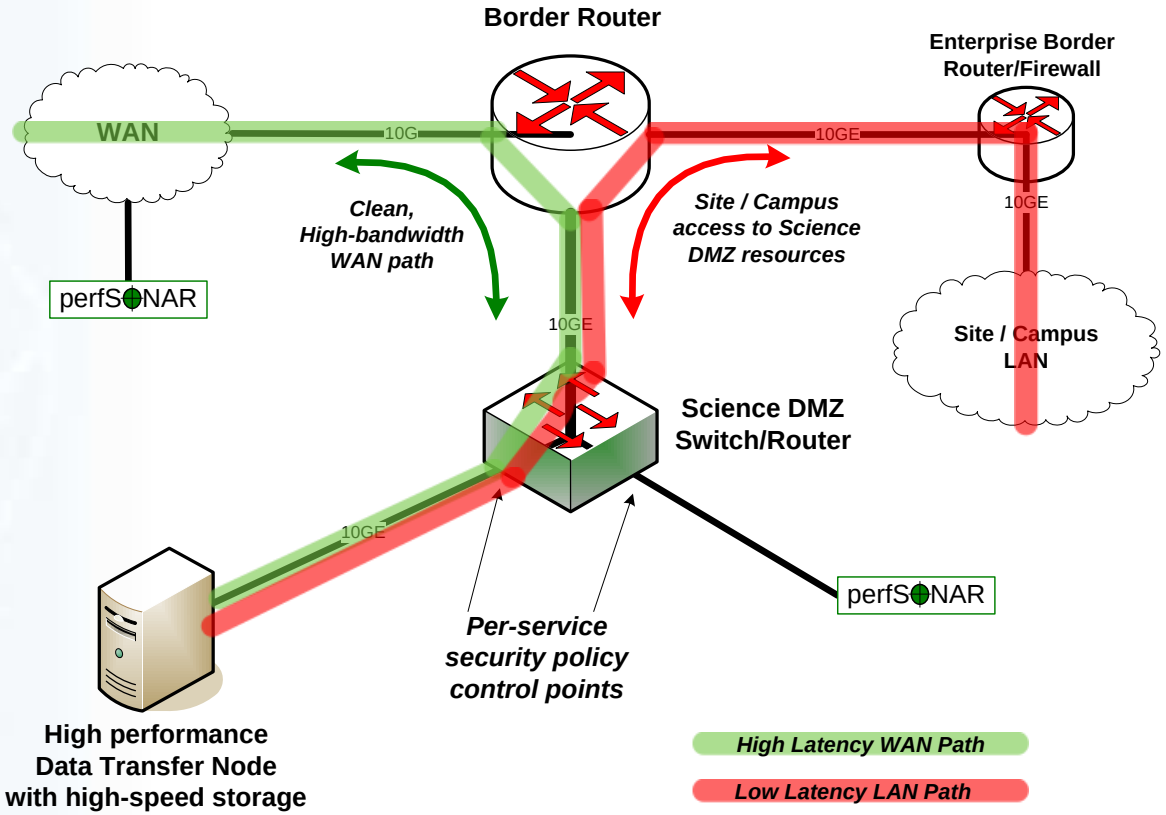
Much more information: http://fasterdata.es.net/science-dmz/

Prototype Science DMZ



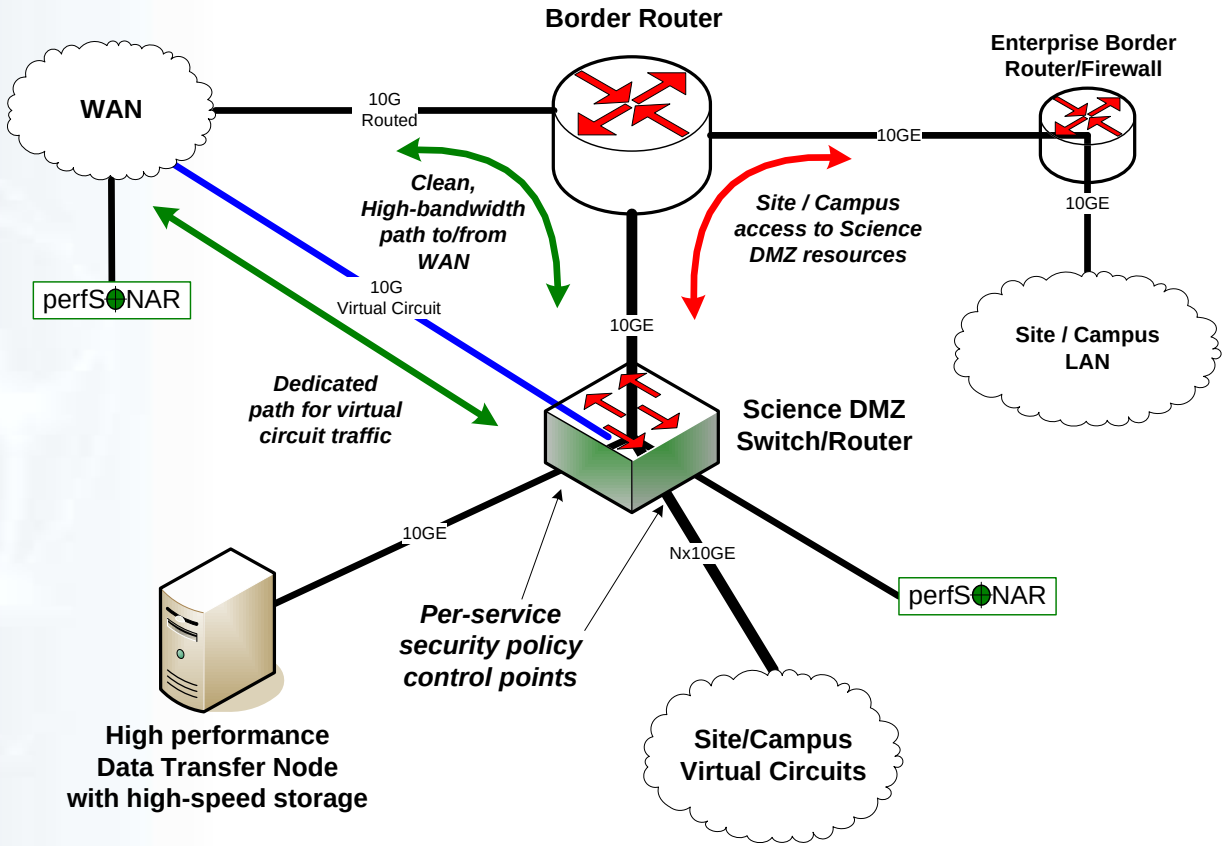
source:
Eli Dart
ESnet

Prototype Science DMZ Data Path



source:
Eli Dart
ESnet

Science DMZ With Virtual Circuits/Openflow



source:
Eli Dart
ESnet

Globus Online: a Tool we Like

ESnet & Globus Online

- Working jointly to integrate data transfer tools with new network capabilities (such as guaranteed bandwidth)
- Developing outreach material, targeting small and medium-sized collaborations.
- Evangelizing science-friendly campus architectures.
- Coordinating applied research (with particular emphasis on integration).

perfSONAR Enables Data Mobility

A platform and protocol for:

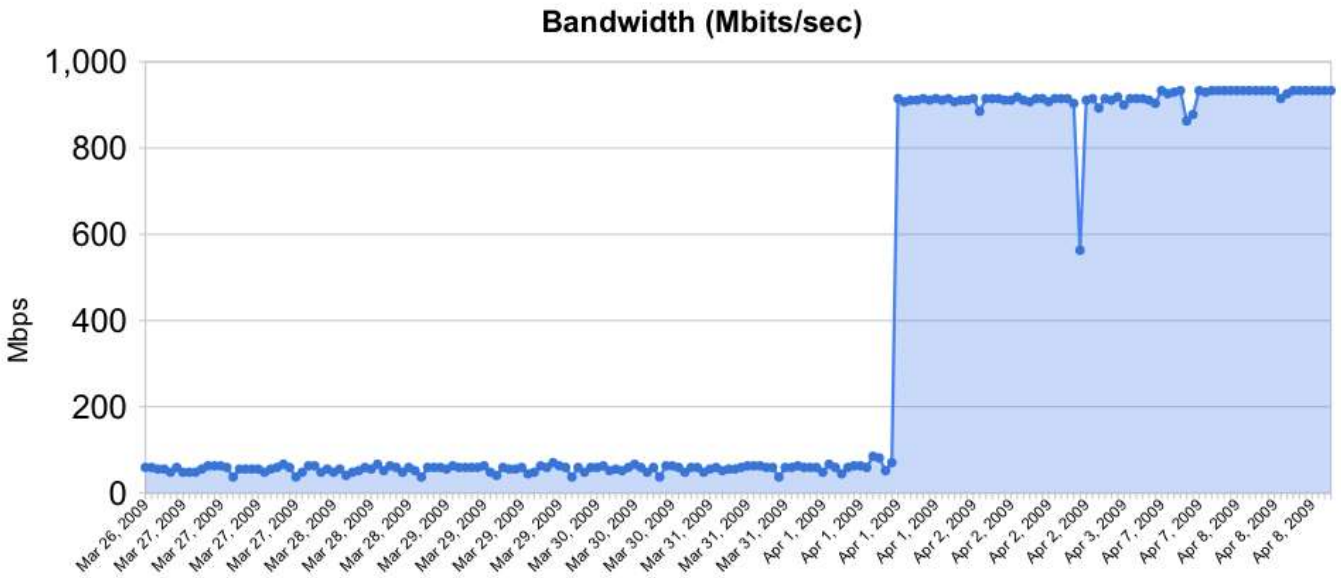
- finding and fixing network problems (“soft failures”)
- setting performance expectations

Particularly important when path of interest crosses multiple domains.

ESnet has deployed 80 perfSONAR nodes on its backbone and site borders:

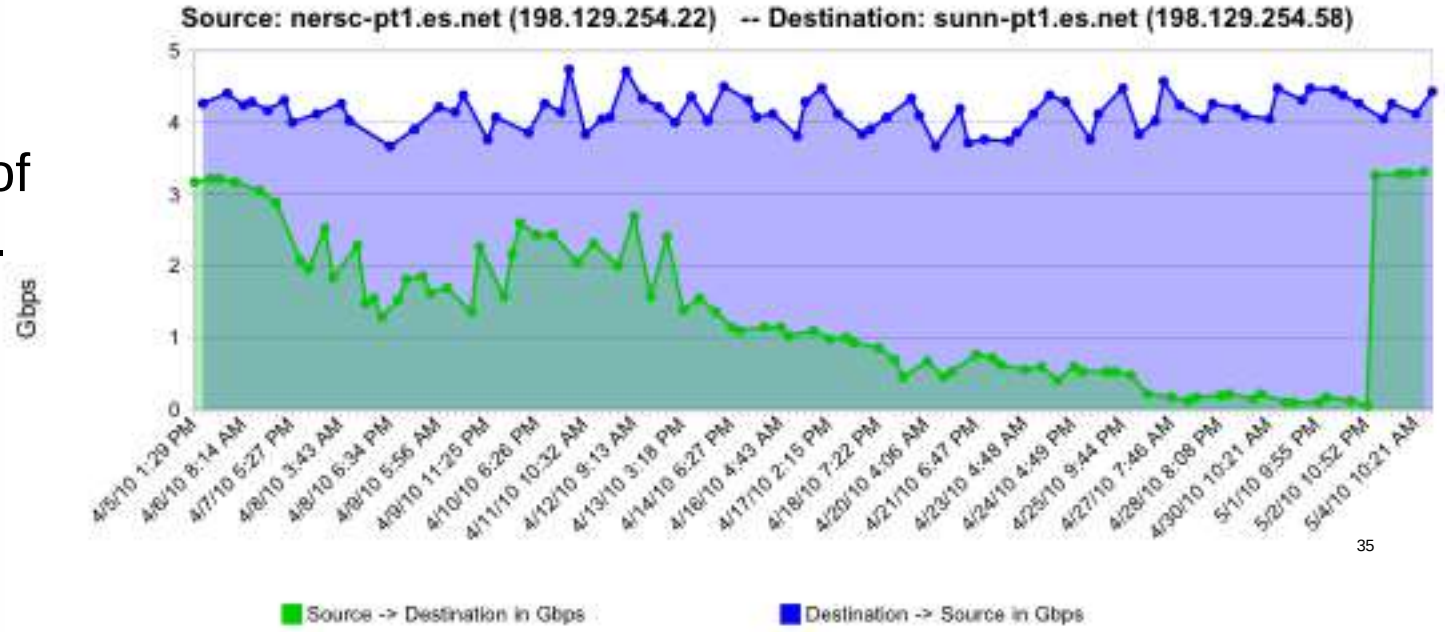
- 36 throughput nodes running regular BWCTL tests
- 34 latency nodes running regular OWAMP tests
- 10 combined hosts running both BWCTL and OWAMP

Finding and Fixing Soft Failures



Full route table, before / after reboot.

Gradual failure of optical line card.

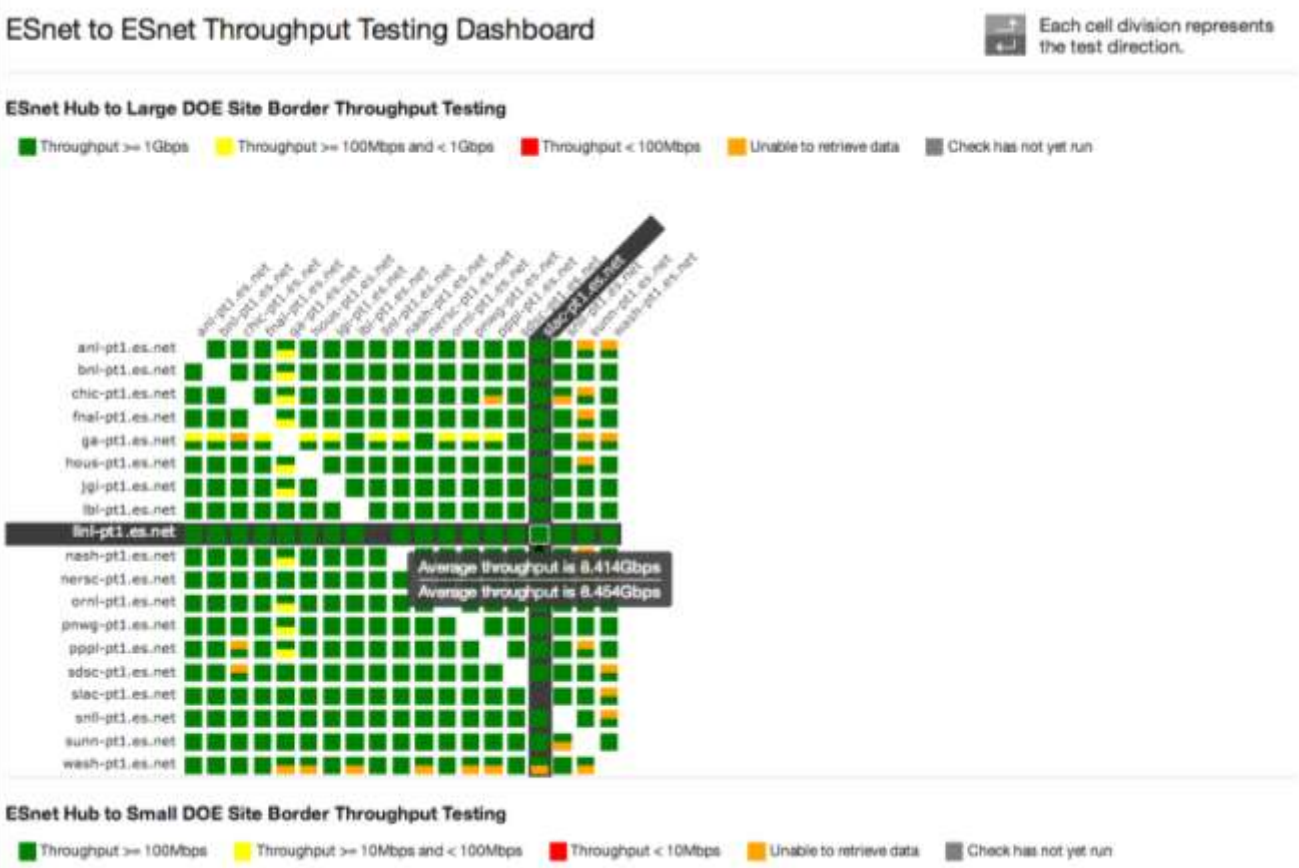


World-Wide perfSONAR-PS Deployments: 536 bwctl nodes, 505 owamp nodes as of Jan 7

Performance Dashboard (my.es.net)

Over 1000 point-to-point tests on ESnet alone!

- cries out for a visual tool



We believe Science DMZ is **critical**. Here's how to get started.

Guide to faster data:

- http://fasterdata.es.net/

Science DMZ explained:

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

When questions about security arise, please start here:

- http://fasterdata.es.net/science-dmz/science-dmz-security/
- http://www.internet2.edu/presentations/tip2013/20130115-dart-science-dmz.pdf
- http://events.internet2.edu/2013/tip/agenda.cfm?go=session&id=10002778

CC-NIE grants for campus infrastructure upgrades in the US:

- http://www.nsf.gov/pubs/2013/nsf13530/nsf13530.htm

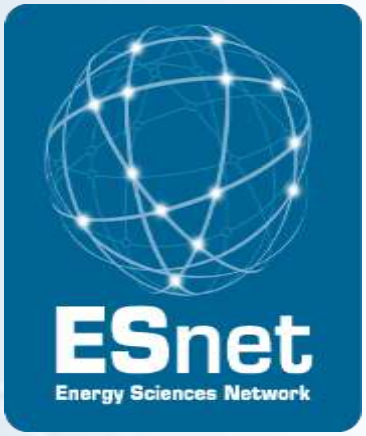
In Summary

Working together, research networks and motivated campuses can impact the world's most pressing problems.

1) Faster data = faster discovery.

2) Science DMZ is the design pattern for faster data.

3) The blueprints: fasterdata.es.net



Thank you!

greg@es.net