ESnet

ENERGY SCIENCES NETWORK

Science Networks: How Data Gets There

Eli Dart, Network Engineer

ESnet Science Engagement

Lawrence Berkeley National Laboratory

ATPESC 2017

Chicago, IL

August 4, 2017

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

Outline

- Science Networks – structure and relationship to the rest of the Internet
- Data transfer at HPC facilities
- Data portals – past, present, and future

NCAR RDA Data Portal

- Let’s say I have a nice compute allocation at the ALCF – climate science
- Let’s say I need some data from NCAR for my project

- https://rda.ucar.edu/

- Data sets (there are many more, but these are two):
- https://rda.ucar.edu/datasets/ds199.1/ (1.5TB)
- https://rda.ucar.edu/datasets/ds313.0/ (430GB)

- Download to ALCF (could also do NCSA or NERSC or OLCF)

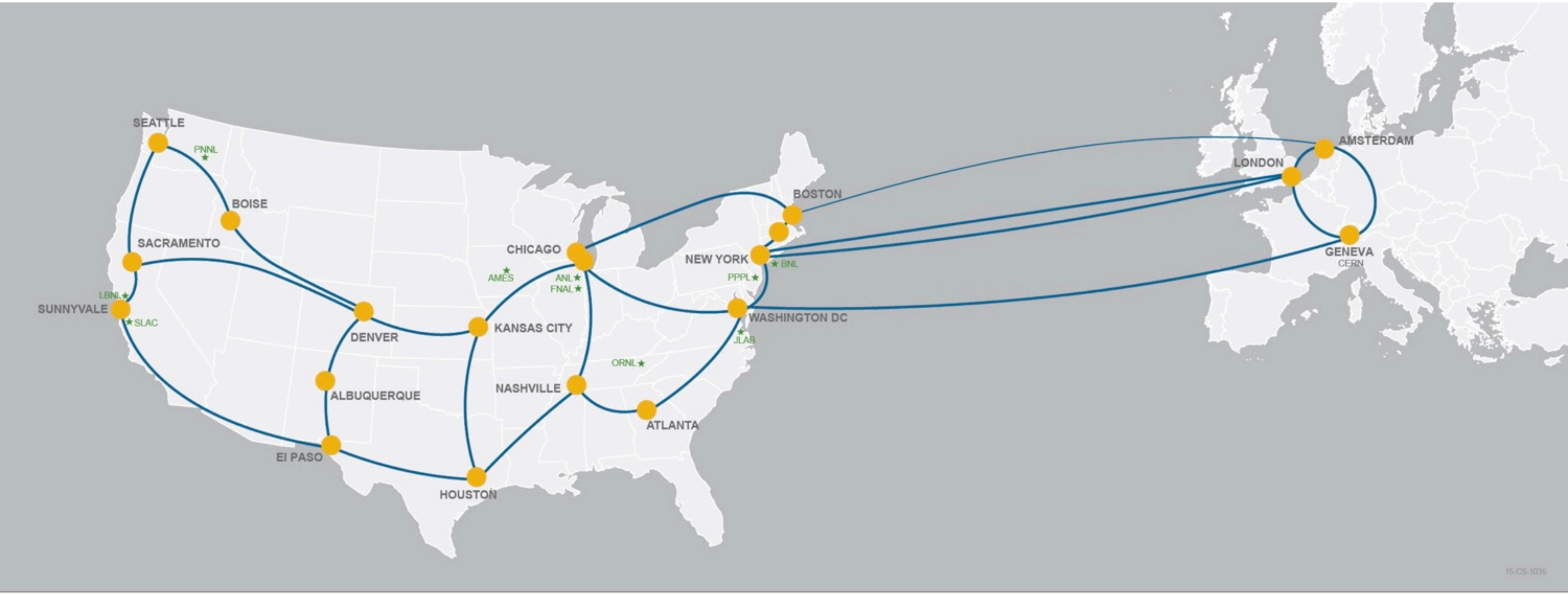
What Is A Science Network?

- Downloading data from a portal happens via the network
- What does "via the network" actually mean?
- What is "the network" anyway?

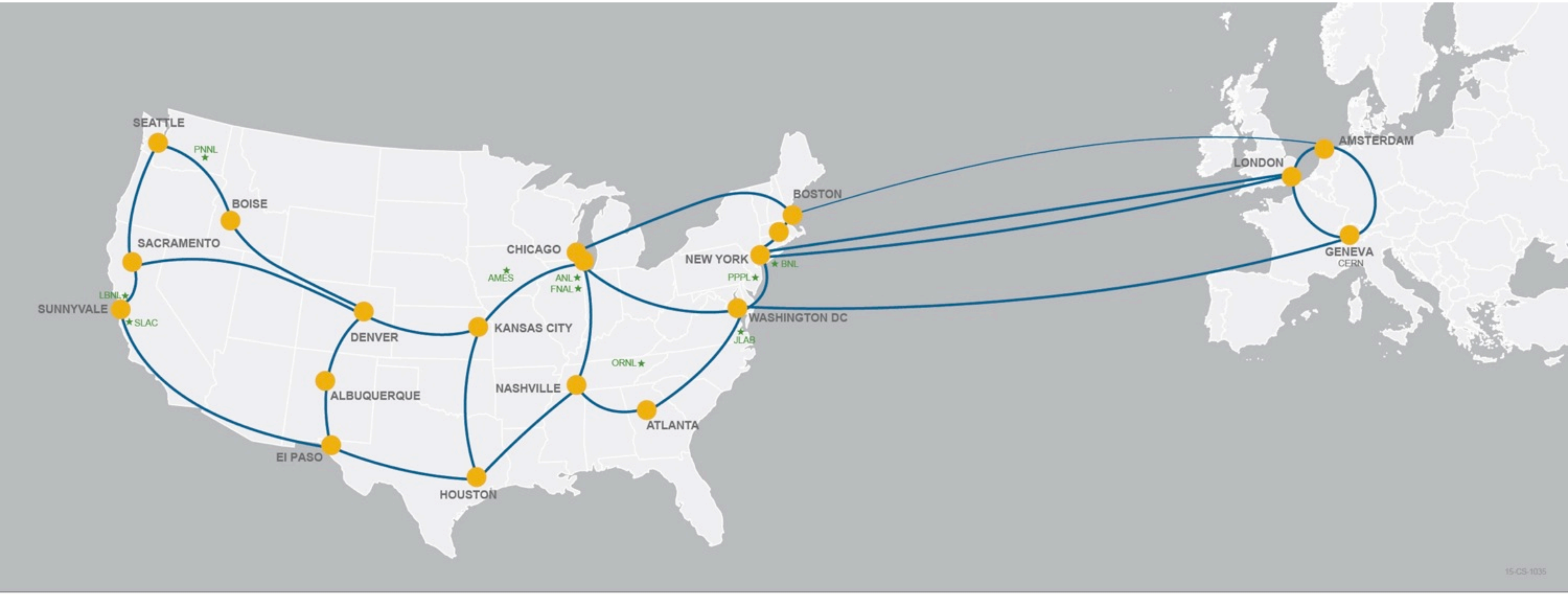
- Most of us are familiar with the notion of an ISP
  - Internet access at home (Netflix, etc.)
  - Data for phones (Facebook, maps, Google, etc.)
  - This is "the Internet" that most people see

- Science networks interconnect scientific sites
  - HPC facilities
  - Particle accelerators (LHC, light sources, …)
  - Data portals
- Science networks use the same protocols as the rest of the Internet
  - They are also connected to the rest of the Internet

This is not an ISP.



It’s a DOE user facility engineered and optimized for Big Data Science

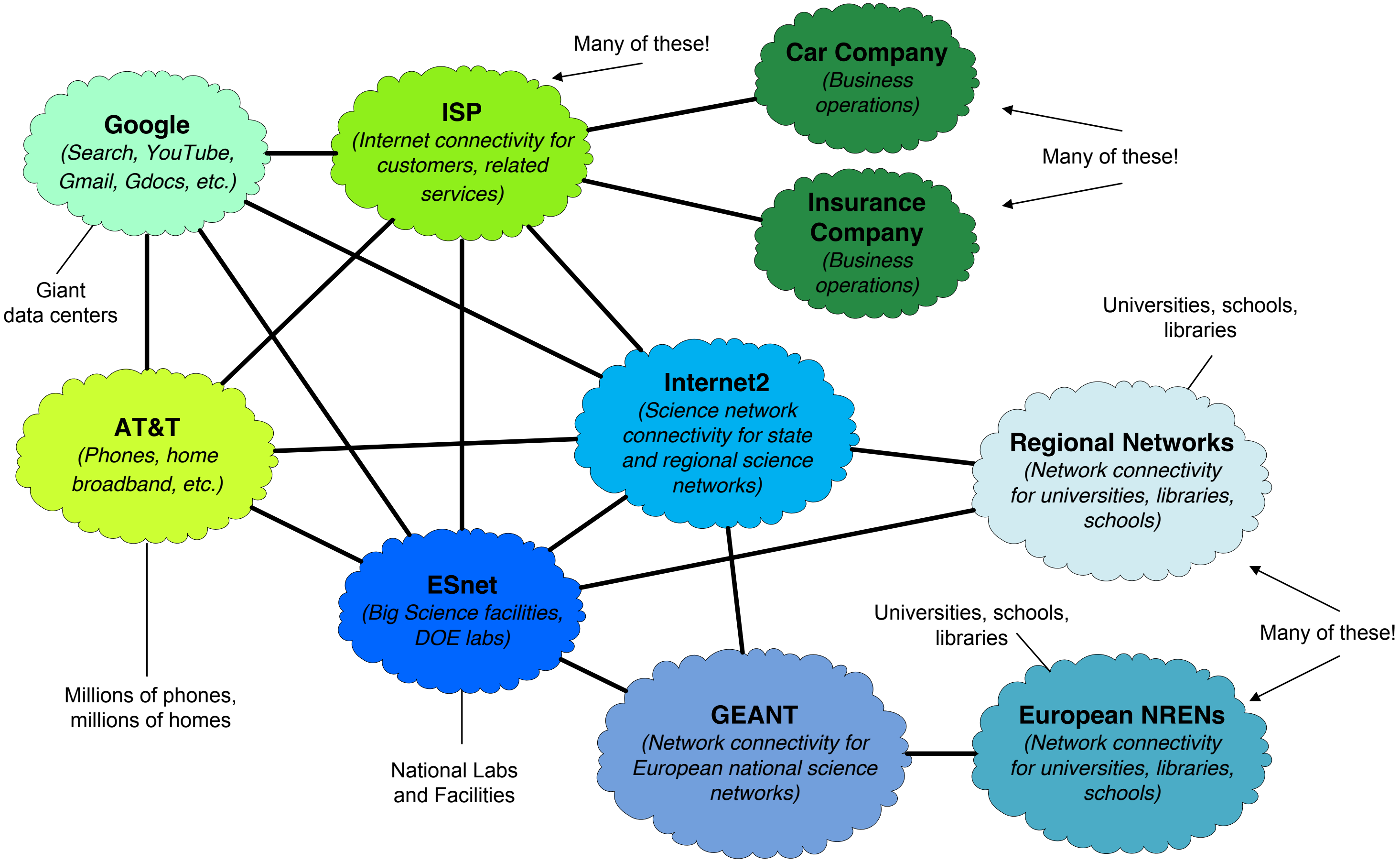


We do this by offering unique capabilities and optimizing the facility for data acquisition, data placement, data sharing, data mobility.

The Internet

- The Internet is composed of a large number of individual networks
  - Each is run by some entity for its own reasons
    - Google
    - US Department of Defense
    - Ford Motor Company
    - US Department of Energy
    - AT&T
  - Each network connects to others for its own reasons
- In general, networks are more valuable when connected to each other
  - But remember – this connectivity happens for selfish reasons
  - Not all networks are the same – each exists for its own reasons

Selected networks and their missions

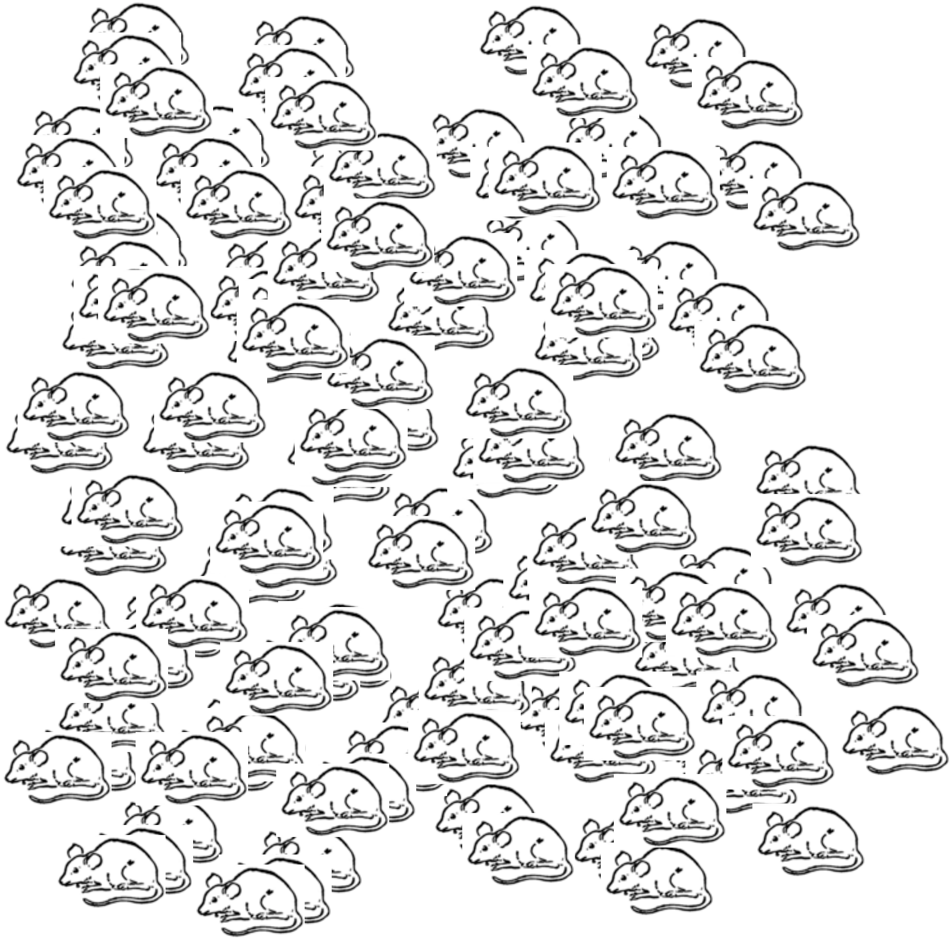

Many of these!
Car Company
(Business operations)
Google
(Search, YouTube, Gmail, Gdocs, etc.)
ISP
(Internet connectivity for customers, related services)
Many of these!
Insurance Company
(Business operations)
Giant data centers
Universities, schools, libraries
Internet2
(Science network connectivity for state and regional science networks)
AT&T
(Phones, home broadband, etc.)
Regional Networks
(Network connectivity for universities, libraries, schools)
ESnet
(Big Science facilities, DOE labs)
Universities, schools, libraries
Many of these!
Millions of phones, millions of homes
GEANT
(Network connectivity for European national science networks)
European NRENs
(Network connectivity for universities, libraries, schools)
National Labs and Facilities

ESnet

Notes about different networks

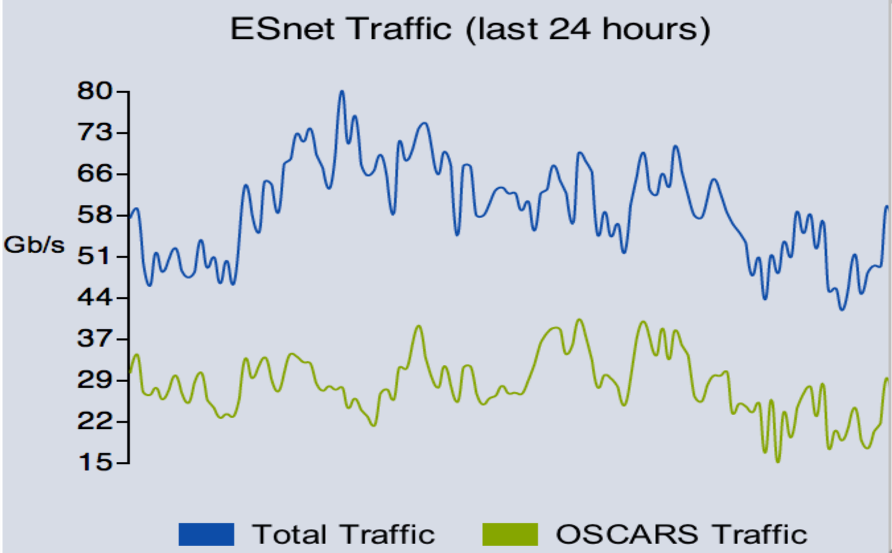
- The previous diagram is a drastic simplification
  - http://www.caida.org/research/topology/as_core_network/2015/
- Key points:
  - All networks exist for a specific reason
    - Some networks provide connectivity between networks
    - Some networks primarily serve their own users
    - Some networks provide services to users who access them via different networks (e.g. Google)
  - These lines are blurry, but it’s a useful way to think about it
- Network mission influences engineering, policy, reliability, etc.
  - Not all networks are built the same way
  - Not all networks can support all use models
  - Science networks have a different traffic profile than commercial networks

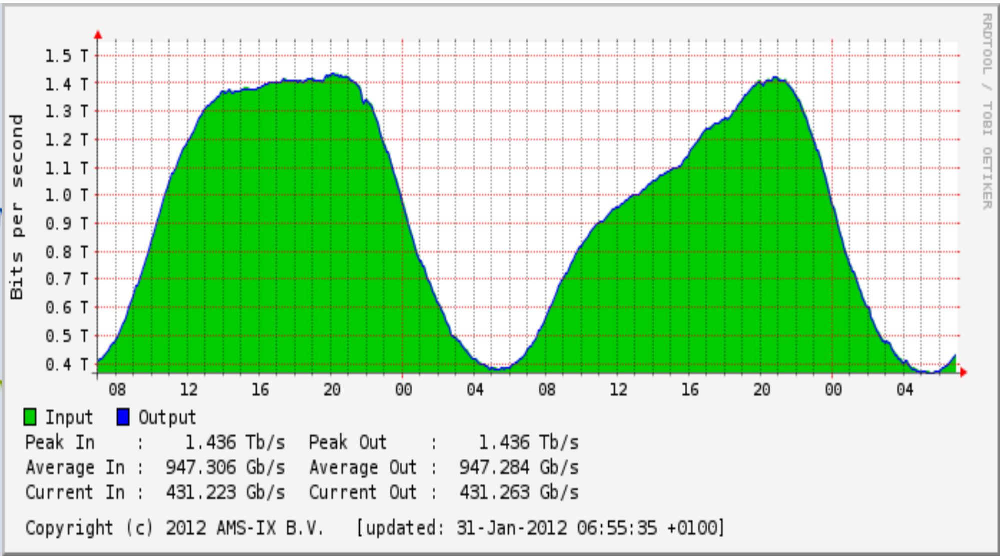
Elephant Data vs. Mice Data



ESnet

Elephant Data vs. Mice Data Behavior





```
Peak In     :    1.436 Tb/s  Peak Out     :    1.436 Tb/s
Average In  :  947.306 Gb/s  Average Out  :  947.284 Gb/s
Current In  :  431.223 Gb/s  Current Out  :  431.263 Gb/s

Copyright (c) 2012 AMS-IX B.V.   [updated: 31-Jan-2012 06:55:35 +0100]
```

Elephant Flows Place Great Demands on Networks

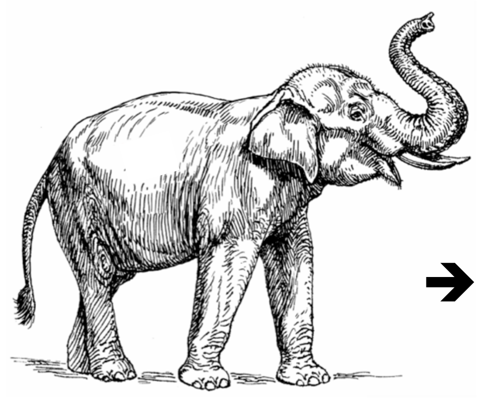
➔

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.

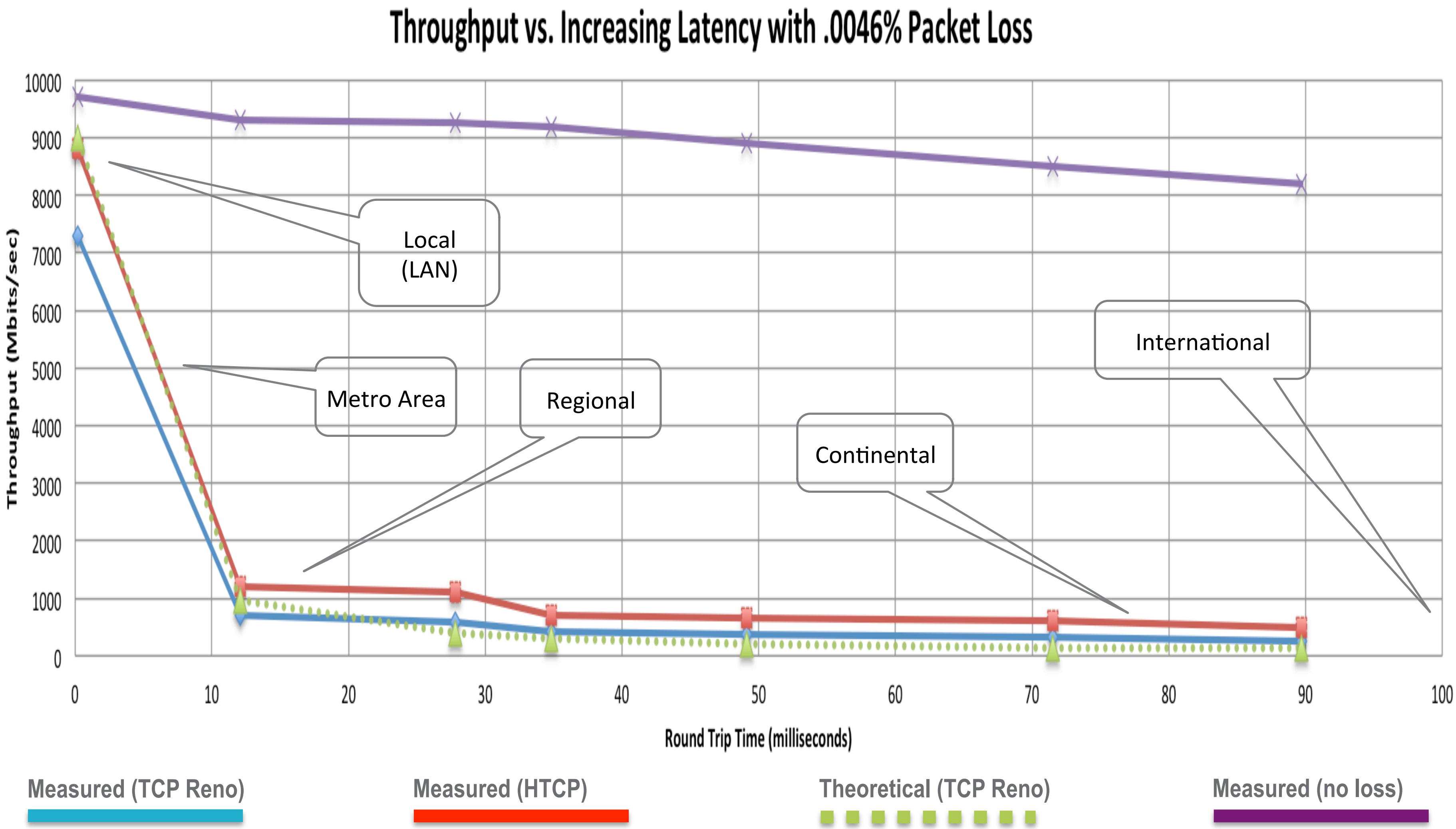
essentially fixed

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.

Elephant flows require essentially *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.

Emerging global consensus around **Science DMZ architecture.**

1. Friction-free network path
2. Dedicated data transfer nodes (DTNs)
3. Performance monitoring (perfSONAR)

- Over 120 universities in the US have deployed this ESnet architecture.
- NSF has invested >>$80M to accelerate adoption.
- Australian, Canadian, British, Brazilian universities following suit.
- **http://fasterdata.es.net/science-dmz/**

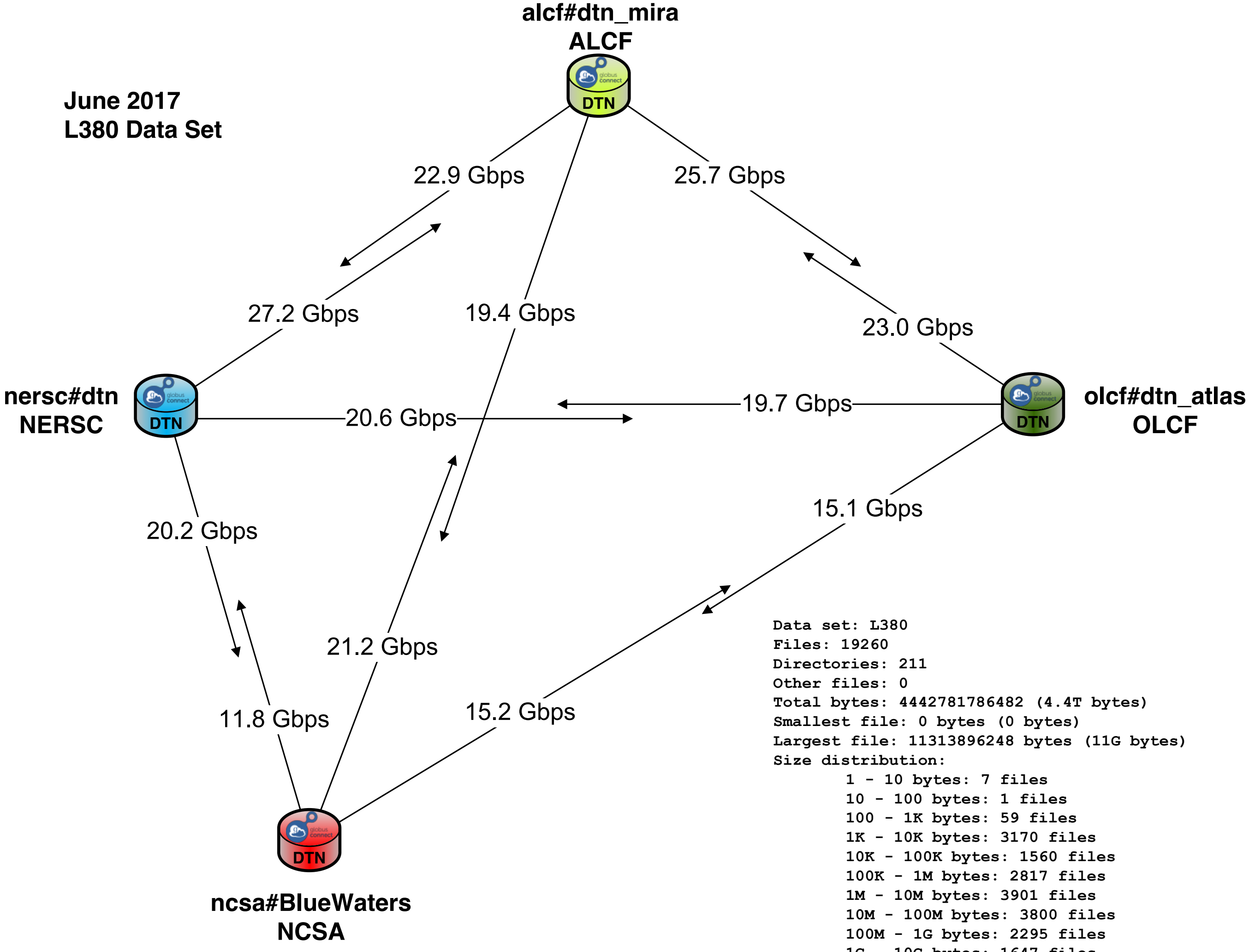
ESnet

The Petascale DTN Project

- Built on top of the Science DMZ model
- Effort to improve data transfer performance between the DOE ASCR HPC facilities at ANL, LBNL, and ORNL, and also NCSA.
  - Multiple current and future science projects need to transfer data between HPC facilities
  - Performance goal is 15 gigabits per second (equivalent to 1PB/week)
  - Realize performance goal for routine Globus transfers without special tuning
- Reference data set is 4.4TB of cosmology simulation data

DTN Cluster Performance – HPC Facilities

June 2017
L380 Data Set



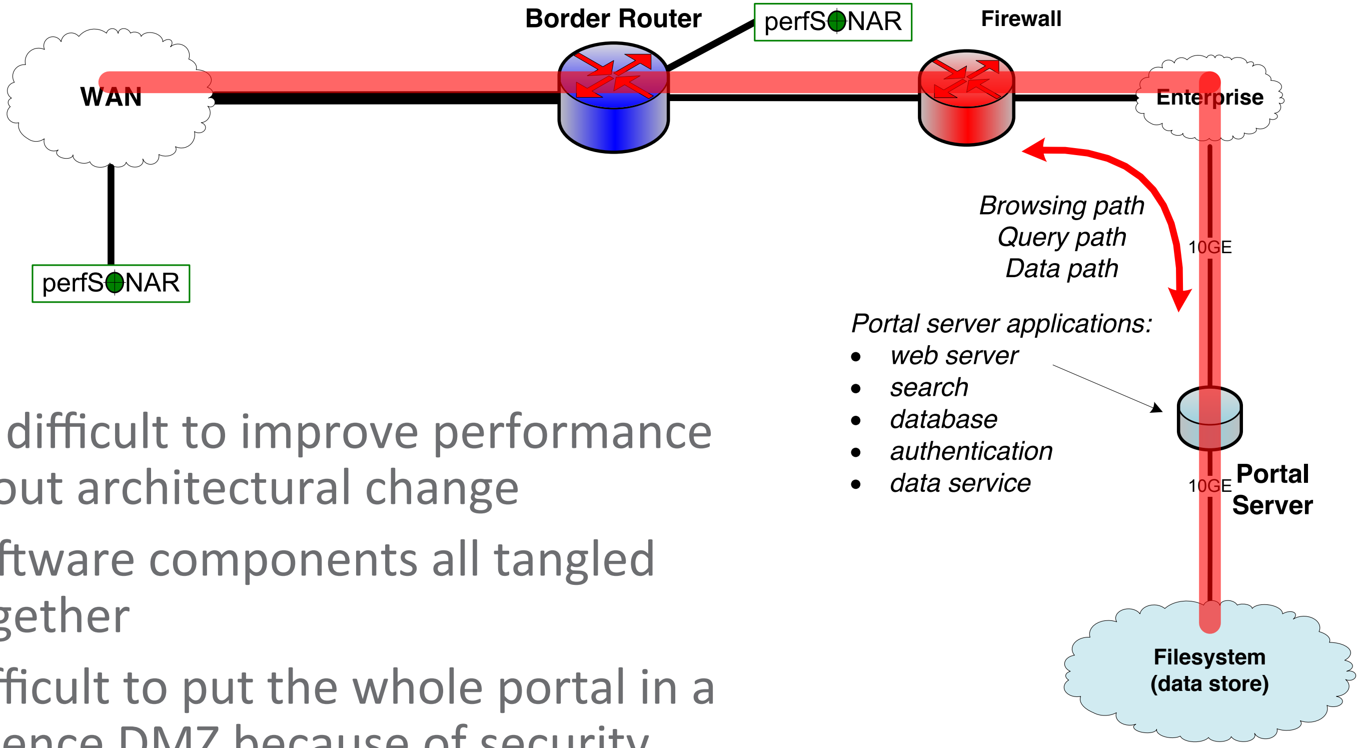
```
Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
        1 - 10 bytes: 7 files
        10 - 100 bytes: 1 files
        100 - 1K bytes: 59 files
        1K - 10K bytes: 3170 files
        10K - 100K bytes: 1560 files
        100K - 1M bytes: 2817 files
        1M - 10M bytes: 3901 files
        10M - 100M bytes: 3800 files
        100M - 1G bytes: 2295 files
        1G - 10G bytes: 1647 files
        10G - 100G bytes: 3 files
```

ESnet

Science Data Portals

- Large repositories of scientific data
  - Climate data
  - Sky surveys (astronomy, cosmology)
  - Many others
  - Data search, browsing, access
- Many scientific data portals were designed 15+ years ago
  - Single-web-server design
  - Data browse/search, data access, user awareness all in a single system
  - All the data goes through the portal server
    - In many cases by design
    - E.g. embargo before publication (enforce access control)

Legacy Portal Design



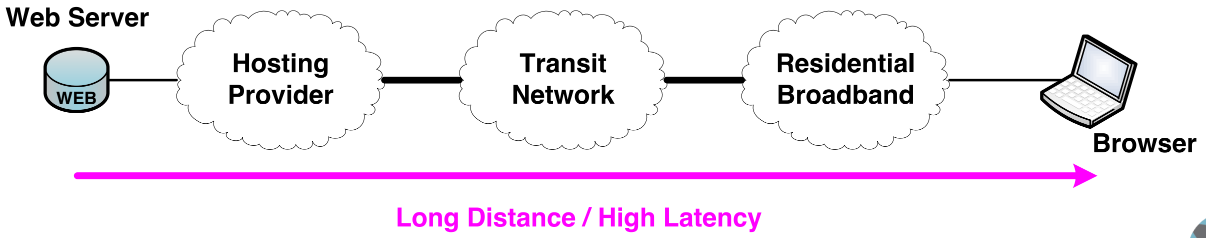
- Very difficult to improve performance without architectural change
  - Software components all tangled together
  - Difficult to put the whole portal in a Science DMZ because of security
  - Even if you could put it in a DMZ, many components aren’t scalable
- What does architectural change mean?

Example of Architectural Change – CDN

- Let’s look at what Content Delivery Networks did for web applications
- CDNs are a well-deployed design pattern (e.g. AirBnB, Olympic Games, etc.)
- What does a CDN do?
  - – Store static content in a separate location from dynamic content
    - Complexity isn’t in the static content – it’s in the application dynamics
    - Web applications are complex, full-featured, and slow
    - Data service for static content is simple − just move the file
  - – Separation of application and data service allows each to be optimized

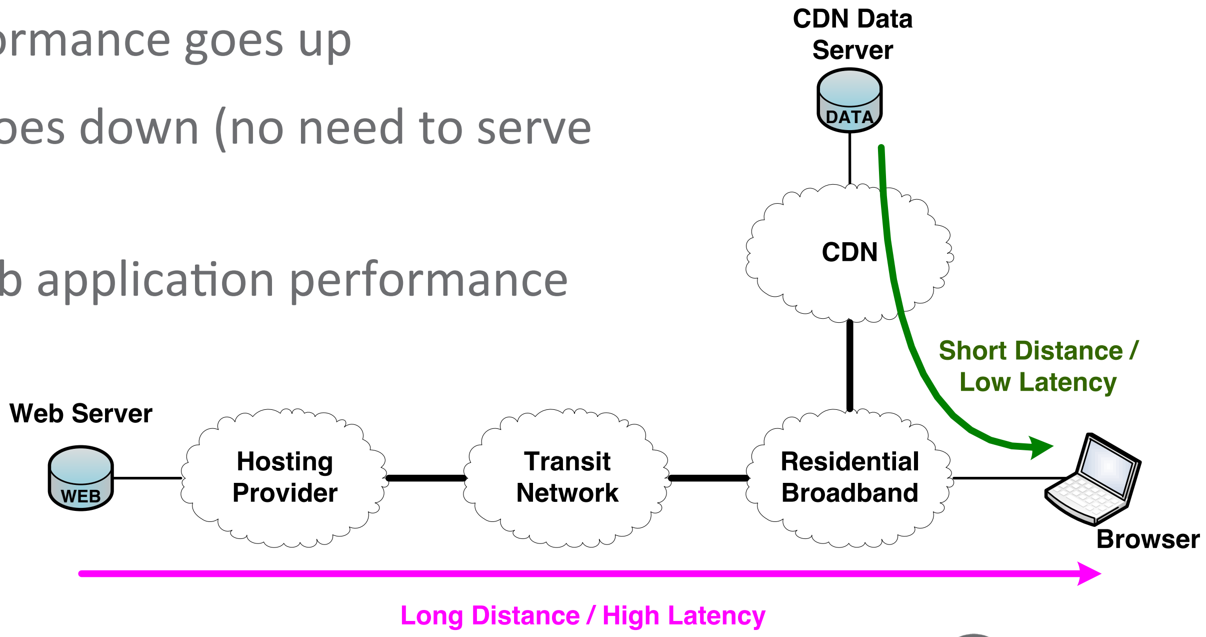
Classical Web Server Model

- Web browser fetches pages from web server
  - All content stored on the web server
  - Web applications run on the web server
  - Web server sends data to client browser over the network
- Perceived client performance changes with network conditions
  - Several problems in the general case
  - Latency increases time to page render
  - Packet loss + latency cause problems for large static objects



Solution: Place Large Static Objects Near Client

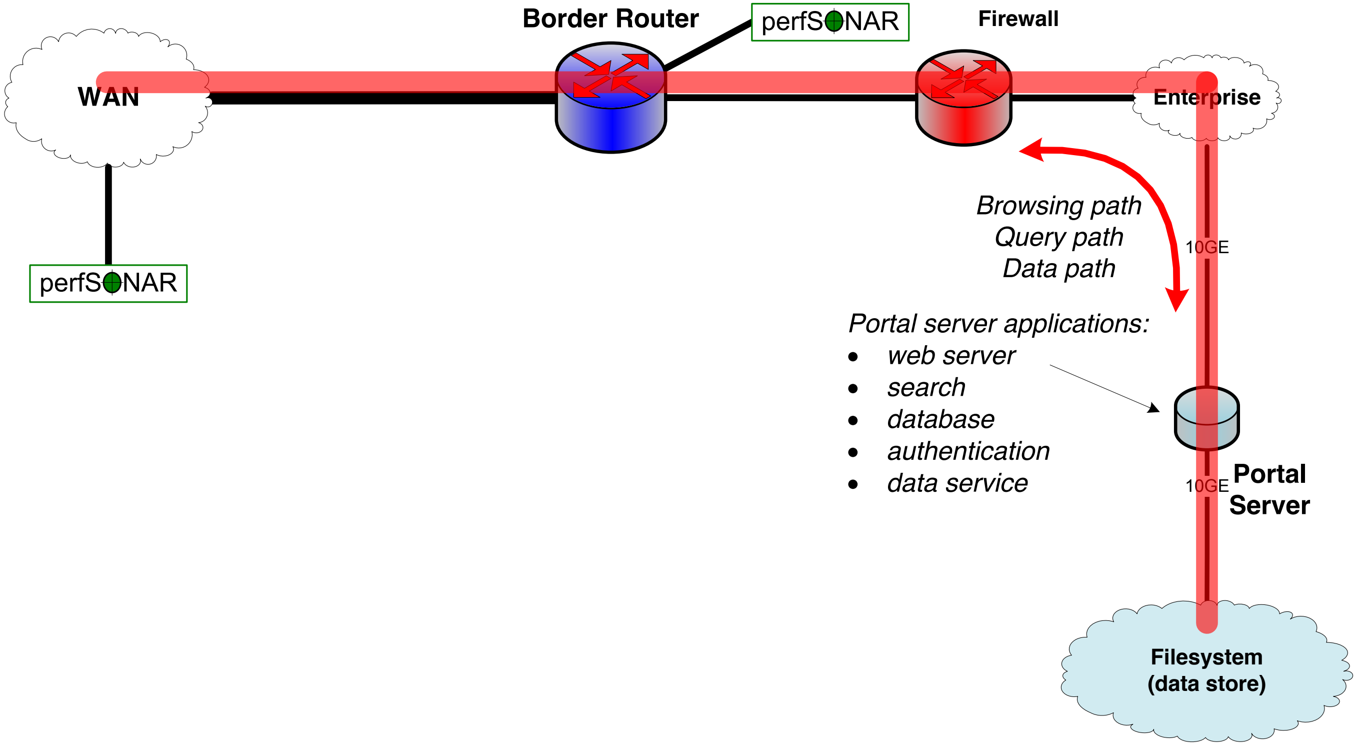
- CDN provides static content “close” to client
- Web server still manages complex behavior
- Latency goes down
  - Time to page render goes down
  - Static content performance goes up
- Load on web server goes down (no need to serve static content)
- Significant win for web application performance



Architectural Examination of Data Portals

- Common data portal functions (most portals have these)
  - Search/query/discovery
  - Data download method for data access
  - GUI for browsing by humans
  - API for machine access – ideally incorporates search/query + download
- Performance pain is primarily in the data handling piece
  - Rapid increase in data scale eclipsed legacy software stack capabilities
  - Portal servers often stuck in enterprise network
- Can we "disassemble" the portal and put the pieces back together better?
  - Use Science DMZ as a platform for the data piece
  - Avoid placing complex software in the Science DMZ

Legacy Portal Design

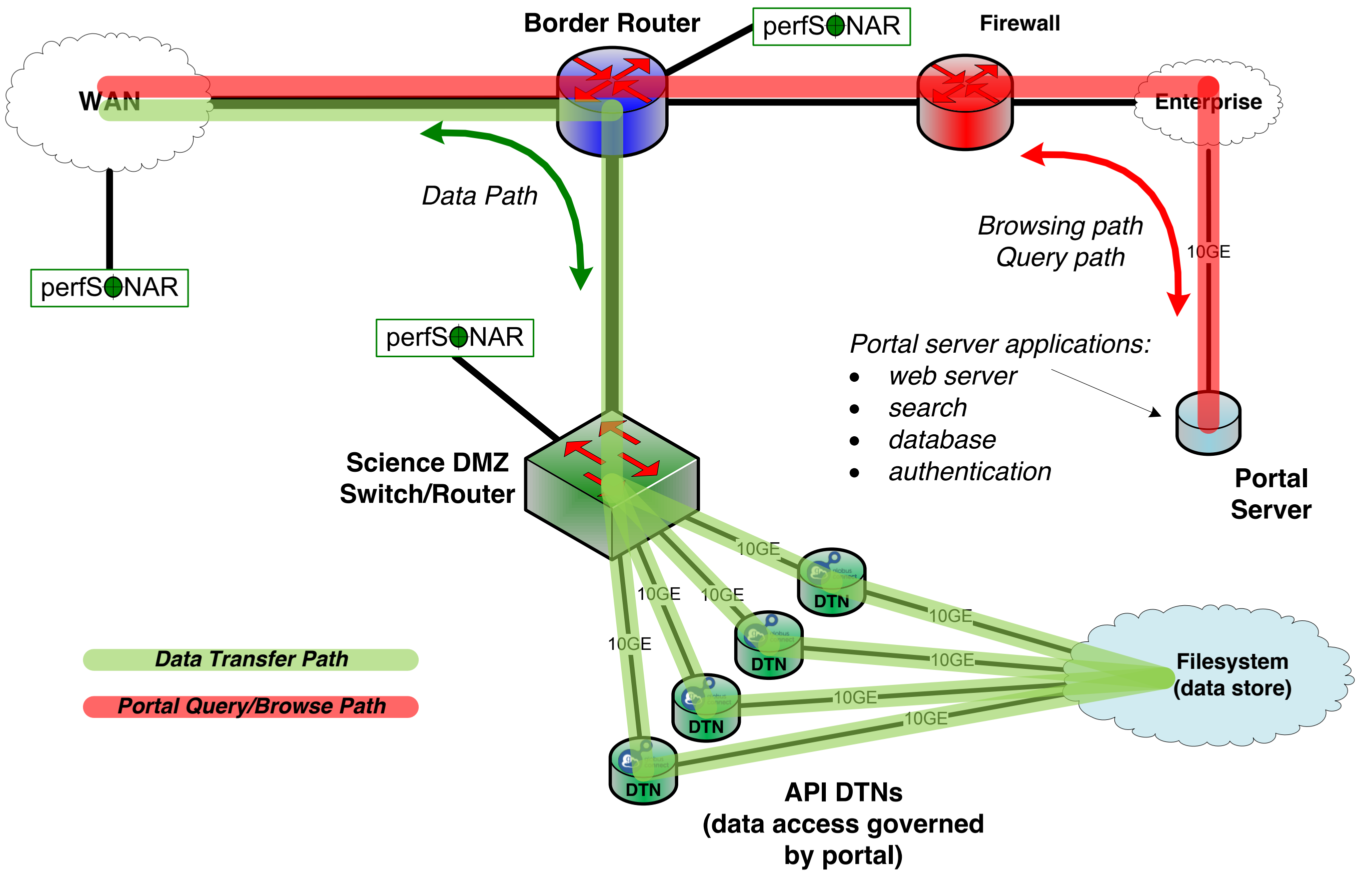


Browsing path
Query path
Data path

Portal server applications:

- *web server*
- *search*
- *database*
- *authentication*
- *data service*

Next-Generation Portal Leverages Science DMZ



Put The Data On Dedicated Infrastructure

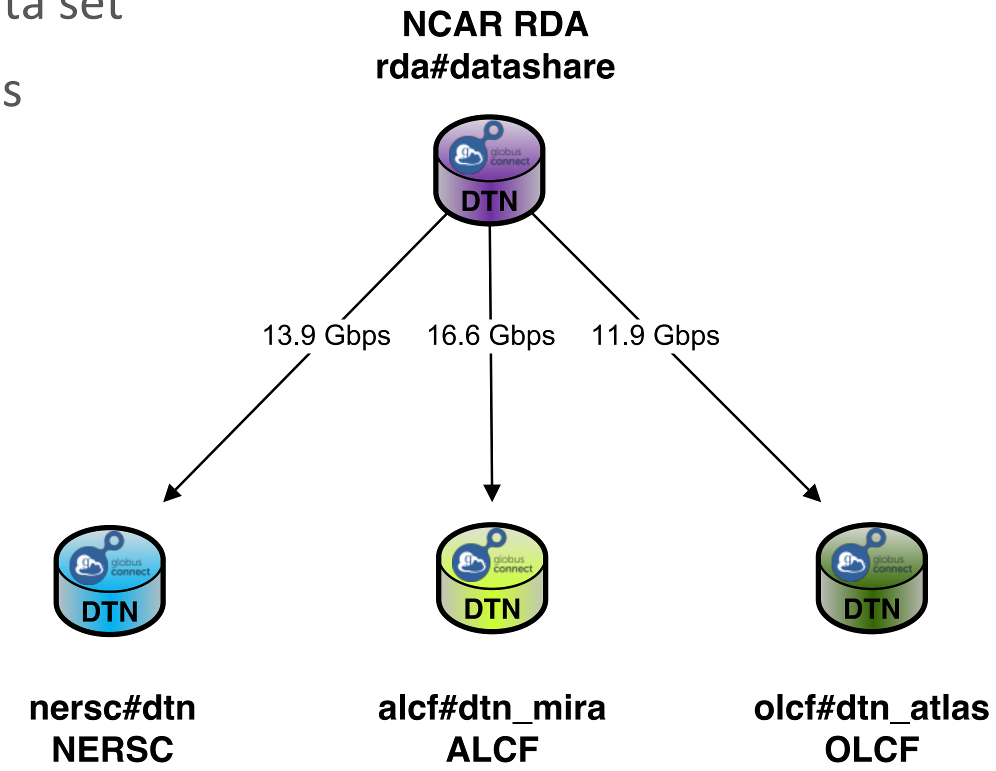
- We have separated the data handling from the portal logic
- Portal is still its normal self, but enhanced
  - Portal GUI, database, search, etc. all function as they did before
  - Query returns pointers to data objects in the Science DMZ
  - Portal is now freed from ties to the data servers (run it on Amazon if you want!)
- Data handling is separate, and scalable
  - High-performance DTNs in the Science DMZ
  - Scale as much as you need to without modifying the portal software
- Outsource data handling to computing centers
  - Computing centers are set up for large-scale data
  - Let them handle the large-scale data, and let the portal do the orchestration of data placement

Data Portal Implications

- Portals hold a lot of valuable data
  - Observations (sky surveys, satellite data, genomes, etc.)
  - Many have been in place for years
- Most are inadequate to support large-scale analysis
  - Legacy search/query interfaces
  - Legacy access protocols/tools
  - This is in the process of changing
- The technology exists to radically improve the utility of data portals
  - What should the performance expectation be?
  - HPC facilities can do 1PB/week – if data portals could do this…

NCAR RDA Performance to DOE HPC Facilities

- 1.5TB data set
- 1121 files



Summary

- Science networks are engineered to support data-intensive science
  - Related to and connected to the rest of the Internet, but different
- Science DMZ model effectively connects data infrastructure to networks
  - If you need to send your sysadmin to me, feel free
- Globus at HPC facilities makes terascale to petascale data transfers possible
  - (more on Globus later today)
- Huge opportunity in upgrading data portals to use Science DMZ, DTNs, advanced tools (e.g. Globus)
  - Make large data repositories available for analysis at HPC facilities

In conclusion – ESnet's vision:

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

ESnet

Links and Lists

- ESnet fasterdata knowledge base
  - http://fasterdata.es.net/
- Science DMZ paper
  - http://www.es.net/assets/pubs_presos/sc13sciDMZ-final.pdf
- Science DMZ email list
  - Send mail to sympa@lists.lbl.gov with subject "subscribe esnet-sciencedmz”
- perfSONAR
  - http://fasterdata.es.net/performance-testing/perfsonar/
  - http://www.perfsonar.net
- Globus
  - https://www.globus.org/

ESnet

ESnet
ENERGY SCIENCES NETWORK

Thanks!

Eli Dart

Energy Sciences Network (ESnet)

Lawrence Berkeley National Laboratory

http://my.es.net/

http://www.es.net/

http://fasterdata.es.net/

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB