



ESnet

ENERGY SCIENCES NETWORK

‘Design Patterns’: Scaling up e-Research

Inder Monga

CTO, Energy Sciences Network

Deputy, Scientific Networking Div.

Lawrence Berkeley National Lab

eResearch NZ 2016,

Queenstown

February 2016



U.S. DEPARTMENT OF
ENERGY
Office of Science



My first visit to New Zealand

 Steve Cotter follows



Pallas Hupé Cotter @pallas_life · 3 Jul 2012

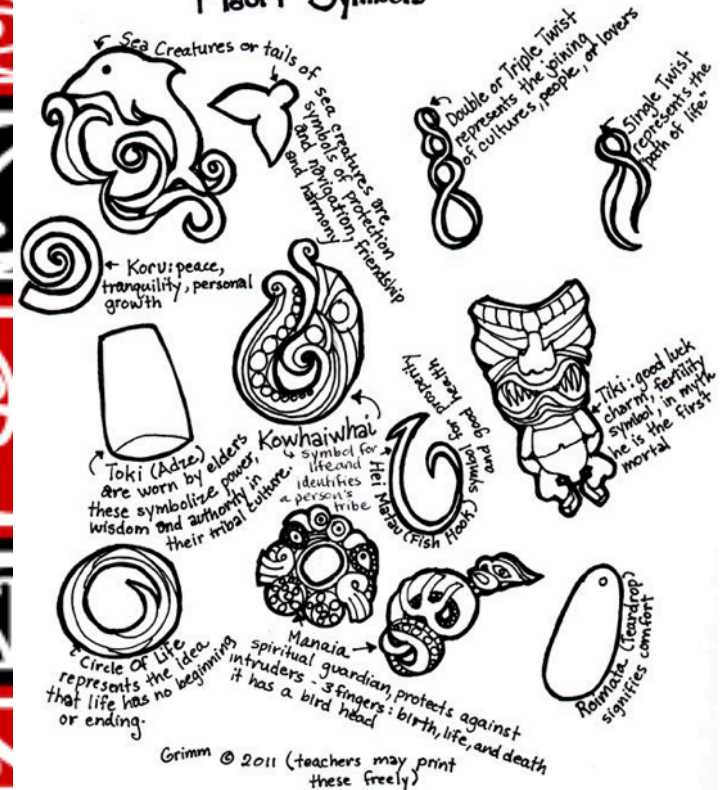
Yup - felt a lot of shaking here in Wellington RT@BreakingNews: USGS: Magnitude 6.3 earthquake hits off New Zealand's North Island @Reuters



Image from NZ, Maori Design Pattern

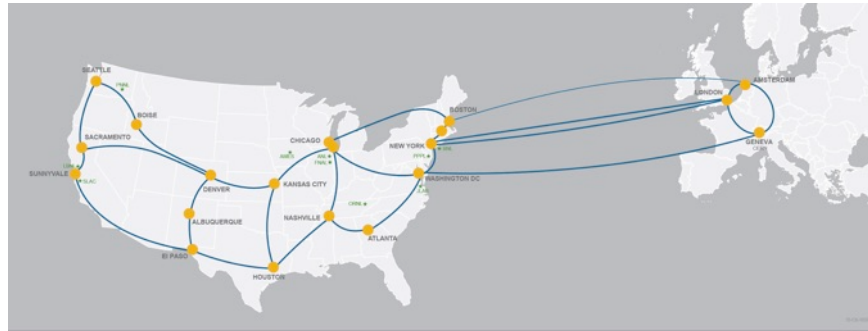


Maori Symbols



Talk

ESnet and
NRENs
Introduction



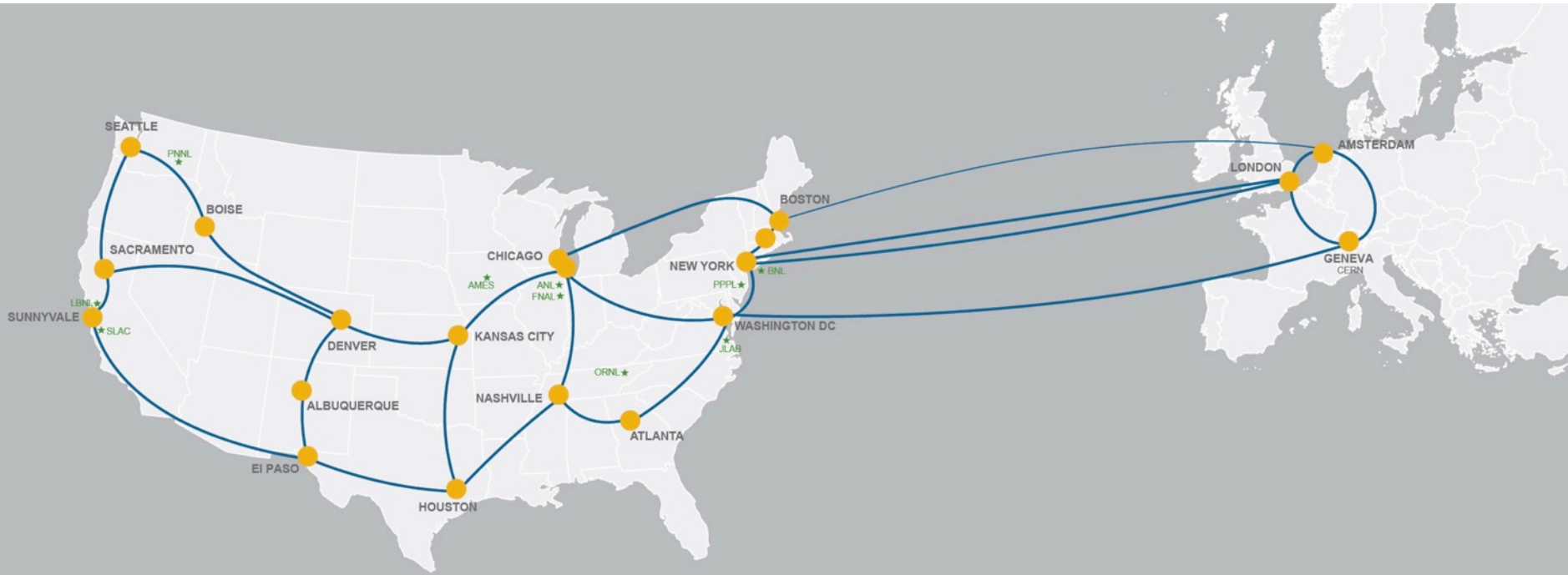
Established
Design
Patterns



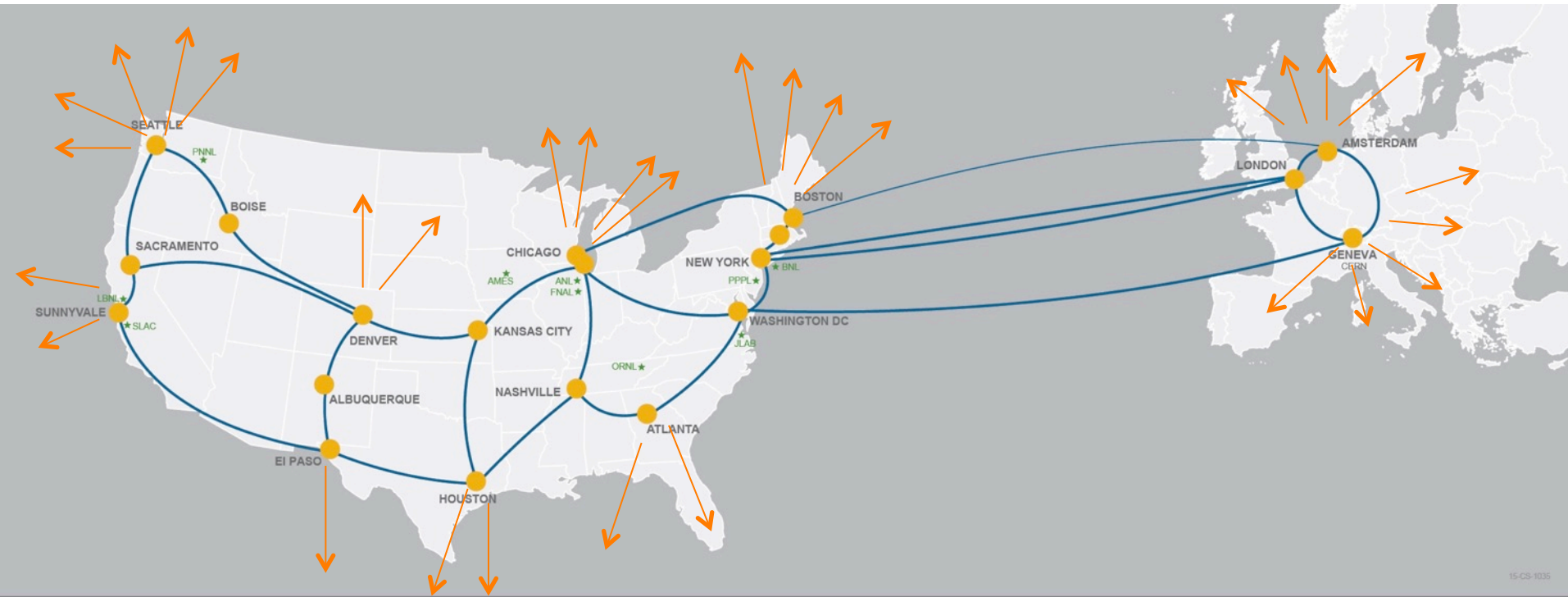
Emerging
Design
Patterns



DOE's Energy Sciences Network (ESnet):



NRENs share fate. No research network can succeed in isolation.

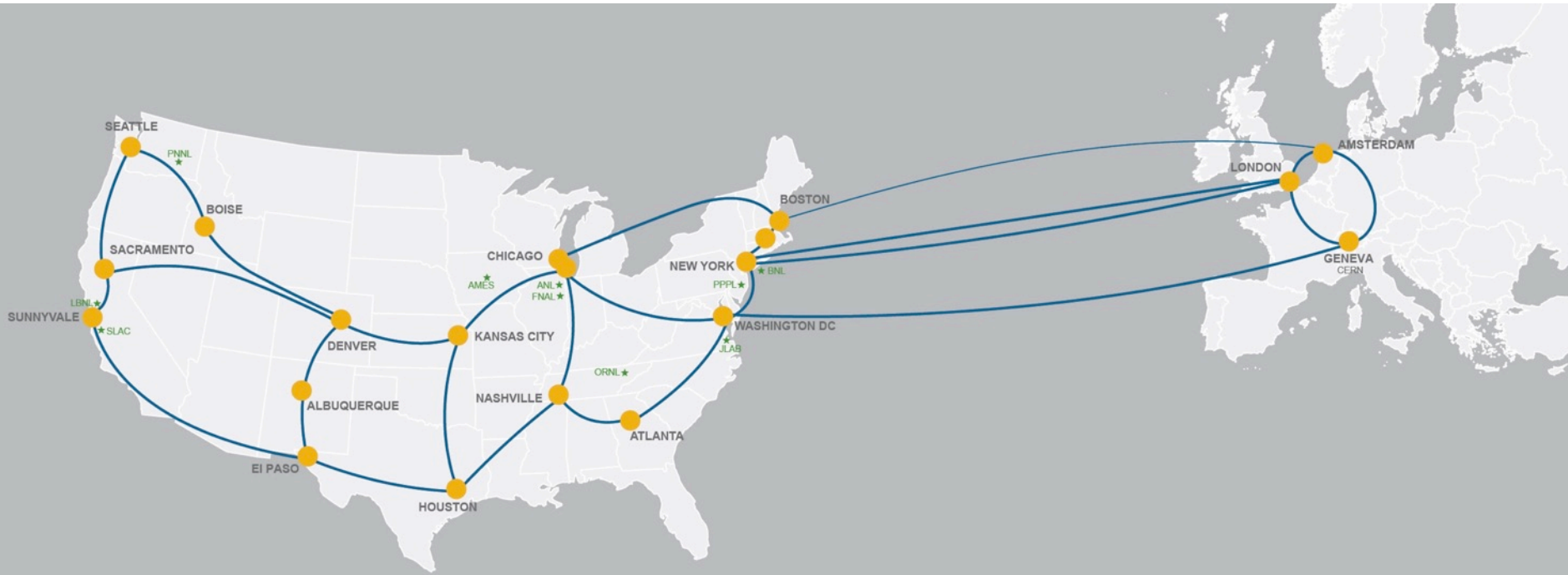


80% of ESnet traffic originates or terminates outside the DOE complex.

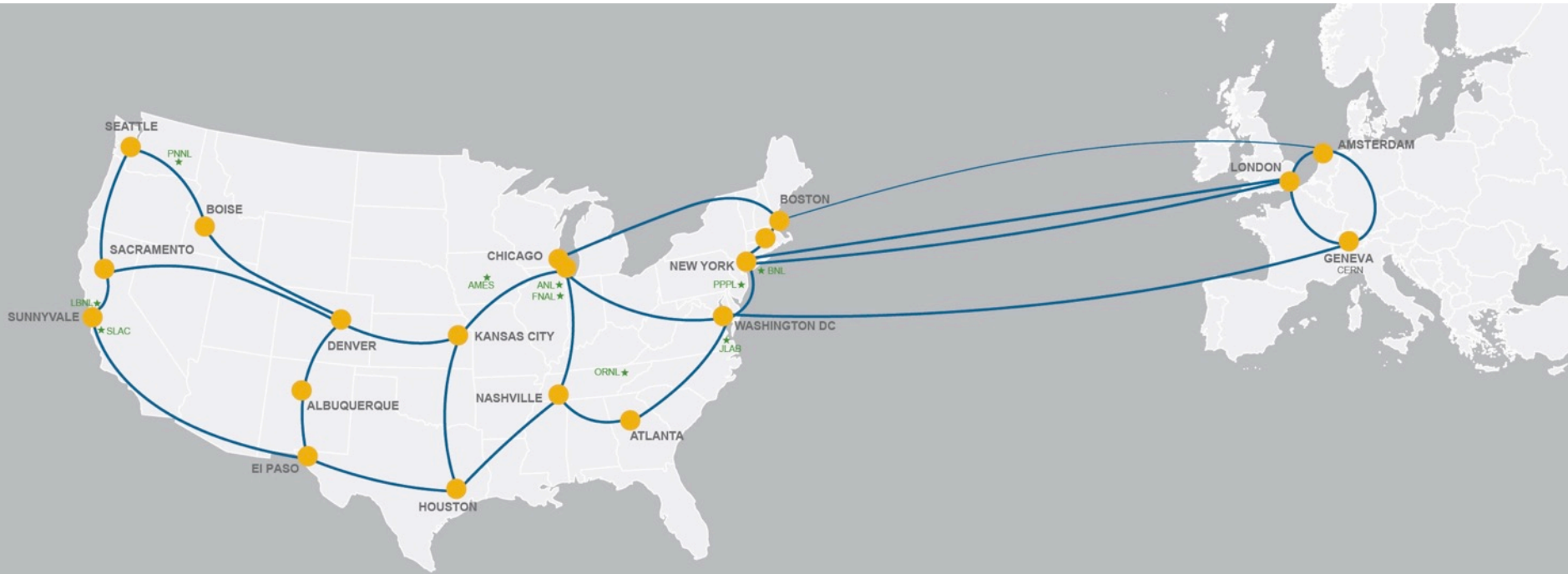
ESnet vision:

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

The most important thing to know about NRENs: they are not ISPs.

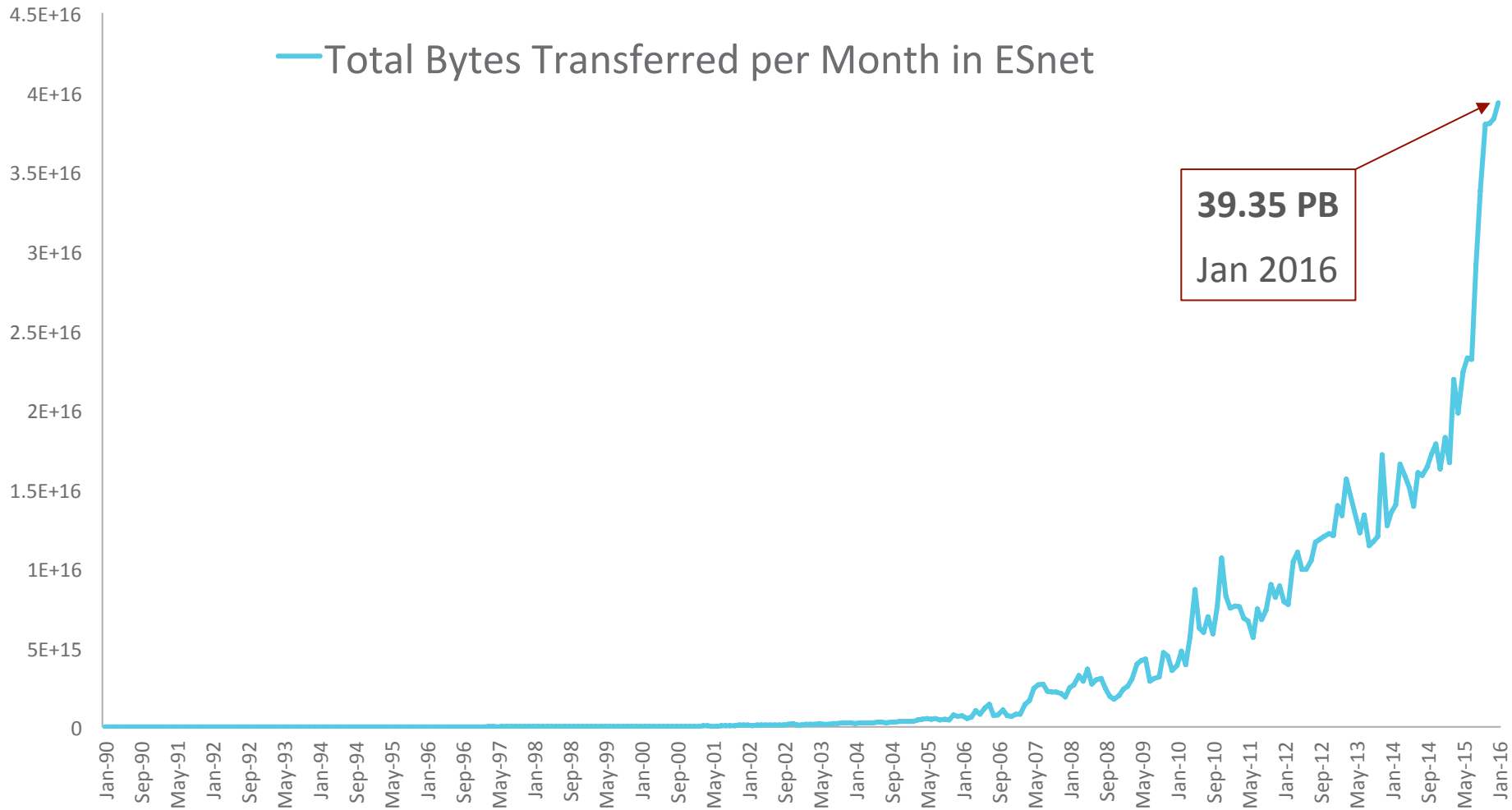


They are instruments for discovery designed to overcome the constraints of geography.

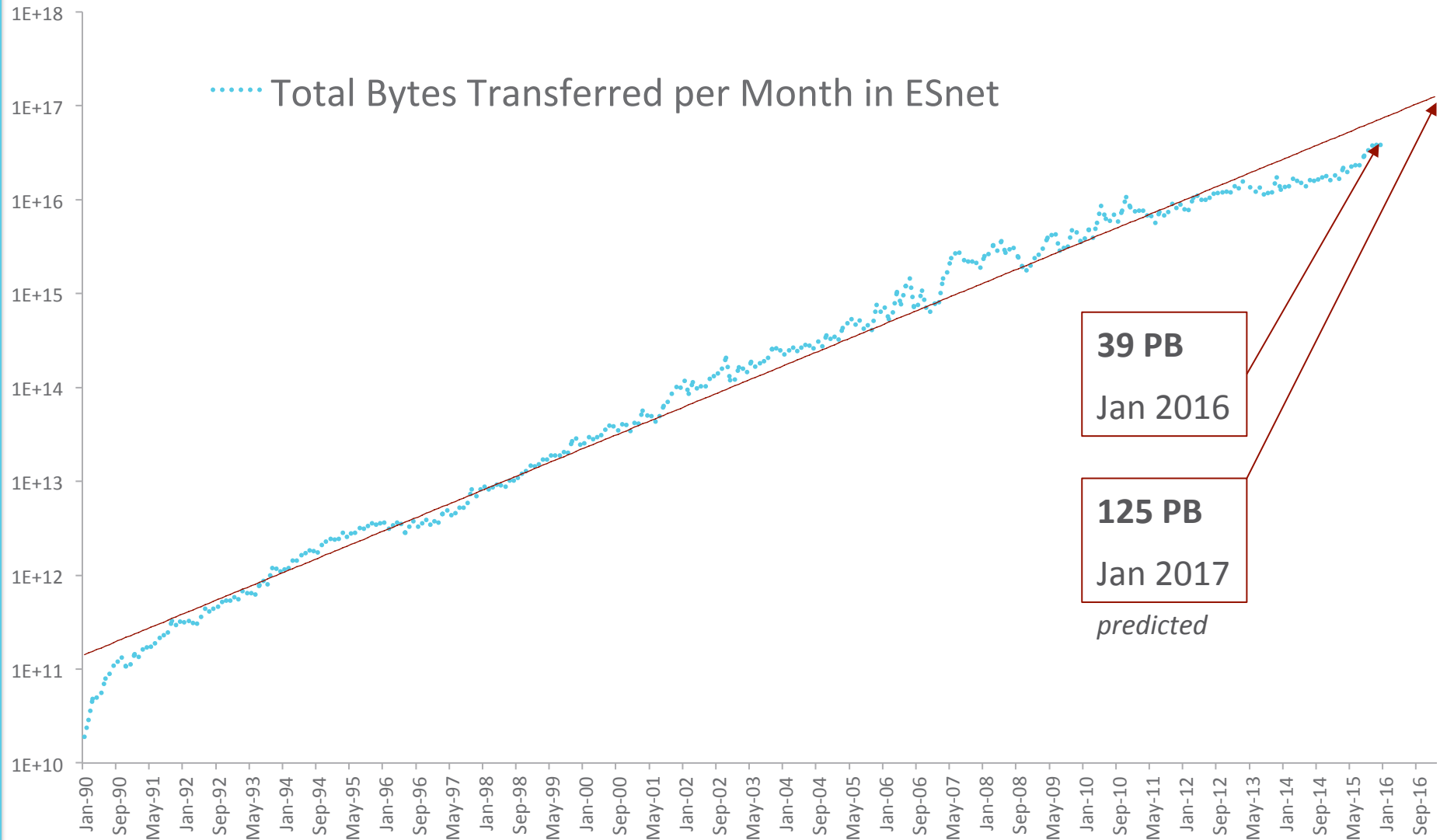


Offering unique capabilities – and optimized for data acquisition, data placement, data sharing, data mobility.

What does success look like?



Planning for growth



Talk

ESnet and
NRENs Intro



Established
Design
Patterns

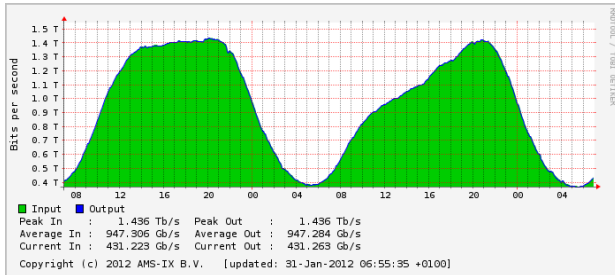
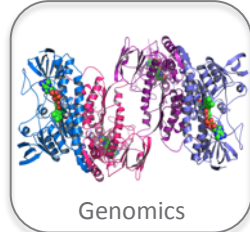
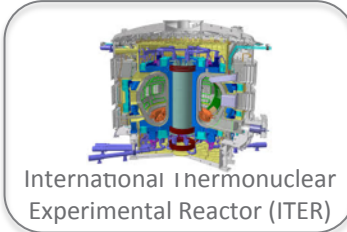
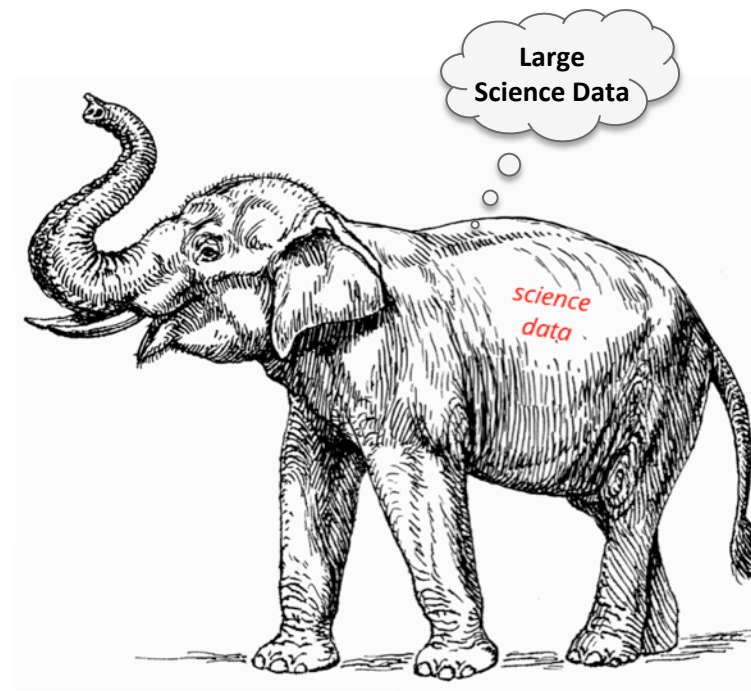
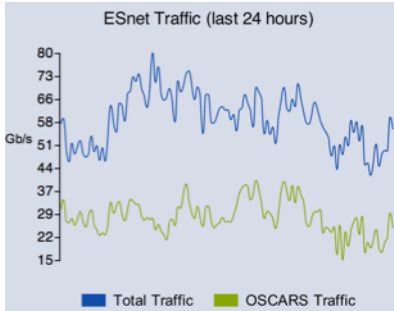
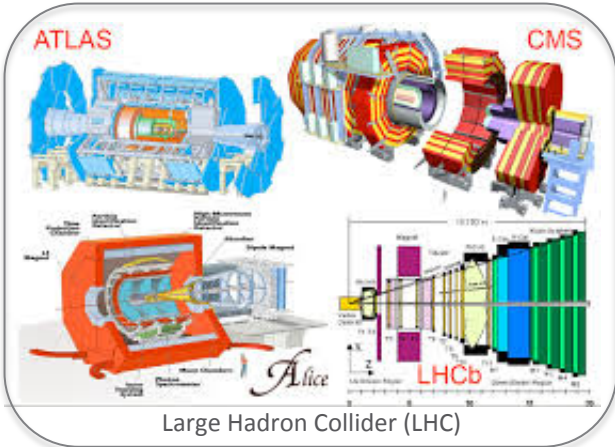


Emerging
Design
Patterns



Design Pattern #1: Protect your *Elephant* Flows

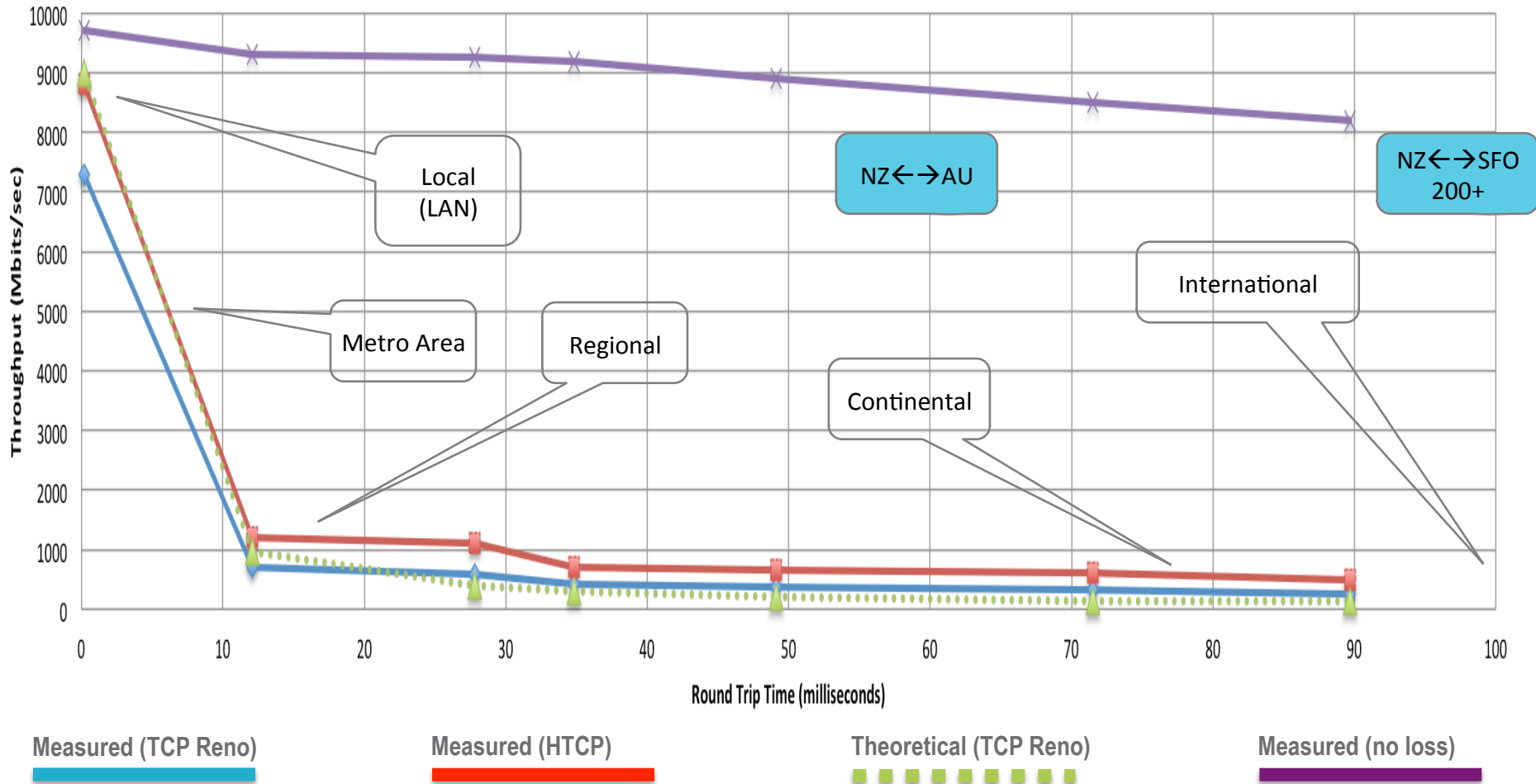




General Internet

Elephant flows require almost *lossless* networks.

Throughput vs. Increasing Latency with .0046% Packet Loss

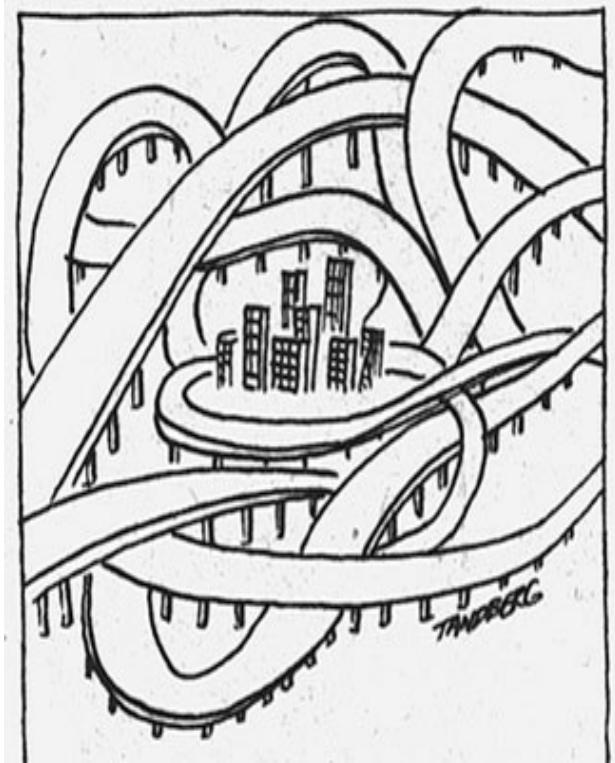


See Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, and Jason Zurawski. The Science DMZ: A Network Design Pattern for Data-Intensive Science. In *Proceedings of the IEEE/ACM Annual SuperComputing Conference (SC13)*, Denver CO, 2013.

Design Pattern #2: Unclog your data taps



Problem and Solution explained illustratively



Big-Data assets **not optimized** for **high-bandwidth access** because of **convoluted campus network and security design**



Science DMZ is a **deliberate, well-designed architecture** to simplify and **effectively on-ramp** 'data-intensive' science to a capable WAN

Data Set Mobility Timeframes - Theoretical

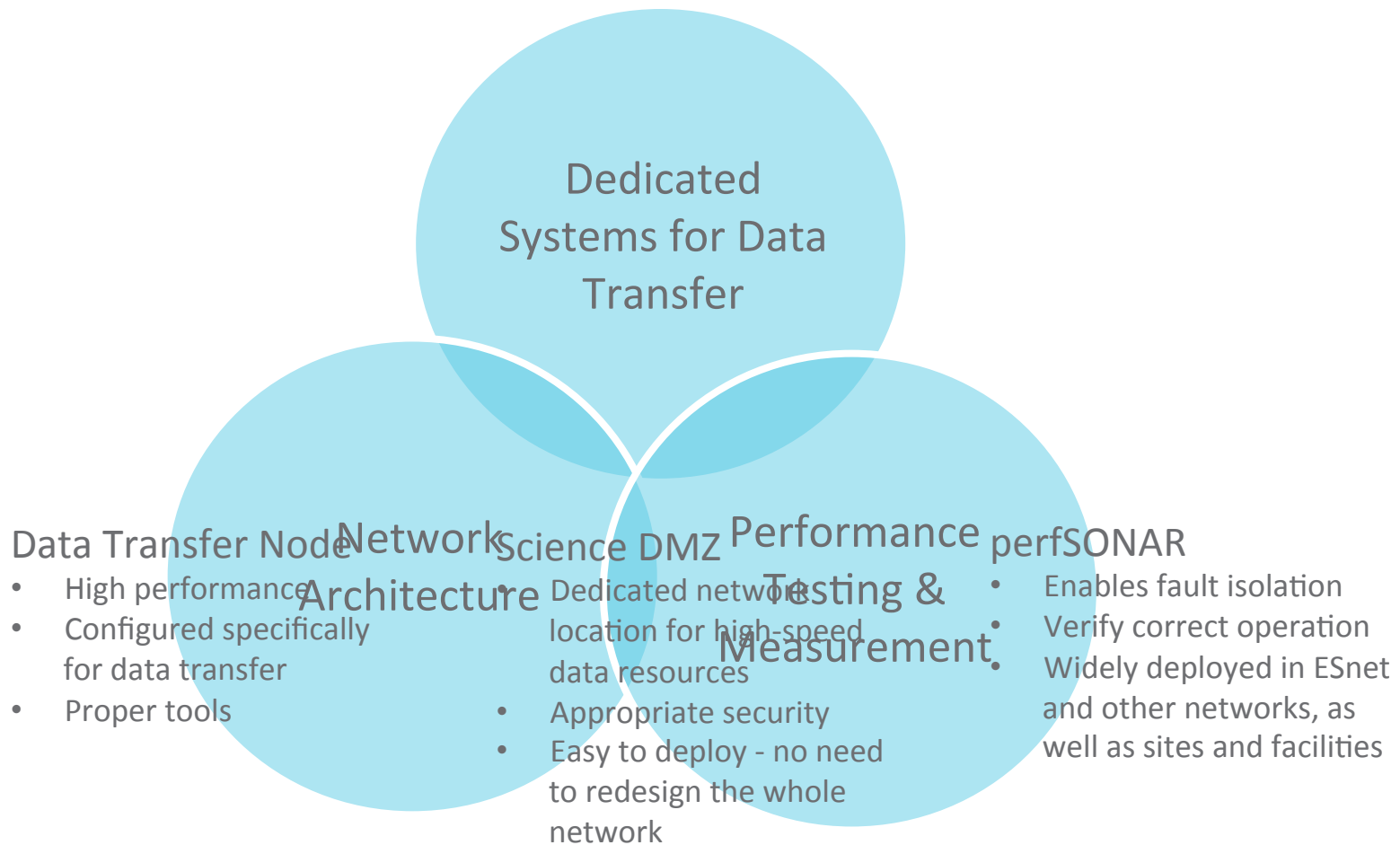
Data set size	1 Minute	5 Minutes	20 Minutes	1 Hour
10PB	1,333.33 Tbps	266.67 Tbps	66.67 Tbps	22.22 Tbps
1PB	133.33 Tbps	26.67 Tbps	6.67 Tbps	2.22 Tbps
100TB	13.33 Tbps	2.67 Tbps	666.67 Gbps	222.22 Gbps
10TB ^{> 100Gbps}	1.33 Tbps	266.67 Gbps	66.67 Gbps	22.22 Gbps
1TB	133.33 Gbps	26.67 Gbps	6.67 Gbps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

This table available at:

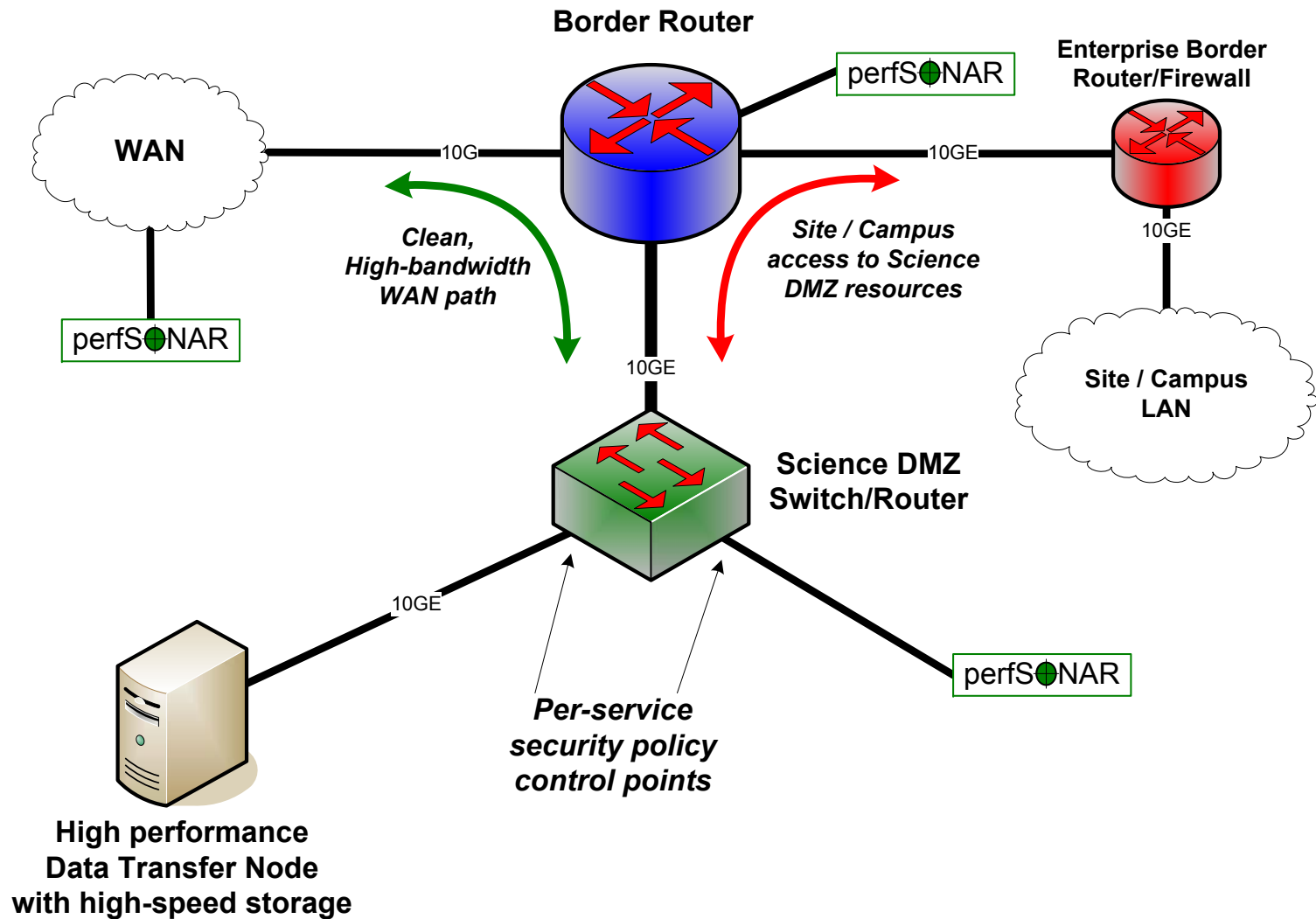
<http://fasterdata.es.net/fasterdata-home/requirements-and-expectations/>



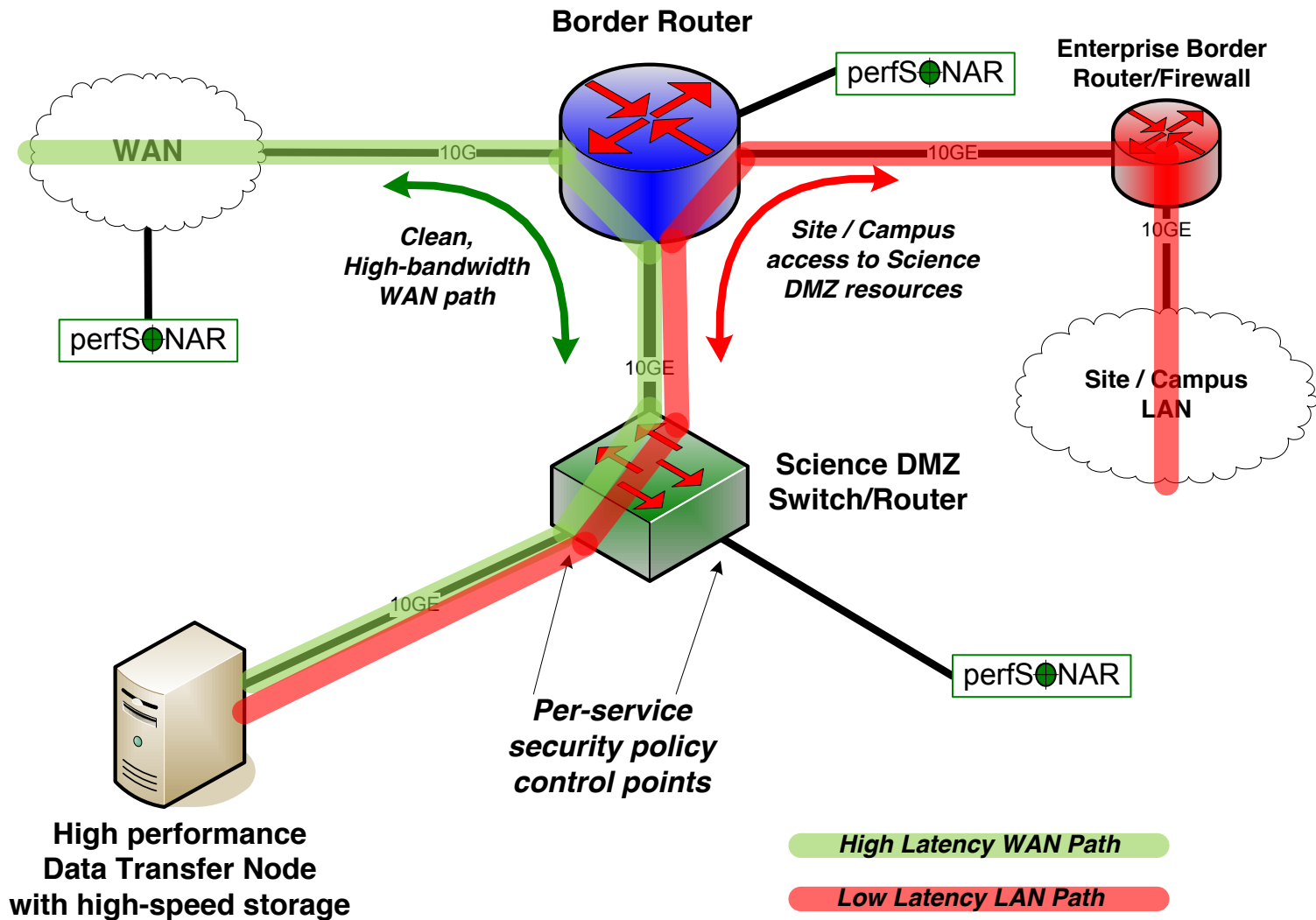
The Science DMZ Design Pattern



Science DMZ Design Pattern (Abstract)

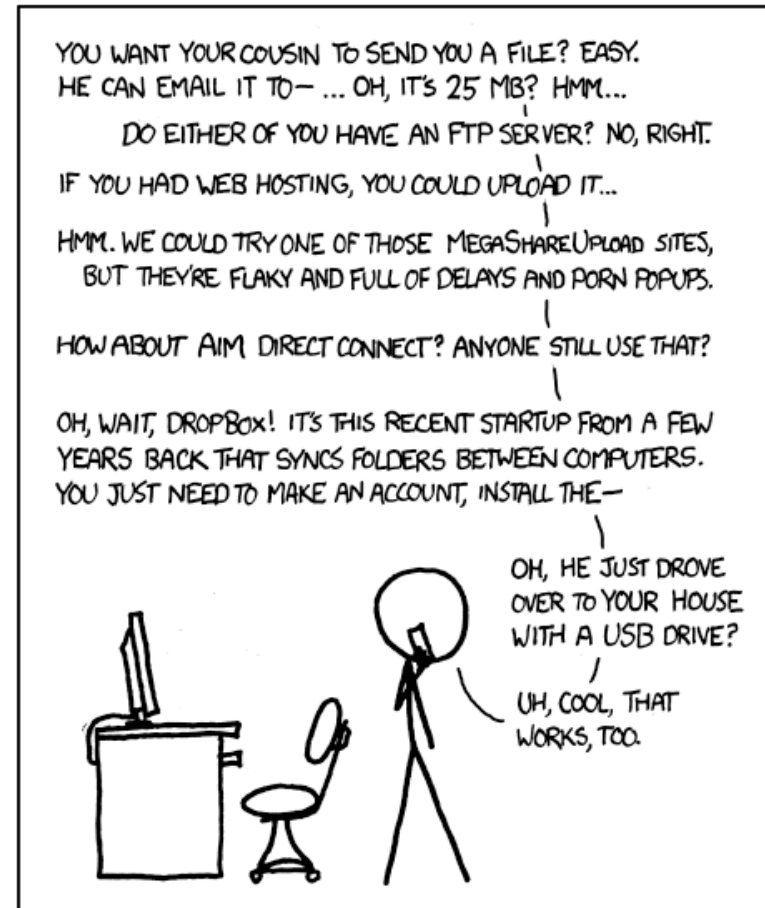


Local And Wide Area Data Flows



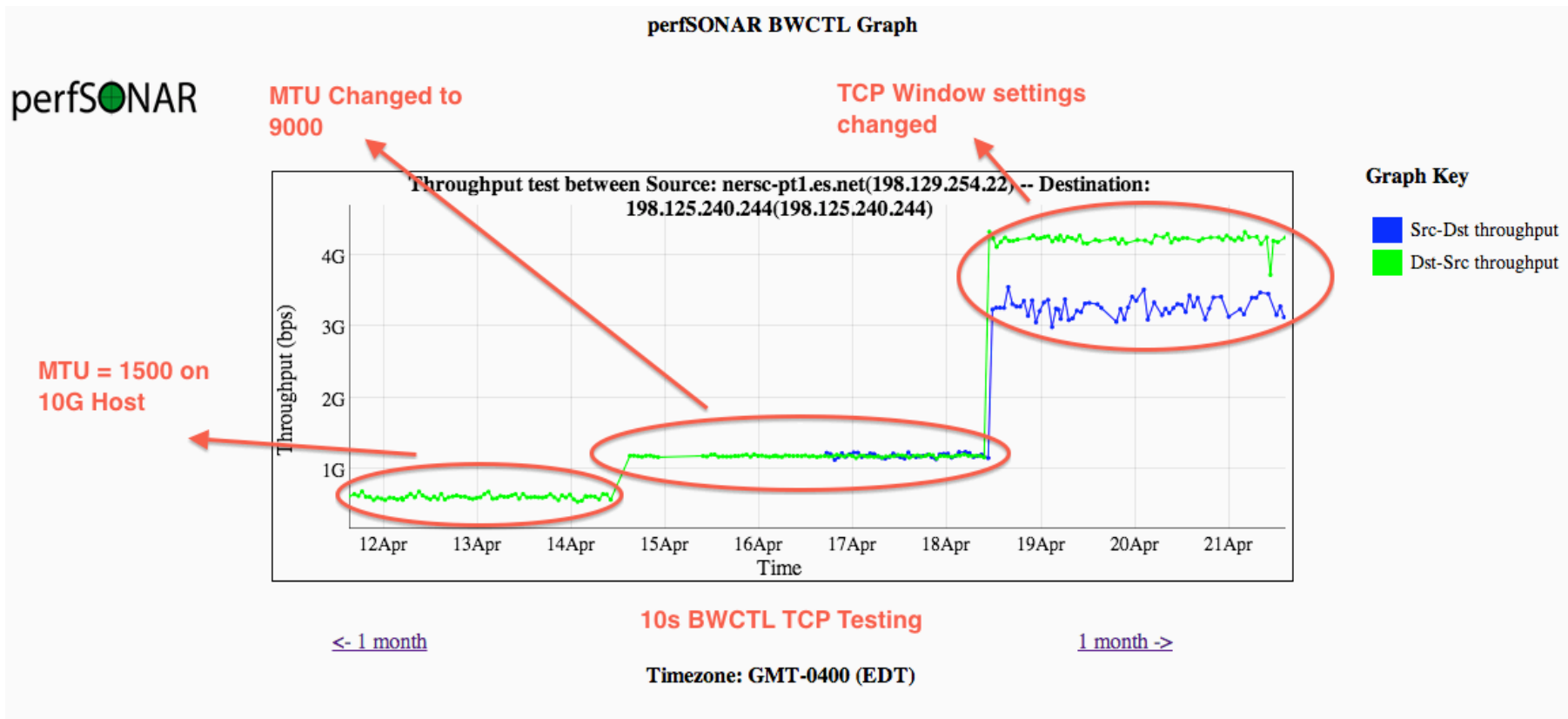
Dedicated Systems – Data Transfer Node

- Set up *specifically* for high-performance data movement
 - System internals (BIOS, firmware, interrupts, etc.)
 - Network stack
 - Storage (global filesystem, Fibrechannel, local RAID, etc.)
 - High performance tools
 - No extraneous software
- Limitation of scope and function is powerful
 - No conflicts with configuration for other tasks
 - Small application set makes cybersecurity easier



I LIKE HOW WE'VE HAD THE INTERNET FOR DECADES, YET "SENDING FILES" IS SOMETHING EARLY ADOPTERS ARE STILL FIGURING OUT HOW TO DO.

Example of perfSONAR monitoring



Improving things, when you don't know what you are doing, is a random walk. Sharing and educating the local community is important



Emerging global consensus around Science DMZ architecture.



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

NSF has invested >>\$60M to accelerate adoption.

Australian, Canadian universities following suit.

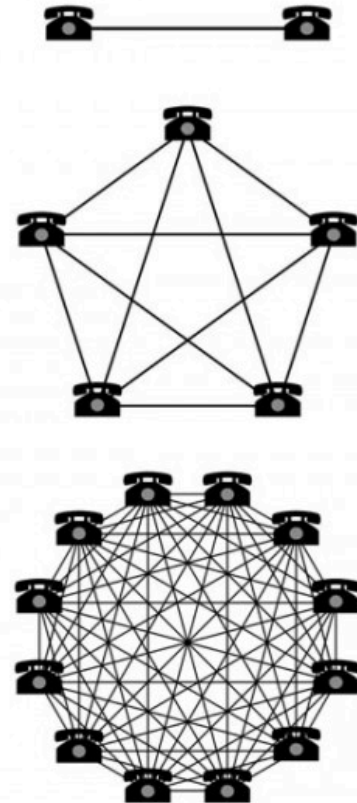


<http://fasterdata.es.net/science-dmz/>

Design Pattern #3: Build a well-tuned end-to-end science infrastructure

Metcalf's Law

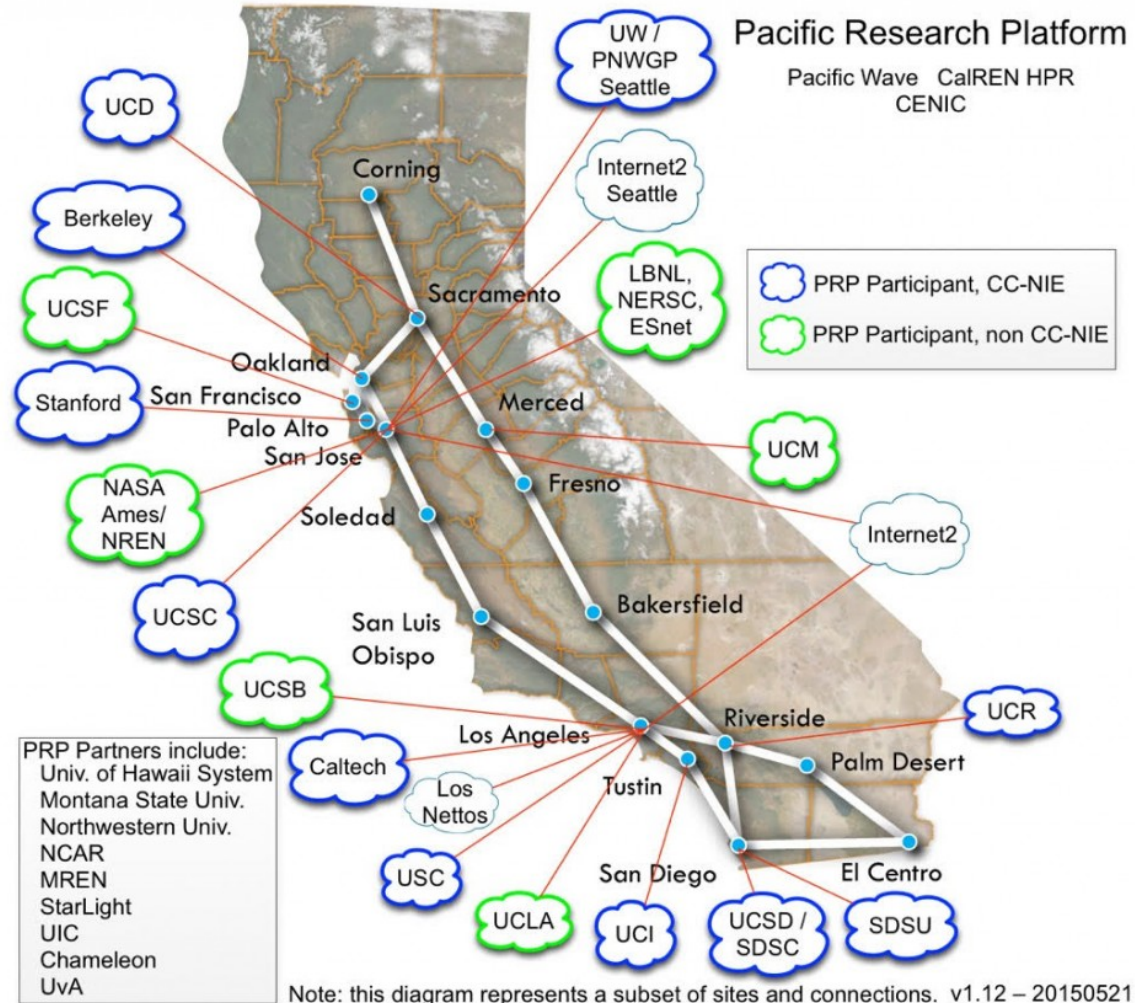
The value of a network is proportional to the square of the number of connected users. As the physical cost of the network grows linearly its value grows exponentially.



Science DMZ as a regional and national 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**
- Many international partners





Psychologists Welcome
Analysis Casting Doubt on
Their Work



NASA's New Horizons
Spacecraft Has Next
Mission After Pluto



Regenerative Medicine
Researcher Cleared of
Scientific Misconduct
Charges

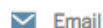


A CONVERSATION WITH
Eric Betzig's Life Over the
Microscope

SCIENCE

Research Scientists to Use Network Much Faster Than Internet

By JOHN MARKOFF JULY 31, 2015



Email



Share



Tweet



Save



More

SAN FRANCISCO — A series of ultra-high-speed fiber-optic cables will weave a cluster of West Coast university laboratories and supercomputer centers into a network called the Pacific Research Platform as part of a five-year \$5 million dollar grant from the [National Science Foundation](#).

The network is meant to keep pace with the vast acceleration of data collection in fields such as physics, astronomy and genetics. It will not be directly connected to the Internet, but will make it possible to move data at speeds of 10 gigabits to 100 gigabits per second among 10 [University of California](#) campuses and 10 other universities and research institutions in several states, tens or hundreds of times faster than is typical now.

The challenge in moving large amounts of scientific data is that the open Internet is designed for transferring small amounts of data, like web pages, said Thomas A. DeFanti, a specialist in scientific visualization at the California Institute for Telecommunications and Information Technology, or Calit2, at the University of California, San Diego. While a conventional network connection might be rated at 10 gigabits per second, in practice scientists trying to transfer large amounts of data often find that the real rate is only a fraction of that capacity.

The new network will also serve as a model for future computer networks in the same way the original NSFnet, created in 1985 to link research institutions, eventually became part of the backbone for the Internet, said Larry Smarr, an astrophysicist who is director of Calit2 and the principal investigator for the new project.

NSFnet connected five supercomputer centers with 56-kilobit-per-second

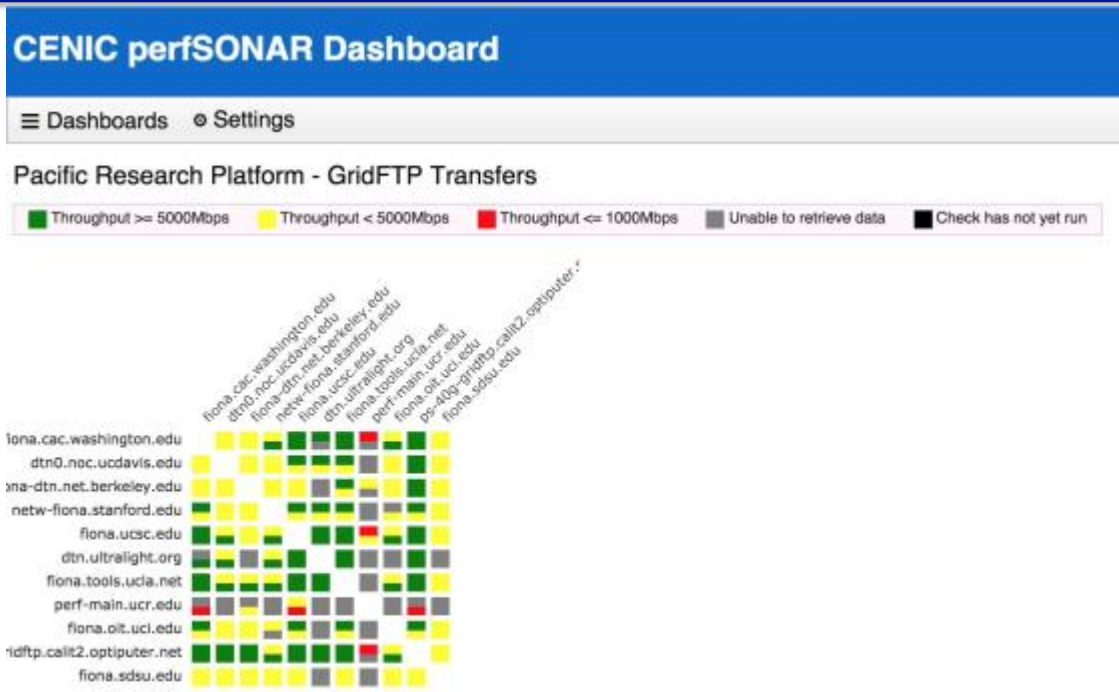
Specialized DTNs: FIONA (Flash I/O Node Appliance)

- Combination of Desktop and Server Building Blocks
 - US\$5K - US\$7K
 - Desktop Flash up to 16TB
 - RAID Drives up to 48TB
 - 10GbE/40GbE Adapter
 - Tested speed 40Gbs
- Developed Under UCSD CC-NIE Prism Award by UCSD's
 - Phil Papadopoulos
 - Tom DeFanti
 - Joe Keefe



PRPv0: Transfer Results from March 2015

- DTNs loaded with Globus Connect Server suite to obtain GridFTP tools.
- cron-scheduled transfers using globus-url-copy.
- ESnet-contributed script parses GridFTP transfer log and loads results in an esmond measurement archive.
- FDT – developed by Caltech in collaboration with Polytechnica Bucharest



As of 3/9/15, the Pacific Research Platform (PRPv0) as a facility, logs rather good performance:

From	To	Measured Bandwidth	Data Transfer Utility
San Diego State Univ.	UC Los Angeles	5Gb/s out of 10	GridFTP
UC Riverside	UC Los Angeles	9Gb/s out of 10	GridFTP
UC Berkeley	UC San Diego	9.6Gb/s out of 10	GridFTP
UC Davis	UC San Diego	9.6Gb/s out of 10	GridFTP
UC Irvine	UC Los Angeles	9.6Gb/s out of 10	GridFTP
UC Santa Cruz	UC San Diego	9.6Gb/s out of 10	FDT
Stanford	UC San Diego	12Gb/s out of 40	FDT
Univ. of Washington	UC San Diego	12Gb/s out of 40	FDT
UC Los Angeles	UC San Diego	36Gb/s out of 40	FDT
Caltech	UC San Diego	36Gb/s out of 40	FDT

Table I.2.1: Bandwidth of flash disk-to-flash disk file transfers shown between several sites for the existing experimental facility “PRPv0.”

PRP Timeline

- **PRPv1**

- A Layer 3 System
- Completed In 2 Years
- Tested, Measured, Optimized, With Multi-domain Science Data
- Bring Many Of Our Science Teams Up
- Each Community Thus Will Have Its Own Certificate-Based Access To its Specific Federated Data Infrastructure.

- **PRPv2**

- Advanced Ipv6-Only Version with Robust Security Features
 - e.g. Trusted Platform Module Hardware and SDN/SDX Software
- Support Rates up to 100Gb/s in Bursts And Streams
- Develop Means to Operate a Shared Federation of Caches

Talk

ESnet and
NRENs Intro



Established
Design
Patterns



Emerging
Design
Patterns



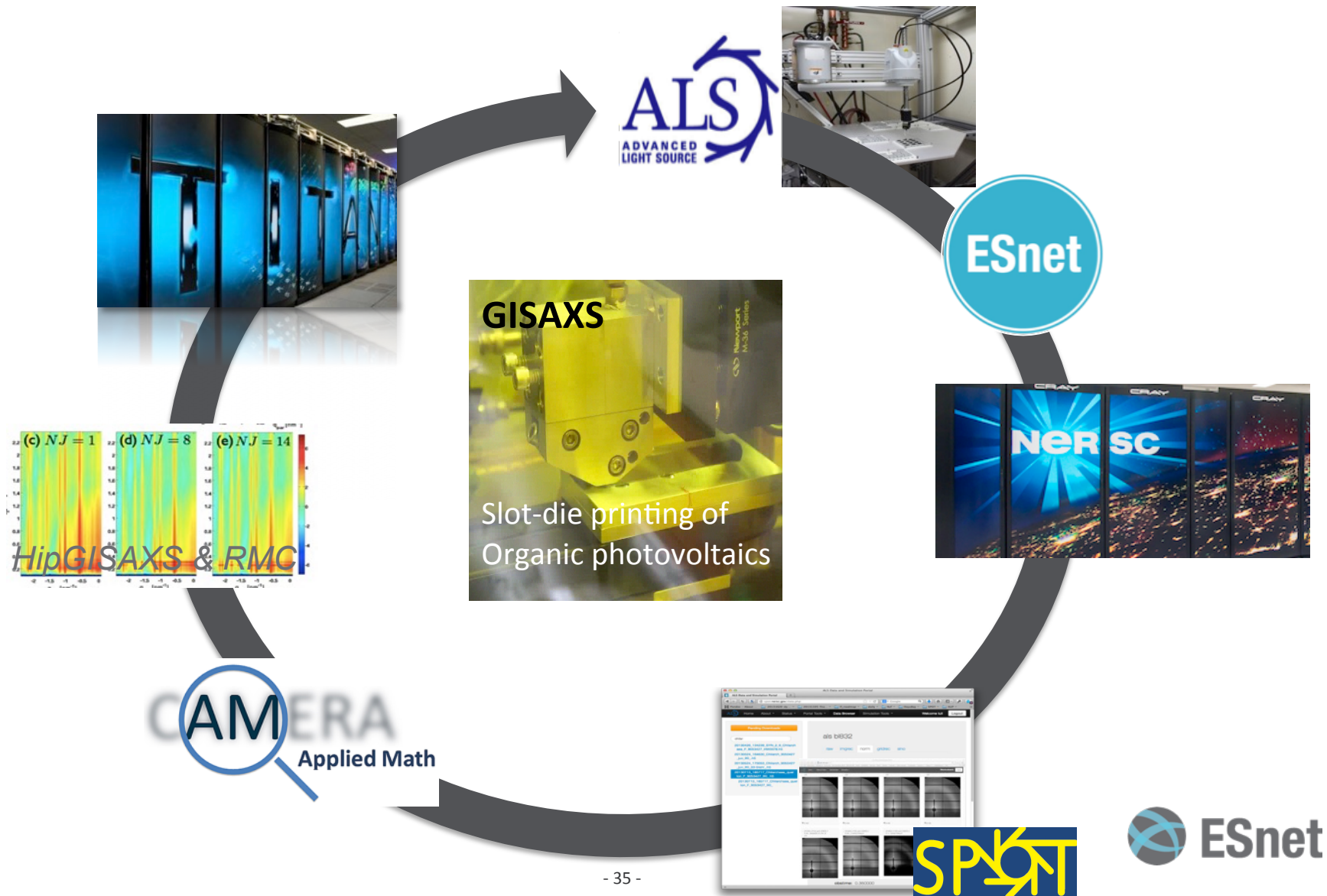
Discovering the next design pattern

- Discovering a new design pattern is not hard, key ingredients
 - **Impatience:** Why is it so hard to get things done?
 - **Experimentation:** Don't be afraid to fail
 - **Observation:** Why did you fail?
 - **Persistence:** Repeat till the model is right
- What does that mean for ESnet?
 - Network Testbeds
 - Collaborate with Network and Systems researchers
 - Listen to our end-users, the scientists (not just the campus IT folks)
- Invest in bridging the gap between good research, prototype and production
 - Write software!
 - We are not staffed to do enough, collaboration and sharing

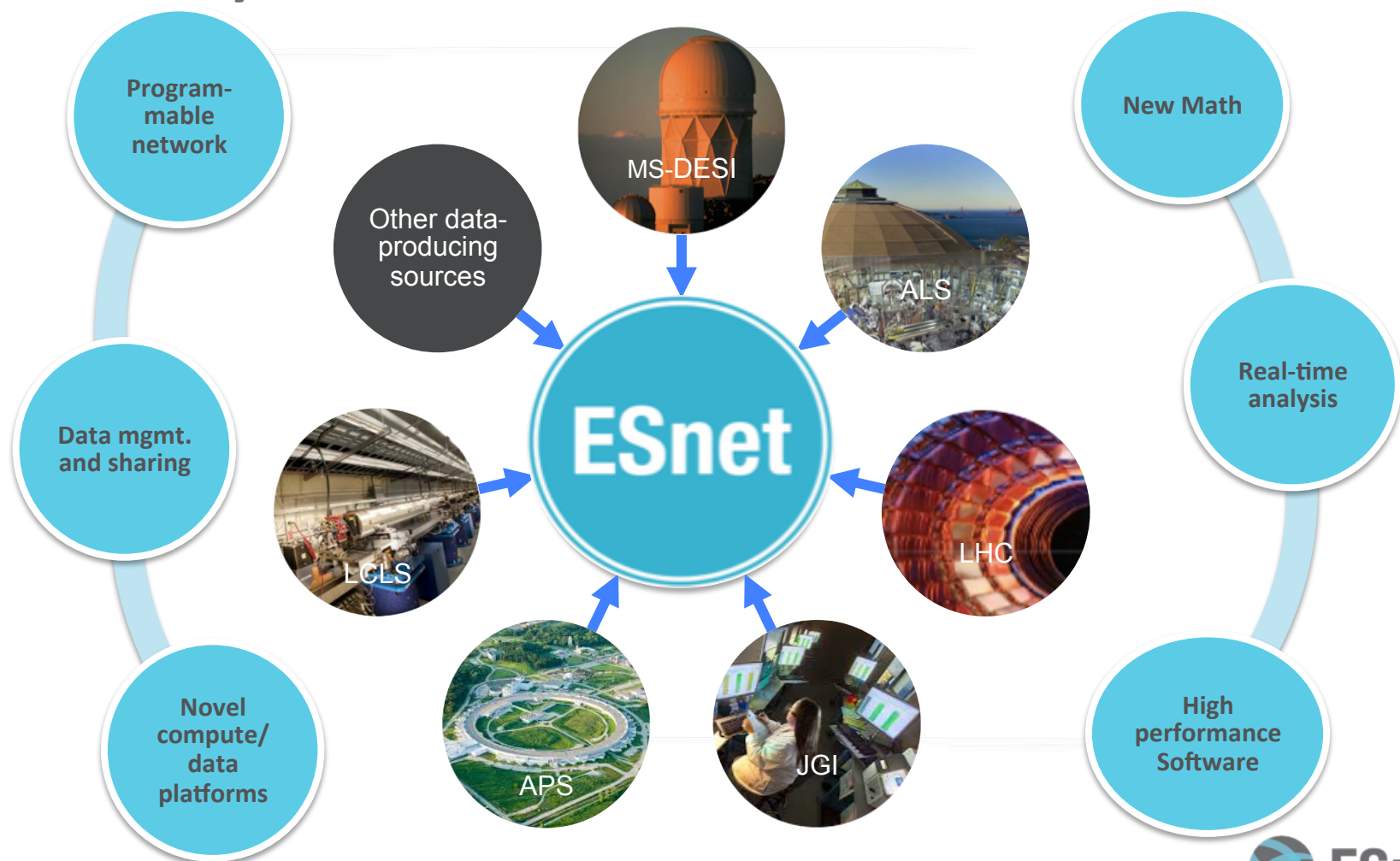


Emerging Design Pattern #4: Rising above the network – the *Superfacility*

Computing, experiments, networking and expertise in a “Superfacility” for Science



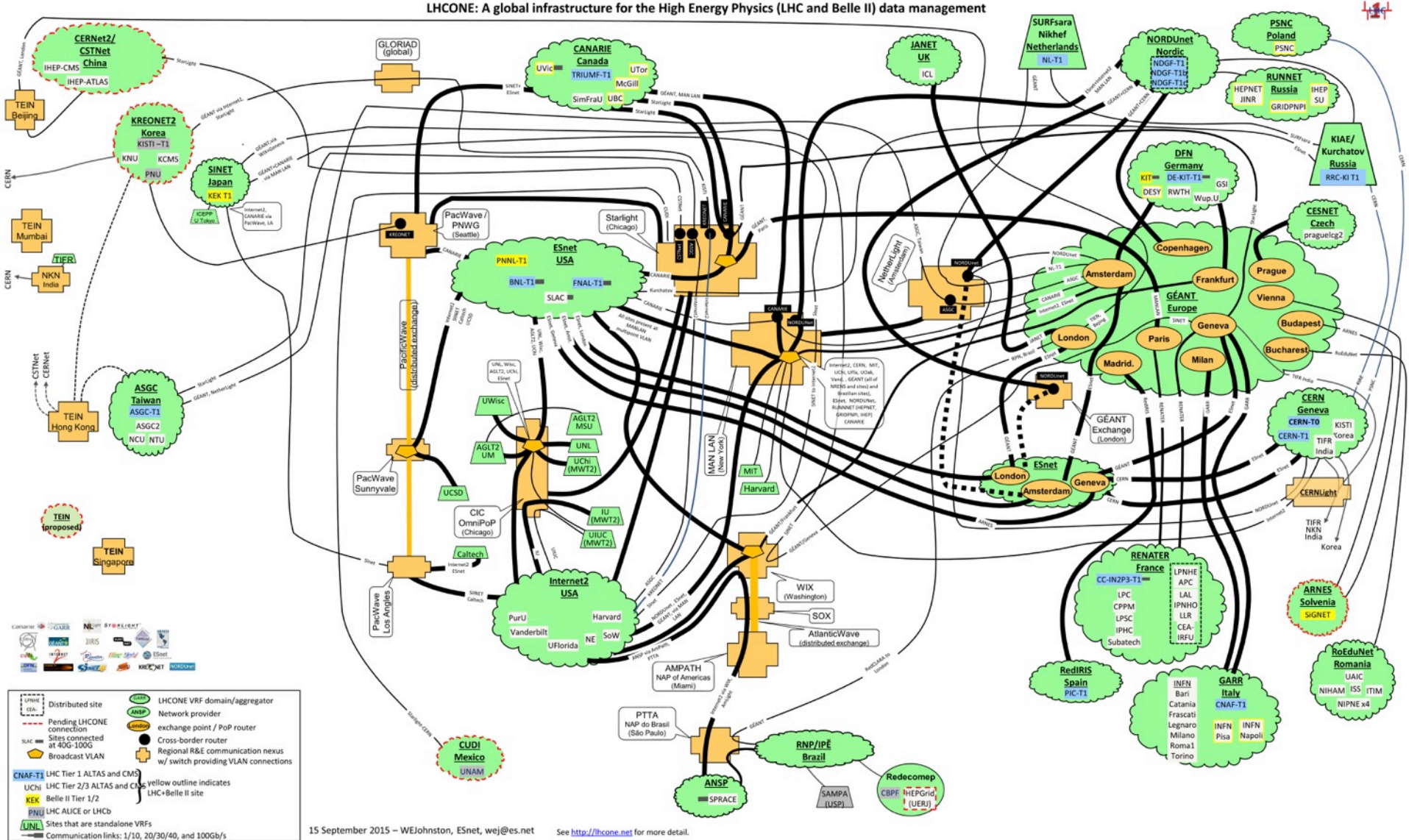
Superfacility Vision: A network of connected facilities, software and expertise to enable new modes of discovery



Emerging Design Pattern #5: Network Operating System

LHCONE Global Virtual Overlay

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



An architecture we foresee: ESnet as *platform* for concurrent, domain-specific network apps.

Requires 'network operating system' for science.

- early-stage project (LBNL LDRD)
- multiple challenges in creating flexible, stable app execution environment

We envision apps that will express high-level *intentions*:

- create and manage virtual networks
- enable programmatic resource allocation
- optimize link utilization

Future apps could support:

- NDN for climate; data management for CMS, ATLAS, Belle-II; security overlay for KBase; replication for ESGF; detector / HPC coupling for light sources
- workflows we haven't imagined

Password authentication

Password:

Welcome to NetShell

```
admin@NetShell> cd /lib/layer2/demo  
changed to: /lib/layer2/demo
```

```
admin@NetShell> vpn create vpn1  
VPN vpn1 is created successfully.
```

```
admin@NetShell> vpn vpn1 addpop amst  
Pop amst is added into VPN vpn1 successfully.
```

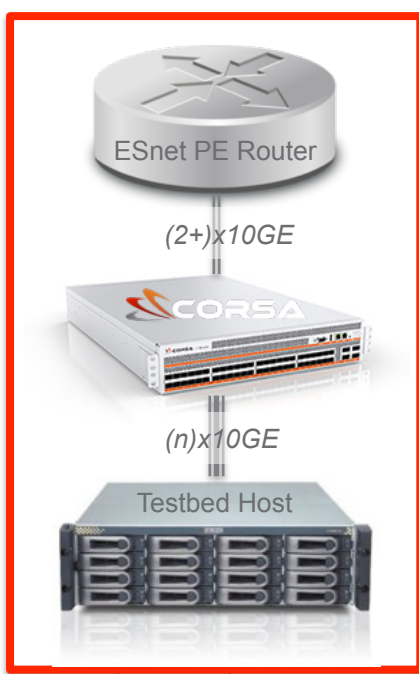
```
admin@NetShell> vpn vpn1 addpop cern  
Pop cern is added into VPN vpn1 successfully.
```



```
admin@NetShell> vpn vpn1 addsite amst  
The site amst is added into VPN vpn1 successfully
```

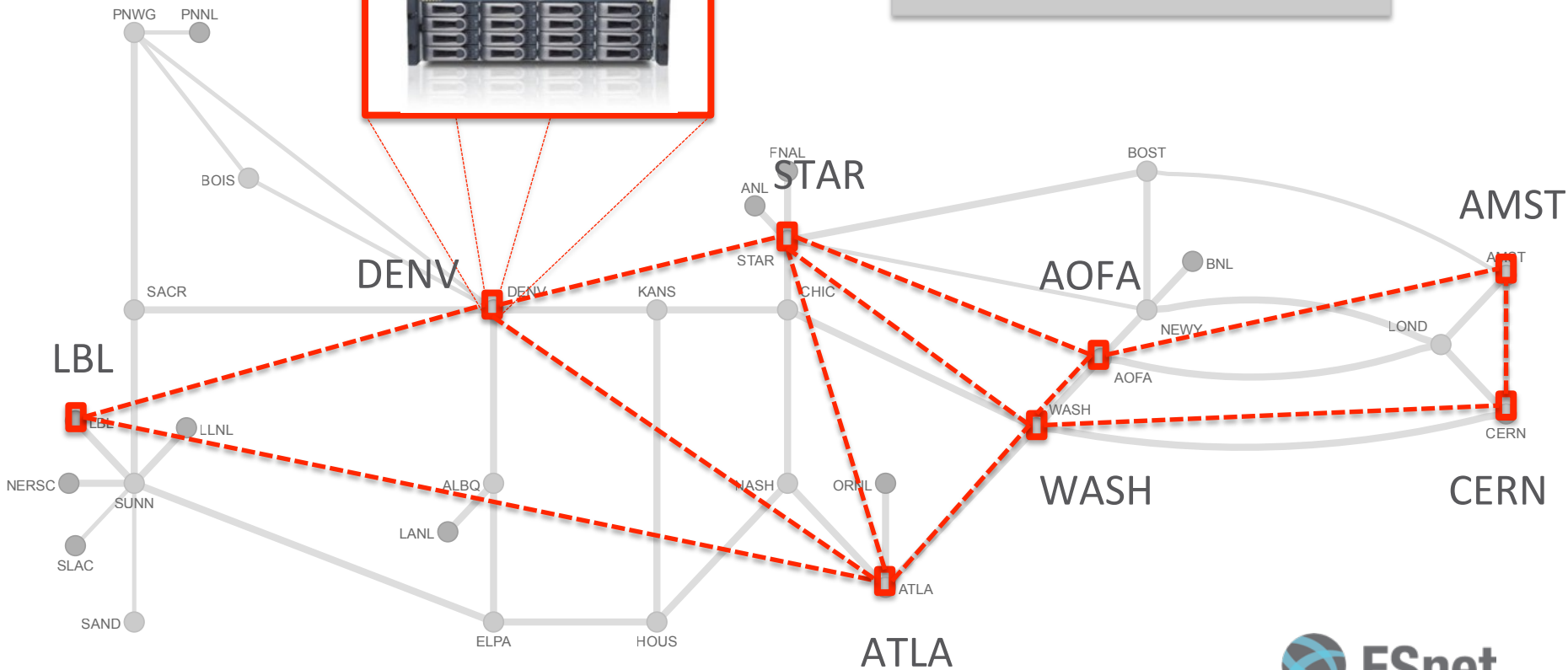
```
admin@NetShell> □
```


ESnet SDN Testbed

Testing SDN Concepts at Scale

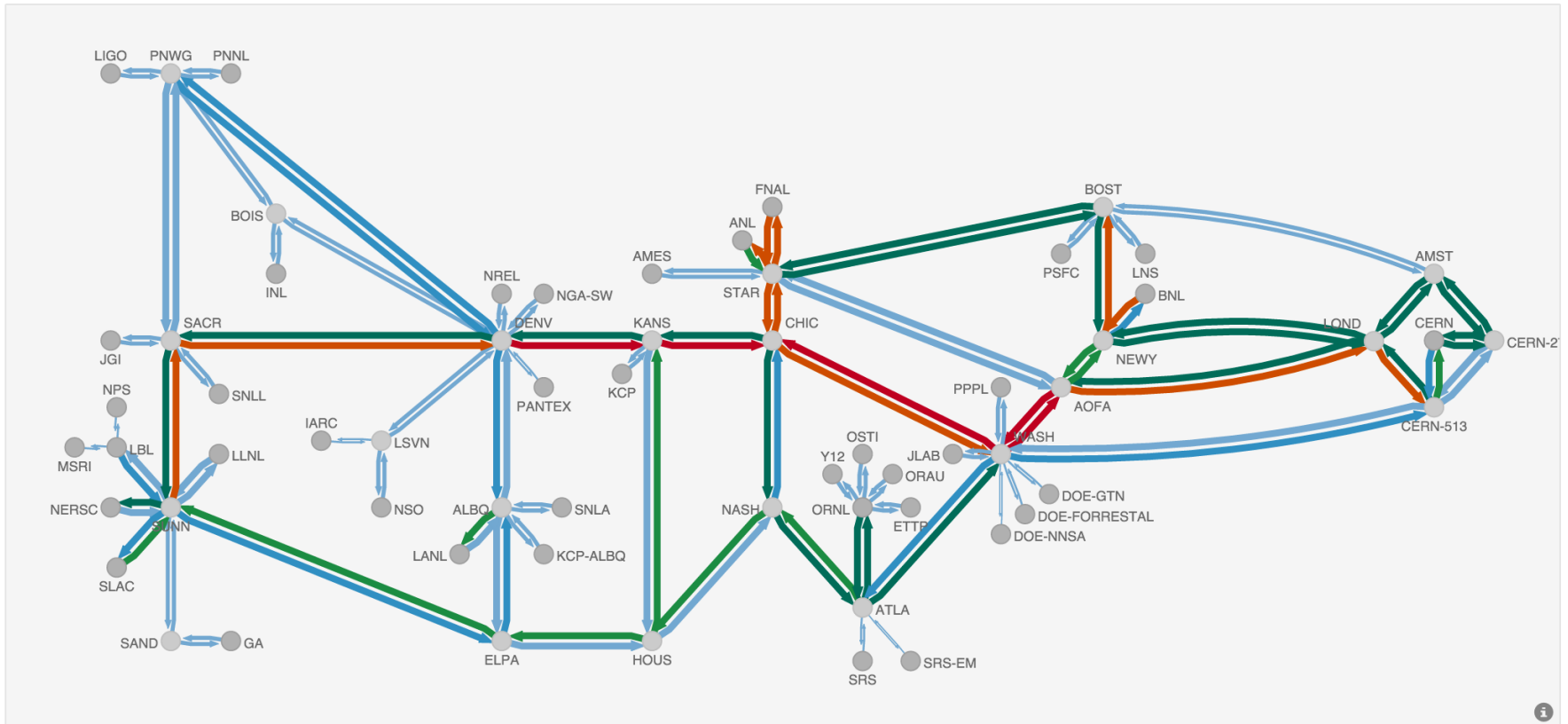


-  Deployed SDN Testbed node locations
-  Deployed SDN Testbed connectivity overlay (using OSCARS circuits)



Emerging Design Pattern #6: Data Driven Decisions

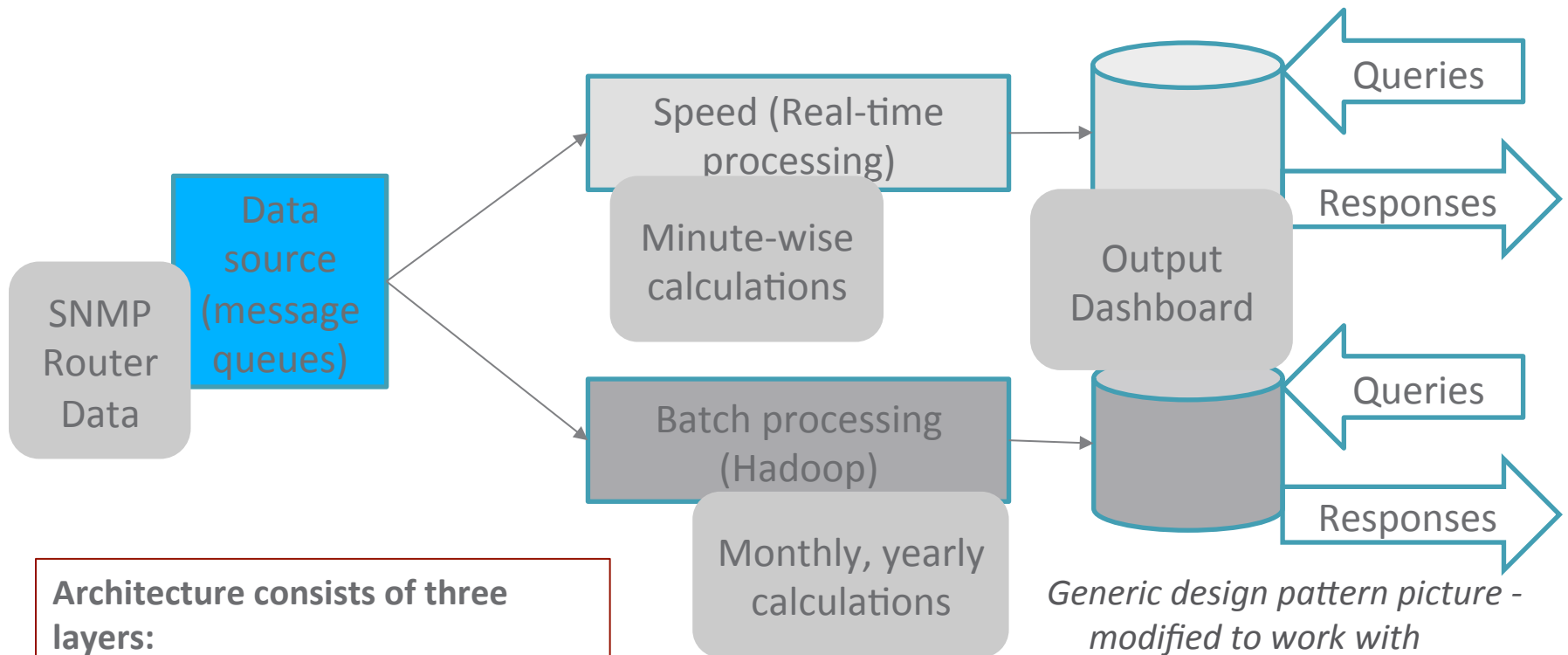
Visualization of data is good, but not enough



Network Analytics

- Data being generated by the network every few seconds but not analysed or available for real-time analysis
 - The ability to ask questions of historical network data, and get answers
 - The answers updated with new data in near real-time
 - SNMP data, Flow data, Topology data, etc..
- Smart Cities, IoT, Smart Grid – have common problems

Leverage cloud computing tools to put together a *network analytics* pipeline



Architecture consists of three layers:

- Batch processing
- Speed or real time
- Layer to respond to queries

Generic design pattern picture - modified to work with specific cloud computing technologies

In conclusion:

- Global science networks are not ISPs – rather, extensions of science discovery instruments.
- Design patterns, architectures, workflows and challenges from science are now crossing over to domains.
- Discovering new design patterns is not rocket-science
 - Requires ingredients of impatience, experimentation, observation and persistence
- Systems approach is needed to identify the emerging design patterns
 - Often compute, data and network experts don't care to look over the boundary and collaborate
- Involve your end users! It is not just about technology but about positively impacting the future

Thank you.

imonga@es.net

