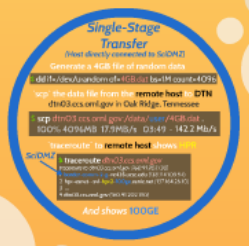
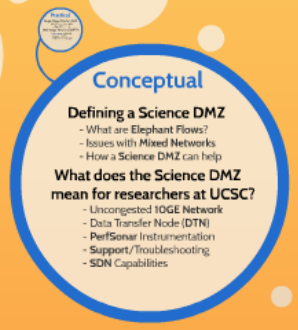




UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Brought to you by the UCSC CCNIE Team

Brad Smith, Jim Warner, Alan Lin,
John Haskins, Mark Boolootian,
George Peek, Shawfeng Dong,
Josh Sonstroem

*Funded by the **NSF CC-NIE***
Campus Cyberinfrastructure -
Network Infrastructure and Engineering Program

The Science DMZ and You

By Josh Sonstroem
UNIX Sysadmin, UC Santa Cruz
jsonstro@ucsc.edu

Conceptual

Defining a Science DMZ

- What are **Elephant Flows**?
- Issues with **Mixed Networks**
- How a **Science DMZ** can help

What does the Science DMZ mean for researchers at UCSC?

- Uncongested **10GE Network**
- Data Transfer Node (**DTN**)
- **PerfSonar** Instrumentation
- **Support**/Troubleshooting
- **SDN** Capabilities

Practical

Single-Stage Transfer (SCP)

- **SCP** from remote to DTN
- **OpenSSH** example

Dual-Stage Transfer (GridFTP)

- **Globus Online Web GUI**
  - Remote to **DTN**
  - **DTN** to laptop via **SCP**

Elephants in the Network?

*An **elephant flow** is a large continuous data flow as measured over a network link. **Elephant flows**, though not numerous, can occupy a disproportionate share of the total bandwidth over any given period of time.*

Both LONG-LIVED *and* LARGE-BANDWIDTH

★**Elephant (long-lived, large bandwidth)**

- Mouse (short-lived, **large bandwidth**)
- Mouse (**long-lived**, small bandwidth)
- Mouse (short-lived, small bandwidth)

A traditional **mixed network** contains both mice and **elephant** flows

Why care?

- Mice are **bursty** and **latency-sensitive,** retransmission is *relatively* cheap but **high-volume** can be limiting factor.
- Elephants are **large transfers** in which **throughput** is generally more important than **latency**, but retranmissions are *expensive* in terms of **bandwidth** and **time**.

Long-lived TCP flows **fill network buffers completely** introducing delays into shared hardware resources. In a **mixed network** the more latency-sensitive mice are affected first, which means **production traffic suffers**.

All types of flows are affected

Science networks
Whats the deal?

Growth trends of data use are EXTREME

***MICE** reproduce VERY fast*
(Internet-of-things, mobile, BYOD, etc.)

Wide-Area Networks that support large, fast transfers end-to-end are costly to build and support

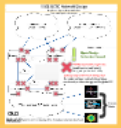
One ELEPHANT in a "room" full of MICE can wreak havoc

Solution?

Move the ELEPHANT flows to the edge of the network

- NO **forklift upgrade** to production path required
- **Business** and **Research** clients both satisfied
- Reduce **cost and time** to deploy/maintain

As demands by MICE increase, the value of moving ELEPHANTS rises as well



Traditional
Mixed Network

Elephant and **mice** flows **share** the same **production paths**. This can cause **contention** on links with **insufficient capacity**, and ultimately, **costly slow downs** for **elephant flows** and **long delays** for **mice**.

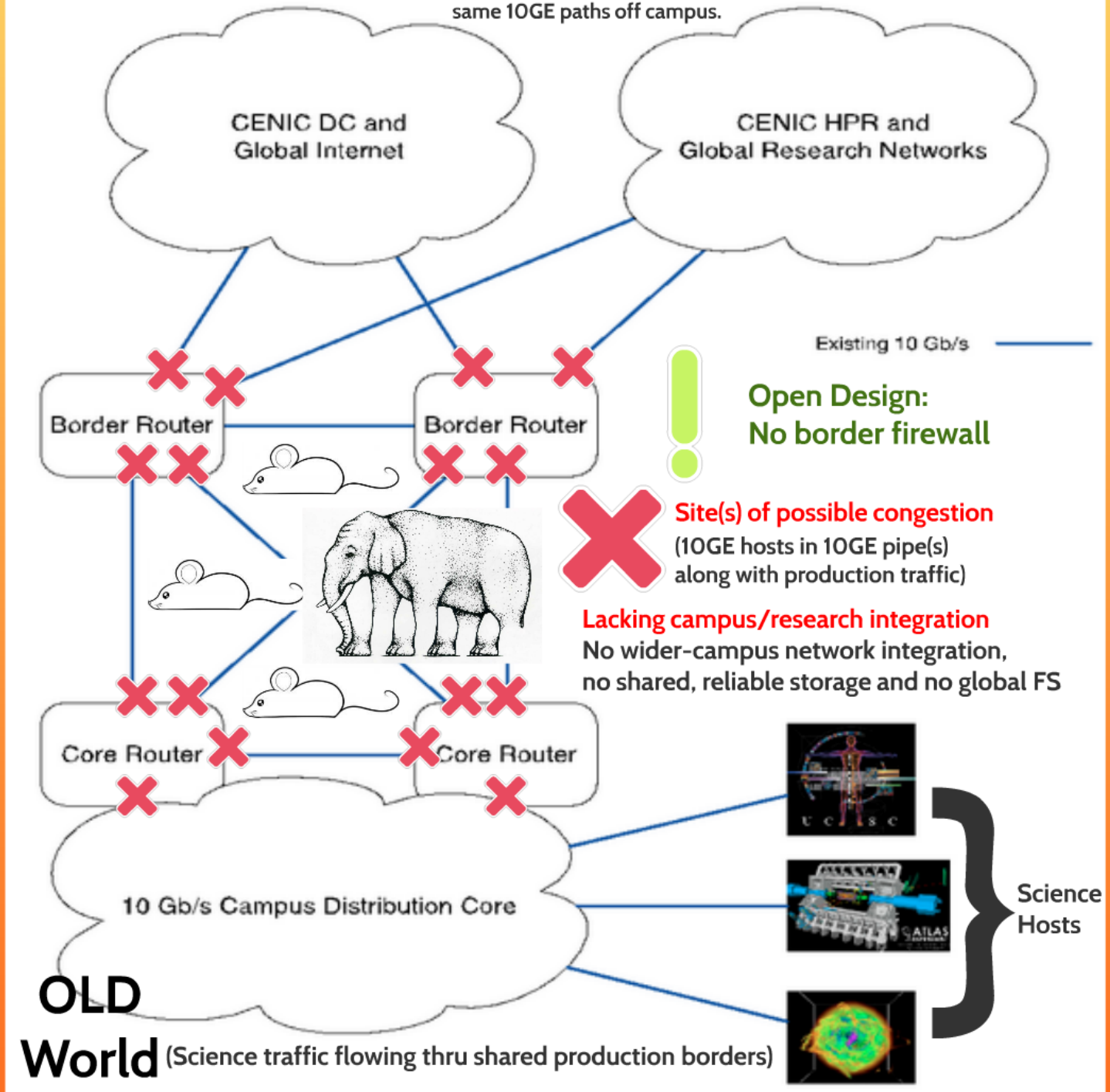
SCIENCE DMZ *To The Rescue!*

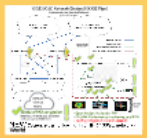
Since **business policies** and **security practices** (firewalls) are tuned to **mice** there are often performance penalties for **elephant flows**, both **during transfer** and **after failure**.

An OPEN NETWORK DESIGN *Can Help Combat These Effects*

10GE UCSC Network Design

Elephant and mice flows both share
same 10GE paths off campus.





What is a Science DMZ?

A **Science DMZ** is a portion of a network, built at or near the local network perimeter that is designed such that the equipment, configuration, and security policies are optimized for **high-performance scientific applications** rather than for general-purpose business systems.

*It is **scalable**, **incrementally deployable**, and **easily adaptable** to emerging technologies*

- **40G/100G Ethernet**
- **Layer 2 Virtual Circuits**
- **SDN Capabilities**

- IPv

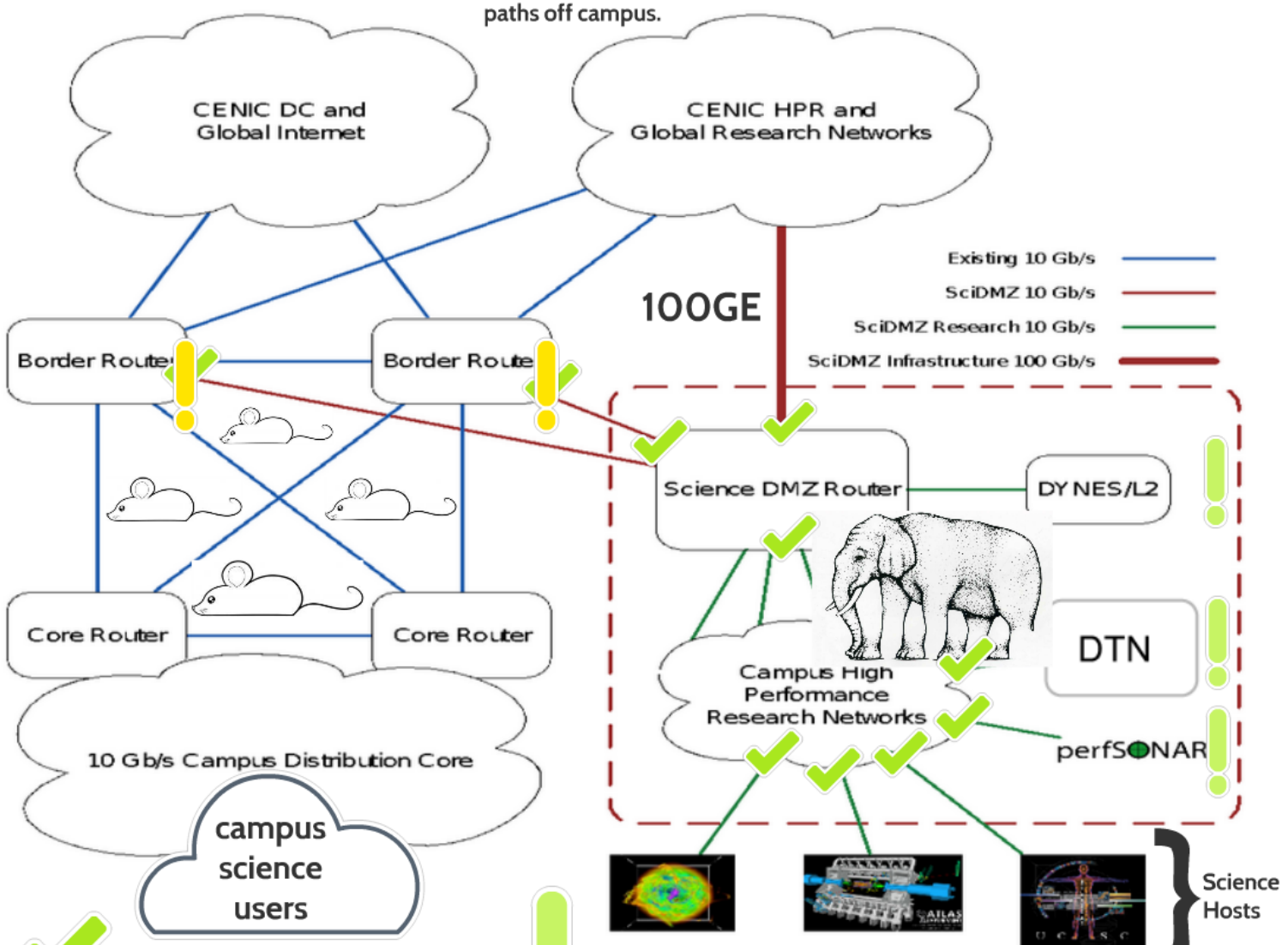
Power of 5

A **Science DMZ** integrates *five key components* into a **unified whole**

- A **network architecture** explicitly designed for high-performance applications, where **science/research** use is distinct from general-purpose use
- The use of **dedicated systems** for data transfer (**DTN**)
- **Performance measurement** and **network testing** systems that are regularly used to characterize the network and are available for troubleshooting (**PerfSonar**)
- **Security policies** and enforcement mechanisms that are tailored for high performance environments
- **Engagement with Network Users** focused on creating partnerships, educating and providing resources/ongoing support

10GE UCSC Network Design (100GE Pipe)

Elephant and mice flows have different paths off campus.



Site(s) of NO congestion

(10GE hosts in 100GE pipe, production traffic on separate 10GE links)

- Open Design: No border FW
- L2/SDN, Performance monitoring, and DTN
- Lacking integrated research support

NEW World (Science traffic flowing thru HPR 100GE network)

Why This Matters

This chart shows **transfers times** for moving **1 Terabyte** of data across **various speed networks**

modem	10 Mbps network	300 hrs (12.5 days)
wireless	100 Mbps network	30 hrs
ethernet	1 Gbps network	3 hrs
SciDMZ	**10 Gbps network**	**20 minutes**

While its *relatively* easy to achieve **line-rate transfers** over short distances **(LAN)**, over larger distances **(WAN)** it can require **special software, hardware,** and **OS tuning**

Data Transfer Node *to the rescue!*

A **DTN** is a specialized file server, typically with **lots of memory** and **many disks**; tuned for **WAN** access, clients on the **LAN** need only be **minimally tuned/configured,** helping reduce both **cost** and **time-to-deployment**.

Great, but what does this all mean for researchers?

SciDMZ@UCSC is available *NOW*

- *Uncongested* **10GE Network**
- **Data Transfer Node** Access
- **PerfSonar** Instrumentation
- **Support/Troubleshooting**
- **SDN Capabilities** for Research/QoS
- **IPv6 enabled**

Traditional
Mixed Network

Elephant and **mice** flows **share** the same **production paths**. This can cause **contention** on links with **insufficient capacity**, and ultimately, **costly slow downs** for **elephant flows** and **long delays** for mice.

SCIENCE DMZ *To The Rescue!*

Since **business policies** and **security practices** (firewalls) are tuned to mice there are often performance penalties for **elephant flows**, both **during transfer** and **after failure**.

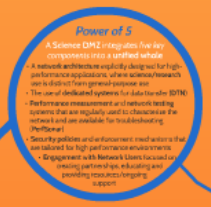
An OPEN NETWORK DESIGN *Can Help Combat These Effects*

What is a Science DMZ?

A **Science DMZ** is a portion of a network, built at or near the local network perimeter that is designed such that the equipment, configuration, and security policies are optimized for **high-performance scientific applications** rather than for general-purpose business systems.

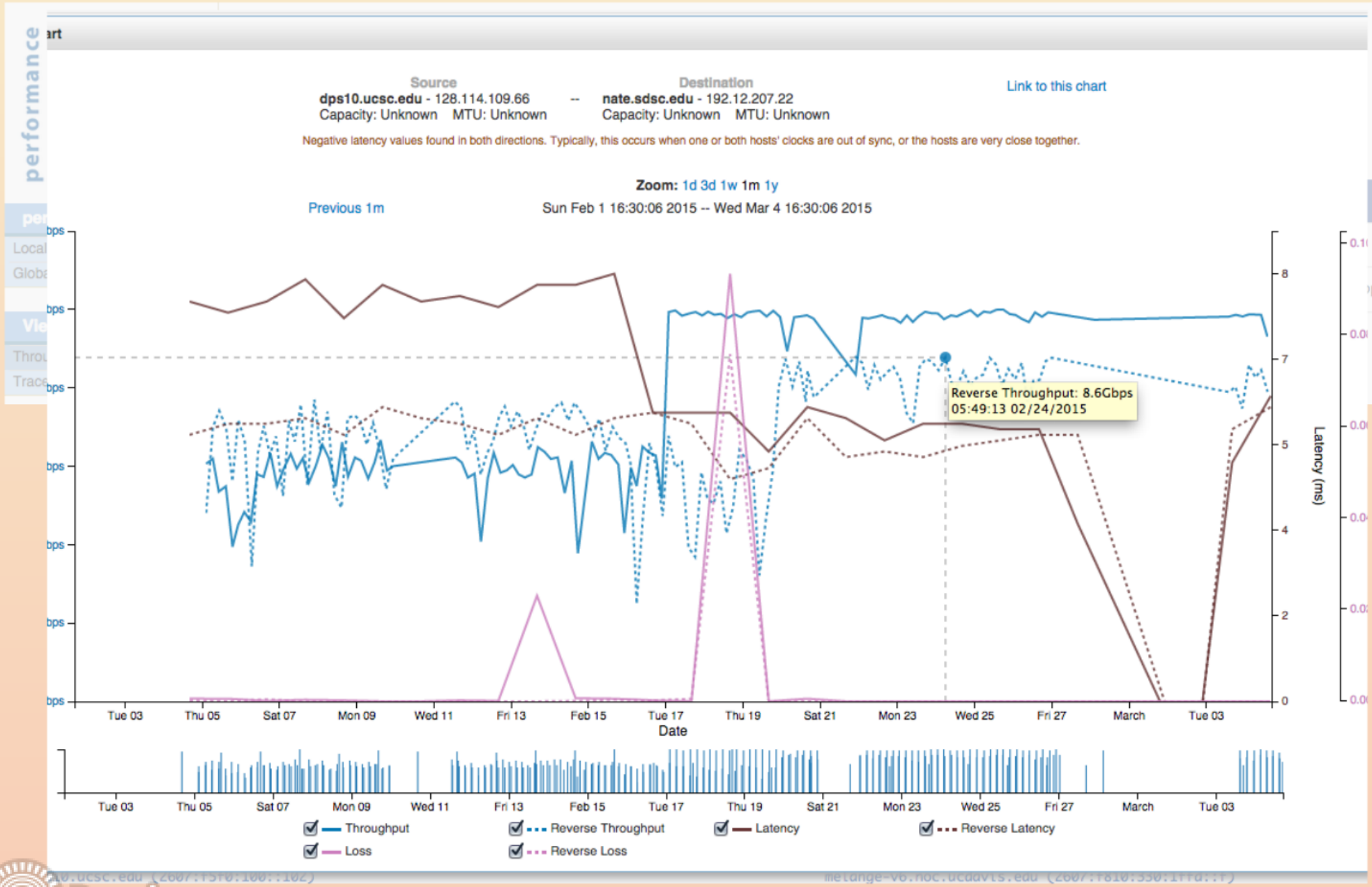
*It is **scalable**, **incrementally deployable**, and **easily adaptable** to emerging technologies*

- 40G/100G Ethernet
- Layer 2 Virtual Circuits
- SDN Capabilities



So what the heck is PerfSonar*?*

Its for automated network bandwidth/latency testing, discovery of soft-failures, and establishing change-from-baseline



What does it take to get connected?

SciDMZ@UCSC is a pilot service

- ***Existing research hosts*** in *campus data center* can connect via the DC switch, but can't be multi-homed behind the data center firewall
- ***New on-campus research hosts*** have to fund connectivity from the nearest **SciDMZ switch** to their lab or building compute location(s)
- ***New large-scale research deployments*** are encouraged to deploy hardware at ***SDSC*** and leverage DTNs to move data between sites

Use-Case Dependent

Limitations: data center power/space and availablility of intra-campus layer-1 resouces

Who is on the SciDMZ@UCSC?

Center for Biomolecular Science & Engineering (*CBSE*)

- **cghub.ucsc.edu** – 2PB dataset (81K files), downloads 1PB/mo
- **genomics.ucsc.edu** - 60TB dataset, dowloads 2TB/day
- **Big Data in Translational Genomics** (**BD2K**) - NIH project '17+

Santa Cruz Institute for Particle Physics (*SCIPP*)

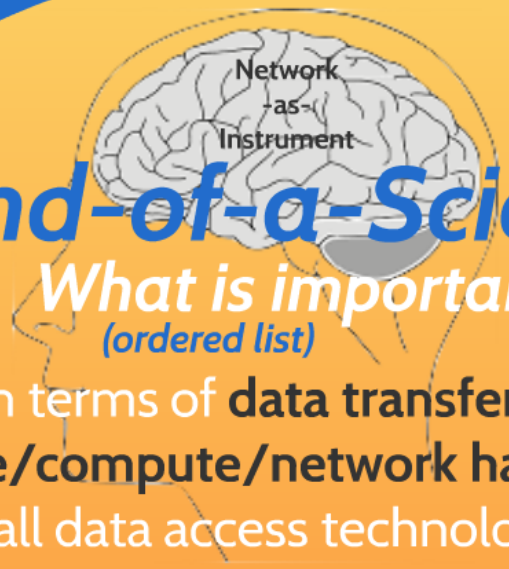
- **ATLAS/LHC** - dataset grows 10-20TB/year

Astrophysics

- **Hyades cluster** – 270TB lustre filesystem, 1PB S3 datastore, 100PB+ simulation data @ national labs

ITS

- **DTN** - GridFTP
- **Fiona** - GridFTP
- **DPS10** - PerfSONAR



Mind-of-a-Scientist

What is important?

(ordered list)

In terms of **data transfer tools, storage/compute/network hardware,** and all data access technologies...

1. **Correctness**
2. **Consistency**
3. **Convenience**
4. **Performance**

Implement and interconnect Science DMZs from around the world

Extend SciDMZ services to the location of end-user resources

How can we this apply this insight?

Make TRANSPARENT DATA PLACEMENT our endgame

Single-Stage Transfer

(Host directly connected to SciDMZ)

Generate a 4GB file of random data

```
$ dd if=/dev/urandom of=4GB.dat bs=1M count=4096
```

`scp` the data file from the remote host to DTN dtn03.ccs.ornl.gov in Oak Ridge, Tennessee

```
$ scp dtn03.ccs.ornl.gov:/data/user/4GB.dat .
  100% 4096MB  17.9MB/s   03:49 ~ 142.2 Mb/s
```

`traceroute` to remote host shows HPR

SciDMZ

```
$ traceroute dtn03.ccs.ornl.gov
traceroute to dtn03.ccs.ornl.gov (160.91.202.130)
1 border-comm-2-g-ve435.ucsc.edu (128.114.109.94)
2 hpr-esnet--svl-hpr2-100ge.cenic.net (137.164.26.10)
3 ...
9 dtn03.ccs.ornl.gov (160.91.202.130)
```

And shows 100GE

Techn

Den

- `scp` from host to clie
- `traceroute host from

Prezi

How is this different than the production pathway?

`scp` the data file from remote host to data center over production path

```
$ scp dtn03.ccs.ornl.gov:/data/user/4GB.dat .
  100% 4096MB    9.3MB/s   7:21
```

~ 74.4 Mb/s

Slower

`traceroute` to remote host shows production border

```
$ traceroute dtn03.ccs.ornl.gov
traceroute to dtn03.ccs.ornl.gov (160.91.202.130)
 2  isb-g-te1-3.ucsc.edu (128.114.1.137)
 3  border-comm-g-te3-2.ucsc.edu (128.114.0.46)
 4  hpr-svl-hpr2--ucsc.cenic.net (137.164.26.93)
 5  hpr-esnet--svl-hpr2-100ge.cenic.net (137.164.26.10)
 6  ...
14 dtn03.ccs.ornl.gov (160.91.202.130)
```

Still shows hpr and 100GE

Dual-Stage Transfer

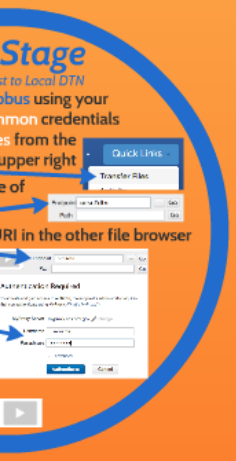
Globus Online - https://www.globus.org

- **DTN** runs their server software
- Uses **GridFTP** under the hood
- **Sign-up** is **FREE**, as is their PC software ***Globus Connect Personal***
- Links up with **InCommon** via **CILogon** (Globus user linked to CILogon CN, CN mapped to local POSIX account)
- Uses **Unix directory perms** on host

Technical

Demo

- **Globus GUI** transfer
  esnet#bnl-diskpt1 to
  ucsc#dtn
- **scp** data from **dtn.uc**
  to **workstation** over

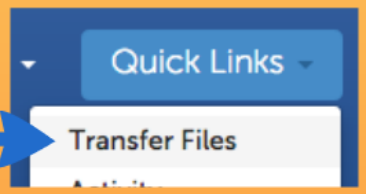


Prezi

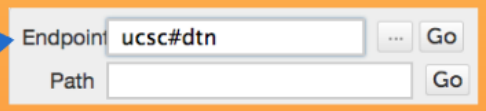
First Stage

Remote Host to Local DTN

1. Login to Globus using your UCSC/InCommon credentials
2. Select Transfer Files from the Quick Links menu at upper right



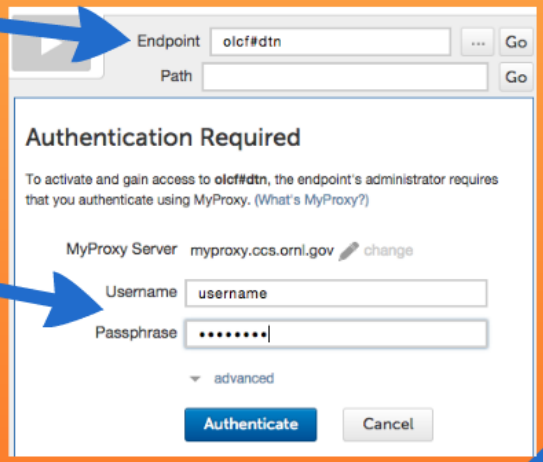
3. Enter ucsc#dtn in one of the two file browsers



4. Enter remote host's URI in the other file browser

If Needed

5. Authenticate via MyProxy and your remote credentials

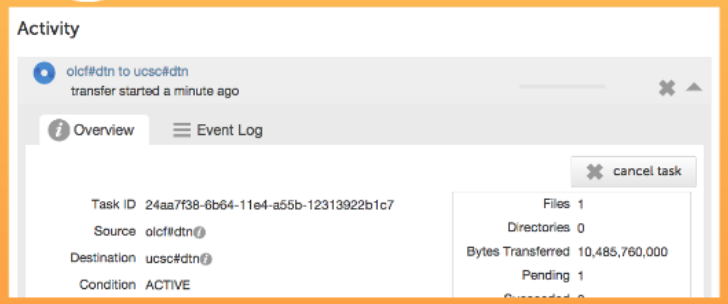


6. select file(s), click transfer

First Stage

Remote Host to Local DTN (continued)

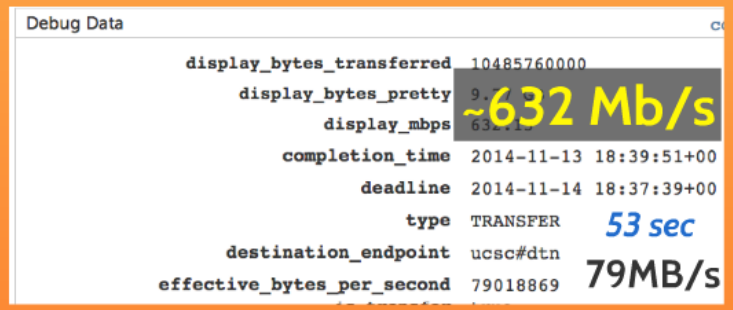
7 check activity window



8 click

view debug data

once complete



Peak of 1.5Gb/s

Now lets move it to our workstation

Prezi

Second Stage

intra-campus

`scp` the data file from dtn to our workstation over production path

```
$ scp user@dtn.ucsc.edu:/data/user/4GB.dat .
100%  4096MB  30.8MB/s  01:45 ~ 246.4 Mb/s
```

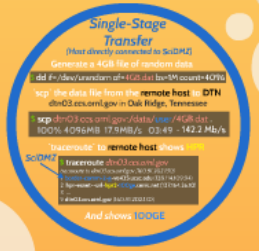
`traceroute` to dtn shows production border

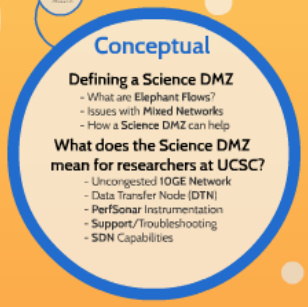
```
$ traceroute dtn.ucsc.edu
traceroute to dtn.ucsc.edu (128.114.109.65)
3 border-comm-g-te3-1.ucsc.edu (128.114.0.66)
4 border-comm-2-g-e1-1-v441.ucsc.edu (128.114.0.27)
5 dtn.ucsc.edu (128.114.109.65)
```

Science DMZ

100G Single File			Dual Stage (10GE, 1GE)						Single Stage (10GE)		
Globus SP	Globus EP	Gb/s	Time (Min)	Stage 2	Mb/s	Time (Min)	Total		Time (Min)	Mb/s	Proto
ANL	UCSC	5.86	2:16:00	*scp*	685	20:23:00	22:39:00	vs.	29:37:00	454	*scp*
BNL	UCSC	5.60	2:23:00	*scp*	595	22:24:00	24:47:00	vs.	32:31:00	410	*scp*
LBL	UCSC	5.65	2:21:00	*scp*	723	18:16:00	20:37:00	vs.	25:19:00	527	*scp*
ANL	UCSC	5.90	2:15:00	*globus*	878	15:10:00	17:25:00				
BNL	UCSC	5.32	2:26:00	*globus*	911	14:38:00	17:04:00				
LBL	UCSC	6.20	2:09:00	*globus*	896	15:01:00	17:10:00				

For more info on data transfer tools see: https://pleiades.ucsc.edu/hyades/Globus_on_dtn





UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Brought to you by the UCSC CCNIE Team

Brad Smith, Jim Warner, Alan Lin,
John Haskins, Mark Booloottian,
George Peek, Shawfeng Dong,
Josh Sonstroem

Funded by the ***NSF CC-NIE***
Campus Cyberinfrastructure -
Network Infrastructure and Engineering Program

The Science DMZ and You

By Josh Sonstroem
UNIX Sysadmin, UC Santa Cruz
jsonstro@ucsc.edu