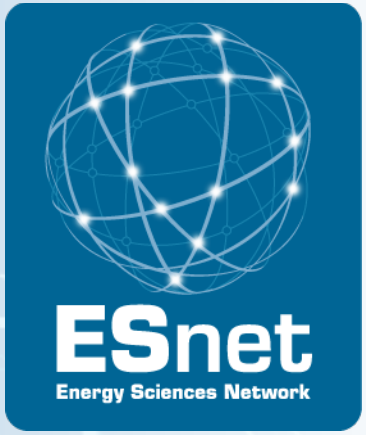
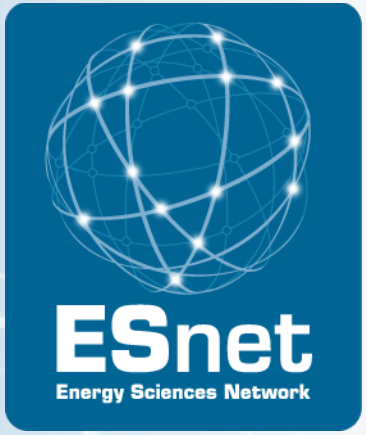


Achieving the Science DMZ

Eli Dart, Brian Tierney and Eric Pouyoul, ESnet

Joe Breen: University of Utah

Joint Techs, Baton Rouge, LA, January, 2012



Achieving the Science DMZ

Section 1: Science DMZ Overview and Architecture

Eli Dart, ESnet

Joint Techs, Baton Rouge, LA, January, 2012

Outline of the Day

Motivation

Services Overview

Science DMZ Architecture

The data transfer node

Performance, measurement, monitoring

Utah deployment

Motivation

Science data increasing both in volume and in value

- Higher instrument performance
- Increased capacity for discovery
- Analyses previously not possible

Lots of promise, but only if scientists can actually work with the data

- Data has to get to analysis resources
- Results have to get to people
- People have to share results

Common pain point – data mobility

- Movement of data between instruments, facilities, analysis systems, and scientists is a gating factor for much of data intensive science
- Data mobility is not the only part of data intensive science – not even the most important part
- However, without data mobility data intensive science is hard

We need to move data – how can we do it consistently well?

Motivation (2)

Networks play a crucial role

- The very structure of modern science assumes science networks exist – high performance, feature rich, global scope
- Networks enable key aspects of data intensive science
  - Data mobility, automated workflows
  - Access to facilities, data, analysis resources

Messing with the network is unpleasant for most scientists

- Not their area of expertise
- Not where the value is (no papers come from messing with the network)
- Data intensive science is about the science, not about the network
- However, it's a critical service – if the network breaks, everything stops

Therefore, infrastructure providers must cooperate to build consistent, reliable, high performance network services for data mobility

Here we describe one blueprint, the Science DMZ model – there are certainly others, but this one seems to work well in a variety of environments

TCP Background

Networks provide connectivity between hosts – how do hosts see the network?

- From an application's perspective, the interface to "the other end" is a socket
- Other similar constructs exist for non-IP protocols
- Communication is between applications – mostly over TCP

TCP – the fragile workhorse

- TCP is (for very good reasons) timid – packet loss is interpreted as congestion
- Packet loss in conjunction with latency is a performance killer
- Like it or not, TCP is used for the vast majority of data transfer applications

TCP Background (2)

It is far easier to architect the network to support TCP than it is to fix TCP

- People have been trying to fix TCP for years – some success
- However, here we are – packet loss is still the number one performance killer in long distance high performance environments

Pragmatically speaking, we must accommodate TCP

- Implications for equipment selection
  - Equipment must be able to accurately account for packets
- Implications for network architecture, deployment models
  - Infrastructure must be designed to allow easy troubleshooting
  - Test and measurement tools are critical – they have to be deployed

A small amount of packet loss makes a huge difference in TCP performance

A Nagios alert based on our regular throughput testing between one site and ESnet core alerted us to poor performance on high latency paths

No errors or drops reported by routers on either side of problem link

- only perfSONAR bwctl tests caught this problem

Using packet filter counters, we saw 0.0046% loss in one direction

- 1 packet in 22000 packets

Performance impact of this: (outbound/inbound)

- To/from test host 1 ms RTT : 7.3 Gbps out / 9.8 Gbps in
- To/from test host 11 ms RTT: 1 Gbps out / 9.5 Gbps in
- To/from test host 51ms RTT: 122 Mbps out / 7 Gbps in
- To/from test host 88 ms RTT: 60 Mbps out / 5 Gbps in
  - More than 80 times slower!

How Do We Accommodate TCP?

High-performance wide area TCP flows must get loss-free service

- Sufficient bandwidth to avoid congestion
- Deep enough buffers in routers and switches to handle bursts
  - Especially true for long-distance flows due to packet behavior
  - No, this isn't buffer bloat

Equally important – the infrastructure must be verifiable so that clean service can be provided

- Stuff breaks
  - Hardware, software, optics, bugs, …
  - How do we deal with it in a production environment?
- Must be able to prove a network device or path is functioning correctly
  - Accurate counters must exist and be accessible
  - Need ability to run tests - perfSONAR
- Small footprint is a huge win – small number of devices so that problem isolation is tractable

Services Overview – Wide Area

Data transfer takes advantage of wide area services

High-performance routed IP with global connectivity

- Bread and butter
- Must be high-bandwidth, verifiably loss-free in general case

Virtual circuit service

- Traffic isolation, traffic engineering
- Bandwidth and service guarantees
- Support for non-IP protocols

Test and measurement

- perfSONAR
- Enable testing, verification of performance, problem isolation
- Understand nominal conditions → what's normal, what's broken

Services Overview – Site/Campus

High performance routed IP

- Well-matched to wide area science service
- Verifiably loss-free

Circuit termination/endpoints

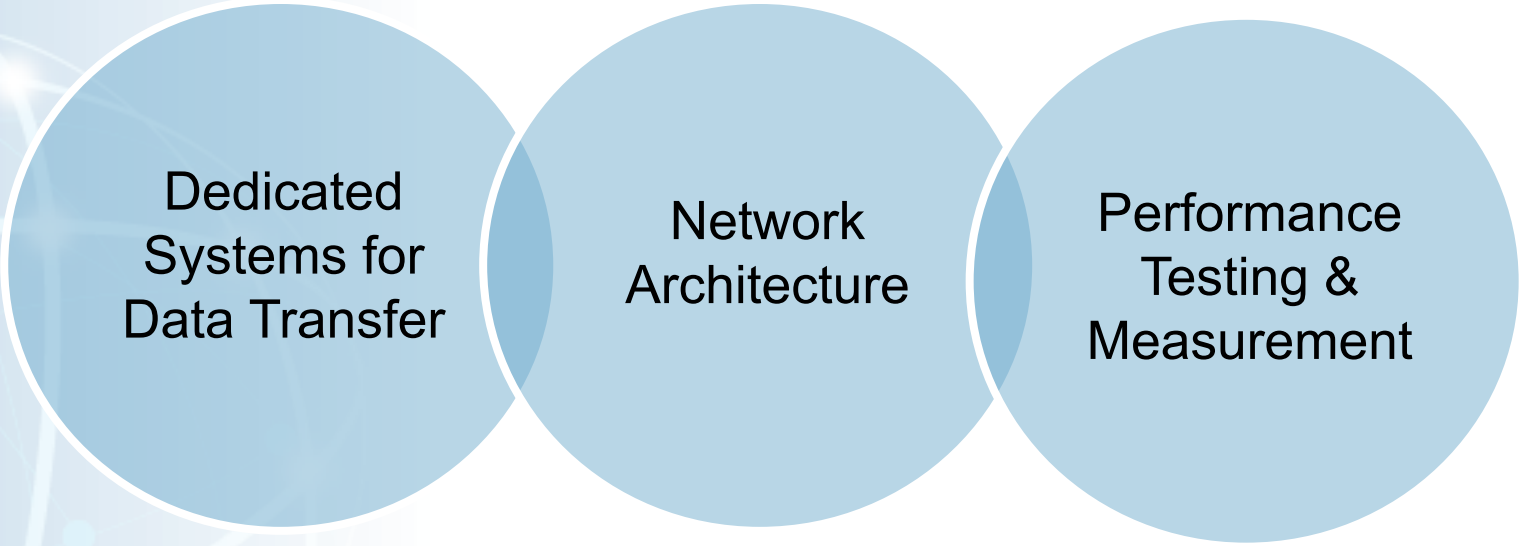
- DYNES, Tier1, …
- Remote filesystem mounts
- Non-IP protocols

Data sources and sinks

- Instruments and facilities
- Analysis resources
- Data systems

It is at the site or campus that it all comes together – scientists, instruments, data, analysis

The Data Transfer Trifecta: The "Science DMZ" Model



Data Transfer Node

- High performance
- Configured for data transfer
- Proper tools

Science DMZ

- Dedicated location for DTN
- Easy to deploy - no need to redesign the whole network
- Additional info: http://fasterdata.es.net/

perfSONAR

- Enables fault isolation
- Verify correct operation
- Widely deployed in ESnet and other networks, as well as sites and facilities

Science DMZ Service Interaction

WAN entry

- How do wide area services enter the site?
- If they don't come to the Science DMZ first, there must be a clean path to the Science DMZ
- Clean wide area path for long-distance flows is key

Circuit services entry

- Virtual circuits support DYNES, LHC experiments, remote filesystem mounts, non-IP protocols, …

Local resources

- Data Transfer Nodes
- Test and measurement (perfSONAR)

Security policy

- Separation of science and business traffic

Science DMZ Takes Many Forms

There are a lot of ways to combine these things – it all depends on what you need to do

- Small installation for a project or two
- Facility inside a larger institution
- Institutional capability serving multiple departments/divisions
- Science capability that consumes a majority of the infrastructure

Some of these are straightforward, others are less obvious

Key point of concentration: High-latency path for TCP

Ad Hoc Deployment

This is often what gets tried first

Data transfer node deployed where the owner has space

- This is often the easiest thing to do at the time
- Straightforward to turn on, hard to achieve performance

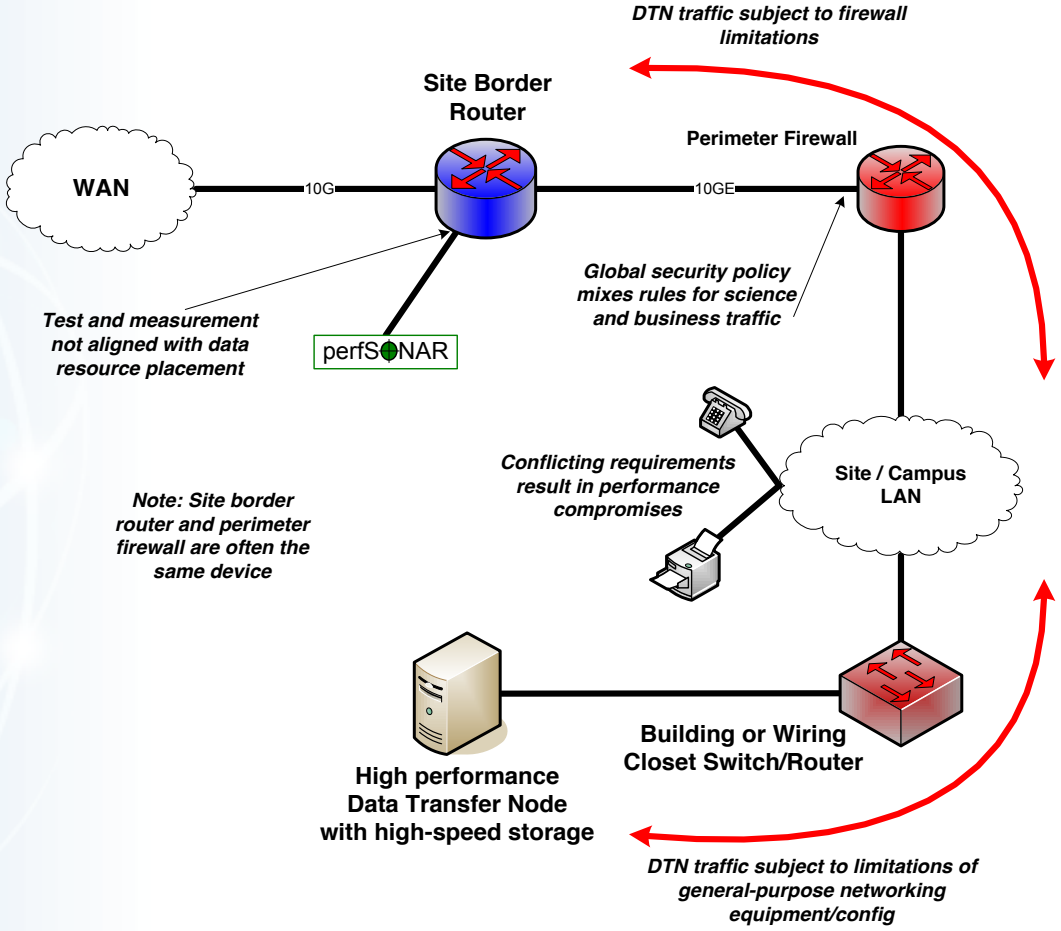
perfSONAR at the border

- This is a good start
- Need a second one next to the DTN

Entire LAN path has to be sized for data flows

Entire LAN path is part of any troubleshooting exercise

Ad Hoc DTN Deployment



Small-scale or Prototype Deployment

Add-on to existing network infrastructure

- All that is required is a port on the border router
- Small footprint, pre-production commitment

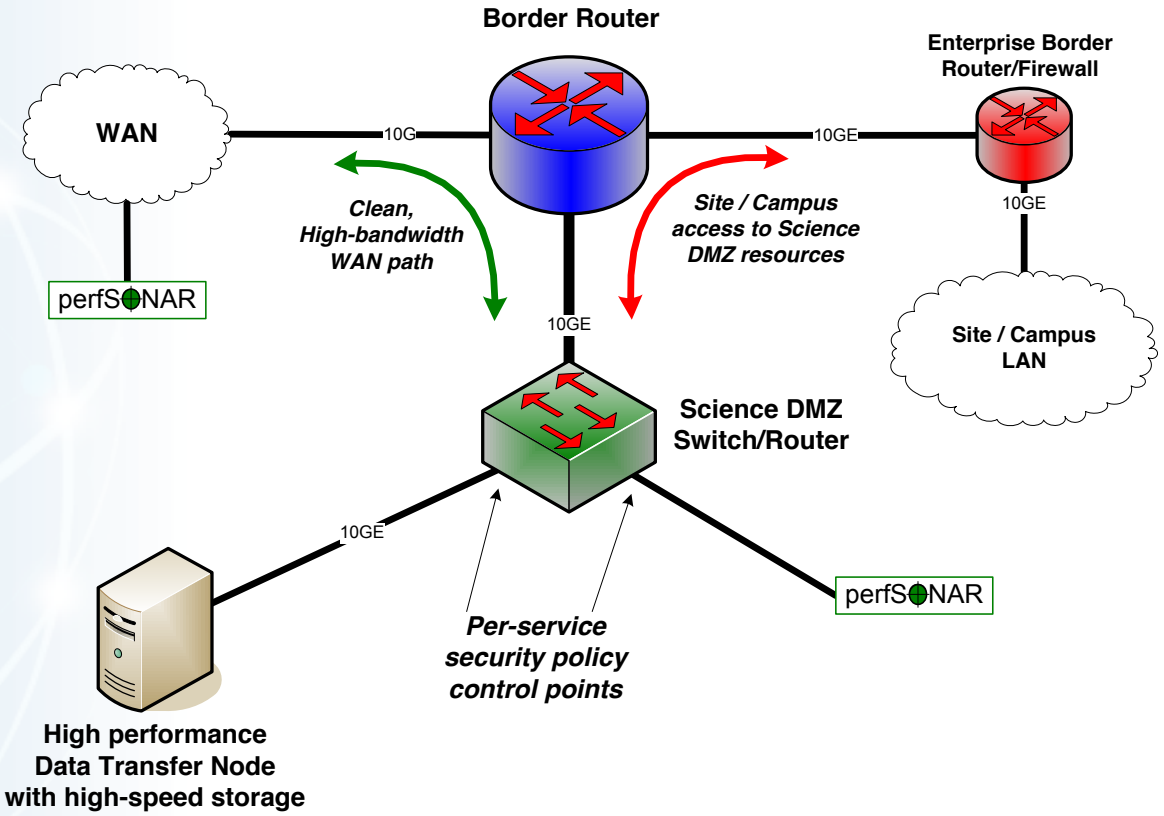
Easy to experiment with components and technologies

- DTN prototyping
- perfSONAR testing

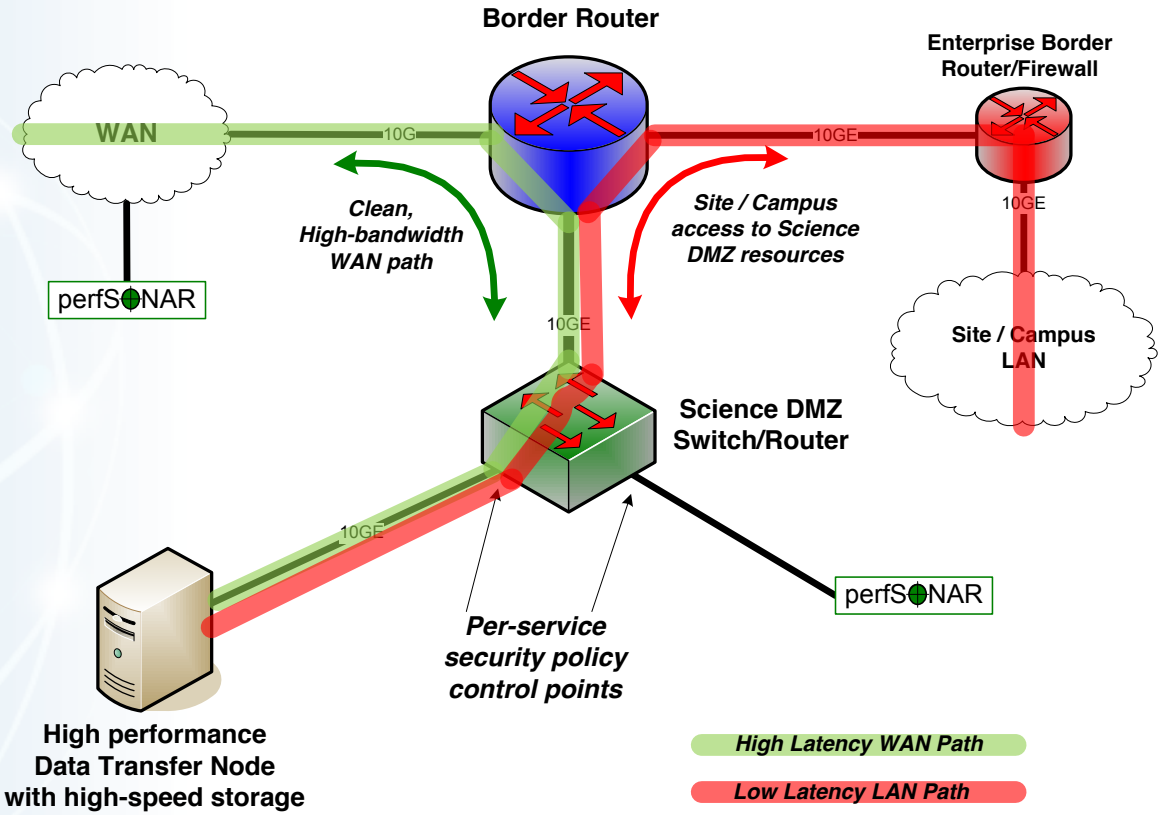
Limited scope makes security policy exceptions easy

- Only allow traffic from partners
- Add-on to production infrastructure – lower risk

Prototype Science DMZ



Prototype Science DMZ Data Path



Prototype With Virtual Circuits

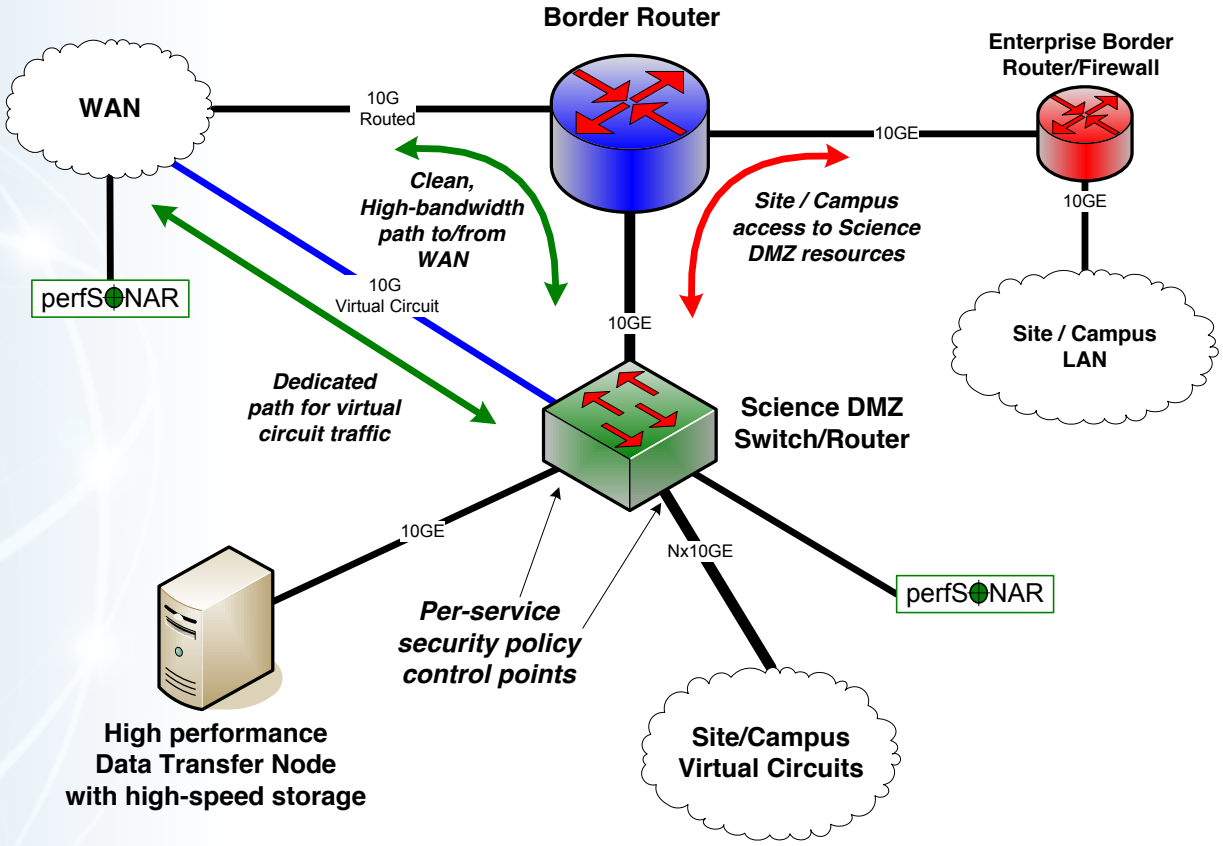
Small virtual circuit prototype can be done in a small Science DMZ

- Perfect example is a DYNES deployment
- Virtual circuit connection may or may not traverse border router

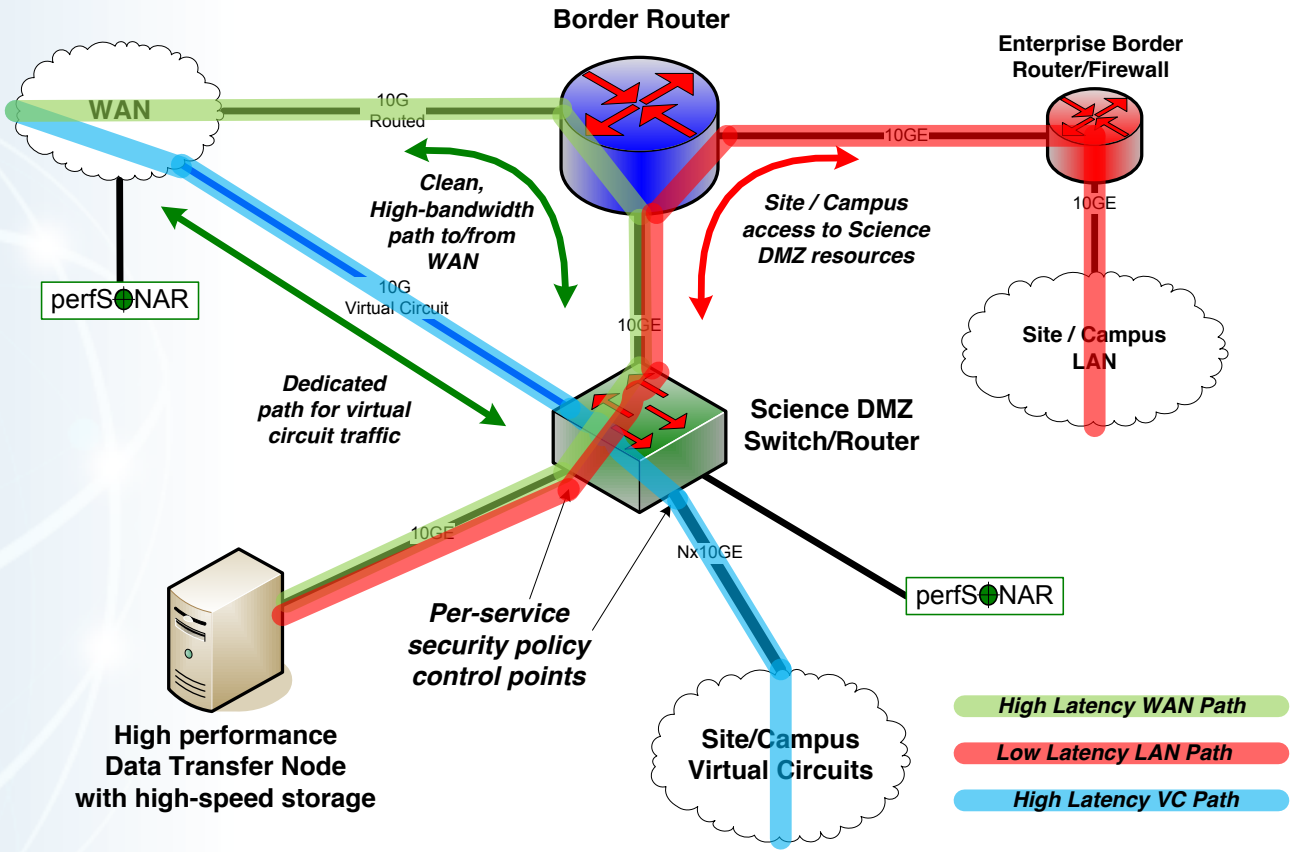
As with any Science DMZ deployment, this can be expanded as need grows

In this particular diagram, Science DMZ hosts can use either the routed or the circuit connection

Virtual Circuit Prototype Deployment



Virtual Circuit Prototype Data Path



Support For Multiple Projects

Science DMZ architecture allows multiple projects to put DTNs in place

- Modular architecture
- Centralized location for data servers

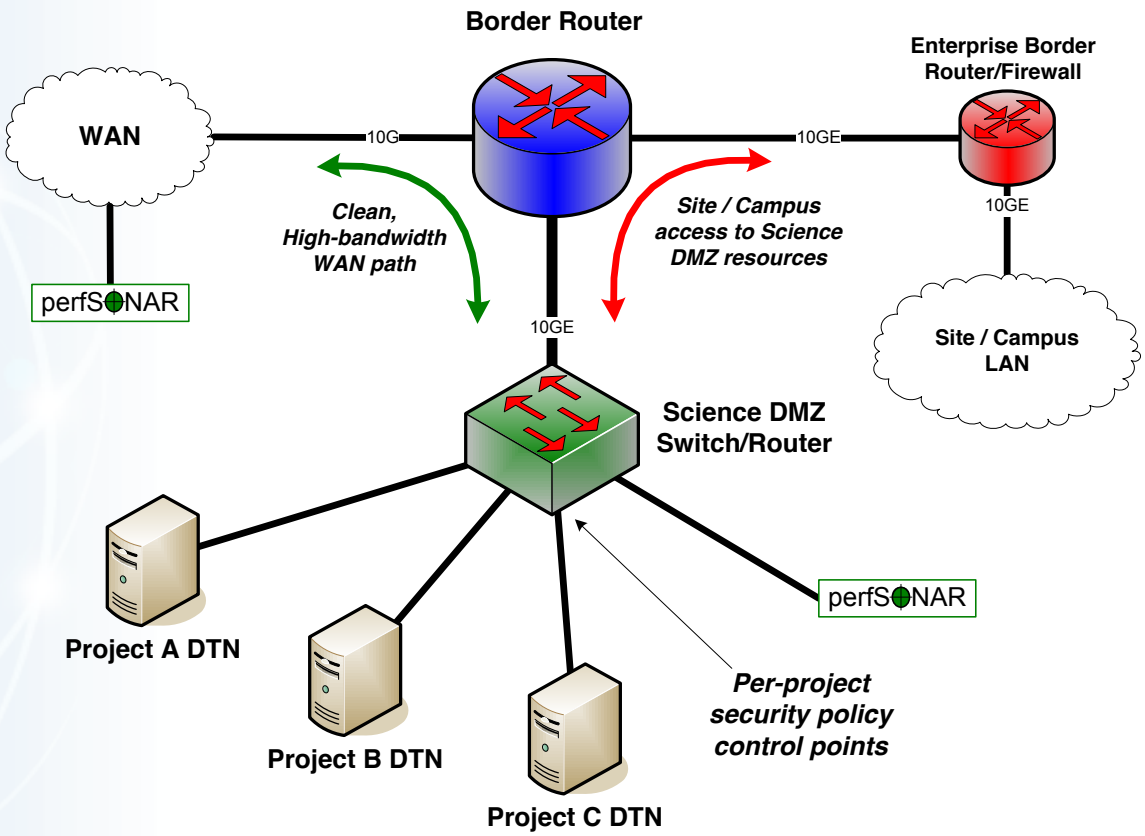
This may or may not work well depending on institutional politics

- Issues such as physical security can make this a non-starter
- On the other hand, some shops already have service models in place

On balance, this can provide a cost savings – it depends

- Central support for data servers vs. carrying data flows
- How far do the data flows have to go?

Multiple Projects



Supercomputer Center Deployment

High-performance networking is assumed in this environment

- Data flows between systems, between systems and storage, wide area, etc.
- Global filesystem often ties resources together
  - Portions of this may not run over Ethernet (e.g. IB)
  - Implications for Data Transfer Nodes

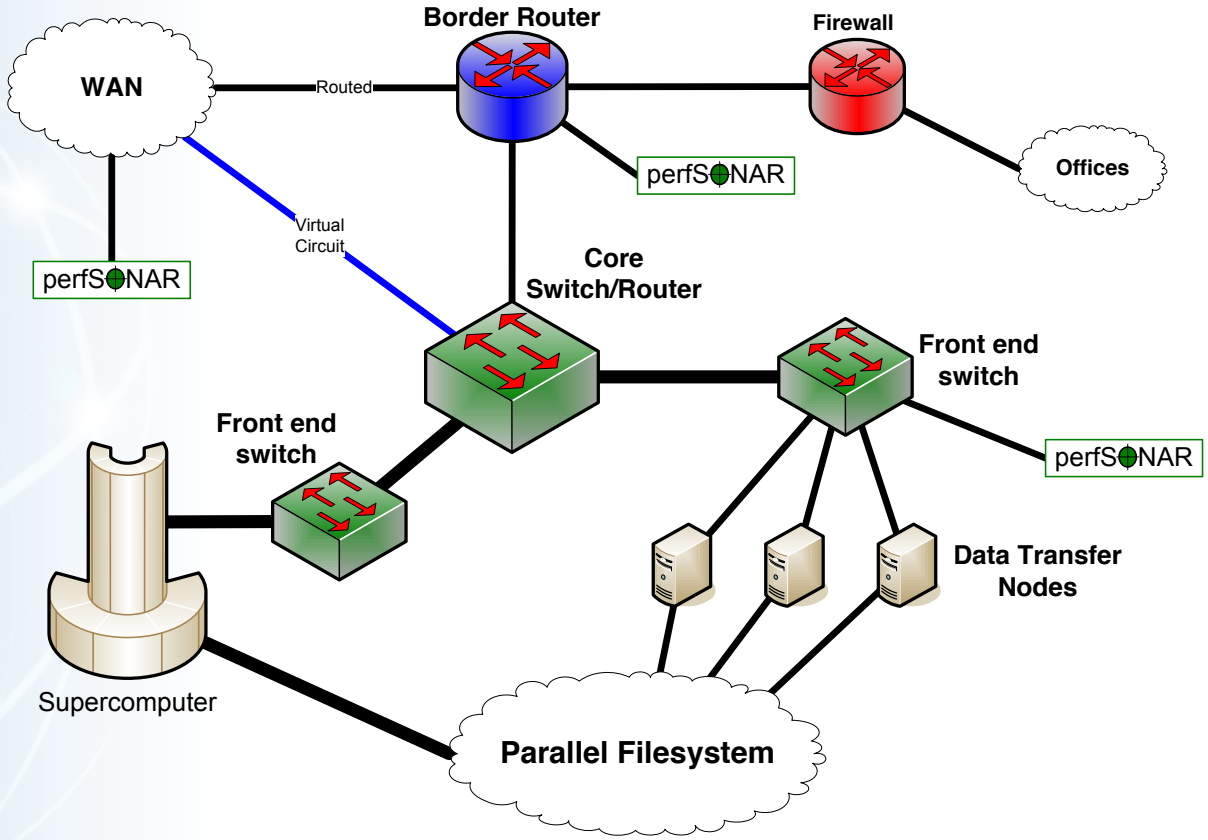
"Science DMZ" may not look like a discrete entity here

- By the time you get through interconnecting all the resources, you end up with most of the network in the Science DMZ
- This is as it should be – the point is appropriate deployment of tools, configuration, policy control, etc.

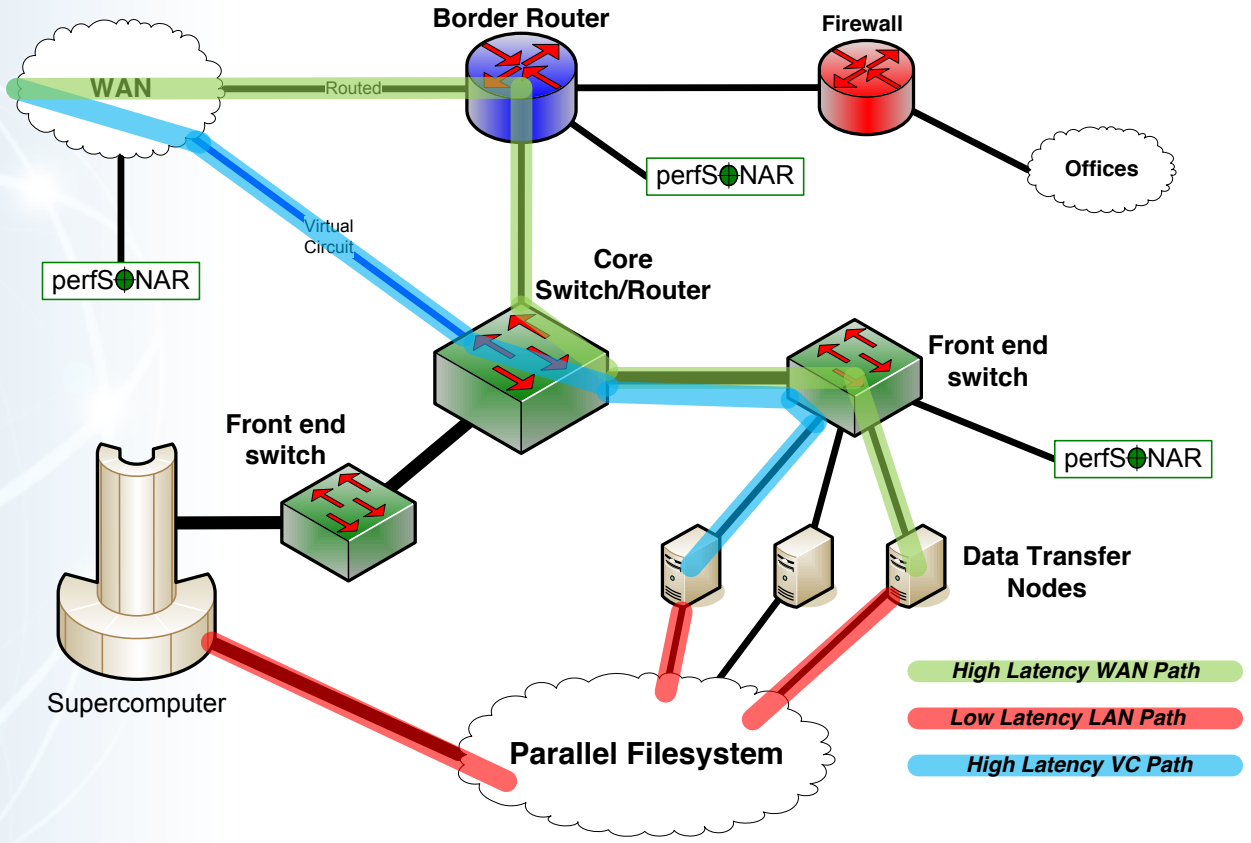
Office networks can look like an afterthought, but they aren't

- Deployed with appropriate security controls
- Office infrastructure need not be sized for science traffic

Supercomputer Center



Supercomputer Center Data Path



Major Data Site Deployment

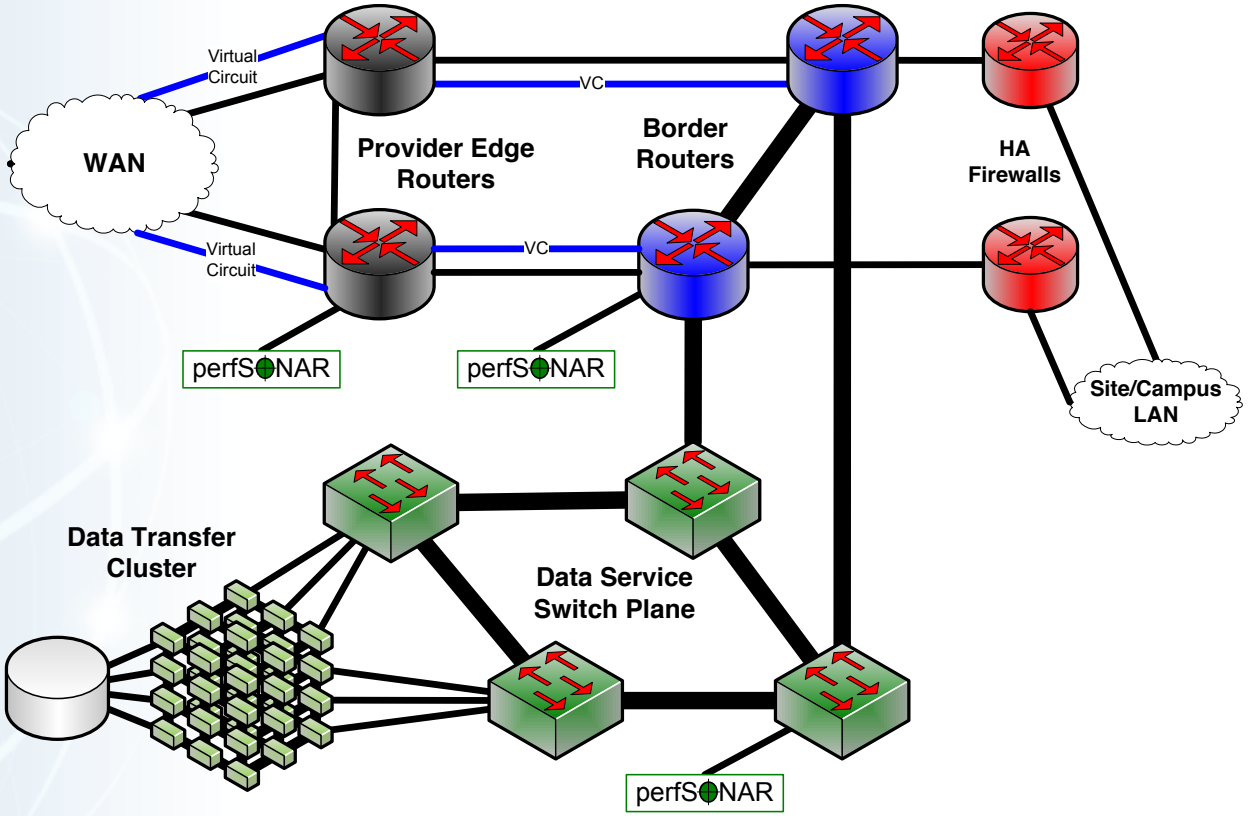
In some cases, large scale data service is the major driver

- Huge volumes of data – ingest, export
- Individual DTNs don't exist here – data transfer clusters

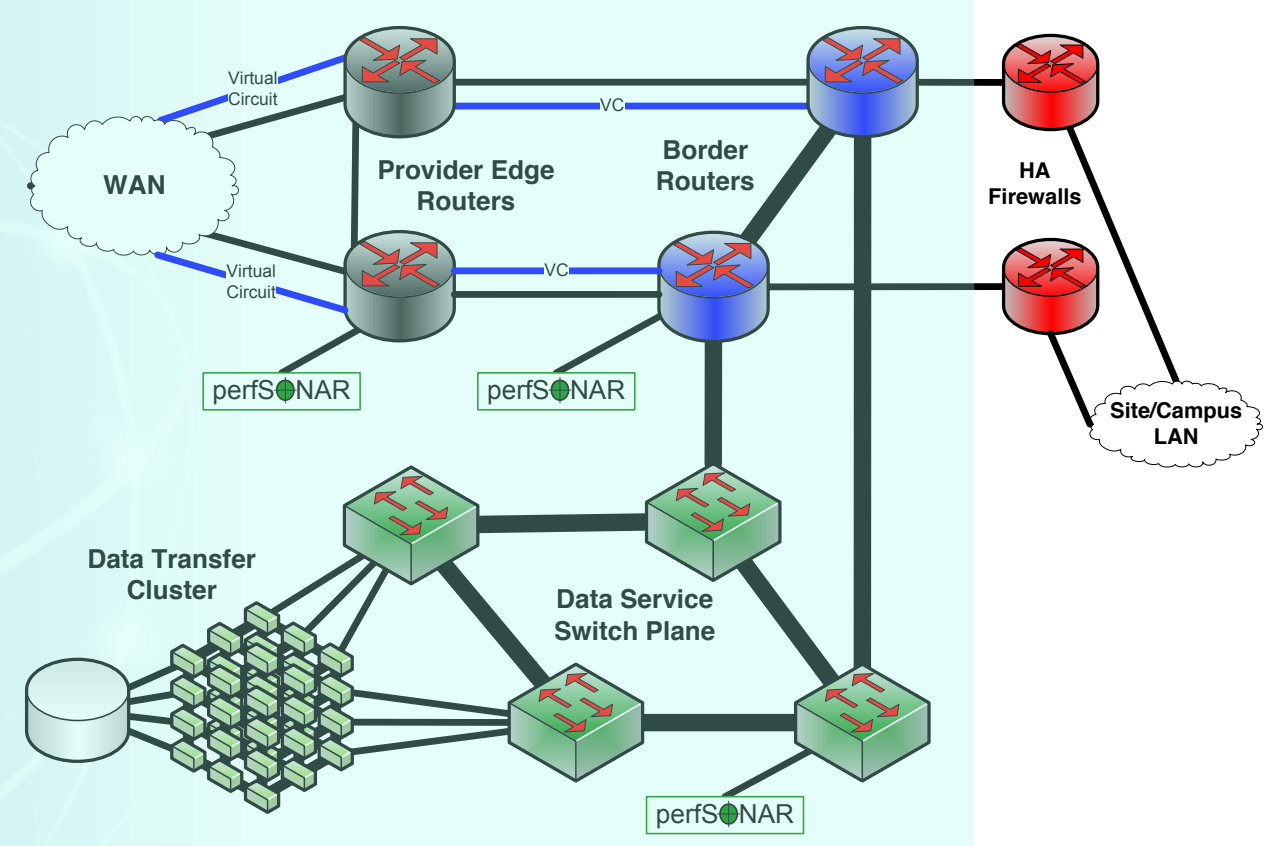
Single-pipe deployments don't work

- Everything is parallel
  - Networks (Nx10G LAGs, soon to be Nx100G)
  - Hosts – data transfer clusters, no individual DTNs
  - WAN connections – multiple entry, redundant equipment
- Choke points (e.g. firewalls) cause problems

Data Site – Architecture



Data Site – Data Path



Distributed Science DMZ

Fiber-rich environment enables distributed Science DMZ

- No need to accommodate all equipment in one location
- Allows the deployment of institutional science service

WAN services arrive at the site in the normal way

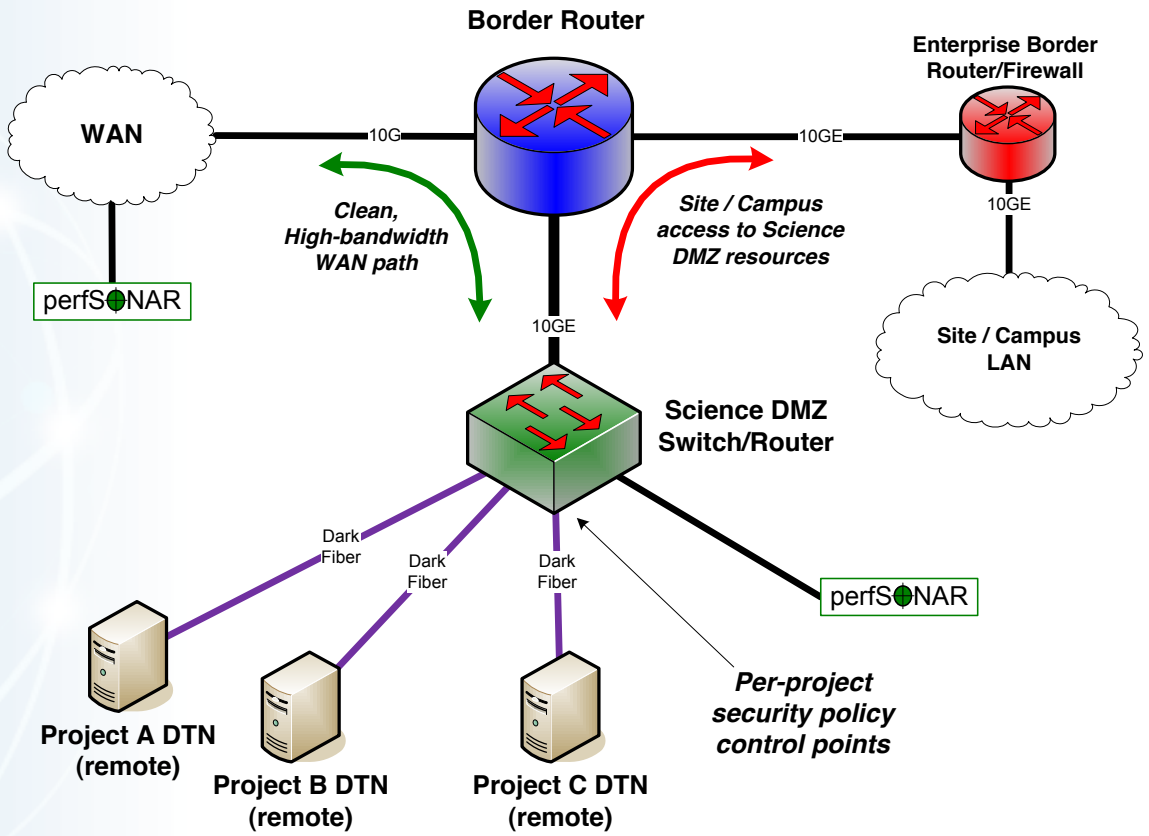
Dark fiber distributes connectivity to Science DMZ services throughout the site

- Departments with their own networking groups can manage their own local Science DMZ infrastructure
- Facilities or buildings can be served without building up the business network to support those flows

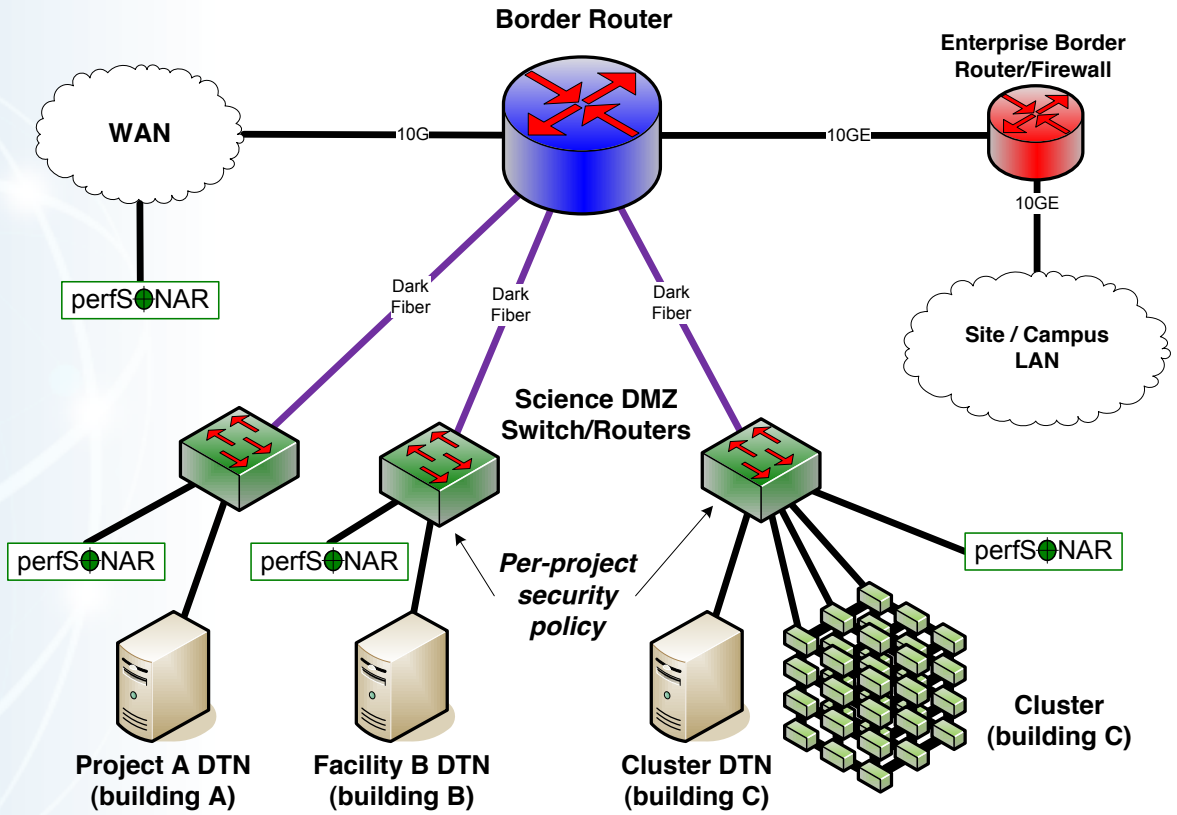
Security is made more complex

- Remote infrastructure must be monitored
- Several technical remedies exist (arpwatch, no DHCP, separate address space, etc)
- Solutions depend on relationships with security groups

Distributed Science DMZ – Dark Fiber



Multiple Science DMZs – Dark Fiber



Common Threads

Two common threads exist in all these examples

Accommodation of TCP

- Wide area portion of data transfers traverses purpose-built path
- High performance devices that don't drop packets

Ability to test and verify

- When problems arise (and they always will), they can be solved if the infrastructure is built correctly
- Small device count makes it easier to find issues
- Multiple test and measurement hosts provide multiple views of the data path
  - perfSONAR nodes at the site and in the WAN
  - perfSONAR nodes at the remote site

Science DMZ Benefits

Better access to remote facilities by local users

Local facilities provide better service to remote users

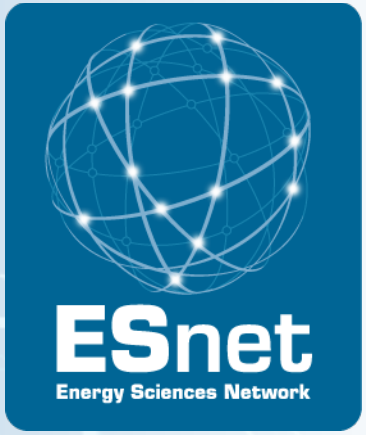
Ability to support science that might otherwise be impossible

Metcalf's Law – value increases as the square of connected devices

- Communication between institutions with functional Science DMZs is greatly facilitated
- Increased ability to collaborate in a data-intensive world

Cost/Effort benefits also

- Shorter time to fix performance problems – less staff effort
- Appropriate implementation of security policy – lower risk
- No need to drag high-speed flows across business network → lower IT infrastructure costs

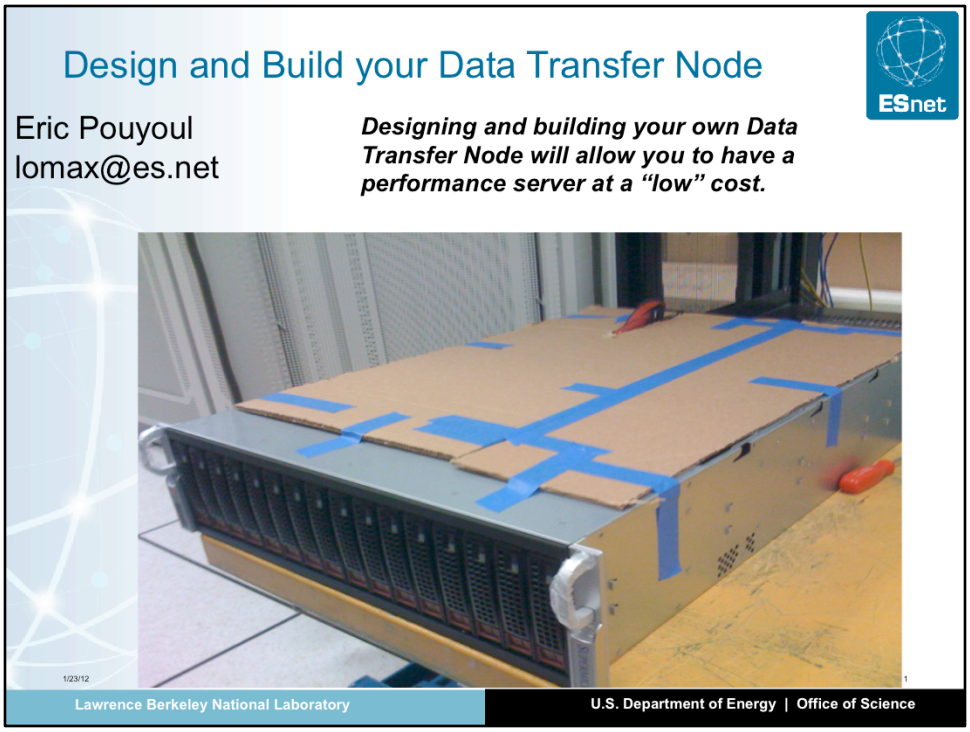


Achieving the Science DMZ

Section 2: Building and Tuning a Data Transfer Node

Eric Pouyoul, ESnet

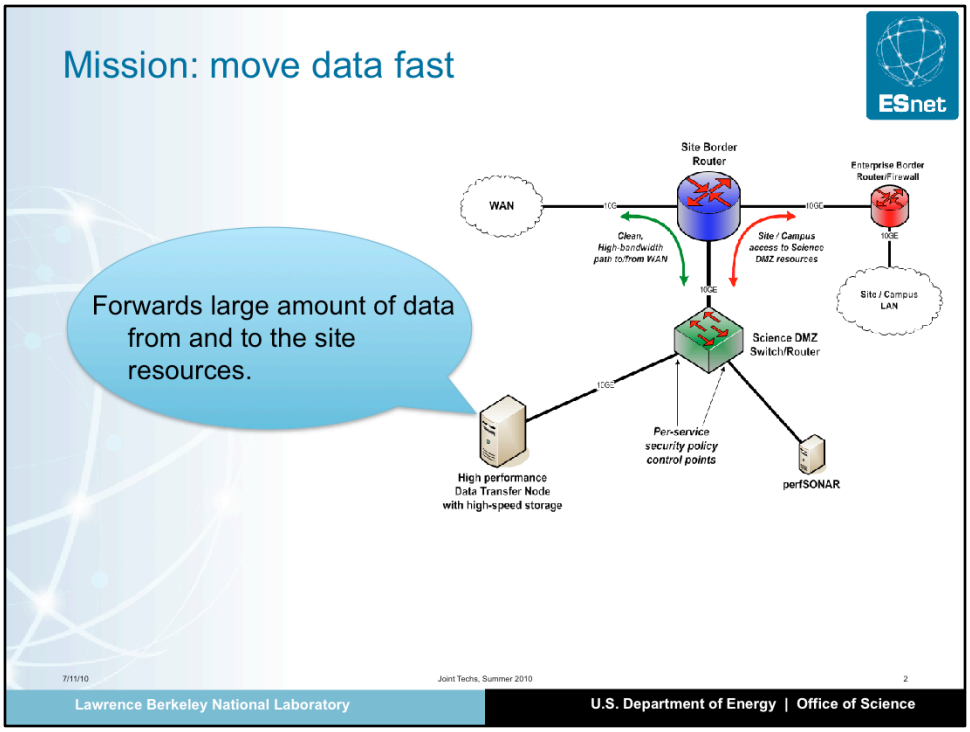
Joint Techs, Baton Rouge, LA, January, 2012



This section of the tutorial focuses on how to design and build a performance server that is dedicated to the Data Transfer function.

First we will look at the various hardware components that makes up a server, and what matters when selecting them. While we will glance over some high end, super computing, hardware, the spirit of this tutorial is have a do it yourself approach, using commodity hardware.

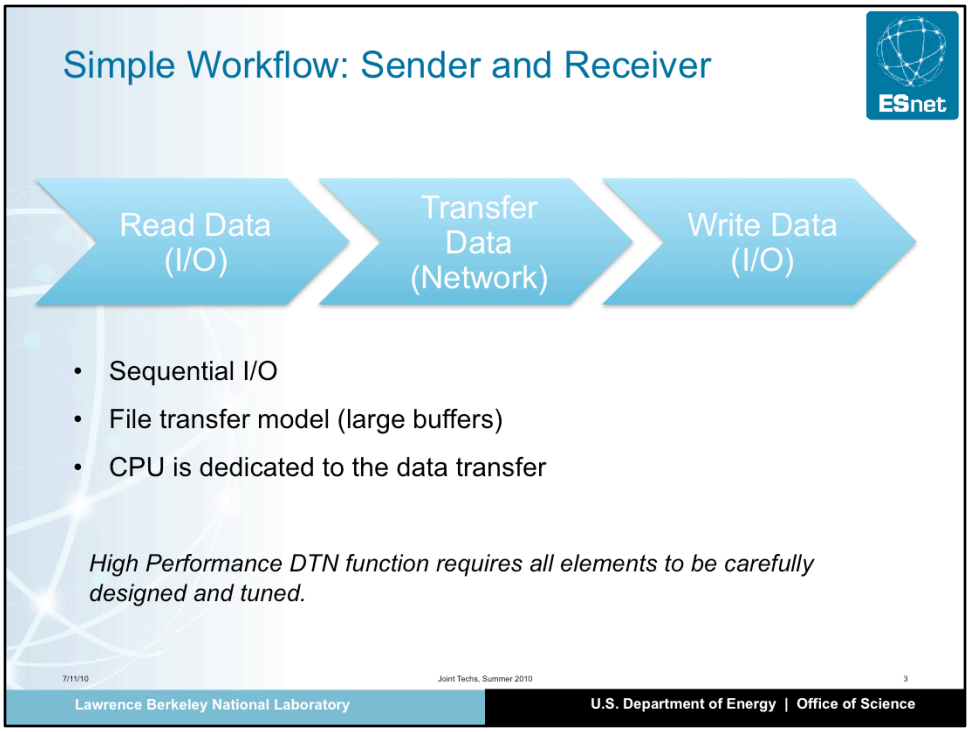
Second, we will discuss the configuration and tuning of the server, so it can perform as expected.



A Data Transfer Node has only one mission to fulfill: send large amount of data across thousands of mile as quickly as possible.

That means that the goal is to fill up the network as closely as possible as line rate.

Another consideration, when designing DTN's is deployment: because of the network topology of a Science DMZ, DTN's sometimes need to be located in racks with very little space available. Density may matter.



A Data Transfer Node is not a processing, rendering, server. Its only workflow is:

Sender host:

1) read data from the storage subsystem
2) send it to the receiver host

Receiver host:

1) read data from the network
2) write it onto the storage subsystem

We will focus only on this workflow: while the Data Transfer Node is tuned to perform at its best for this workflow, it may perform poorly for other workflows: the DTN is a dedicated host.

Hardware matters

The true sources and destinations of data are often large data centers and super computers.

A slice can/could be used to perform the DTN function.

Most efficient but difficult, but not impossible to deploy.

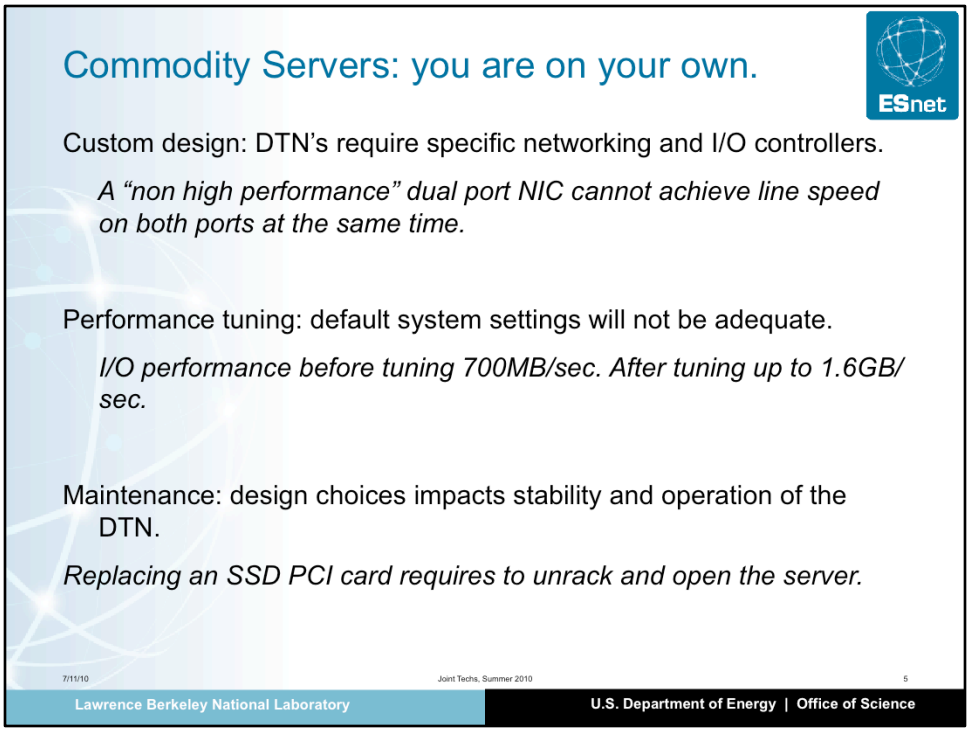
Dedicated, commodity, servers can make excellent front-end of larger systems.

7/11/10 Joint Techs, Summer 2010 4

Lawrence Berkeley National Laboratory | U.S. Department of Energy | Office of Science

The size of science data is huge. But, depending on the type of science, huge may mean terrabytes, petabytes, or more. Some Data Transfer Nodes may have to be deployed as a slive of a datacenter or supercomputer for the very large datasets. Those super DTN's are outside the scope of this tutorial: we will focus on DTN's that can scale up to a dozen of terrabyte: scaling up means adding more servers, not increasing the capacity of it.

Typically, a 6TB system, with a 20G network capability costs about $10,000.



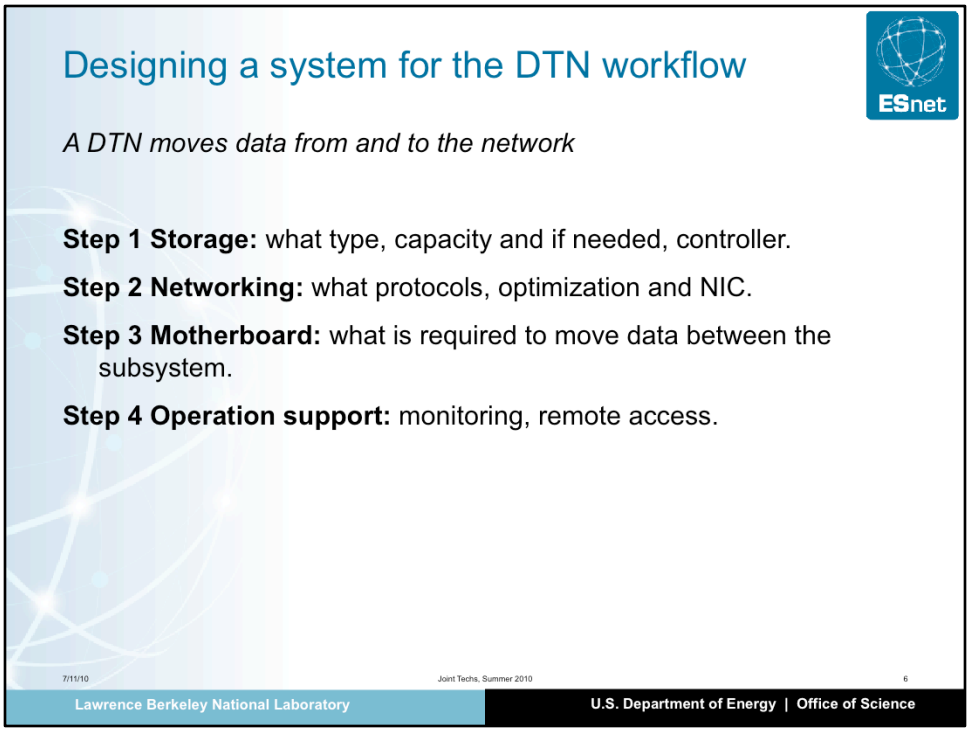
Very few vendor sale high end servers that are capable of being a DTN right out of the box, and those servers are very expensive.

A more typical way is to "design" the server, by selecting all the elements to put together (motherboard, cpu, raid controller, NIC's…).

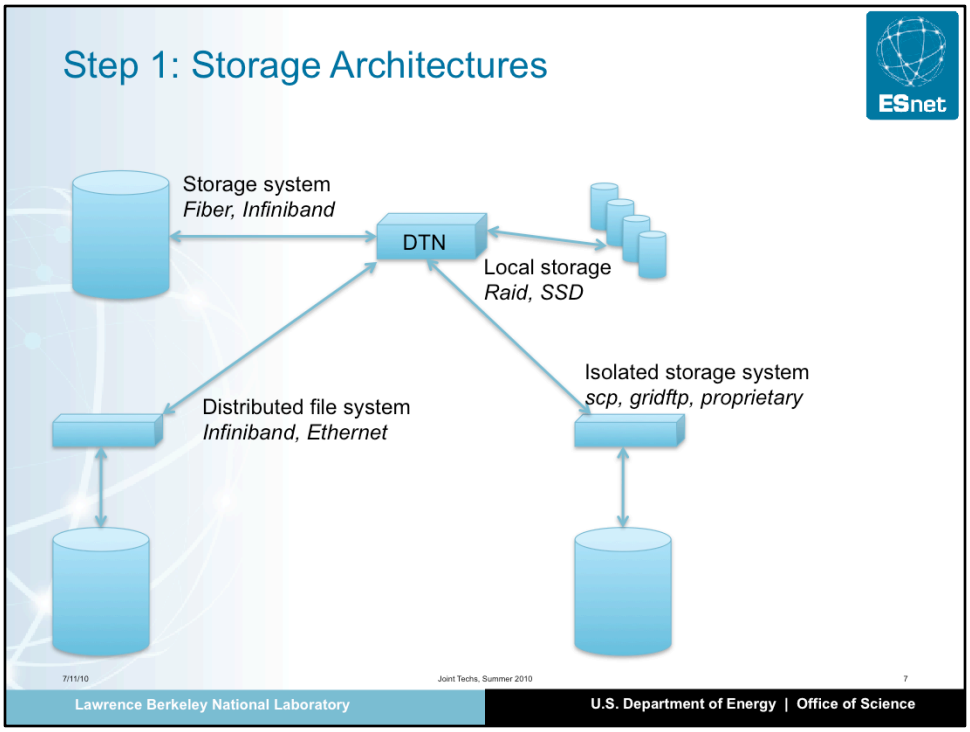
This allows to build exactly what is needed: you can get what you need for less.

However, custom design means that there is little support, especially if the system does not work as expected. Lot of time and effort will have to be spent to design the first server.

When designing a DTN, it is also important to keep in mind that it will deployed. Remote access, power, cooling and maintenance needs to be thought about early on: the server, eventually, may have to be vetted before being racked.



In a nutshell, the motherboards moves data from/to the storage to/from the network. It is critical that this can be accomplished efficiently.

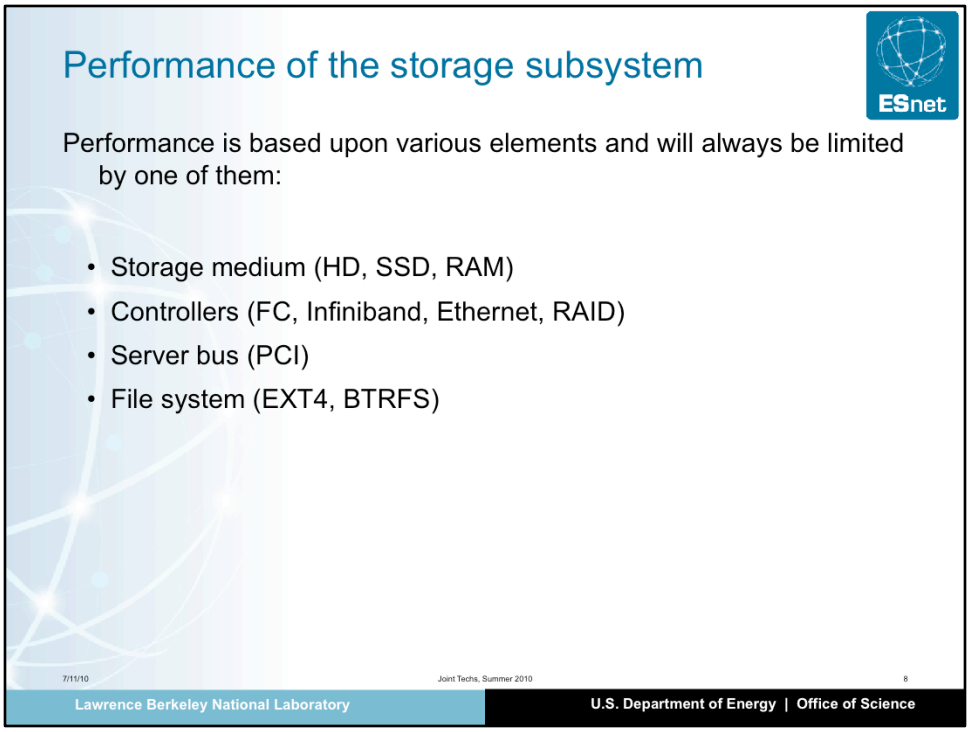


Depending on the deployment environment, a DTN may have different type of storages.

Storage Systems: Those are usually more massive systems (EMC, Netapp, DDN, etc) providing raw volumes to servers. The connectivity is usually fiber, but it can also be infiniband and even ethernet. Typically this architecture benefits storage capacity.
However, it requires, usually, a dedicated HCA and sometimes, a special software stack (OFED)

Dsitributed FileSystem: this is similar to the storage system, except that the exported volumes are not RAW but file system (PNFS, Lustre, GPFS…). This set up is typical of a tiered system: data is acquired and processed, and stored. Then the DTN read from the shared storage.

Local storage: the storage system (just a bunch of drives / JBOD) is packaged with the server. That can from 12 drivers up to 48 drives depending on vendor/ model. In addition, external chassis with more drives can be added, connected to the server with SAS or FC. This is ideal for standalone servers since it does not require plumbing for the storage subsystem. However, maintenance is typically more difficult since it does not have all the tooling that usually comes with storage systems.



Performance of a storage subsystem varies depending on its type and architecture.

Type:
Hard drives are cheapwith high capacity. However, their performance is low. Good drives, on average, can do 130MB/sec read or write.
SSD are expensive and have low capacity but they are fast.

Architecture
RAID controller (disk controllers) can be a bottleneck. Some controllers are optimized.

File System: using a file system typically introduces an overhead in the I/O performance. Bad file system such ext3 may use up to 40% overhead. Good file systems (EXT4, BTRFS, ZFS) can almost reach bare metal performance.

Storage Systems

Pros

- May already be deployed (legacy)
- Can scale up
- Highly Available

Cons

- Controller Bottleneck
- Expensive
- Large footprint

7/11/10

Networked Storages

Ethernet (iSCSI) or Infiniband (SRP) provide networked access to raw storage:

- Flexible storage scalability
- Virtualization
- Especially with SRP, high controller performance

- Requires specific controller
- Requires specific software stack

Local storage: RAID

Capacity of hard drive has increased

Form factor has decreased

Reliability has increased

Hardware RAID is efficient

Inexpensive

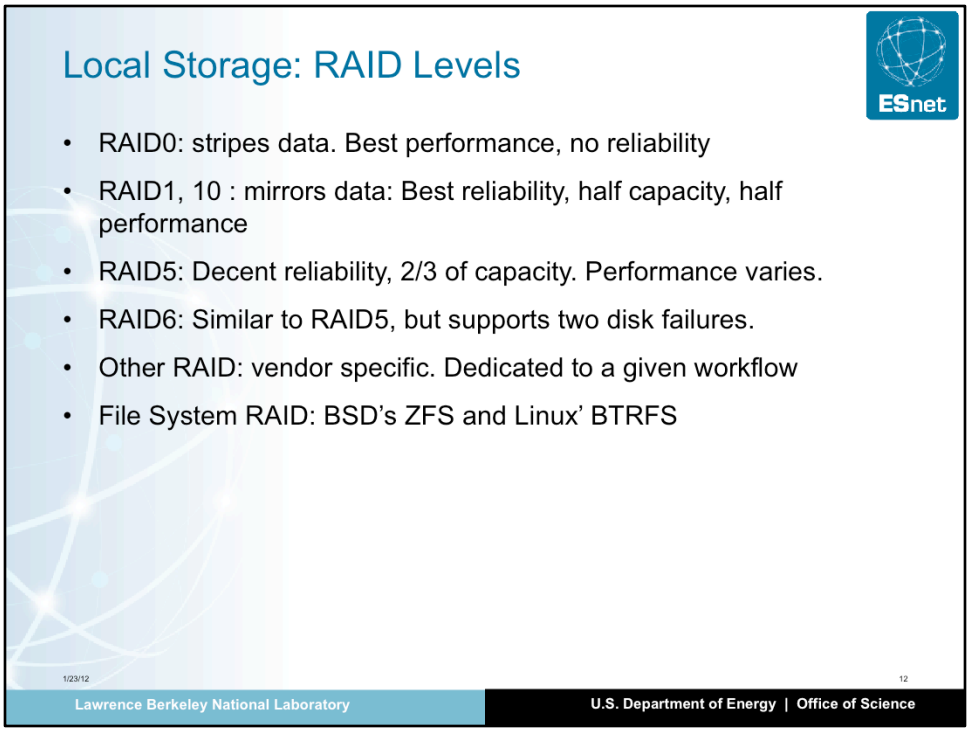
Local RAID storage is ideal for custom design DTN

7/11/10 Joint Techs, Summer 2010 11

Lawrence Berkeley National Laboratory | U.S. Department of Energy | Office of Science

RAID controllers are capable of high performance while offloading the CPU with the disk operations.

Of course the choice of controller matters. Look at reviews and experiment with loaners when possible.



RAID0 is the right choice when try to get the maximum storage performance at the lowest cost (the lowest number of drives). A single drive failure will cause the entire volume to be lost.

RAID10 is the best choice for reliability (a single disk failure is fully recoverable) but is twice as expensive (it needs twice as many disks as RAID0)

RAID 5,6 and other specialty levels: those levels are compromises between performance, reliability and cost. Often, those are the right choices but quality and power of the RAID controller impacts more performance. In other words, a decent but not exceptional RAID controller may perform very well in RAID0 and poorly in RAID5. Performance RAID5.6 do exist, however, but are typically in the $2,000 price range while good RAID controller (good at RAID0) typically cost less than $1,000.

Note that some file systems, namely ZFS and BTRFS implements RAID in software and are good at it. If the server is powerful (enough cores, at least 16), those file system may perform better than a RAID controller. But they will use much more CPU on the server.

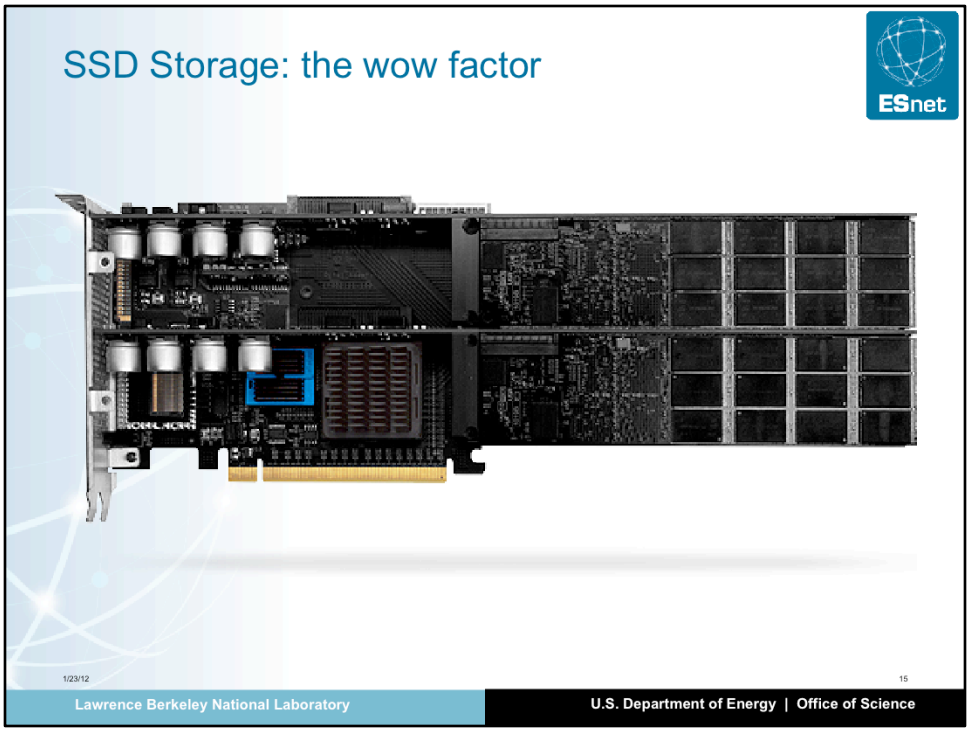
RAID: Great (cheap) but Experiment First

- RAID is a bottleneck !
- Performance depends on RAID engine

- Select the right RAID Level
  - RAID0: need best I/O performance but can afford losing all dataset.
  - RAID5/6: need reliability, can afford to only have 2/3 of capacity and performance of RAID0.
- Select the right RAID Controller
- Plan for expansion
- Experiment on prototype first.

RAID Controllers

- Often optimized for a given workload, rarely for performance.
- RAID0 requires less CPU than other RAID levels.
- The CPU required to process queries is a factor of the number of drives.
- Each controller has its own best configuration forcing to make compromises.



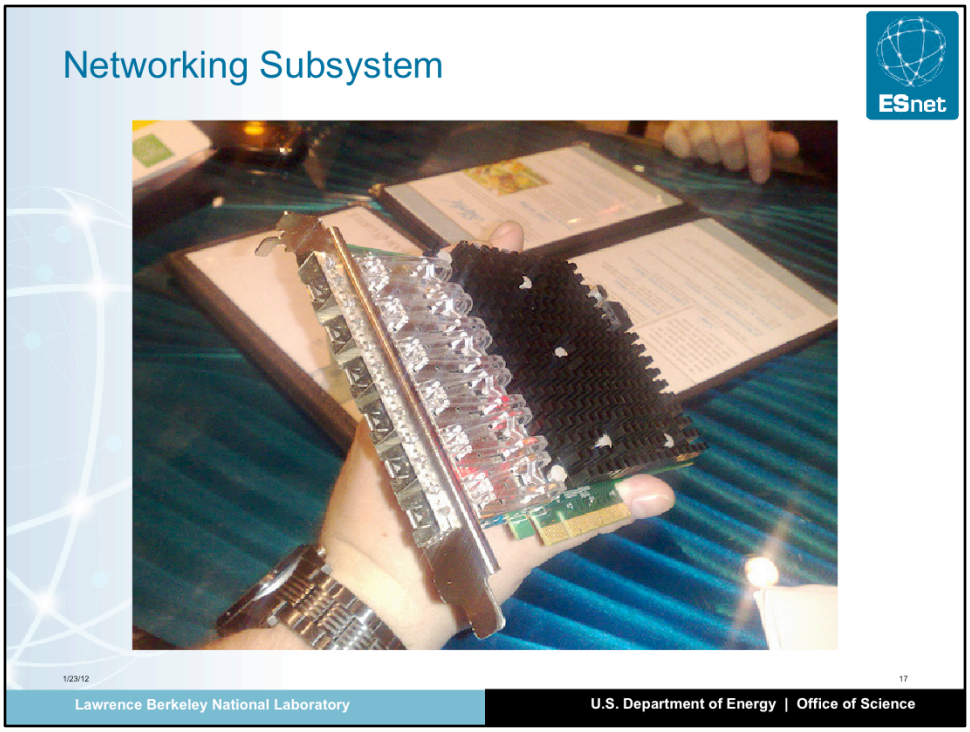
SSD, cost much more than HD, but a much faster. They come in different packages:

-PCIe card: some vendors (Fusion I/O) build PCI cards with SSD. The current maximum capacity is 1TB per card. Since those cards are PCIe, the data transfer between the main memory and the SSD is just limited by the SSD speed and the PCIe speed: in other words, it is really fast (several GB/sec per card). The drawbacks are that 1TB uses a PCIe slot. This design often means that a PCI extender is needed, but if space and performance is an issue, this is the best solution. Those SSD cards can also be an deployment issue: replacing a failed card means that the server must be open.

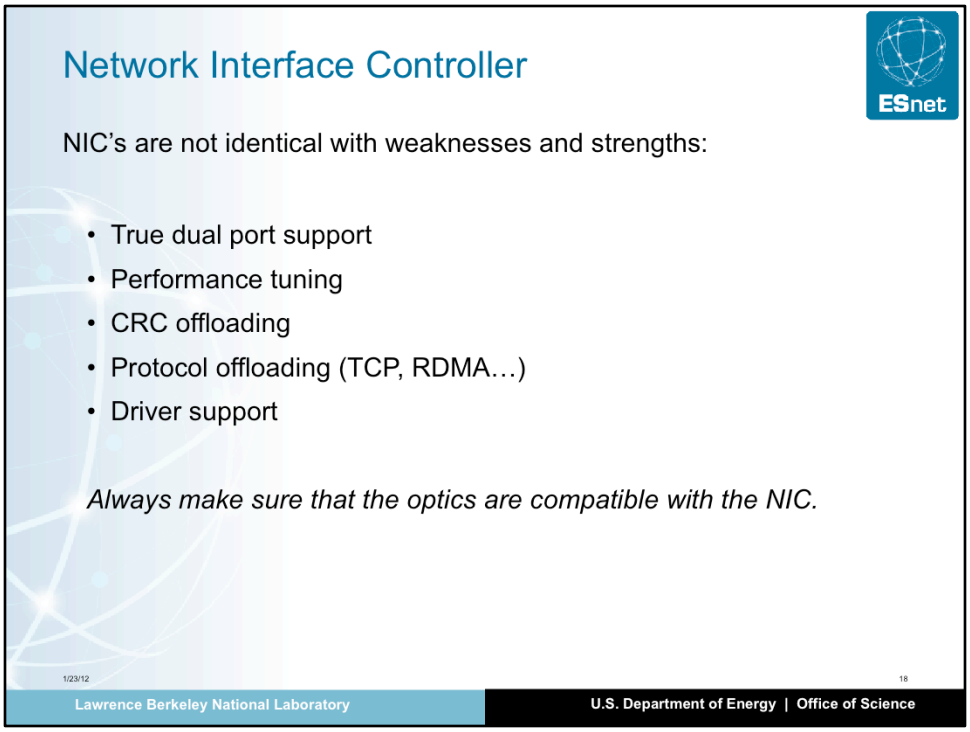
-- HD replacement: some vendors (IBM, WD, etc) have product that a physical replacement for HD: same form factor, same connectivity (SAS, SATA…). Thi allows for easier migration path from HD to SSD, but the performance is limited by the controller. Also, not all controllers are good at controlling SSD drives: always make sure that the controller is “SSD capable”.

SSD: Current State

- 6 GB/sec read (PCIe 2.0 x16) ! More to expect with PCIe 3.0.
- Acceptable/Excellent MTBF
- Still more expensive
- Potentially harder to deploy within standard IT
- Migration path / Hybrid – Needs high end RAID Controllers



The networking subsystem is the second subsystem after the storage that is critical and will be a bottle neck. The choice of NIC will impact performance.



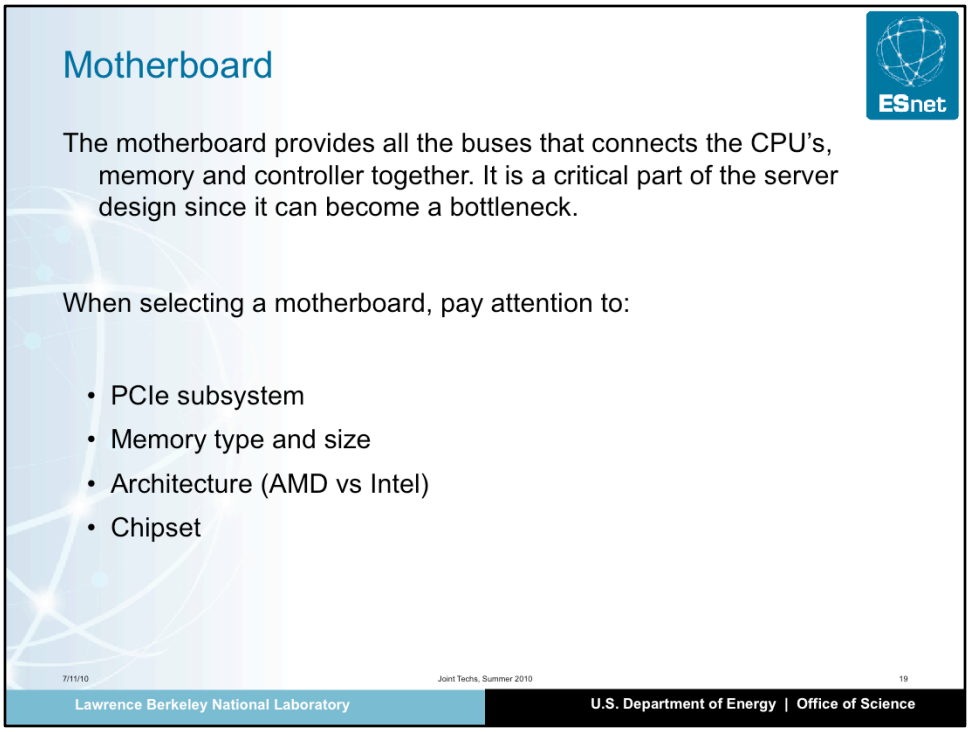
NIC vendors seems to specialize in a given market:

Myricom: some of best performance per port, simple controller. Limited support for exotic protocols.

Chelsio: very large support for protocols (iwarp), and protocol offloading.

Intel: robust driver, supports some third party

Mellanox: somewhat a new player in Ethernet. Converges Infiniband and Ethernet. Excellent support for OFED. Support Layer 2 RDMA (RoCE)



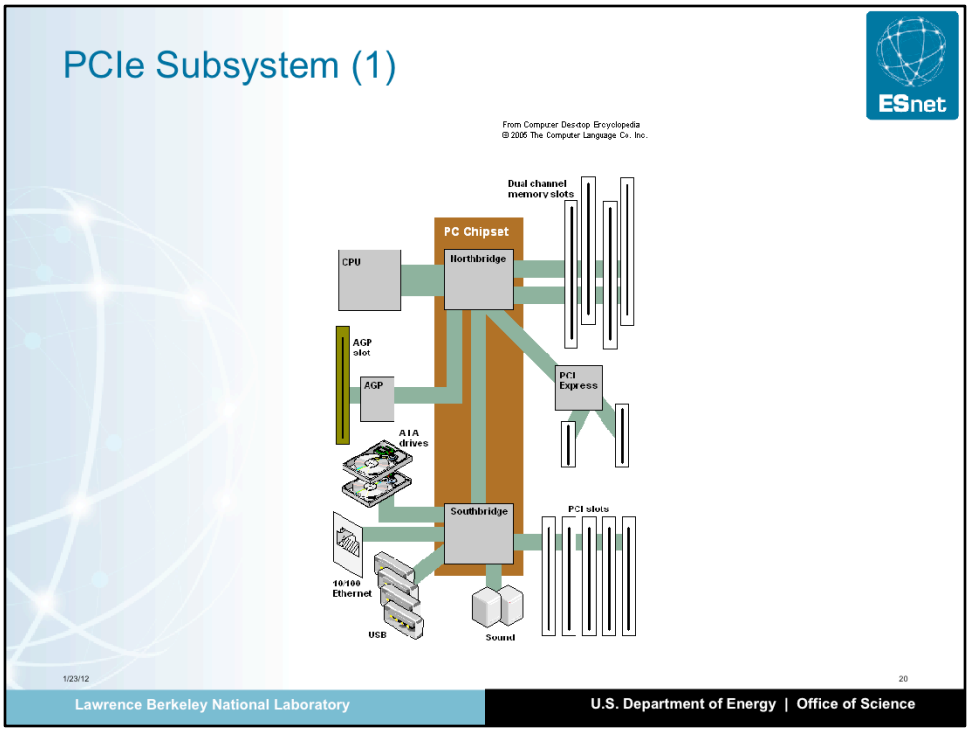
The motherboard not only incorporates the CPU's and memory, it is also providing all the busses between the various component of the server. An inadequate motherboard can become a major bottleneck. It is, then, very important to correctly select it. The questions to ask while selecting are:

How many PCI cards will I need ? How many lanes each ?

What is the aggregate throughput I need on my PCI cards ?

How many cores do I need ? At what speed ?

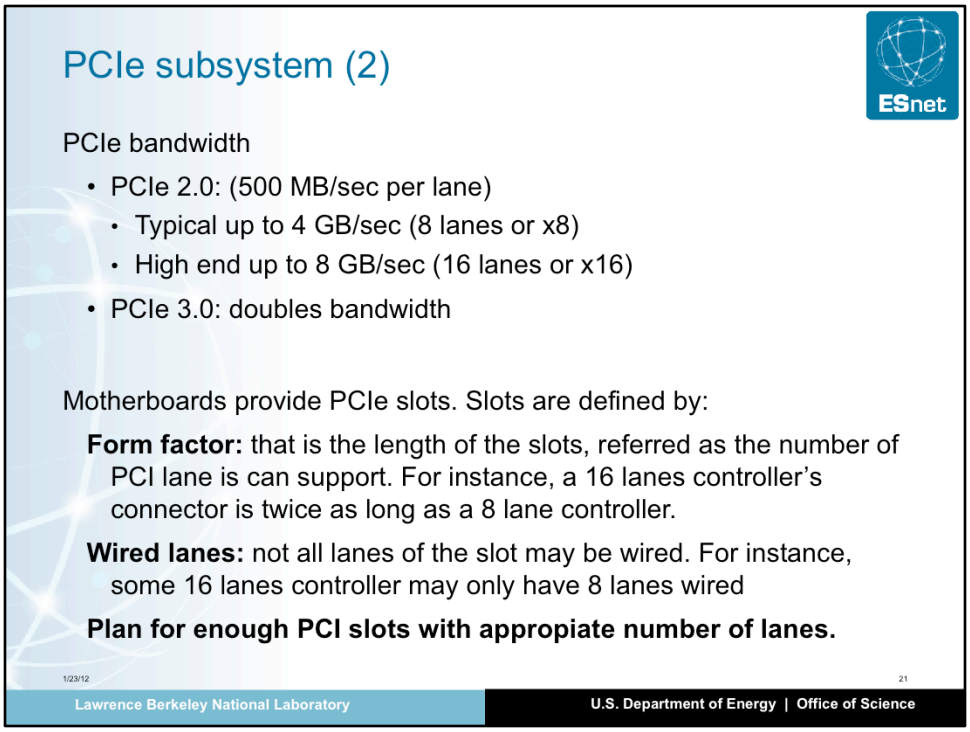
What kind of remote access (maintenance) do I need.



The Chipset is the component in the server that handles all the I/O. In other words, it is responsible for moving data from the PCI cards and the CPU.

Some chipset are better than others (read review), but the most important part of the chipset is the maximum number of PCI lanes it can handle.

Also, depending on how the chipset and the PCI bus is wired, the architecture may or may not fit your needs. It is then important to look at the schematic of the motherboard to see if the I/O subsystem can provide the required performance.



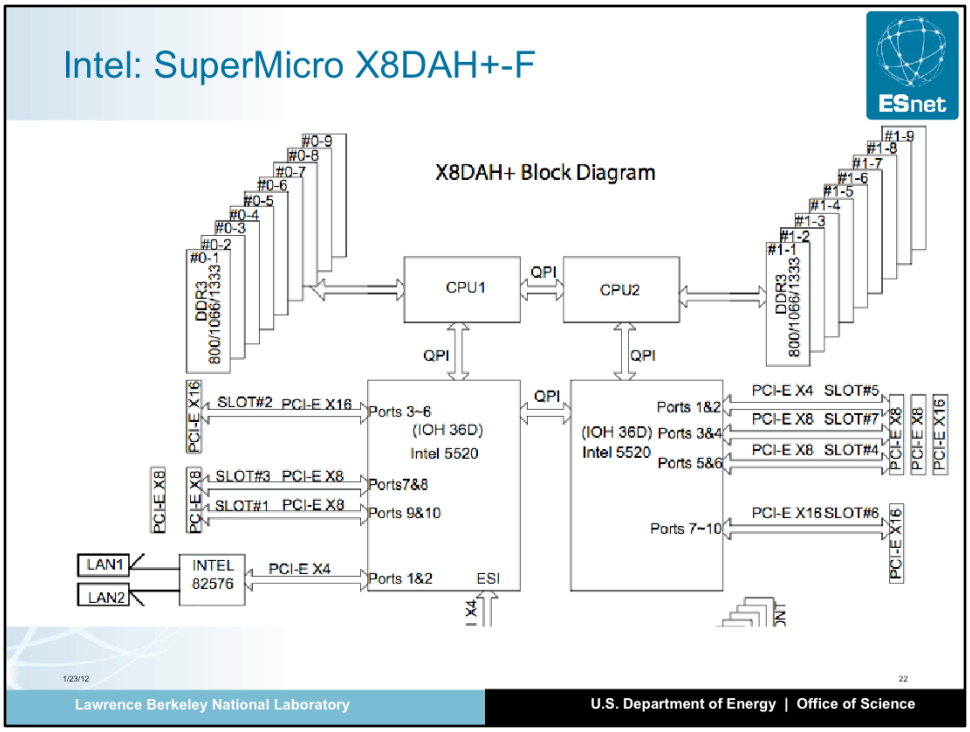
In order for the overall system to performance to the specifications, it is critical to set the card into the appropriate slot:

The slot must have wired the correct number of lanes. With a PCIe Gen2 system, most cards are x8 (8 lanes). Some cards such as a 6 x 10GE port or a SSD Fusion I/O card are 16 lanes. Be careful when selecting a PCI slot:

some motherboards have slot that look like x8 or x16, but a fewer number of lanes are really wired. Typically the board will say something like "x4 in a x8 slot"

If you are running out of slots, there are products that adds an external chassis with just an array of PCI slots. Those are named "PCI extender"

PCIe Gen3 is coming ! This will multiply by 2 the PCI throughput. While this is very exiting (DTN do need PCIe Gen3), wait until the second generation of Gen3 motherboards come out: you do not want to hit bios bugs or hardware bugs. But again, Gen3 PCI is much needed considering the data size and the modern WAN capability (100G fiber)

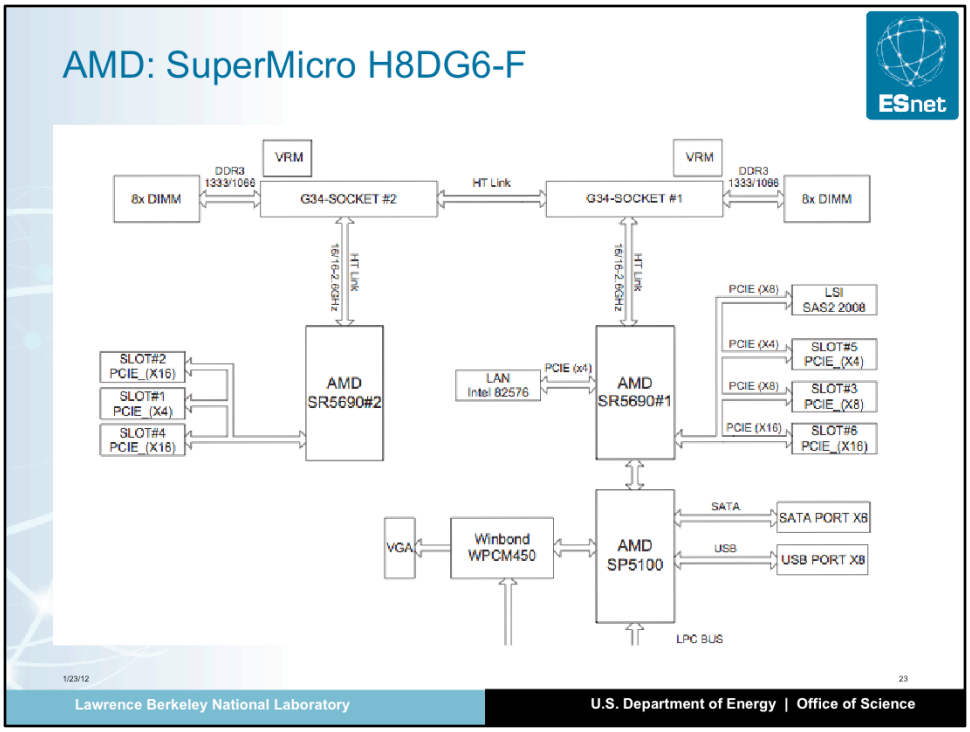


Notice the two independent I/O path:

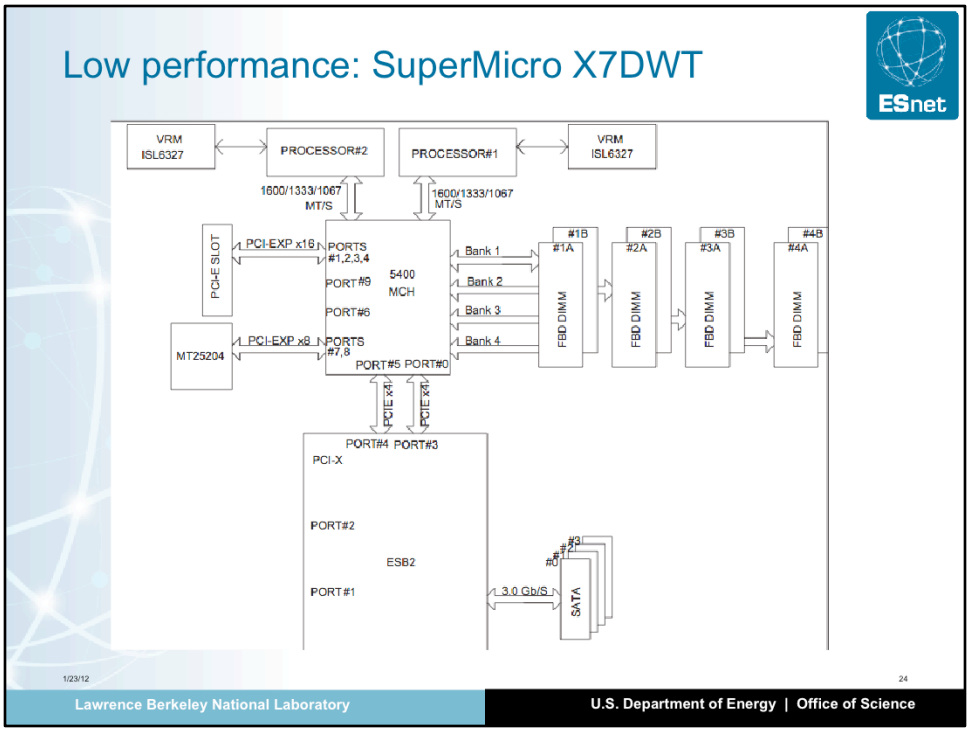
Memory <-> CPU <-> Chipset <-> PCI card

This architecture is good because it allows two split the I/O and networking cleaning without congestion point.

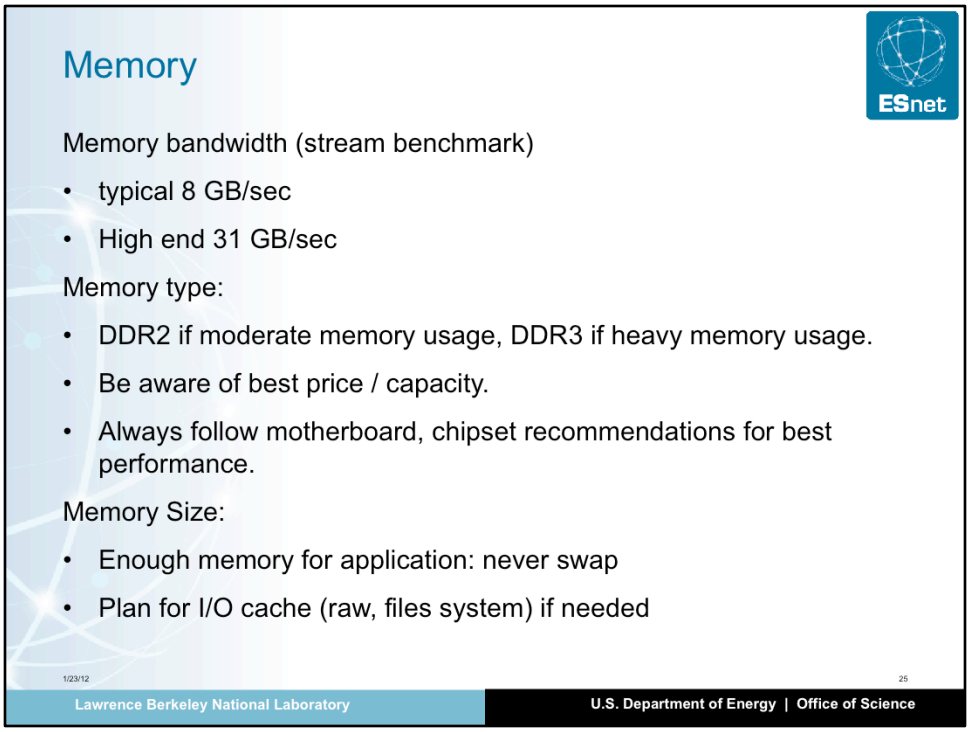
Note the number of lanes of each of the PCI slots



Similar architecture than the previous board but for Intel



This is not a bad motherboard, just not designed for performance. It has only one chipset (but still two processors). It also has a single memory bank.

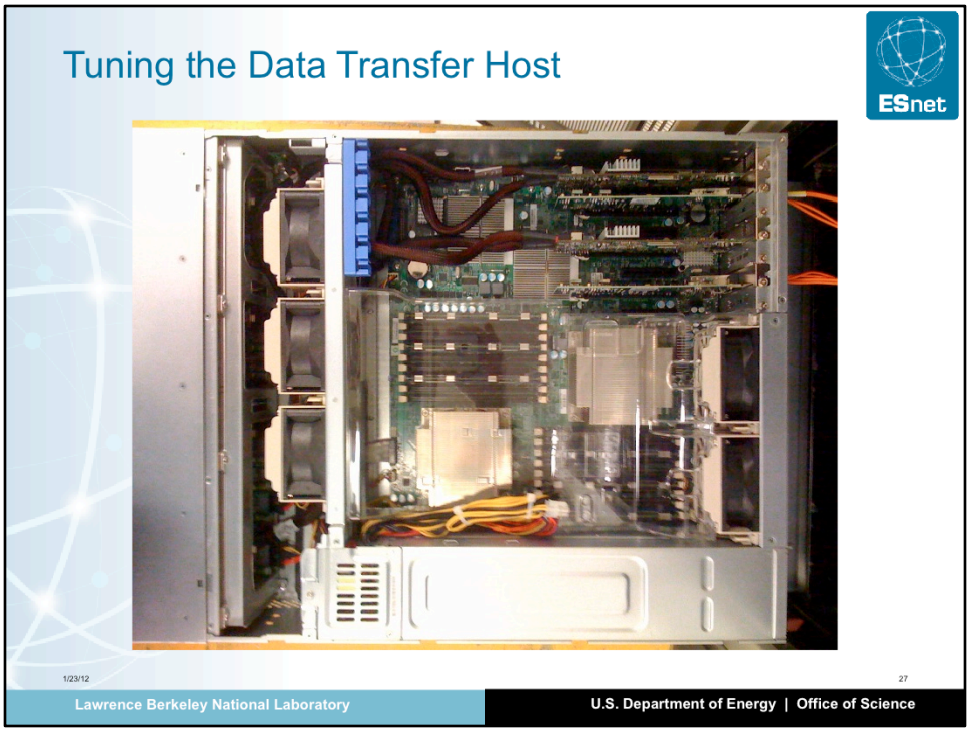


Remember that memory is used for several functions:
1) Application
2) I/O Write/Read cache (the more memory for the cache, the better the system will handle performance spikes. A good DTN would typically have 10G of write cache
3) Network buffers.

AMD or Intel ?

- Currently, Intel has a faster bus (QPI) than AMD's HT's
- Faster clock on Intel
- More cores on AMD
- Memory can be cheaper on AMD (AMD support DDR2)
- AMD typically supports architecture much longer than Intel (backward compatibility).

AMD and Intel alternates as the leader in performance computing (look at manufacturing problems, etc)



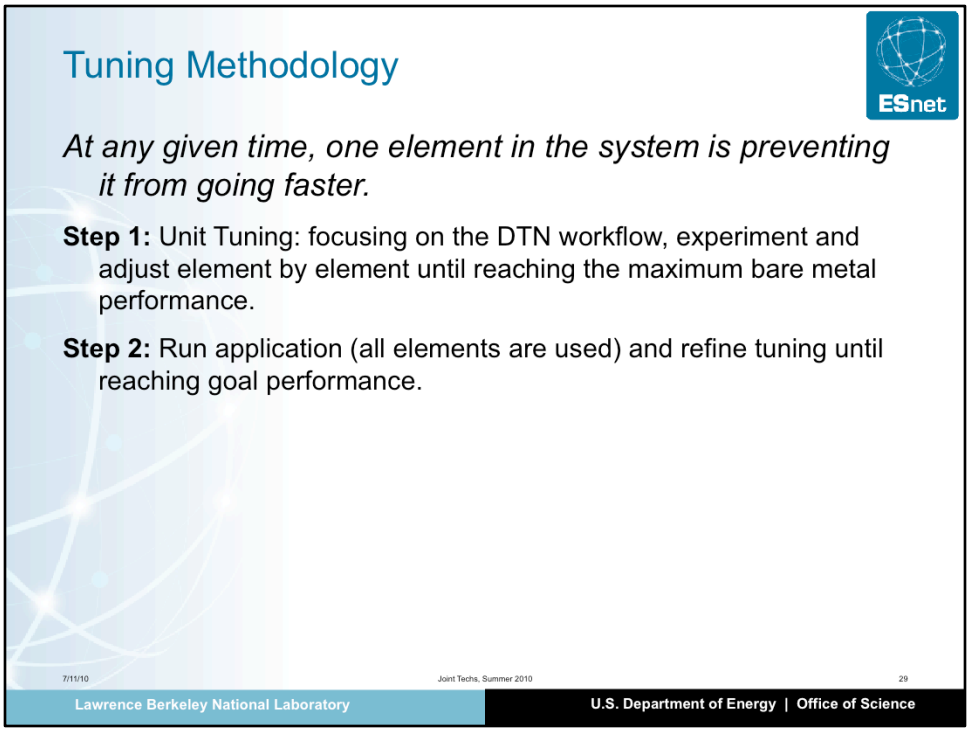
At this point the DTN is designed and assembled to your specification. The next step is to configure, tune the entire system, so it performs as expected. If the DTN is correctly designed, in other words, the hardware is capable of delivering the required performance, with patience and methodology,

Tuning

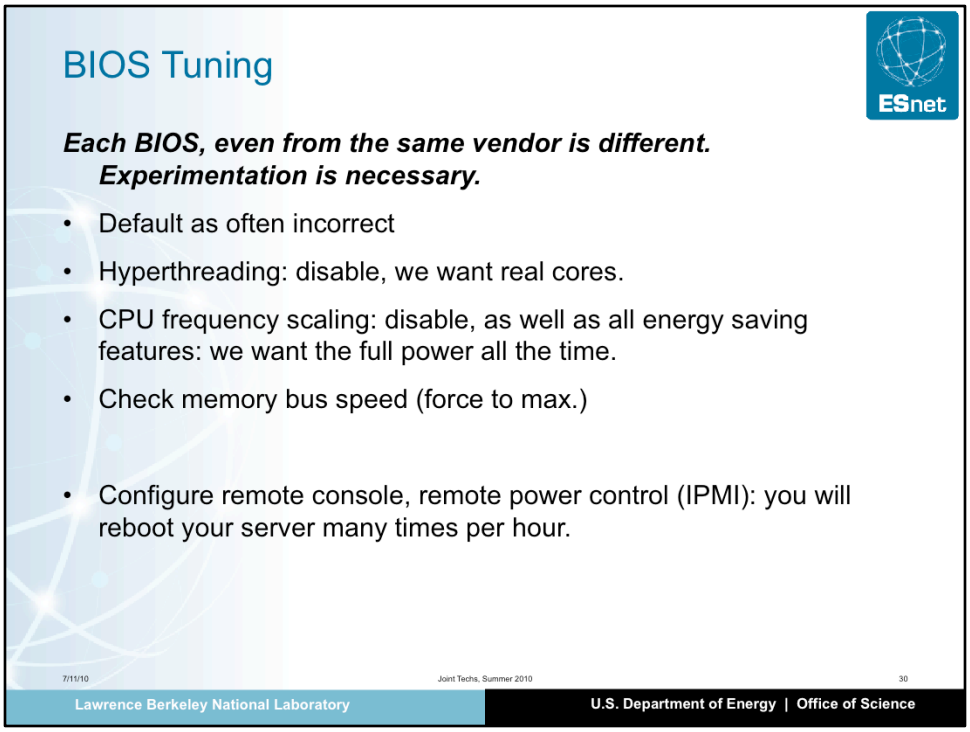
Defaults are usually not appropriate for performance.

What needs to be tuned:

- BIOS
- Firmware
- Device Drivers
- Networking
- File System
- Application

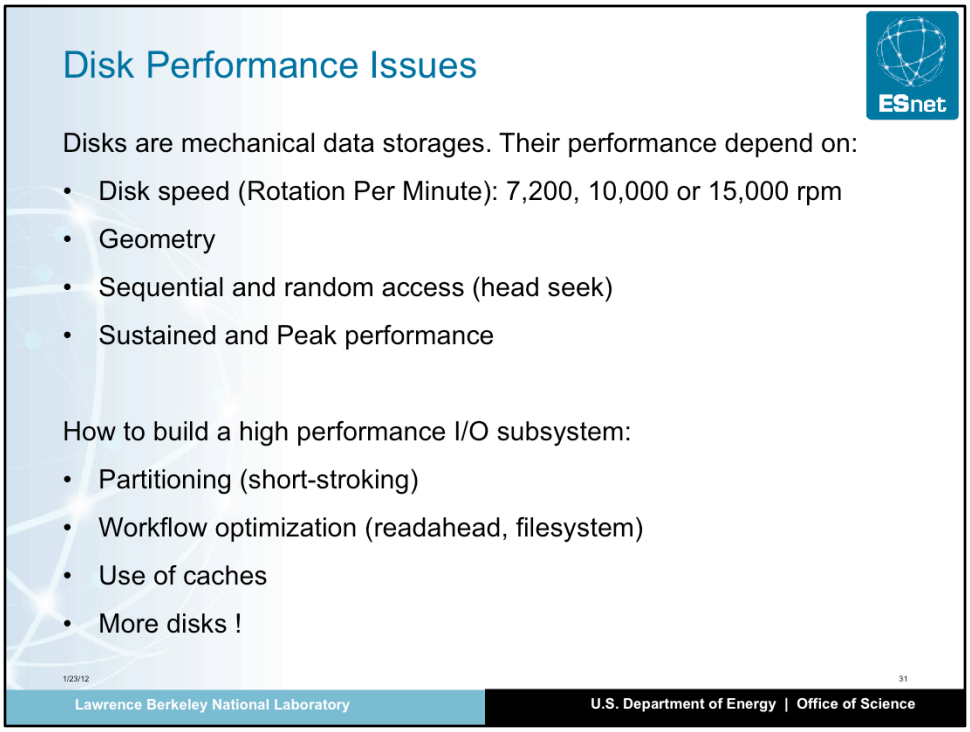


A lot of time can be spent tuning a system and it is easy to not make progress. It helps to use a methodology which is based working on one element of the system at the time, gathering and recording measurements.



BIOS tuning can be painful. A wrong setting can have dramatic effect on performance, but also on stability of the system. The goal is to make the behavior of the hardware as predictable as possible and to run it at maximum performance.

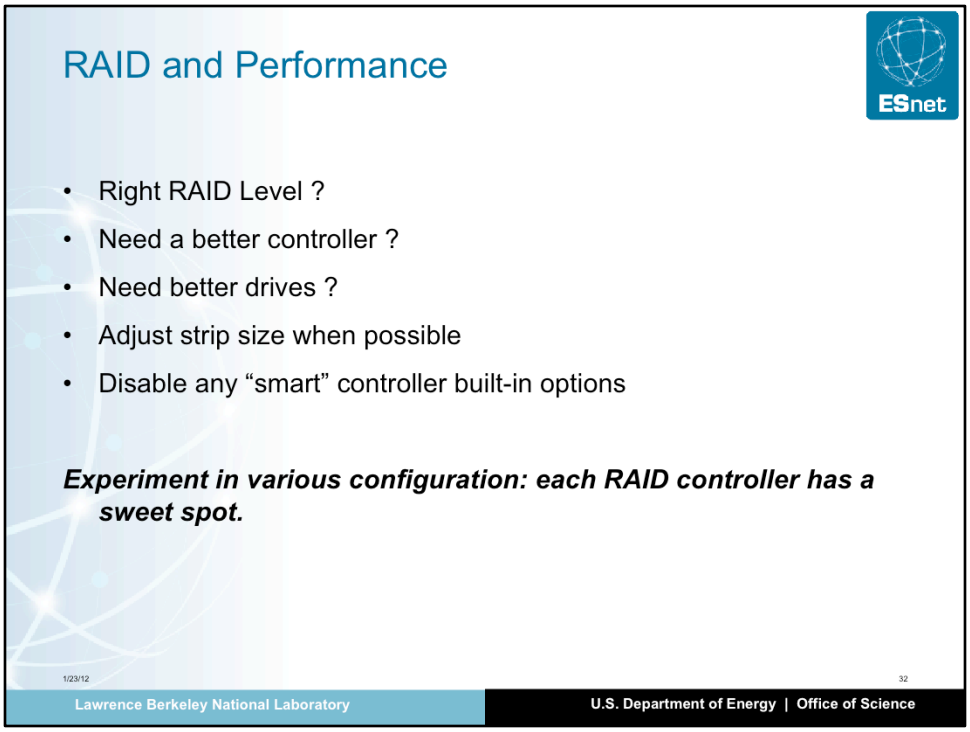
Before changing a BIOS setting, always note what it the current state: you may need to return to a previous state of the BIOS if you make an error.



Designing a RAID system is almost an art: there are so many constraints that while it is almost always possible to optimize the storage subsystem, it is almost always impossible to get what you really want. When working on the storage subsystem, ask yourself the following questions:

- how many files do I need to send or receive at the same time
- based on the maximum network throughput, how fast a file must be read or written ?
- how large is the average file ?
- how reliable the storage must really be ?
- do the files compress well ?
- how will you answer to those same questions in one year, two yers, four years ?

Fortunately optimizing storage is perhaps one of the performance work that is the most publically documented (blogs, storage vendors…). A rule of thumbs is when using hard drives, you should get at least about 130MB/sec per disk..



Disk controllers, RAID or not, are usually designed for “entreprise”. This usually means that the controller is often configured with RAID 5 or 6. As a consequence, controller are most of the times, not capable of running all the drives at full speed: in entreprise context, there are always a few drives that are hot replacement. A rule of thumb is that if a controller is said to handle up to X drives, it can handle up to 2X/3 drivers at full speed.

Some high end controller (Areca for instance) are specifically designed for a workflow similar to a DTN workflow: sequential read/write. They may have Gigabytes of SRAM for internal buffering, PCIe x16..

Finally, some RAID controller are specifically designed to scale up. In addition to wire internal drives, they can control external drives, directly or in a daisy chain manner

Tool: vmstat

From man page:

reports information about processes, memory, paging, block IO, traps, and cpu activity.

- Shown true I/O operation
- Shows CPU bottlenecks
- Shows memory usage
- Shows locks

```
$ vmstat 1
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu------
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0  0      0 22751248 192800 1017000    0    0     0     0    4    7  0  0 100  0  0
```

I/O testing tool: dd

From man page: *"convert and copy a file"*

- Generate I/O traffic
- Control over block size, seek
- Input and output agnostic (raw or file)
- Can be used in parallel

```
$ dd if=/storage/data1/test-file1 of=/dev/null bs=4k
13631488+0 records in
13631488+0 records out
55834574848 bytes (56 GB) copied, 54.1224 seconds, 1.0 GB/s
```

34

Example of a “dd test”


```
# dd of=/dev/null if=/storage/data1/test-file1 bs=4k &
# dd of=/dev/null if=/storage/data1/test-file2 bs=4k &

# dd of=/dev/null if=/storage/data2/test-file1 bs=4k &
# dd of=/dev/null if=/storage/data2/test-file2 bs=4k &

# dd of=/dev/null if=/storage/data3/test-file1 bs=4k &
# dd of=/dev/null if=/storage/data3/test-file2 bs=4k &
```

Example vmstat / dd


```
# vmstat 1
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu------
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 6  0      0 150132 215204 23428260    0    0     0     0 16431 2245  0 13 86  0  0
 2  3      0 1692948 218924 21920000    0    0  4428 499712 24599 5341  1 29 65  6  0
 2  5      0 1610216 222512 22001264    0    0  3532 725012 25230 5363  0 15 75 10  0
 4  5      0 720020 224532 22865412    0    0  2048 847296 24566 4277  0 13 65 22  0
 3  7      0 1917556 225440 21686980    0    0  1672 1099036 27333 4314  0 17 60 23  0
 6  7      0 1419324 225496 22180252    0    0     0 1312704 29410 25386  0 24 45 31  0
 3  6      0 391860 225560 23182336    0    0     4 1261536 25797 27532  0 20 48 32  0
 8  4      0  80624 224672 23486864    0    0     0 1296932 26799 3373  0 22 52 26  0
 3  6      0  88860 224184 23475516    0    0     0 1322248 28338 3529  0 22 51 27  0
```

I/O Testing Tips

- Two windows, one with dd, one with vmstat
- Influence of the read and write caches
- Flush caches before running tests:

  # echo 3 > /proc/sys/vm/drop_caches
- Discussion on data size: three times the memory size
- Influences of the block size: use block size that matches application's pattern
- Remote Console (IPMI)

Linux I/O Scheduler

- I/O scheduler: different policies. Default policy is “fair” meaning bad for performance. Typically deadline scheduler is better for performance, but favors the most I/O hungry application.

  In /boot/grub/grub.conf:

  title CentOS (2.6.35.7)

  root (hd0,0)

  kernel /vmlinuz-2.6.35.7 ro root=/dev/VolGroup00/LogVol00 rhgb quiet elevator=deadline

  initrd /initrd-2.6.35.7.img

I/O Tuning: readahead

- Necessary optimization when workload is mostly sequential read
- Needs to be experimented with
- Does not always play nice with hardware optimization (but is often better than hardware optimization)
- Needs to be done at each boot of the server (add configuration in /etc/rc.local)
- Interesting reading http://www.kernel.org/doc/ols/2004/ols2004v2-pages-105-116.pdf

/sbin/blockdev --setra 262144 /dev/sdb

/sbin/blockdev --setra 262144 /dev/sdc

/sbin/blockdev --setra 262144 /dev/sdd

File Systems Performance

- Very few file systems are designed for high performance
- EXT4 is currently the fastest production file system for Linux.
- ZFS provides “smart” software RAID and compression on Solaris
- BTRFS: bleeding edge, integrates RAID and compression on Linux
- File systems must be tuned for performance
- Compromise performance versus data reliability: be careful for what you ask for !

File System Optimization (1)

- File System independent optimization (in /etc/fstab)

```
/dev/sdb1      /storage/data1      ext4  noatime,nodiratime 0 0
```

- File System specific optimization (EXT4)

```
/dev/sdb1   /storage/data1     ext4
      inode_readahead_blks=64,data=writeback,nobh,barrier=0,commit=300,noatime,nodiratime 0 0
```

Inode_readahead: useful when directories have lots of files

Data=writeback: metadata is written onto the disk in a “lazy” mode

barrier=0: does no longer enforce journal write ordering.

File System Optimization (2)

- Necessary in order to get performance close to bare metal
- Be careful what you ask for: some of the optimization may render the file system less reliable in case of crashes

NIC Tuning

As the bandwidth of a NIC goes up (10G, 40G), it becomes critical to fine tune the NIC.

- Lot of interrupts per second (IRQ)
- Network protocols requires to process data (CRC) and copy data from/to application's space.
- NIC may interfere with other components such as the RAID controller.

Do not forget to tune TCP and other network parameters as described previously.

Handling flood of frames: Interrupt Affinity

- Interrupts are triggered by I/O cards (storage, network). High performance means lot of interrupts per seconds
- Interrupt handlers are executed on a core
- By default, core 0 gets all the interrupt, or, interrupt are dispatched in a round-robin fashion among the core: both is bad for performance:
  - Core 0 get all interrupt: with very fast I/O, the core is overwhelmed and becomes a bottleneck
  - Round-robin dispatch: very likely, the core that executes the interrupt handler will not have the code in its L1 cache.
  - Two different I/O channels may end up on the same core.

A simple solution: interrupt binding

- Each interrupt is statically bound to a given core (network -> core 1, disk -> core 2)
- Works well, but can become an headache and does not fully solve the problem: one very fast card can still overwhelm the core.
- Needs to bind application to the same cores for best optimization: what about multi-threaded applications, for which we want one thread = one core ?

PCI optimization: MSI-X

- Extension to MSI (Message Signaled Interrupts)
- Increases the number of interrupt “pins” per card
- Associates rx/tx queues to a given core
- Allows to stitch together on the same core, the thread that runs the program and the asynchronous event it may receive (incoming network packets, asynchronous I/O…), resulting in maximizing L1 cache hit.
- Requires Chipset, card, and operating support.
- Optimized for Linux’ kernel > 2.6.26
- This is a major optimization: on a system with 4 x 10G ethernet, performance gain can be up to 20%

/proc/interrupts

 # cat /proc/interrupts displays interrupts statistics on which core each of the interrupts are being executed.

- Find cores that are overloaded with interrupts
- Find the interrupts number of given queues (per interface)

By default, the Linux distribution is configured for automatically balance IRQ's across cores. This must be disable:

The linux service *irqbalance* must be turned off:

chkconfig irqbalanced off

/proc/irq/<number>/smp_affinity

For a given interrupt, it is possible to know on which core(s) it is bound:

 # cat /proc/irq/32/smp_affinity will return a cpu mask in hex. Examples of masks are:

- 2 (hex) = 10 (bin) = core 1 (first core is 0)
- 3 (hex) = 11 (bin) = core 0 and core 1
- ffff (hex) = 11111111 11111111 = all cores

Binding IRQ 32 to core 7 is done by:

echo 80 > /proc/irq/32/smp_affinity

Core 7 = 10000000 (bin) = 80 (hex)

Interrupt Coalescence

Avoid flooding the host system with too many interrupts, packets are collected and one single interrupt is generated for multiple packets.

- Not all NIC support it
- 75-100 micro-seconds timeout
- Can be critical for high performance NIC (10Gb, 40Gb…)

TCP Autotuning Settings: http://fasterdata.es.net/TCP-Tuning/

Linux 2.6: add to /etc/sysctl.conf

```
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
# autotuning min, default, and max number of bytes to use
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
```

FreeBSD 7.0+: add to /etc/sysctl.conf

```
net.inet.tcp.sendbuf_max=16777216
net.inet.tcp.recvbuf_max=16777216
```

Mac OSX: add to /etc/sysctl.conf

```
kern.ipc.maxsockbuf=16777216
net.inet.tcp.sendspace=8388608
net.inet.tcp.recvspace=8388608
```

Windows XP

- use “DrTCP” (http://www.dslreports.com/drtcp/) to modify registry settings to increase TCP buffers

Windows Vista/Windows 7: autotunes by default, 16M Buffers

Selecting TCP Congestion Control in Linux

To determine current configuration:

- sysctl -a | grep congestion
- net.ipv4.tcp_congestion_control = cubic
- net.ipv4.tcp_available_congestion_control = cubic reno

Use /etc/sysctl.conf to set to any available congested congestion control.

Supported options (may need to enabled by default in your kernel):

- CUBIC, BIC, HTCP, HSTCP, STCP, LTCP, more..
- E.g.: Centos 5.5 includes these:
  - CUBIC, HSTCP, HTCP, HYBLA, STCP, VEGAS, VENO, Westwood

Use modprobe to add:

- /sbin/modprobe tcp_htcp
- /sbin/modprobe tcp_cubic

Additional Host Tuning for Linux

Linux by default caches ssthresh, so one transfer with lots of congestion will throttle future transfers. To turn that off set:

```
net.ipv4.tcp_no_metrics_save = 1
```

Also should change this for 10GE

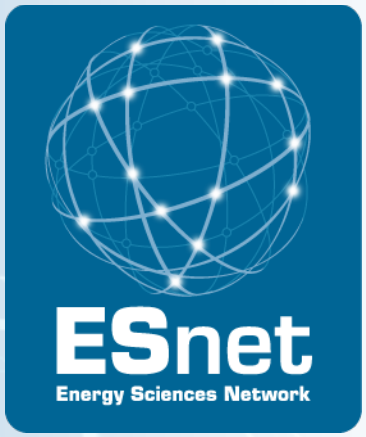
```
net.core.netdev_max_backlog = 250000
```

Warning on Large MTUs:

- If you have configured your Linux host to use 9K MTUs, but the MTU discovery reduces this to 1500 byte packets, then you actually need 9/1.5 = 6 times more buffer space in order to fill the pipe.
  - Some device drivers only allocate memory in power of two sizes, so you may even need 16/1.5 = 11 times more buffer space!

Application Tuning

- Depends on application
- Bind the application threads to the right core: the thread that is sending or receiving from the network should be running on the same core as the IRQ for that network interface. (the unix command *taskset* is useful)
- Threads that are doing disk I/O should be running on the same core as where the RAID controller IRQ is bound.
- Applications must not use more memory than what is physically available (no swap).



Achieving the Science DMZ

Section 3: Bulk Data Transfer Tools

Brian Tierney, ESnet

Joint Techs, Baton Rouge, LA, January, 2012

Section Outline

Setting expectations

What makes a fast data transfer tool

Just say no to scp

GridFTP

Commercial Tools

Tool Tuning

Time to Copy 1 Terabyte

10 Mbps network : 300 hrs (12.5 days)

100 Mbps network : 30 hrs

1 Gbps network : 3 hrs (are your disks fast enough?)

10 Gbps network : 20 minutes (need really fast disks and filesystem)

These figures assume some headroom left for other users

Compare these speeds to:

- USB 2.0 portable disk
  - 60 MB/sec (480 Mbps) peak
  - 20 MB/sec (160 Mbps) reported on line
  - 5-10 MB/sec reported by colleagues
  - 15-40 hours to load 1 Terabyte

Bandwidth Requirements

Bandwidth Requrements to move Y Bytes of data in Time X

	Bits per Second Requirements				
10PB	25,020.0 Gbps	3,127.5 Gbps	1,042.5 Gbps	148.9 Gbps	34.7 Gbps
1PB	2,502.0 Gbps	312.7 Gbps	104.2 Gbps	14.9 Gbps	3.5 Gbps
100TB	244.3 Gbps	30.5 Gbps	10.2 Gbps	1.5 Gbps	339.4 Mbps
10TB	24.4 Gbps	3.1 Gbps	1.0 Gbps	145.4 Mbps	33.9 Mbps
1TB	2.4 Gbps	305.4 Mbps	101.8 Mbps	14.5 Mbps	3.4 Mbps
100GB	238.6 Mbps	29.8 Mbps	9.9 Mbps	1.4 Mbps	331.4 Kbps
10GB	23.9 Mbps	3.0 Mbps	994.2 Kbps	142.0 Kbps	33.1 Kbps
1GB	2.4 Mbps	298.3 Kbps	99.4 Kbps	14.2 Kbps	3.3 Kbps
100MB	233.0 Kbps	29.1 Kbps	9.7 Kbps	1.4 Kbps	0.3 Kbps
	1H	**8H**	**24H**	**7Days**	**30Days**

This table available at http://fasterdata.es.net

Sample Data Transfer Results

Using the right tool is very important

Sample Results: Berkeley, CA to Argonne, IL (near Chicago).
RTT = 53 ms, network capacity = 10Gbps.

Tool	Throughput
– scp:	140 Mbps
– HPN patched scp:	1.2 Gbps
– ftp	1.4 Gbps
– GridFTP, 4 streams	5.4 Gbps
– GridFTP, 8 streams	6.6 Gbps

– Note that to get more than 1 Gbps (125 MB/s) disk to disk requires RAID.

Data Transfer Tools

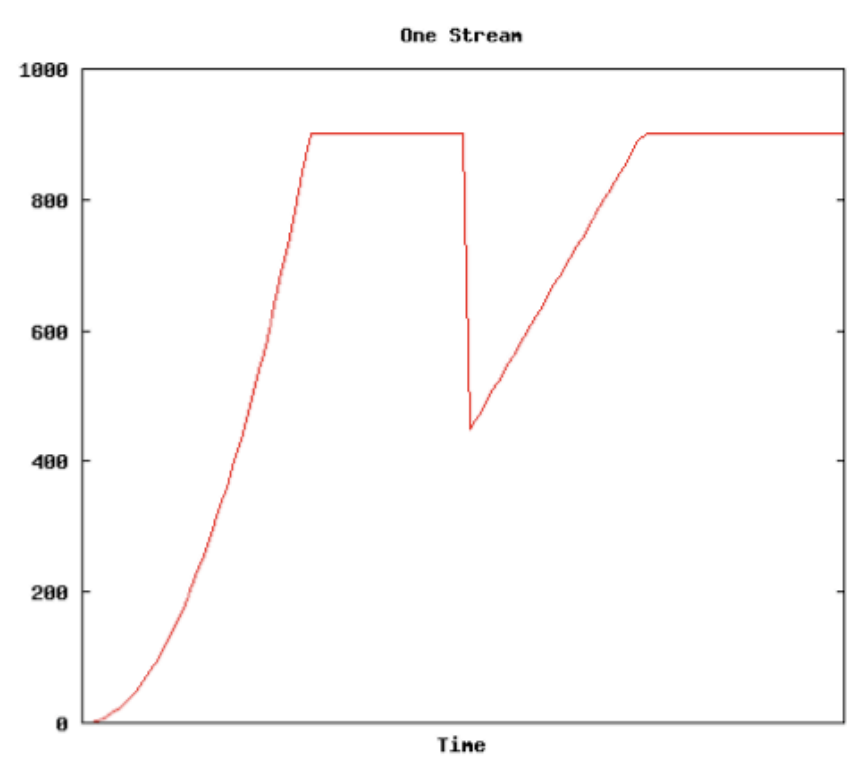
Parallelism is key

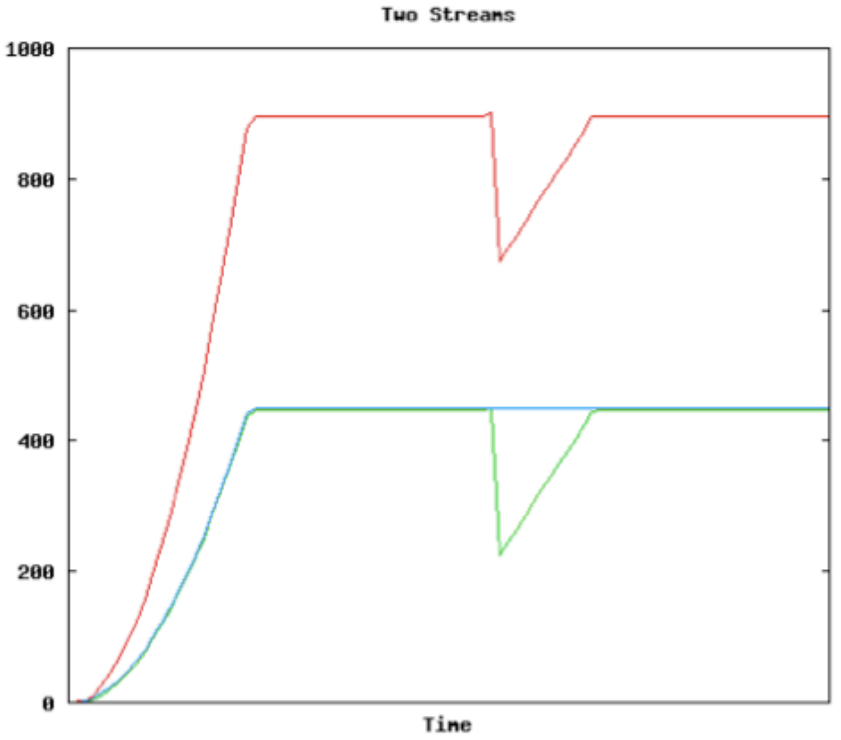
- It is much easier to achieve a given performance level with four parallel connections than one connection
- Several tools offer parallel transfers

Latency interaction is critical

- Wide area data transfers have much higher latency than LAN transfers
- Many tools and protocols assume a LAN
- Examples: SCP/SFTP, HPSS mover protocol

Parallel Streams Help With TCP Congestion Control Recovery Time





Why Not Use SCP or SFTP?

Pros:

- Most scientific systems are accessed via OpenSSH
- SCP/SFTP are therefore installed by default
- Modern CPUs encrypt and decrypt well enough for small to medium scale transfers
- Credentials for system access and credentials for data transfer are the same

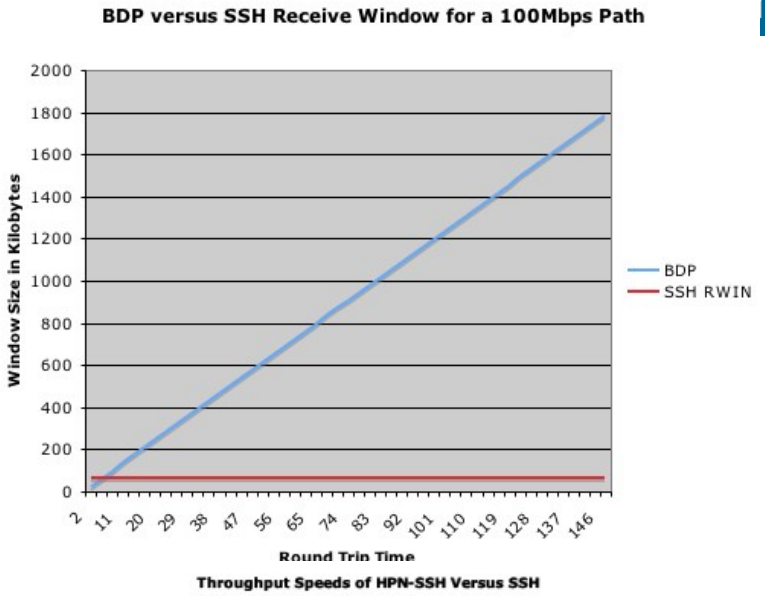
Cons:

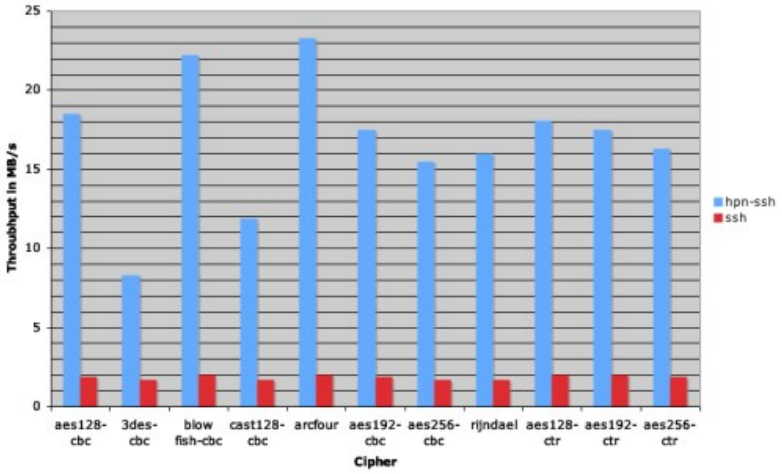
- The protocol used by SCP/SFTP has a fundamental flaw that limits WAN performance
- CPU speed doesn't matter – latency matters
- Fixed-size buffers reduce performance as latency increases
- It doesn't matter how easy it is to use SCP and SFTP – they simply do not perform

Verdict: Do Not Use Without Performance Patches

A Fix For scp/sftp

- PSC has a patch set that fixes problems with SSH
  - http://www.psc.edu/networking/projects/hpn-ssh/
- Significant performance increase
- Advantage – this helps rsync too





sftp

Uses same code as scp, so don't use sftp WAN transfers unless you have installed the HPN patch from PSC

But even with the patch, SFTP has yet another flow control mechanism

- By default, sftp limits the total number of outstanding messages to 16 32KB messages.
- Since each datagram is a distinct message you end up with a 512KB outstanding data limit.
- You can increase both the number of outstanding messages ('-R') and the size of the message ('-B') from the command line though.

Sample command for a 128MB window:

- sftp -R 512 -B 262144 user@host:/path/to/file outfile

FDT

FDT = Fast Data Transfer tool from Caltech

- http://monalisa.cern.ch/FDT/
- Java-based, easy to install
- used by US-CMS project
- being deployed by the DYNES project

GridFTP

GridFTP from ANL has features needed to fill the network pipe

- Buffer Tuning
- Parallel Streams

Supports multiple authentication options

- Anonymous
- ssh
- X509

Ability to define a range of data ports

- helpful to get through firewalls

Sample Use:

- globus-url-copy -p 4 sshftp://data.lbl.gov/home/mydata/myfile file://home/mydir/myfile

Available from: http://www.globus.org/toolkit/downloads/

Some newer GridFTP Features

ssh authentication option

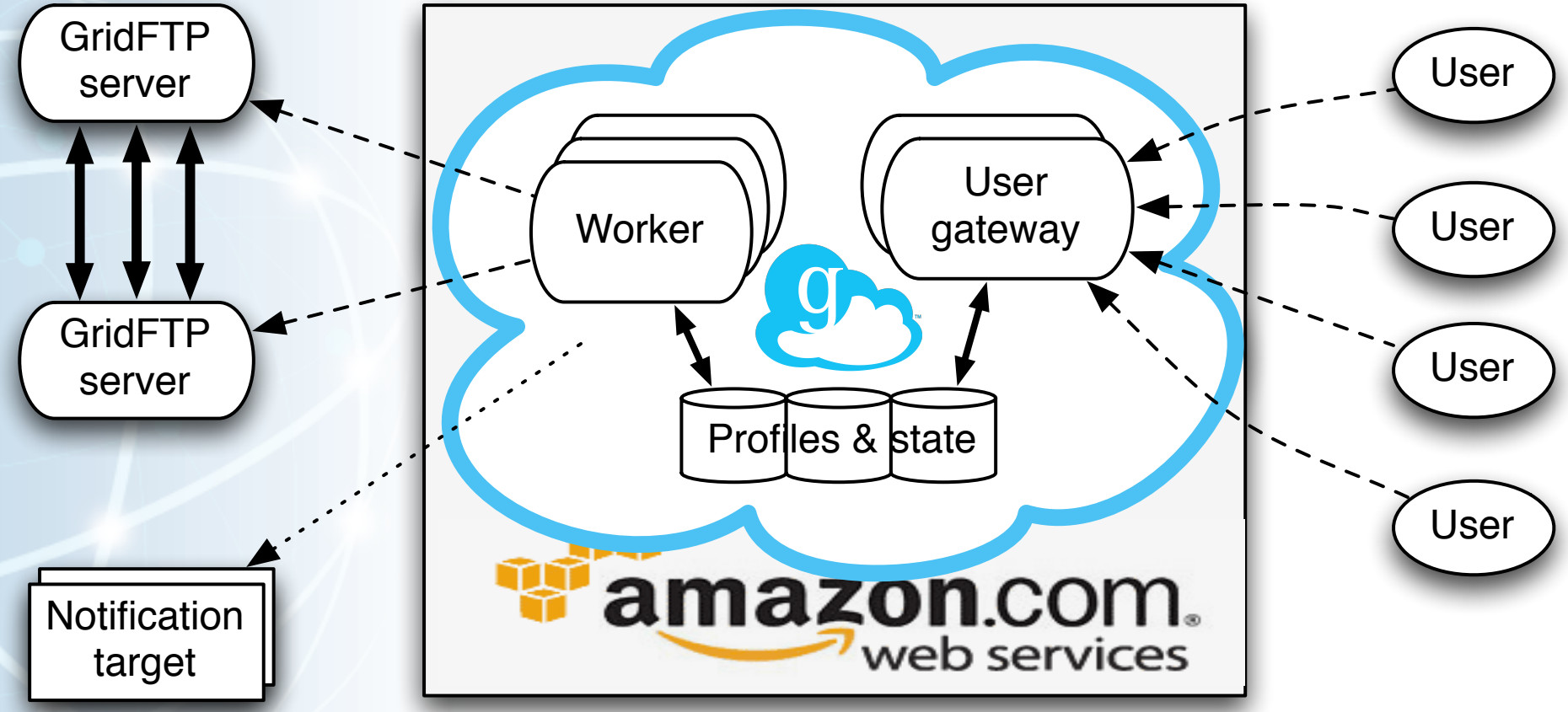
- Not all users need or want to deal with X.509 certificates
- Solution: Use SSH for Control Channel
  - Data channel remains as is, so performance is the same
- see http://fasterdata.es.net/gridftp.html for a quick start guide

Optimizations for small files

- Concurrency option (-cc)
  - establishes multiple control channel connections and transfer multiple files simultaneously.
- Pipelining option for multi-file transfers (-pp):
  - Client sends next request before the current completes
- Cached Data channel connections
  - Reuse established data channels (Mode E)
  - No additional TCP or GSI connect overhead

Support for UDT protocol

Globus Online: An easy to use wrapper for GridFTP:



Globus Online highlights

Command line interface
ls alcf#dtn:~
scp alcf#dtn:~/myfile \
nersc#dtn:~/myfile

HTTP REST interface
POST https://transfer.api.
globusonline.org/ v0.10/
transfer <transfer-doc>

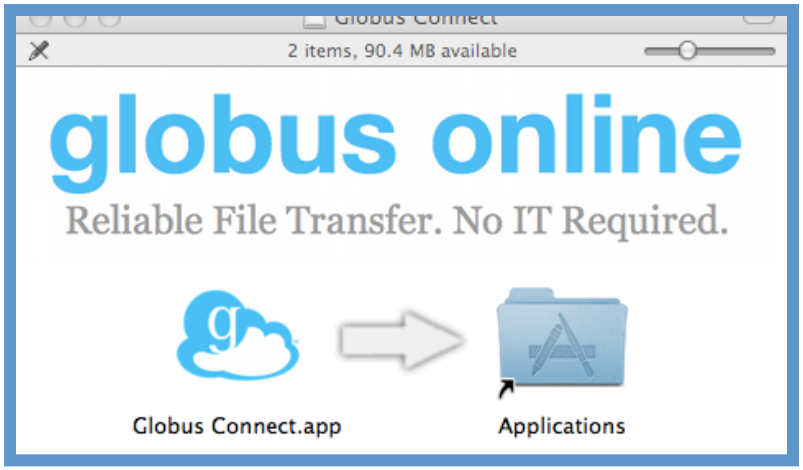
Fire-and-forget data movement
Many files and lots of data
Third-party transfers
Performance optimization
Across multiple security domains
Expert operations and support

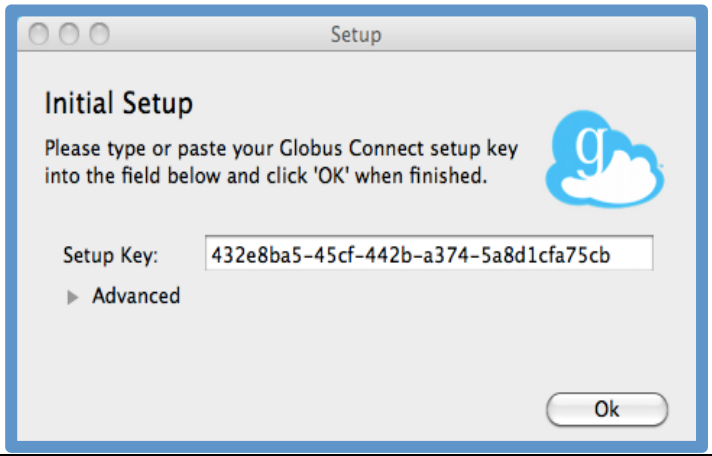
GridFTP servers
FTP servers

High-performance
data transfer nodes

Globus Connect
on local computers

Globus Connect to/from your laptop





Globus Connect Multi-User

Use Globus Connect Multi-User (GCMU) to:

- Create transfer endpoints in minutes
- Enable multi-user GridFTP access for a resource
- GCMU packages a GridFTP server, MyProxy server and MyProxy Online CA pre-configured for use with Globus Online
  - Avoids the fairly complex GridFTP server installation process

See: http://www.globusonline.org/gcmu/

Globus Connect Multi-User Installation

GridFTP finally comes as an easy to install RPM wrapped in a shell script

Installation steps:

```
wget http://connect.globusonline.org/linux/stable/
  globusconnect-multiuser-latest.tgz

tar -xvzf globusconnect-multiuser-latest.tgz

cd gcmu*

sudo ./install

  (And answer a couple simple questions)
```

Other Data Transfer Tools

bbcp: http://www.slac.stanford.edu/~abh/bbcp/

- supports parallel transfers and socket tuning
- bbcp -P 4 -v -w 2M myfile remotehost:filename

lftp: http://lftp.yar.ru/

- parallel file transfer, socket tuning, HTTP transfers, and more.
- lftp -e 'set net:socket-buffer 4000000; pget -n 4 [http|ftp]://site/path/file; quit'

axel: http://axel.alioth.debian.org/

- simple parallel accelerator for HTTP and FTP.
- axel -n 4 [http|ftp]://site/file

Commercial Data Transfer Tools

There are several commercial UDP-based tools

- Aspera: http://www.asperasoft.com/
- Data Expedition: http://www.dataexpedition.com/
- TIXstream: http://www.tixeltec.com/tixstream_en.html

These should all do better than TCP on a congested, high-latency path

- advantage of these tools less clear on an uncongested path

They all have different, fairly complicated pricing models

Next Generation Tools/Protocols

RDMA-based tools:

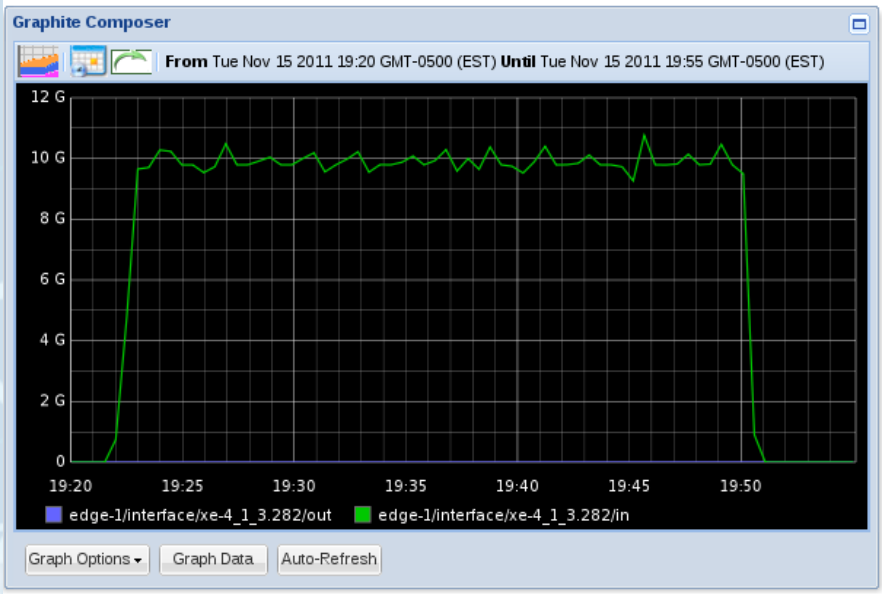
- Several groups have been experimenting with RDMA over the WAN
  - XIO driver for GridFTP (UDEL, OSU)
  - RFTP: BNL
- Over a dedicated layer-2 circuit, performance is the same as TCP, with **much** less CPU
- Requires hardware support on the NIC (e.g.: Mellanox)
  - Software version exists, but requires custom kernel and is slower
- RDMA tuning can be quite tricky to get right

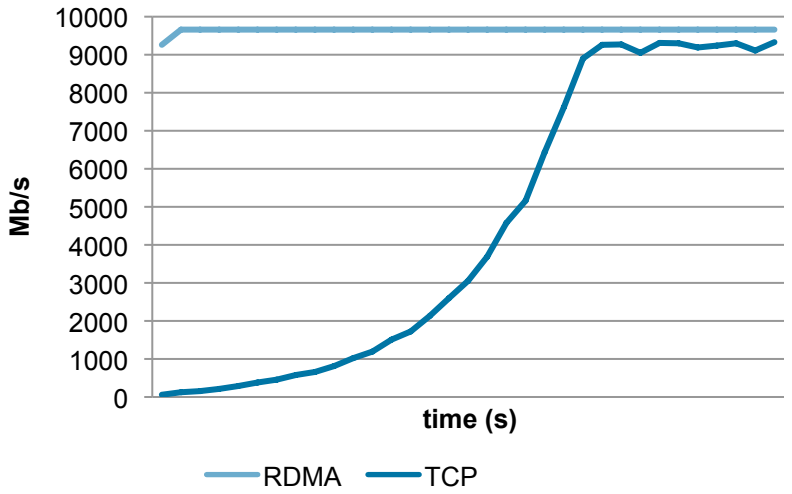
Session Layer Networking / Phoebus:

Phoebus Gateway can by used to translate being the LAN protocol (e.g. TCP) and a more efficient WAN protocol (e.g.: RDMA)

Sample RDMA Results: 10G dedicated layer-2 circuit, Long Island NY to Seattle

- 9.9G for both TCP and RDMA
  - 80% CPU for TCP
  - 3-4% CPU load for RDMA
- RDMA ramps up much faster than TCP

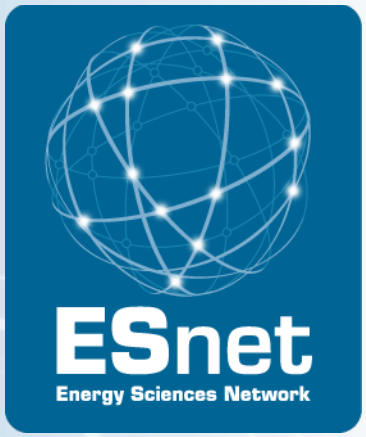




Tuning your Data Transfer Tools

Be sure to check the following:

- What is your host's maximum TCP window size?
  - 32M is good for most many environments
  - More for jumbo frames or very long RTT paths
- Which TCP congestion algorithm are you using?
  - Cubic or HTCP are usually best
- How many parallel streams are you using?
  - Use as few as possible that fill the pipe, usually 2-4 streams
  - Too many streams usually end up stepping on each other
    - May need more streams in cases of:
      - Very high RTT paths
      - Traversing slow firewalls
      - Paths without enough switch buffering



Section 4: Network Performance Monitoring and Troubleshooting using perfSONAR

Section Outline

Problem definition

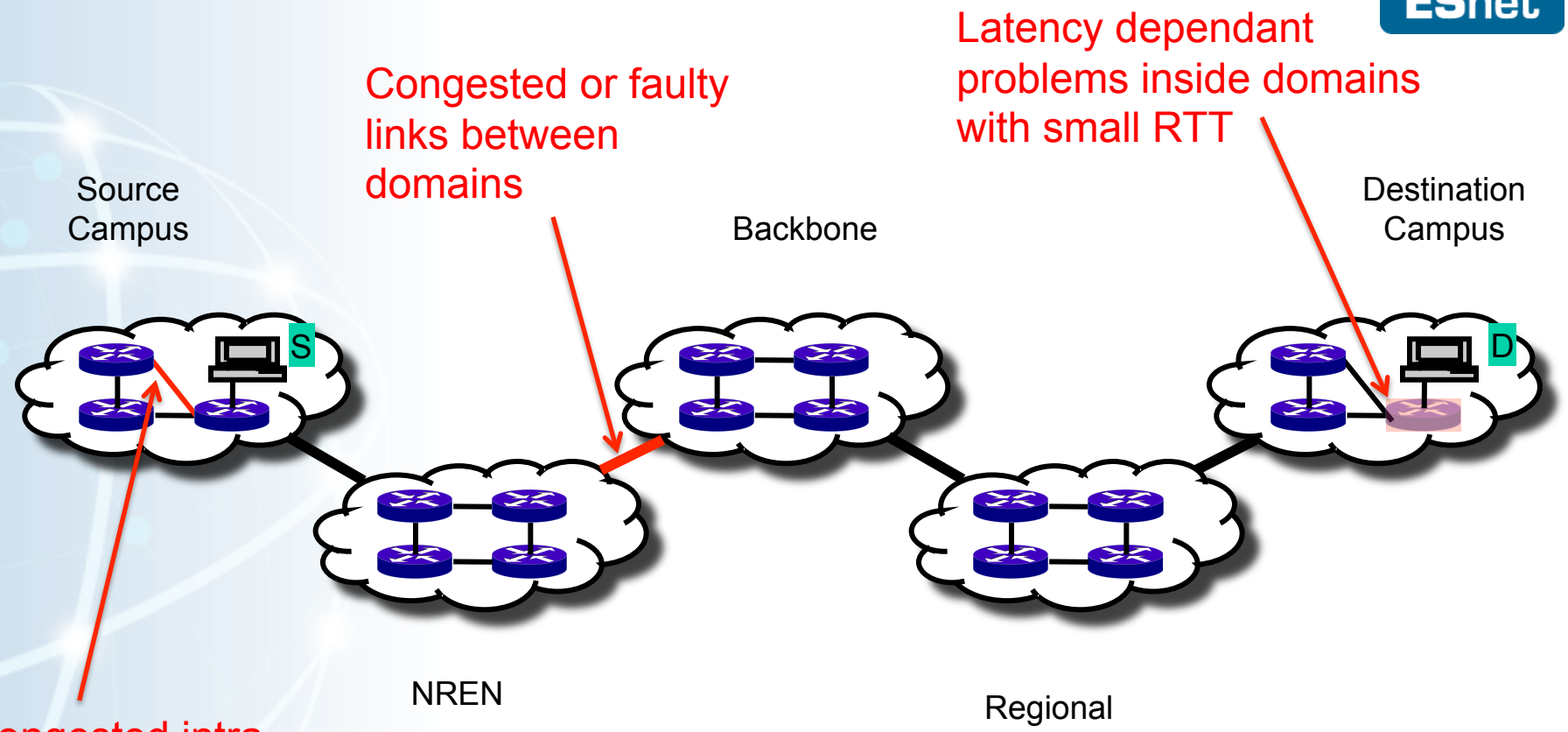
perfSONAR overview

Case studies

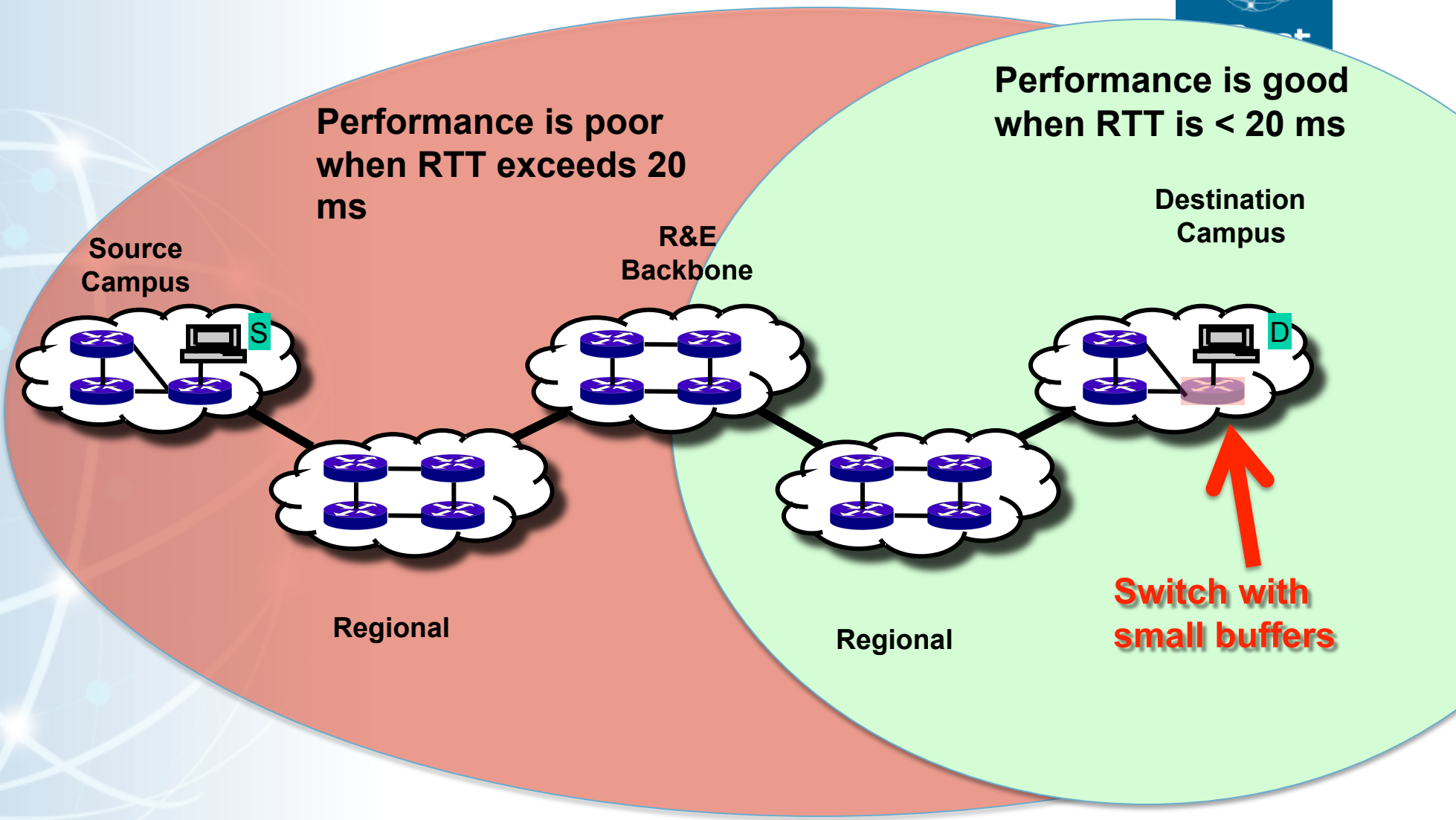
Site deployment recommendations

perfSONAR host recommendations

Where are common problems?



Local testing will not find all problems



Soft Network Failures

Soft failures are where basic connectivity functions, but high performance is not possible.

TCP was intentionally designed to hide all transmission errors from the user:

- "As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users." (From IEN 129, RFC 716)

Some soft failures only affect high bandwidth long RTT flows.

Hard failures are easy to detect & fix

- soft failures can lie hidden for years!

One network problem can often mask others

A small about of packet loss makes a huge difference in TCP performance

A Nagios alert based on our regular throughput testing between one site and ESnet core alerted us to poor performance on high latency paths

No errors or drops reported by routers on either side of problem link

- only perfSONAR bwctl tests caught this problem

Using packet filter counters, we saw 0.0046% loss in one direction

- 1 packets out of 22000 packets

Performance impact of this: (outbound/inbound)

- To/from test host 1 ms RTT : 7.3 Gbps out / 9.8 Gbps in
- To/from test host 11 ms RTT: 1 Gbps out / 9.5 Gbps in
- To/from test host 51ms RTT: 122 Mbps out / 7 Gbps in
- To/from test host 88 ms RTT: 60 Mbps out / 5 Gbps in
  - More than 80 times slower!

Common Soft Failures

Random Packet Loss

- Bad/dirty fibers or connectors
- Low light levels due to amps/interfaces failing
- Duplex mismatch

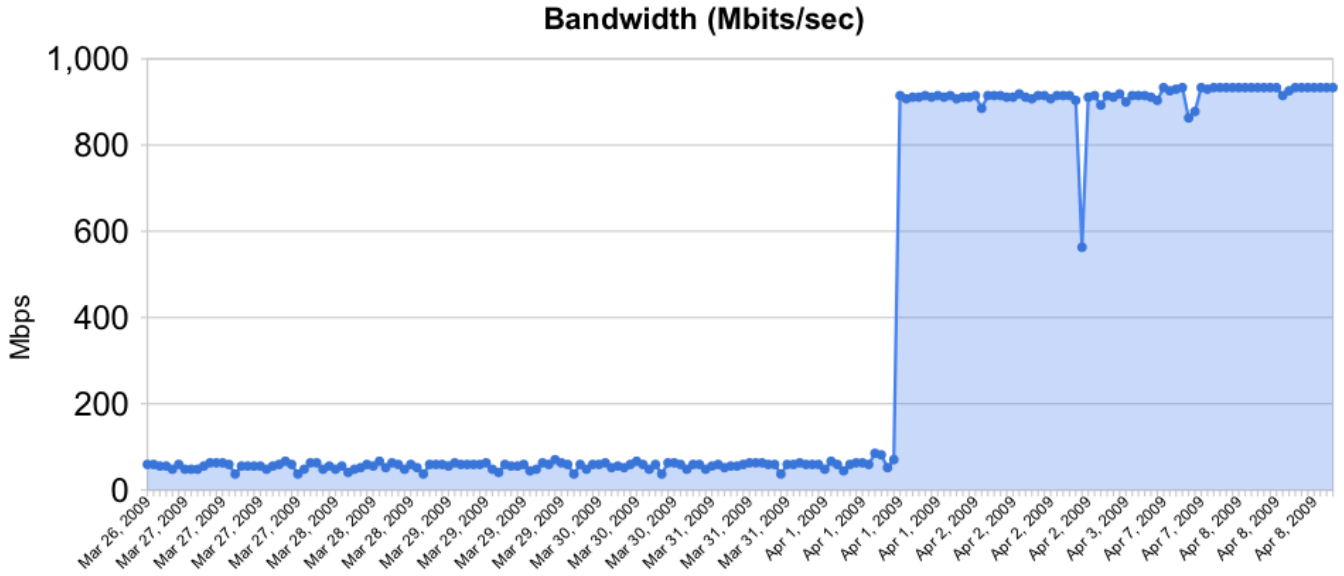
Small Queue Tail Drop

- Switches not able to handle the long packet trains prevalent in long RTT sessions and local cross traffic at the same time

Un-intentional Rate Limiting

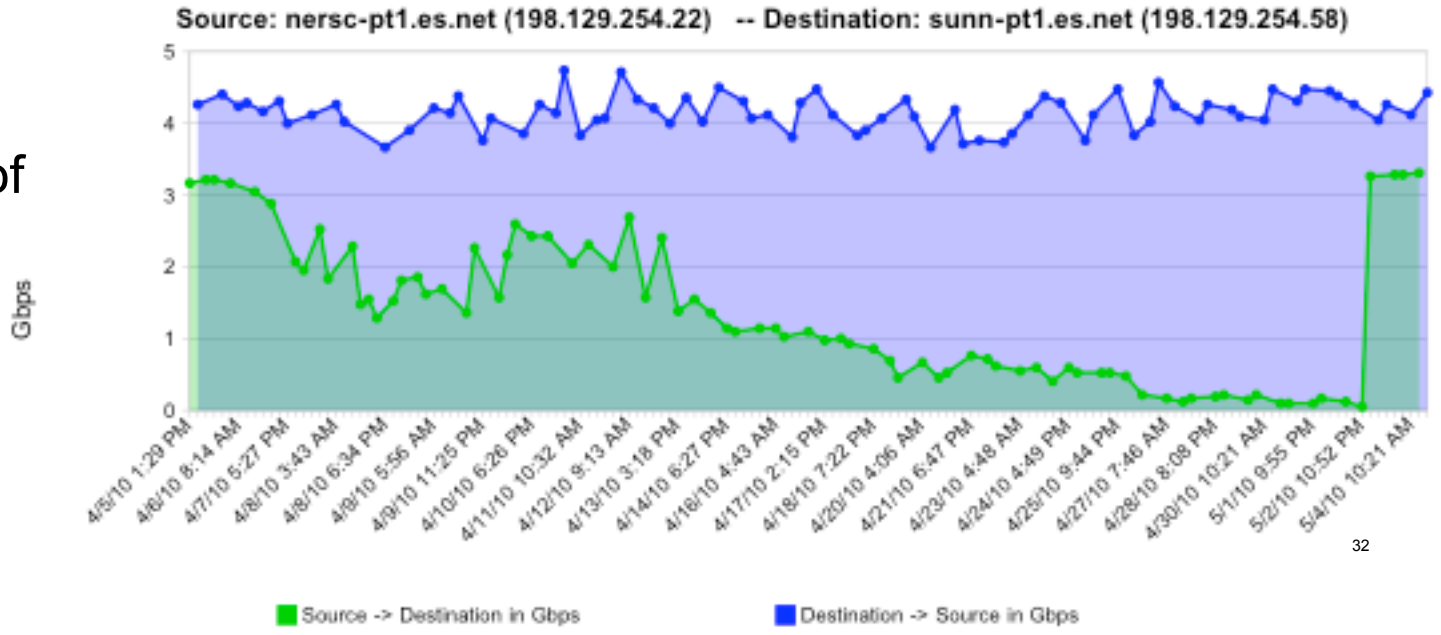
- Processor-based switching on routers due to faults, acl's, or mis-configuration
- Security Devices
  - E.g.: 10X improvement by turning off Cisco Reflexive ACL

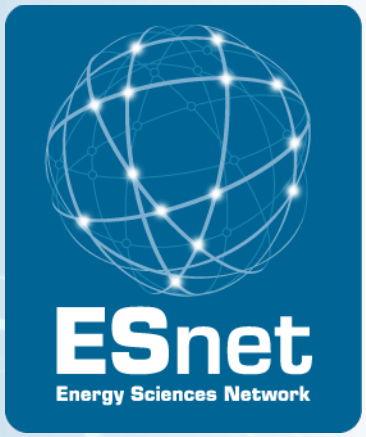
Sample Results: Finding/Fixing soft failures



Rebooted router with full route table

Gradual failure of optical line card





perfSONAR Overview

Addressing the Problem: perfSONAR

perfSONAR - an open, web-services-based framework for:

- running network tests
- collecting and publishing measurement results

ESnet and Internet2 are:

- Deploying the framework across the science community
- Encouraging people to deploy ‘known good’ measurement points near domain boundaries
  - “known good” = hosts that are well configured, enough memory and CPU to drive the network, proper TCP tuning, clean path, etc.
- Using the framework to find and correct soft network failures.

Who is perfSONAR?

The perfSONAR Consortium is a joint collaboration between

- ESnet
- Géant
- Internet2
- Rede Nacional de Ensino e Pesquisa (RNP)

Decisions regarding protocol development, software branding, and interoperability are handled at this organization level

There are at least two independent efforts to develop software frameworks that are perfSONAR compatible.

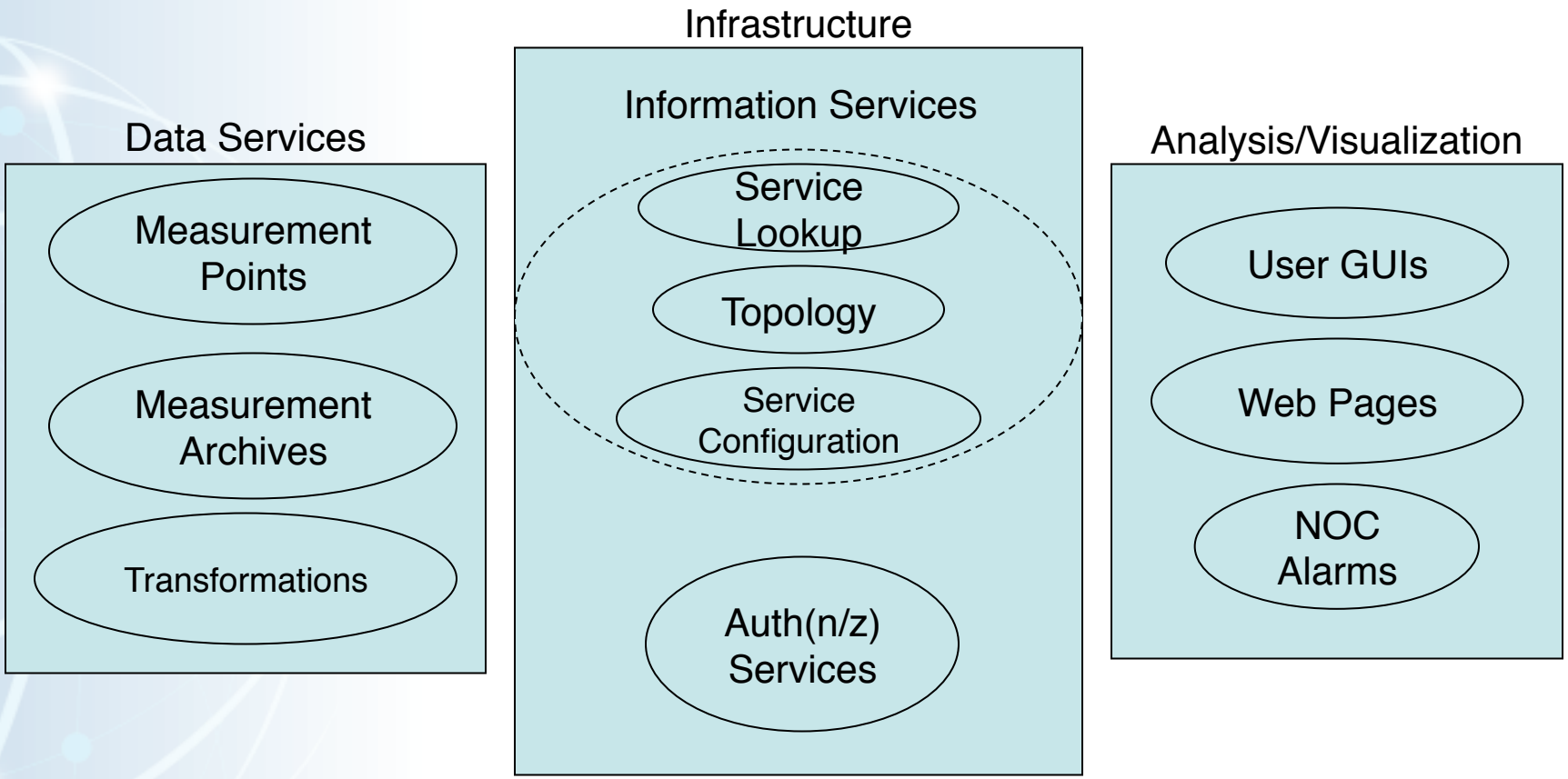
- perfSONAR-MDM
- perfSONAR-PS

Each project works on an individual development roadmap and works with the consortium to further protocol development and insure compatibility

perfSONAR Terminology

- perfSONAR: standardized schema, protocols, APIs
- perfSONAR-MDM: GÉANT Implementation and deployment
  - aimed at NRENS
- perfSONAR-PS: ESnet/Internet2 implementation and deployment
  - aimed at end-users and network admins (site and backbone)
- perfSONAR Performance Toolkit
  - Easy to install Packaging of perfSONAR-PS
  - “network install” and “LiveCD” versions

perfSONAR Architecture Overview



perfSONAR Services

PS-Toolkit includes these measurement tools:

- BWCTL: network throughput
- OWAMP: network loss, delay, and jitter
- PINGER: network loss and delay

Measurement Archives (data publication)

- SNMP MA – Interface Data
- pSB MA -- Scheduled bandwidth and latency data

Lookup Service

- gLS – Global lookup service used to find services
- hLS – Home lookup service for registering local perfSONAR metadata

PS-Toolkit includes these Troubleshooting Tools

- NDT (TCP analysis, duplex mismatch, etc.)
- NPAD (TCP analysis, router queuing analysis, etc)

perfSONAR-PS Utility

perfSONAR-PS appeals to both network users and operators:

- Operators:
  - Easy deployment
  - Minimal maintenance
  - Results relevant to common problems (e.g. connectivity loss, equipment failure, performance problems)
- Users:
  - Immediate access to network data
  - Cross domain capabilities

Adoption is spreading to networks of all sizes

The perfSONAR-PS framework has two primary high level use cases:

- Diagnostic (e.g. on-demand) use
- Monitoring Infrastructure

perfSONAR-PS Utility - Diagnostics

The pS Performance Toolkit was designed for diagnostic use and regular monitoring

- All tools preconfigured
- Minimal installation requirements
- Can deploy multiple instances for short periods of time in a domain

perfSONAR-PS Utility - Monitoring

Regular monitoring is an important design consideration for perfSONAR-PS tools

- perfSONAR-BUOY and PingER provide scheduling infrastructure to create regular latency and bandwidth tests
- The SNMP MA integrates with COTS SNMP monitoring solutions

The pSPT is capable of organizing and visualizing regularly scheduled tests

NAGIOS can be integrated with perfSONAR-PS tools to facilitate alerting to potential network performance degradation

Global PerfSONAR-PS Deployments

Based on “global lookup service” (gLS) registration, Dec 2011: currently deployed in over 150 locations

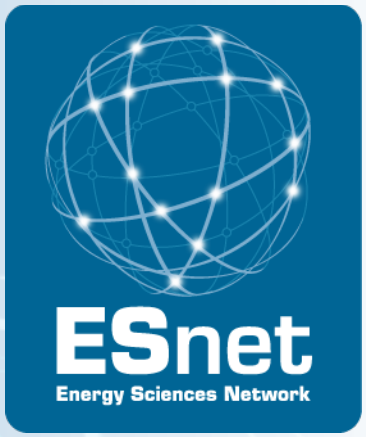
- ~ 275 bwctl and owamp servers
- ~ 230 active probe measurement archives
- ~ 25 SNMP measurement archives
- Countries include: USA, Australia, Hong Kong, Argentina, Brazil, Japan, China, Canada, Netherlands, Switzerland, Finland, Sweden, Italy, France, Pakistan

US Atlas Deployment

- Monitoring all “Tier 1 to Tier 2” connections

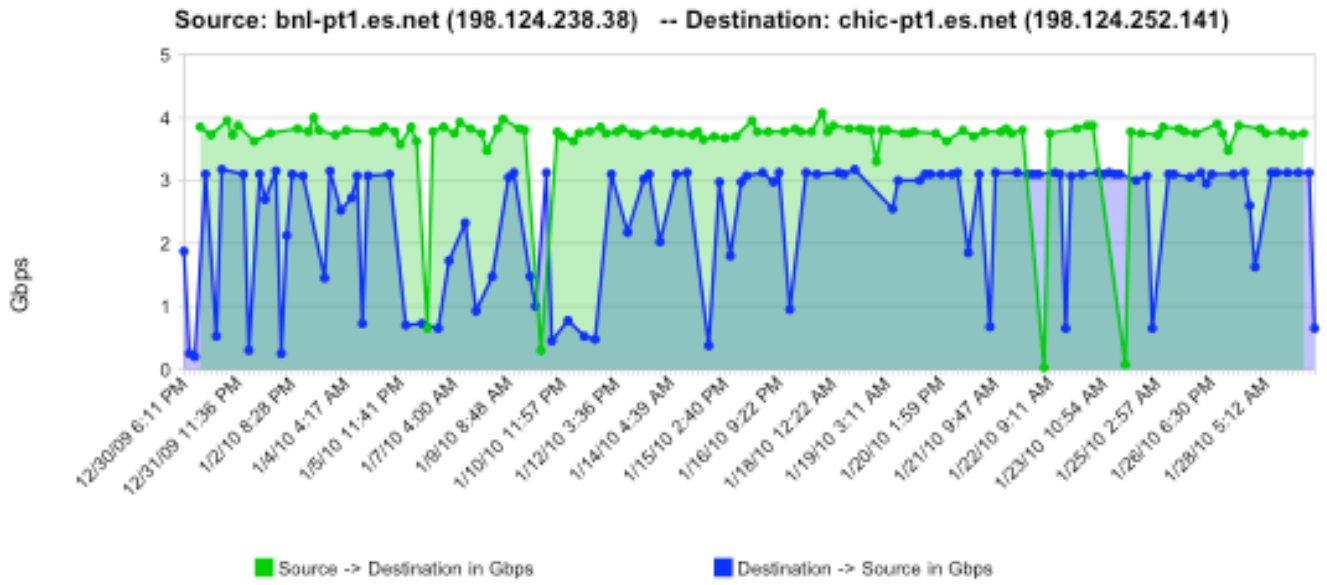
For current list of public services, see:

- http://stats.es.net/perfSONAR/directorySearch.html
- Many more “private” perfSONAR nodes deployed



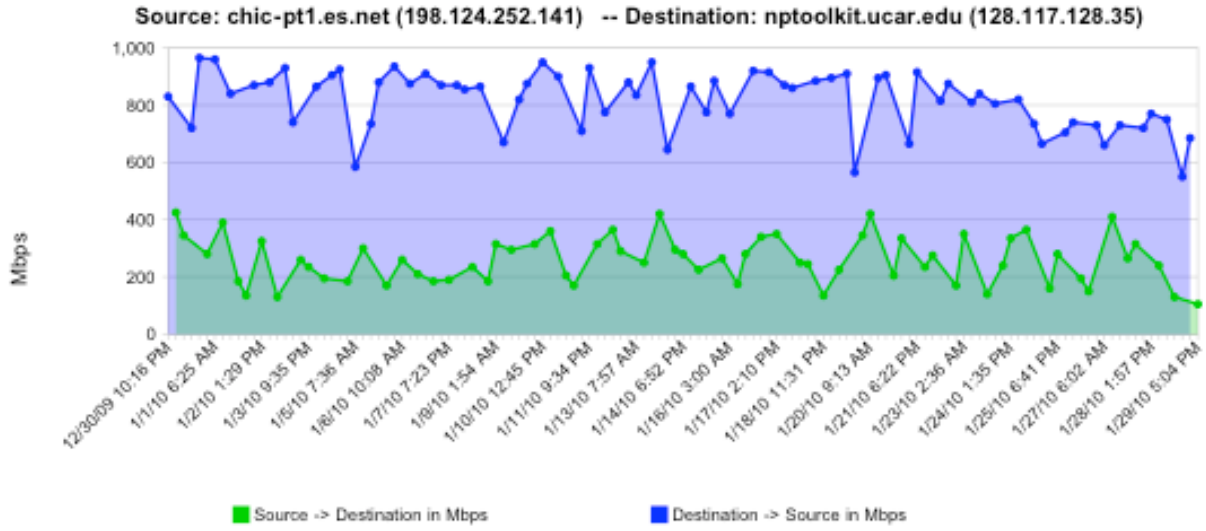
perfSONAR Case Studies

Sample Results: Throughput tests

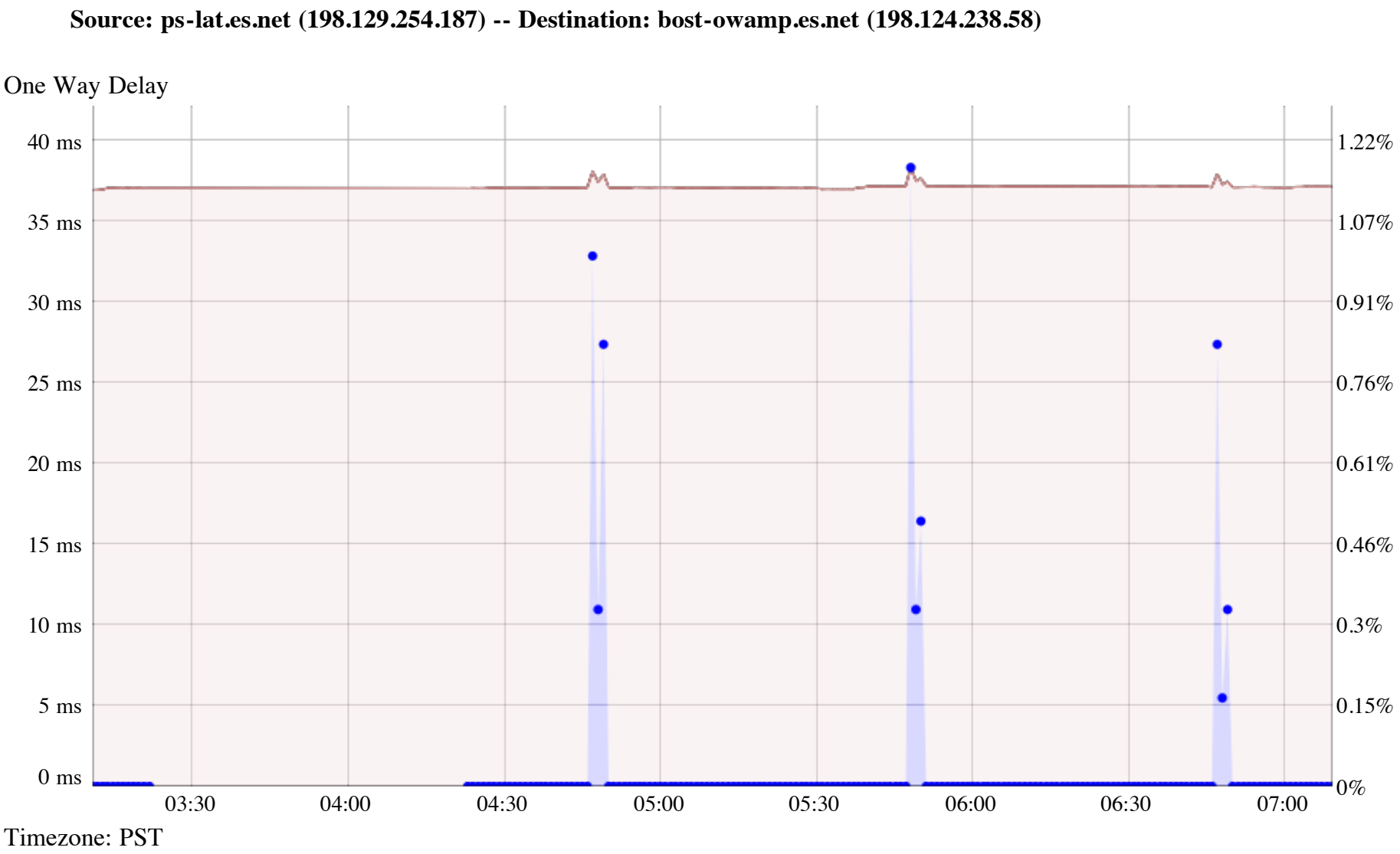


Heavily used path: probe traffic is "scavenger service"

Asymmetric Results: different TCP stacks?



Sample Results: Latency/Loss Data



Common Use Case

Trouble ticket comes in:

"I'm getting terrible performance from site A to site B"

If there is a perfSONAR node at each site border:

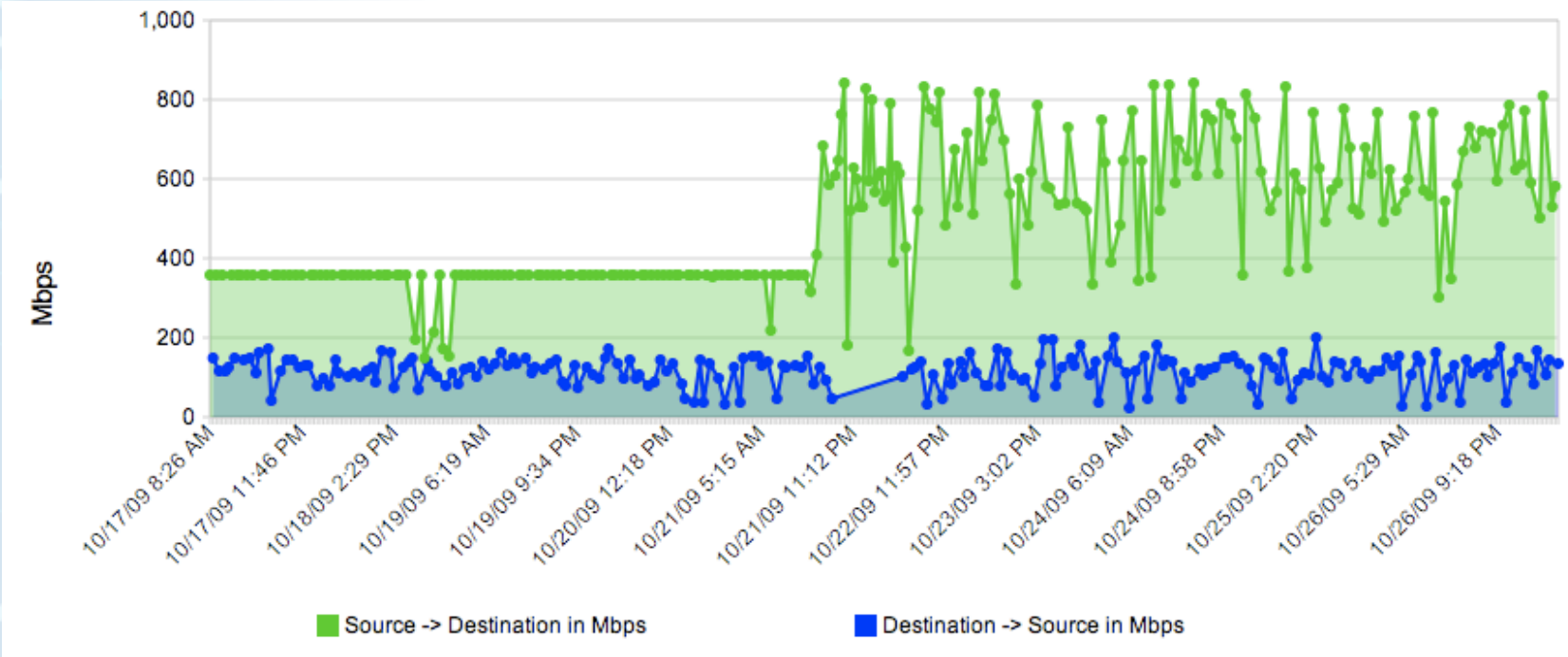
- Run tests between perfSONAR nodes
  - performance is often clean
- Run tests from end hosts to perfSONAR host at site border
  - Often find packet loss (using owamp tool)
  - If not, problem is often the host tuning or the disk

If there is not a perfSONAR node at each site border

- Try to get one deployed
- Run tests to other nearby perfSONAR nodes

REDDnet Use Case – Host Tuning

- Host Configuration – spot when the TCP settings were tweaked…



- N.B. Example Taken from REDDnet (UMich to TACC, using BWCTL measurement)
- Host Tuning: http://fasterdata.es.net/fasterdata/host-tuning/linux/

Troubleshooting Example: LLNL to BADC (Rutherford Lab, UK)

User trying to send climate data from LLNL (CA, USA) to BADC (U.K.) reports terrible performance (< 30 Mbps) in 1 direction, good performance (700 Mbps) in the other direction

Network Path used:

ESnet to AofA (aofa-cr2.es.net): bwctl testing from llnl-pt1.es.net to aofa-pt1.es.net:

- 5 Gbps both directions

GÉANT2 to UK via Amsterdam: bwctl tests llnl-pt1.es.net to london.geant2.net:

- 800 Mbps both directions
- Testing to GÉANT perfSONAR node in London critical to rule out trans-Atlantic issues

JANET to Rutherford lab

- no bwctl host ☹, but used router filter packet counters to verify no packet loss in JANET

Suspect router buffer issue at RL, but very hard to prove without a perfSONAR hosts at Rutherford lab and in JANET

Problems finally solved once test hosts temporarily deployed in JANET and at RL (just-in-time deployment of test hosts makes troubleshooting *hard*)

Troubleshooting Example: Bulk Data Transfer between DOE Supercomputer Centers

Users were having problems moving data between supercomputer centers, NERSC and ORNL

- One user was: “waiting more than an entire workday for a 33 GB input file” (this should have taken < 15 min)

perfSONAR-PS measurement tools were installed

- Regularly scheduled measurements were started

Numerous choke points were identified & corrected

- Router tuning, host tuning, cluster file system tuning

Dedicated wide-area transfer nodes were setup

- Now moving 40 TB in less than 3 days

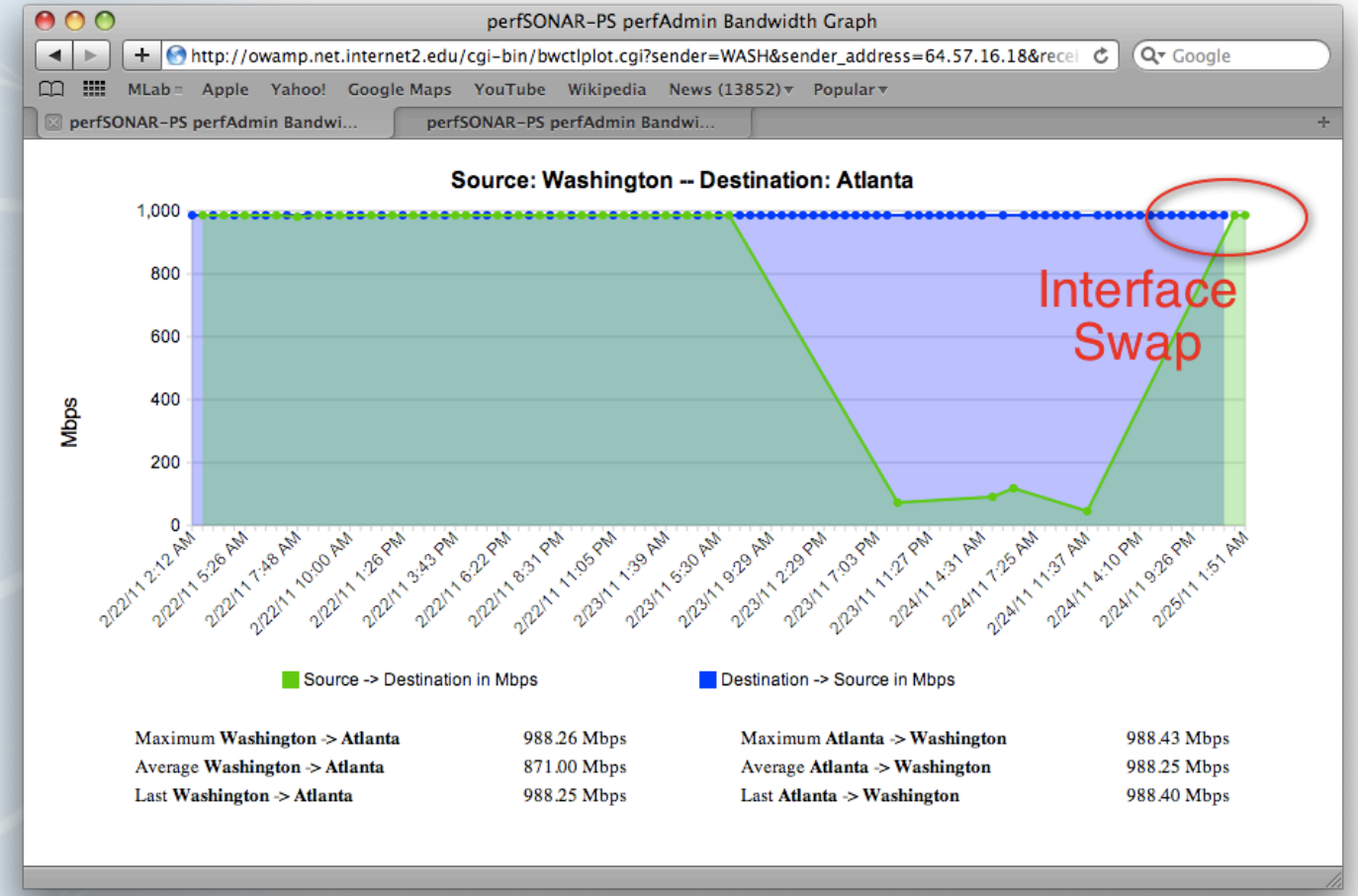
Troubleshooting Example: China to US

Difficulty getting science data moved from neutrino detectors in China to analysis in US

- Multiple difficulties (host config, packet loss, etc.)
- Installed perfSONAR-PS host in Hong Kong
  - Regular tests were started
  - Over time, multiple issues discovered and corrected, and performance improved
  - Performance went from 3Mbps to 200Mbps
- Automated testing over time provided several advantages
  - Performance problems can be correlated with network events
    - Path changes; Hardware failures; Host-level changes
  - Sometimes difficult to convince some entities that they have problems to fix without proof

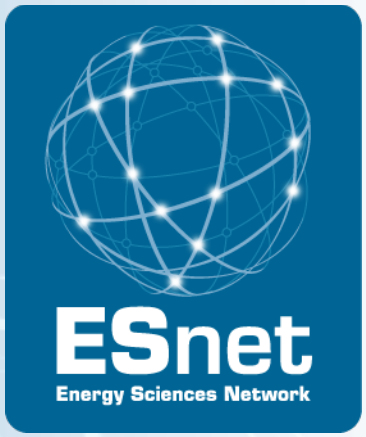
Internet2 Backbone Example

bwctl results



Internet2 Backbone Example

Owamp data plot



Effective perfSONAR Deployment Strategies

Levels of perfSONAR deployment

ESnet classifies perfSONAR deployments into 3 "levels":

Level 1: Run a bwctl server that is registered in the perfSONAR Lookup Service.

- This allows remote sites and ESnet engineers to run tests to your site.

Level 2: Configure "perfSONAR BOUY" to run regularly scheduled tests to/from your host.

- This allows you to establish a performance baseline, and to determine when performance changes.

Level 3: Full set of perfSONAR services deployed (everything on the PS Performance Toolkit)

perfSONAR-PS Software

perfSONAR-PS is an open source implementation of the perfSONAR measurement infrastructure and protocols

- written in the perl programming language

http://software.internet2.edu/pS-Performance_Toolkit/

All products are available as RPMs.

The perfSONAR-PS consortium supports CentOS (version 5).

RPMs are compiled for i386 architecture, but work w/ x86 64 bit too

Functionality on other platforms and architectures is possible, but not supported.

- Should work: Red Hat Enterprise Linux and Scientific Linux (v5)
- Harder, but possible:
  - Fedora Linux, SuSE, Debian Variants

Deploying perfSONAR-PS Tools In Under 30 Minutes

There are two easy ways to deploy a perfSONAR-PS host

"Level 1" perfSONAR-PS install:

- Build a Linux machine as you normally would (configure TCP properly! See: http://fasterdata.es.net/TCP-tuning/)
- Go through the Level 1 HOWTO
- http://fasterdata.es.net/ps_level1_howto.html
  - Includes bwctl.limits file to restrict to R&E networks only
- Simple, fewer features, runs on a standard Linux build

Use the perfSONAR-PS Performance Toolkit netinstall CD

- Most of the configuration via Web GUI
- http://psps.perfsonar.net/toolkit/
- Includes more features (perfSONAR level 3)

Measurement Recommendations for end sites

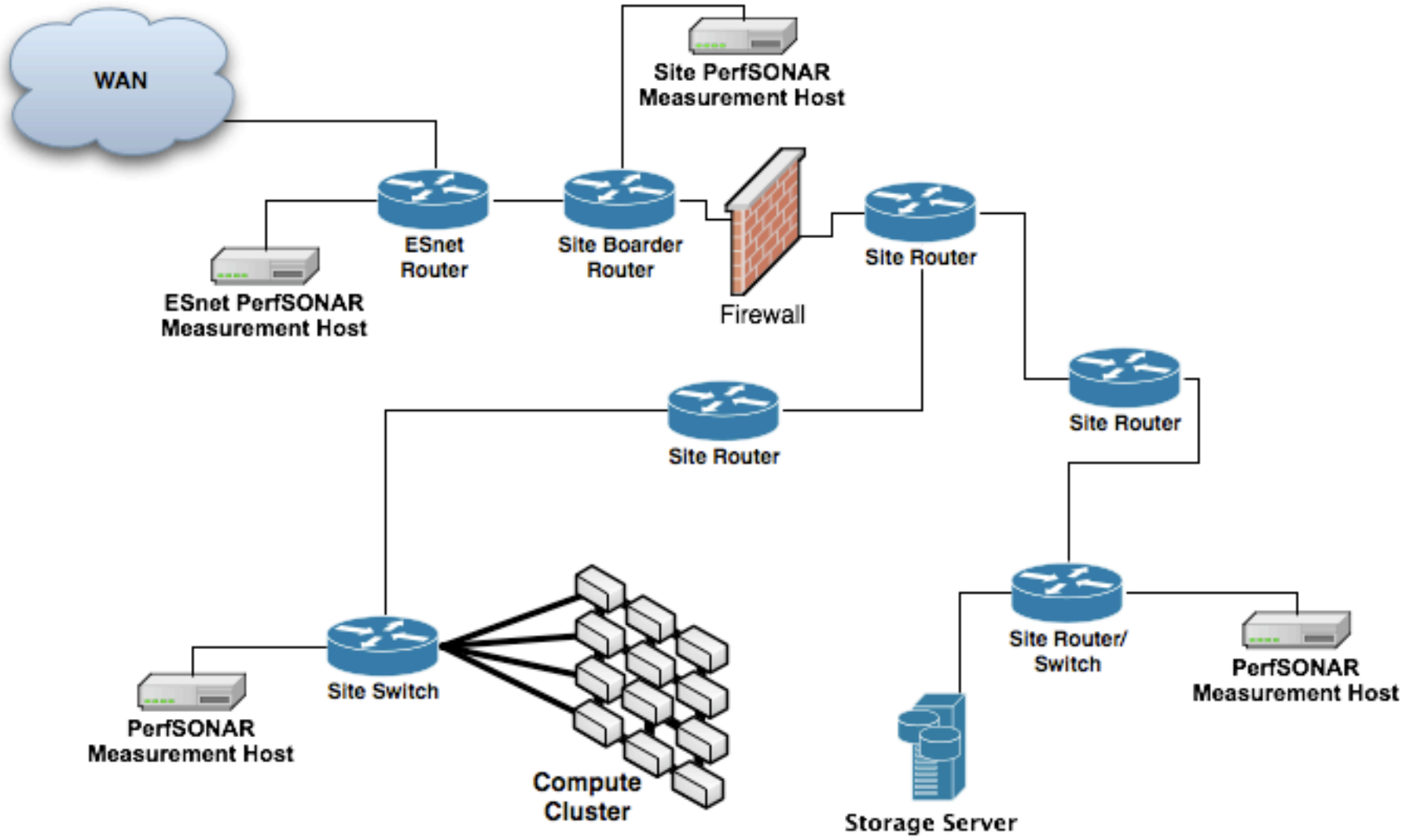
Deploy perfSONAR-PS based test tools

- At Site border
  - Use to rule out WAN issues
- Near important end systems and all DTNs
  - Use to rule out LAN issues

Use it to:

- Find & fix current local problems
- Identify when they re-occur
- Set user expectations by quantifying your network services

Sample Site Deployment



Importance of Regular Testing

You can't wait for users to report problems and then fix them (soft failures can go unreported for years!)

Things just break sometimes

- Failing optics
- Somebody messed around in a patch panel and kinked a fiber
- Hardware goes bad

Problems that get fixed have a way of coming back

- System defaults come back after hardware/software upgrades
- New employees may not know why the previous employee set things up a certain way and back out fixes

Important to continually collect, archive, and alert on active throughput test results

Developing a Measurement Plan

What are you going to measure?

- Achievable bandwidth
  - 2-3 regional destinations
  - 4-8 important collaborators
  - 4-10 times per day to each destination
  - 20 second tests within a region, longer across the Atlantic or Pacific
- Loss/Availability/Latency
  - OWAMP: ~10 collaborators over diverse paths
  - PingER: use to monitor paths to collaborators who don't support owamp
- Interface Utilization & Errors

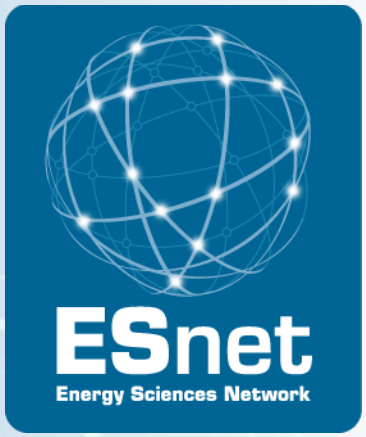
What are you going to do with the results?

- NAGIOS Alerts
- Reports to user community
- Post to Website

Sample tool: Atlas perfSONAR Dashboard

Status of perfSONAR Throughput Matrix

-	0	1	2	3	4	5	6	7	8
0:atlas-npt2.bu.edu	-	OK OK	OK OK	OK OK	OK OK	OK OK	UNKNOWN OK	OK OK	OK OK
1:lhcmon.bnl.gov	OK OK	-	OK OK	OK OK	OK OK	OK OK	OK OK	OK UNKNOWN	OK OK
2:ps2.ochep.ou.edu	OK OK	OK OK	-	OK OK	OK OK	OK OK	OK UNKNOWN	OK OK	OK OK
3:psmsu02.aglt2.org	OK OK	OK OK	OK OK	-	OK OK	OK OK	UNKNOWN UNKNOWN	OK OK	OK OK
4:netmon2.atlas-swt2.org	OK UNKNOWN	UNKNOWN OK	OK OK	OK OK	-	OK UNKNOWN	OK UNKNOWN	OK OK	OK OK
5:iut2-net2.iu.edu	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK OK	OK OK	OK OK
6:psnr-bw01.slac.stanford.edu	OK UNKNOWN	OK OK	UNKNOWN OK	UNKNOWN UNKNOWN	UNKNOWN UNKNOWN	OK OK	-	OK OK	UNKNOWN UNKNOWN
7:uct2-net2.uchicago.edu	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK OK
8:psum02.aglt2.org	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	UNKNOWN UNKNOWN	OK OK	-



perfSONAR Security models

Security and Privacy Issues with perfSONAR

The ESnet viewpoint is that perfSONAR services should be as open as possible

We make all of the following publically accessible via perfSONAR:

- all SNMP data on utilization, errors, drops
- All topology data

Anyone from an R&E network anywhere in the world can run bwctl tests to our servers

- TCP tests limited to 120 seconds
- UDP tests limited to 200 Mbps, 600 seconds

ESnet has had no security related issues since we deployed perfSONAR 5 years ago.

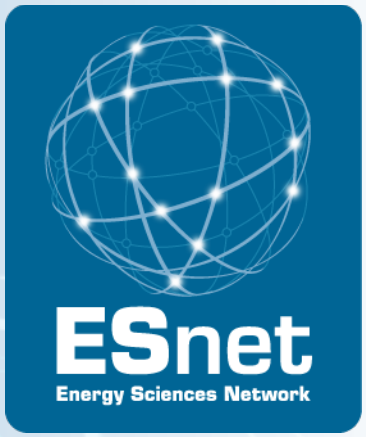
Commonly heard Security Concerns

DDOS attack using bwctl:

- bwctl has controls to limit test duration, UDP rates, allow subnets
- ESnet provides a bwctl control file with only R&E networks, updated nightly

SNMP utilization data is sensitive information

- maybe for the military, but we don't think so for R&E



perfSONAR Host Recommendations

Host Considerations

Dedicated perfSONAR hardware is best

Other applications will perturb results

Separate hosts for throughput tests and latency/loss tests is preferred

- Throughput tests can cause increased latency and loss
- Latency tests on a throughput host are still useful however

1Gbps vs 10Gbps testers

- There are a number of problem that only show up at speeds above 1Gbps

Virtual Machines do not work well for perfSONAR hosts

- clock sync issues
- throughput is reduced significantly for 10G hosts
- caveat: this has not been tested recently, and VM technology and motherboard technology has come a long way

Sample Host Configuration #1

10G throughput host: 1U, RAID disk and dual power supplies for reliability, on board IPMI: ($3000 USD)

- Intel Xeon 2.66GHz 4 Cores Processor
- (2) 4GB Modules Kingston Brand DDRIII 1333 ECC
- (2) 500GB WD SATA II Drive Enterprises
- 3Ware 9650SE-4LP 4 Ports with BBU Installed
- Myricom 10G-PCIE-8B-S

Sample Host Configuration #2

1G Host deployed by the US Atlas project in 2008:

- Intel Pentium DC E2200 2.4GHz 1MB 800MHz Processor
- Intel 945GC/ICH7 Chipset Main Board
- Onboard Marvel 8056 GbE LAN Controller
- 2GB DDR2-5300 RAM 667MHz Non-ECC Unbuffered
- 160GB SATA 7200RPM Hard Drive
- $650 USD

Perfect for a latency host or a 1G tester, no redundancy however

perfSONAR Summary

Soft failures are everywhere

We all need to look for them, and not wait for users to complain

perfSONAR is MUCH more useful when its on every segment of the end-to-end path

Ideally all networks and high BW end sites to deploy at least a “level 1” host

10G test hosts are needed to troubleshoot 10G problems

perfSONAR is MUCH more useful when its open

locking it down behind firewalls/ACLs defeats the purpose

perfSONAR-PS Community

perfSONAR-PS is working to build a strong user community to support the use and development of the software.

perfSONAR-PS Mailing Lists

- Announcement List: https://mail.internet2.edu/wws/subrequest/perfsonar-ps-announce
- Users List: https://mail.internet2.edu/wws/subrequest/performance-node-users
- Announcement List: https://mail.internet2.edu/wws/subrequest/performance-node-announce

More Information

Download the perfSONAR performance Toolkit:

- http://software.internet2.edu/pS-Performance_Toolkit/

ESnet network performance troubleshooting guide:

- http://fasterdata.es.net/fasterdata/troubleshooting/overview/

Information on downloading/installing perfSONAR

- http://fasterdata.es.net/fasterdata/perfSONAR/

Graphs of ESnet perfSONAR data:

- http://stats.es.net/

Slides from recent full day perfSONAR workshop from Internet2

- http://www.internet2.edu/workshops/npw/roster/learn-11.cfm

email: BLTierney@es.net

CASE STUDY: UNIVERSITY OF UTAH RESEARCH/ SCIENCE AND PERFORMANCE DMZ NETWORK

JOE BREEN
JOE.BREEN @UTAH.EDU
UNIV. OF UTAH

UNIVERSITY OF UTAH RESEARCH DRIVERS AND NETWORK ASSETS

Computational Science

- Scientific Computing and Imaging Institute (SCI)
- Institute for Clean and Secure Energy (ISCE)
- Center for the Simulation of Accidental Fires and Explosions (C-Safe)
- Pharmacy modeling (AMBER)
- High Energy Physics
- Computational Chemistry

Network Research

- FLUX/EMULAB – GENI infrastructure

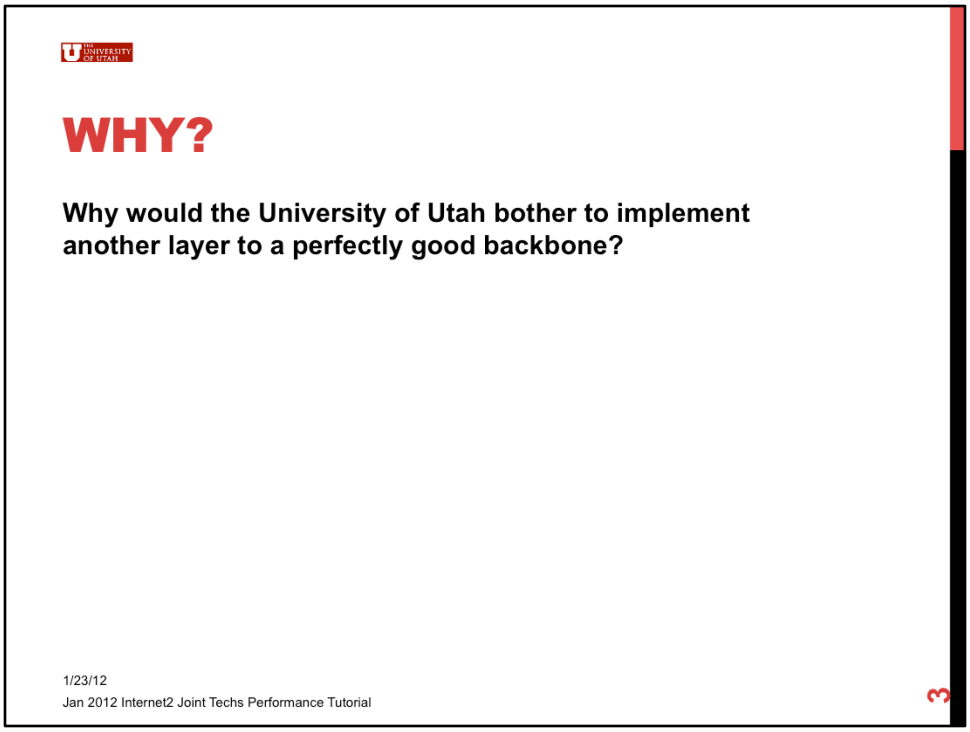
Medical Research

- University Hospital and Clinics
- Huntsman Cancer Institute (HCI)
- Strong genetics research – Mario Capecchi (Nobel Prize)
- Strong genomics research
- Utah Population DB

Local Network Assets

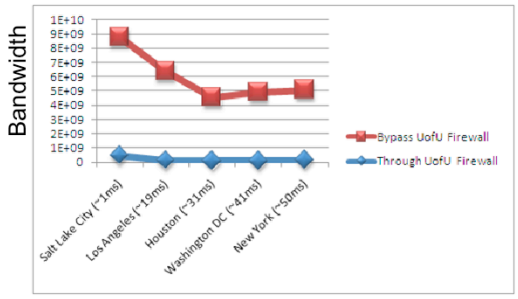
- Utah Education Network
- Metro Optical Network
- New Data Center (building is 75000 sq ft)
- Internet2 and Level 3 PoP conveniently located by airport (within 9 fiber miles)
- Good partnerships with local transportation entities UTA, UDOT
- Consolidated, Redundant 10+Gb/s campus/hospital backbone

* Dave Pershing



- University of Utah backbone is fully redundant with one or more 10Gb/s connecting each distribution node to a redundant core which connects to a redundant WAN which connects to redundant firewalls which connect to redundant Internet Border routers which connect to the Utah Education Network with a 10Gb/s connection apiece.

WHY? … PAIN!



WITH DMZ

WITHOUT DMZ

1/23/12

Jan 2012 Internet2 Joint Techs Performance Tutorial

4

- Starting for a moment with some of the results quickly highlights the pain points…
- Univ of Utah has 2 10Gb/sec links to the Utah Education Network which has 10Gb/s to Internet2
- Red line denotes performance without UofU firewall
- Blue line denotes performance THROUGH UofU firewall

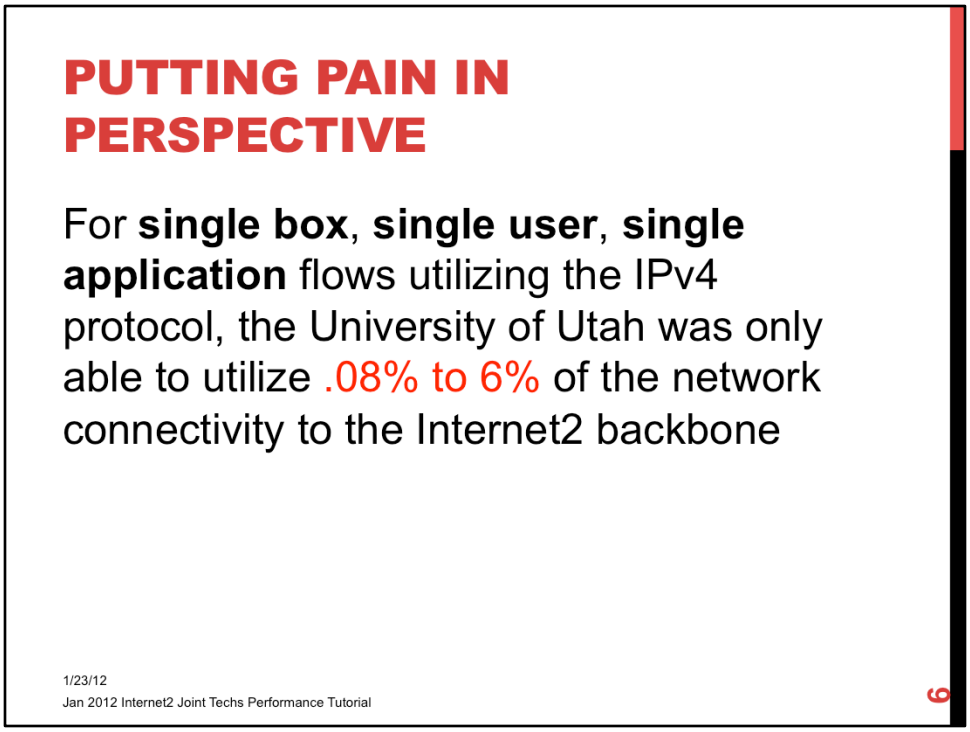
WHY? … PAIN!

50Megabit/second transfers from the Texas Advanced Computer Center (10Gig connectivity)

12Mbit/sec transfers from Fermi National Labs

6.7Mb/s transfers from Oak Ridge National Labs

- Started out looking at connections from UEN to the outside world and then moved back into the campus.
- Saw dramatic drop once within the campus border.

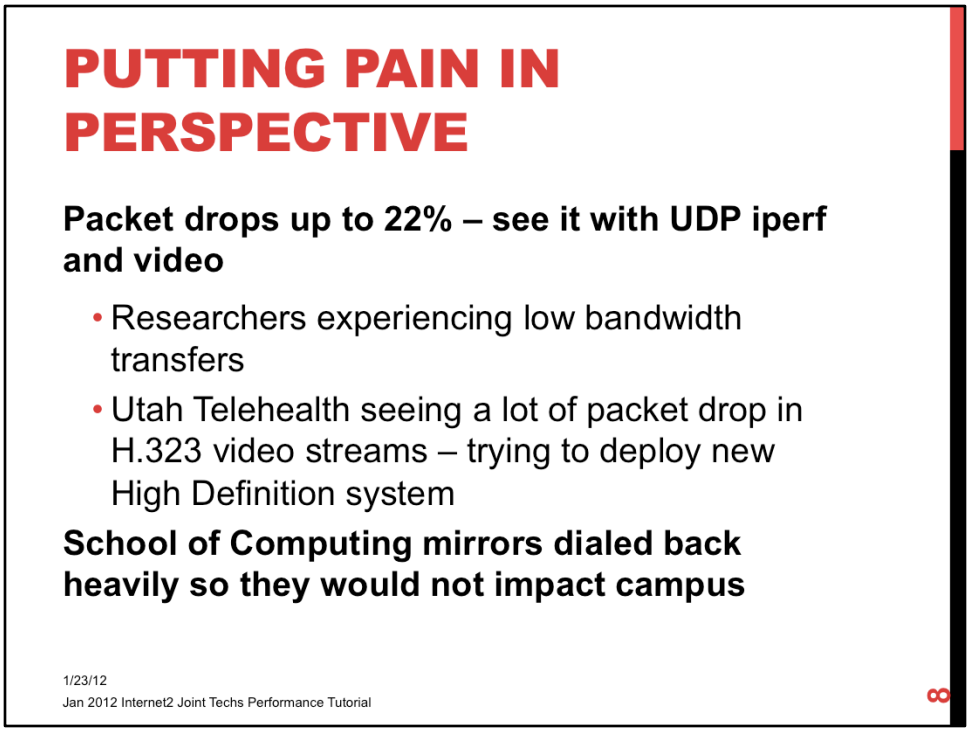


- Used iperf and FDT to test the baseline network and then file transfers.

PUTTING PAIN IN PERSPECTIVE

For multiple box, multiple user, multiple application flows, the Univ. of Utah was hitting ceilings of 20-30% of the available network bandwidth

- Used iperf and FDT to test the baseline network and then file transfers.
- Created multiple parallel flows, both from UEN's perspective and from within the University.



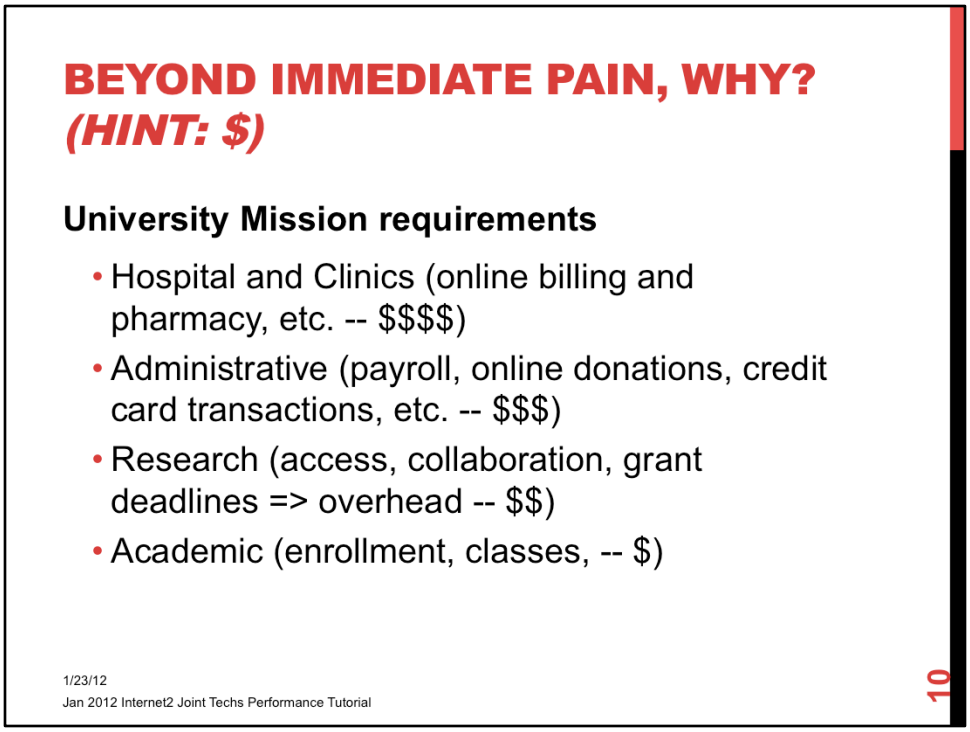
- We didn't start with all of this info at the beginning, we had to dig it up by looking at a lot of aspects of the network.
- Started with pain of large research transfers and kept digging. Utah Telehealth started researching their own issues in parallel.
- Campus saw School of Computing bury the existing firewalls when some of the Linux distros released another distribution. School of Computing wanted 10Gb/s but funding and a bit of concern held campus back from allowing the connectivity.

PUTTING PAIN IN PERSPECTIVE -- $

UofU/BYU/USU/UEN/Montana maintains 2x10Gb/s connection to Internet2 at $525k/yr

The performance issues were preventing the University of Utah from fully realizing the significant investments it is making in the network

- UofU has 10Gb/s+ backbone
- UofU has two 10Gig connections to UEN



- All billing and drug orders, medical records, etc. now handled online. When the network loses connectivity, the hospital has tangible records of $/min. loss of revenue. People get more than a little grumpy.
- Access to administrative payroll, online billing, online donations, credit card billing, etc. is all online. Less tangible records of lost revenue but still very visible.
- Access to research collaborators, ability to access national labs, ability to move data, ability to submit grants by deadlines, all rely on network stability. Tangible and intangible impacts to research overhead revenue.
- Academics rely on students finding a welcoming online presence. Online classes, online enrollment, online grading, homework submittal, etc. Most of these topics are intangible impacts to the University revenue but still impact it.

BEYOND IMMEDIATE PAIN, WHY? *(HINT: $)*

Diverging Business rule sets

- Research == Openness and Collaboration especially with data movement to national labs
- University Hospital and University Administrative business == closed and protected
  - PCI compliance -
  - HIPAA compliance - hospital, clinics
  - Compliance acronym of the week…

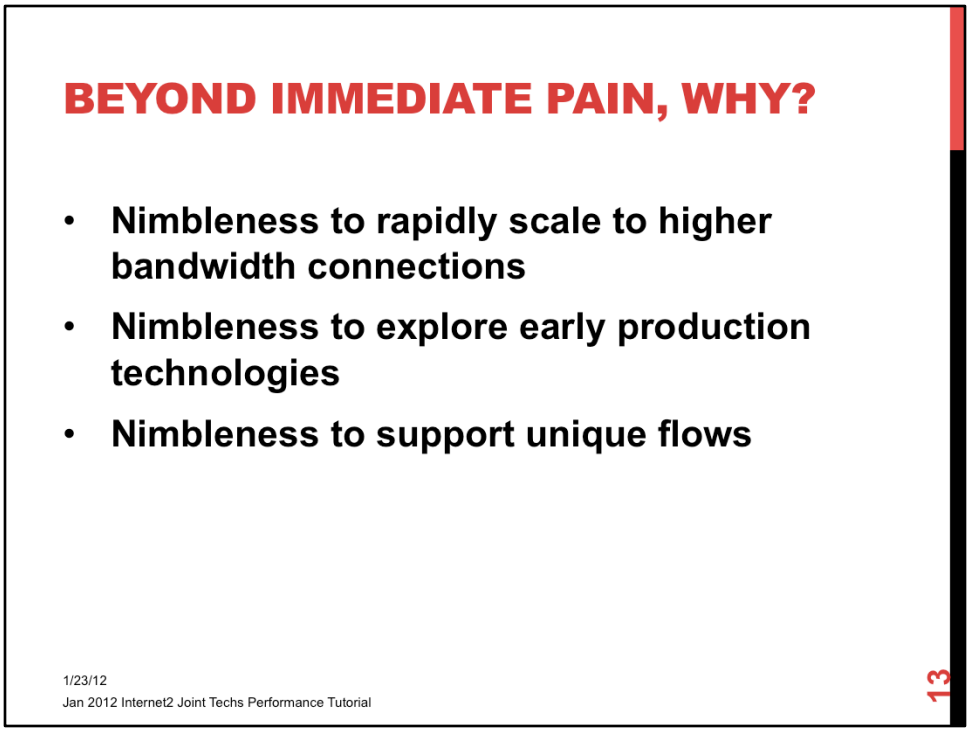
1/23/12
Jan 2012 Internet2 Joint Techs Performance Tutorial
11

The “station wagon” effect still rules – faster to send wagon full of DVDs, thumb drives or disks than to use the network.

BEYOND IMMEDIATE PAIN, WHY? *(HINT: $)*

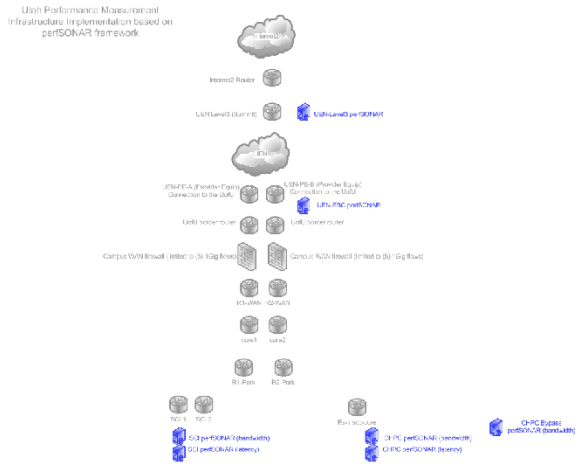
- **Operations:**
  - Longer amortization of redundant WAN equipment and redundant WAN firewalls → $

- Want to be able to connect to Internet2 at 100Gb/s within the next 1.5-3yrs. Amortization on the firewalls will be approximately 5yrs.



- Want to be able to connect to Internet2 at 100Gb/s within the next 1.5-3yrs. Amortization on the firewalls will be approximately 5yrs.
- Ability to prototype gear, i.e. new security gear, new network technology (think OpenFlow), in a pseudo-production environment. Past a development lab scenario but not quite prime-time for the main production network.
- Try to support unique flows, i.e. GENI implementations, that could pose a higher risk than the production environment is comfortable.

NOW THAT WE HAVE ESTABLISHED THE WHY, HOW?
FIRST, INSTRUMENT THE NETWORK.

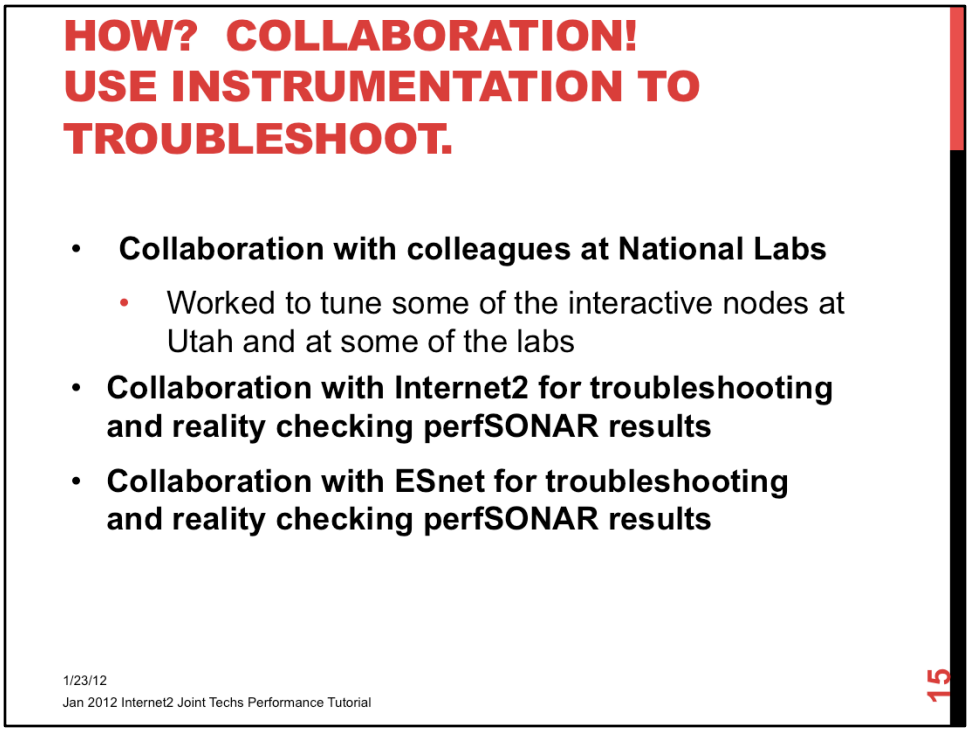


1/23/12

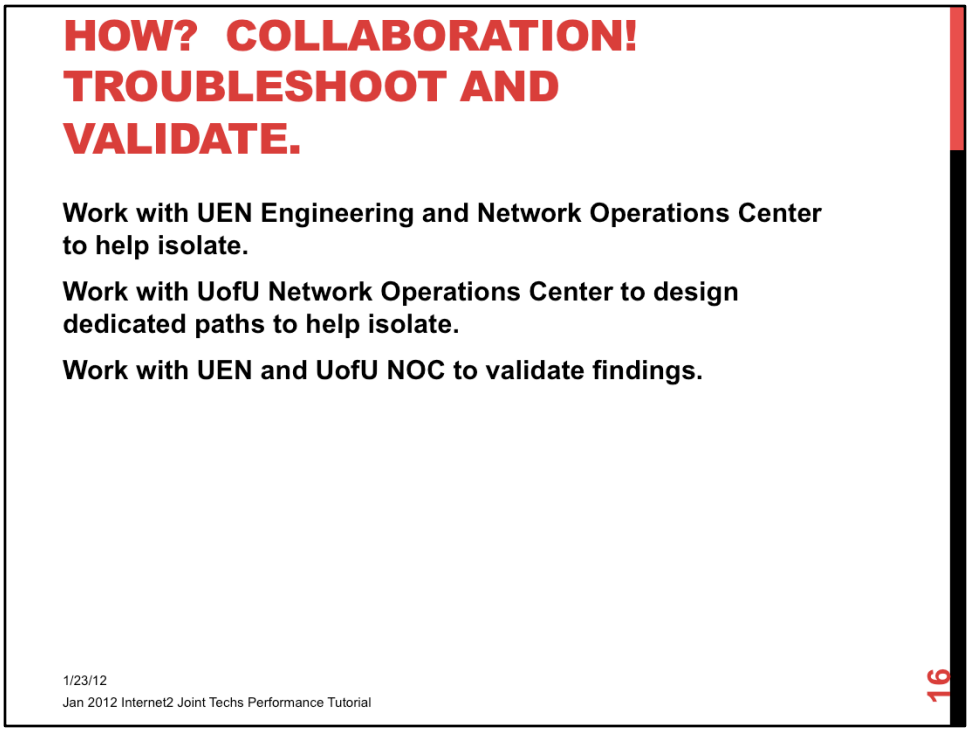
Jan 2012 Internet2 Joint Techs Performance Tutorial

14

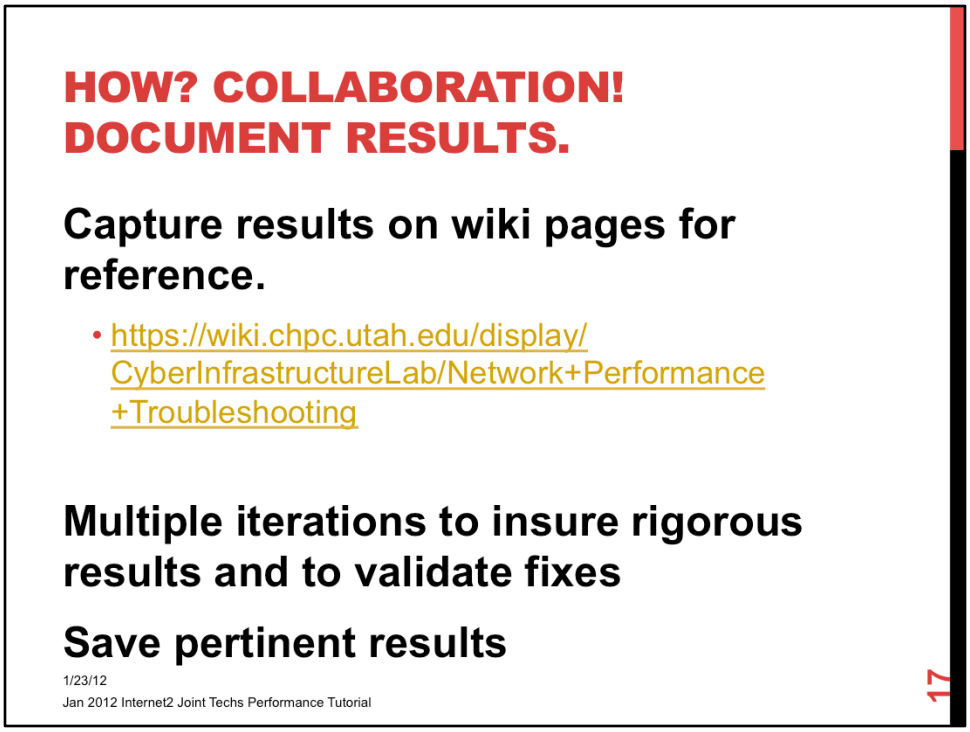
- Deployed perfsonar nodes in UEN
  - Immediately outside campus
  - Immediately before Internet2
- Deployed in campus space
  - Within CHPC
  - Within SCI
  - On bypass network



- Collaboration within the R&E community and leveraging the perfSONAR instrumentation is key to successful troubleshooting.



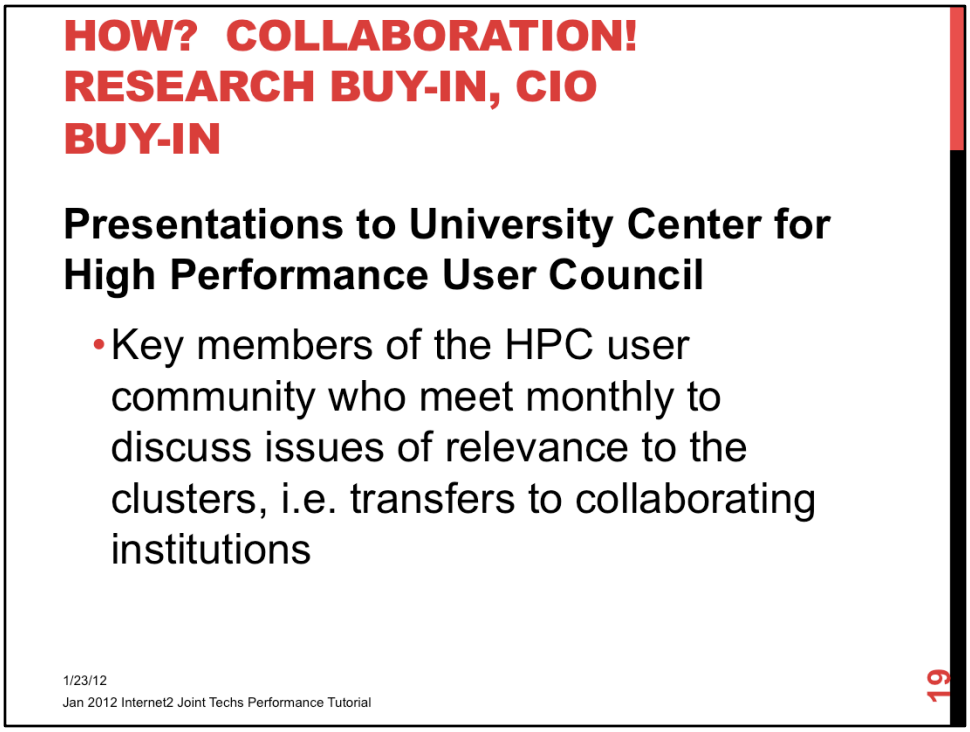
- Collaboration with campus entities and the regional network were key to localized troubleshooting of the campus and regional networks. The feedback from the various engineers and the multiple sets of eyes helped in faster isolation of issues.



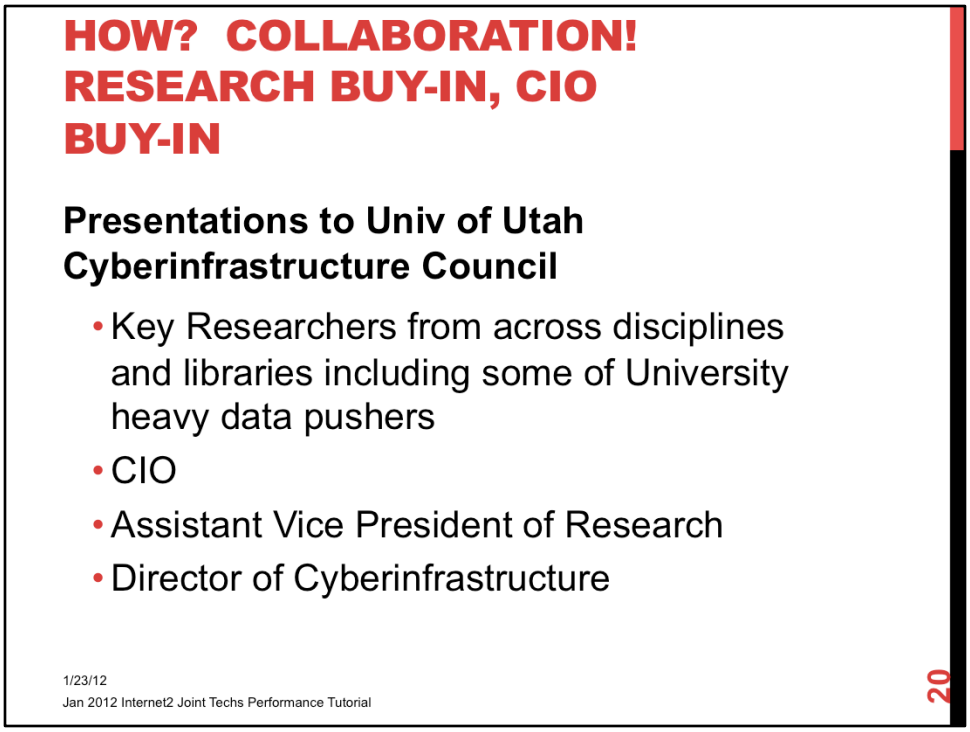
- Documenting notes on wiki really helped in putting together results that we could look back on and see improvement. Also helped when we saw things go worse. For example, we found out the firewalls were affecting IPv6 packets worse than IPv4 quite by accident. We did not realize that Internet2 had fixed some DNS records and our tests were utilizing DNS names instead of IP addresses. The traffic started using IPv6 instead of IPv4 because we had a full IPv6 path. Traffic took a dive.

HOW? COLLABORATION! DOCUMENT RESULTS.

- Leveraged info from http://fasterdata.es.net and slides from ESnet group
- UofU and UEN Team wrote up collaborative white paper - http://www.chpc.utah.edu/~jbreen/network/performance/2011-05-30_Network_Performance_Issues_at_the_University_of_Utah.pdf



* The HPC community is always looking for ways to improve data flow and get more from cycles. Several of the UofU researchers account for significant use of the national lab cycles. They were particularly sensitive to moving their data effectively.



- CI council includes following representation
  - head of Eccles Medical Library
  - head of university Marriott Library
  - dean of School of Architecture
  - School of Computing
  - Communications
  - Chair of Geography
  - University Information Technology Faculty representative
  - University Information Technology CIO
  - University Information Technology Director of Operations/Assistant CIO for hospital
  - College of Pharmacy
  - Chemical Engineering/ Institute for Clean & Secure Energy
  - Physics
  - Assistant Vice President Information Technology Health Sciences and Biomedical Informatics
  - Huntsman Cancer Institute
  - Vice President of Research
  - College of Engineering/Electrical Engineering/Assistant Vice President of Research
  - University Information Technology Director for Cyberinfrastructure

HOW? COLLABORATION! DESIGN POTENTIAL SOLUTIONS. MITIGATE RISKS. COMMUNICATE!

Now that the buy-in exists, how do we start putting the pieces together?

- Collaborate with team to identify solutions
- Collaborate with team to identify and mitigate risks
- Communicate!

HOW? COLLABORATION! DESIGN POTENTIAL SOLUTIONS. MITIGATE RISKS. COMMUNICATE!

- Work with UofU Information Security Office (ISO) to review thoughts and vulnerabilities
- Work with UofU Architecture to make adjustments to campus backbone directions
- Work with UofU NOC to design and implement campus backbone
- Work with UofU Compliance office to review and validate risk mitigation

- UEN and UofU are collaborating on metro optical network which will mitigate the single 10G link but it exists for now and is a bottleneck. Always important to work with the upstream provider and keep them in the loop regarding activities in which you may be experimenting. Otherwise, your local fast pipe may become an itty, bitty straw above you. PerfSONAR instrumentation helps in identifying some things. Lots of communication helps mitigate them.
- Having a good policy helps with clarification and understanding of all concerned. The policy also helps to give the security team some teeth and protection so they can work closely with the research community.

HOW? EDUCATE.

Educate community regarding tools, i.e. FDT, bbcp, GridFTP, etc.

- Continual process
- Still heavy use of scp, rsync, etc.

Implement optimized tools and make easy

- i.e. HPN ssh

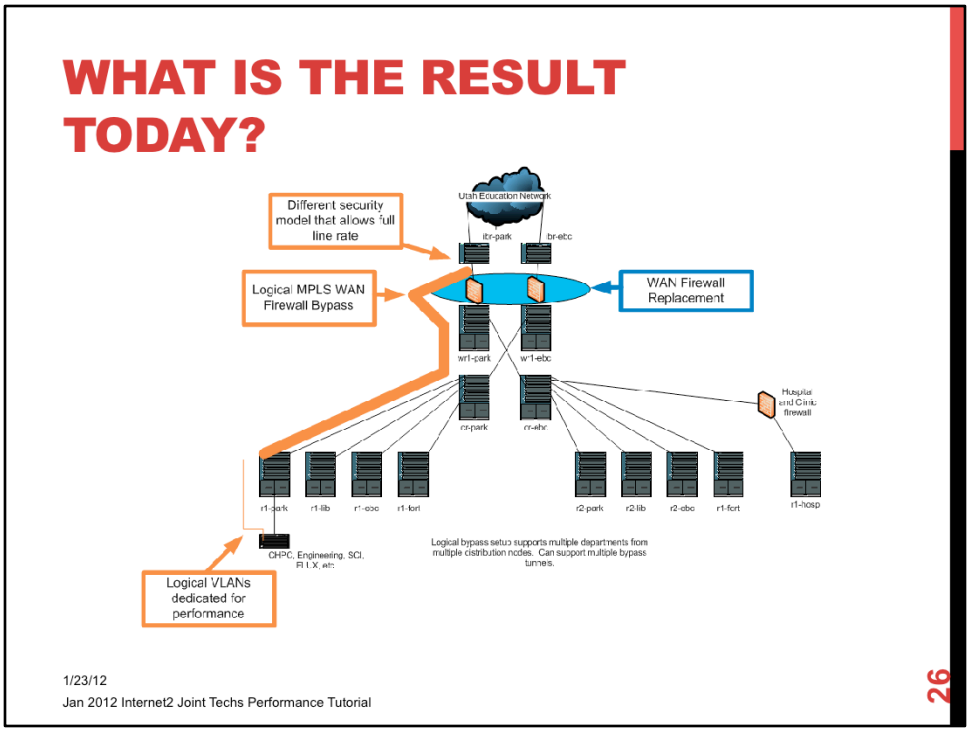
Use of parallel rsync streams somewhat effective

HOW? COLLABORATION! IMPLEMENT POTENTIAL SOLUTIONS AND PROTOTYPE ADDITIONAL TOOLS.

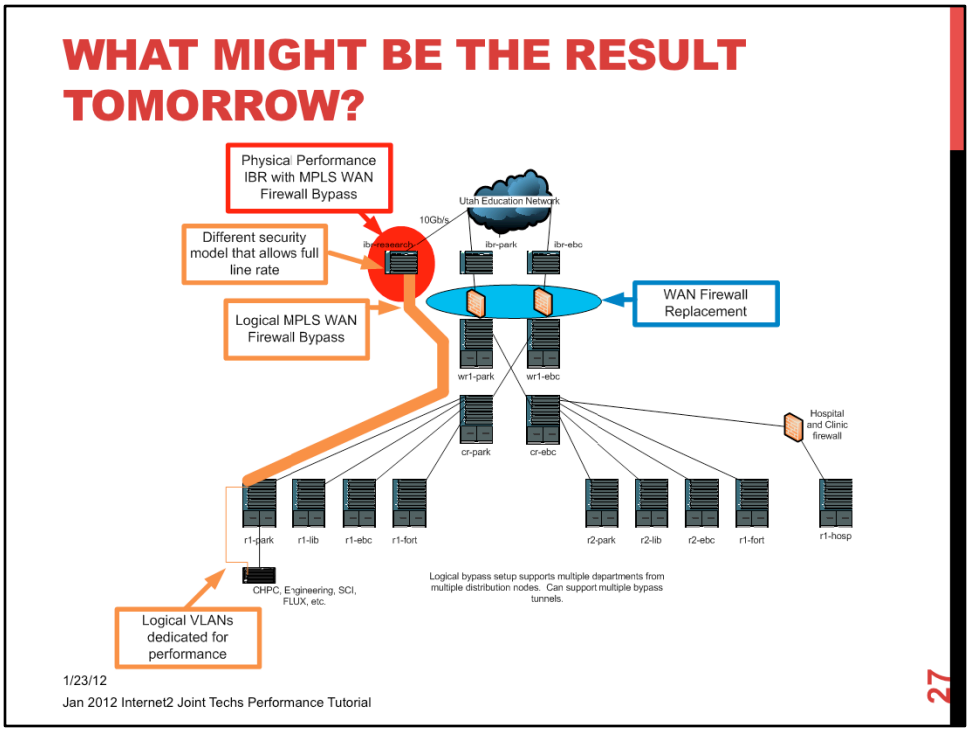
BGP Null Routing – scripting based on Netflow triggers by UofU security team and NOC

Out of band security – Bro prototype project happening now by UofU security team

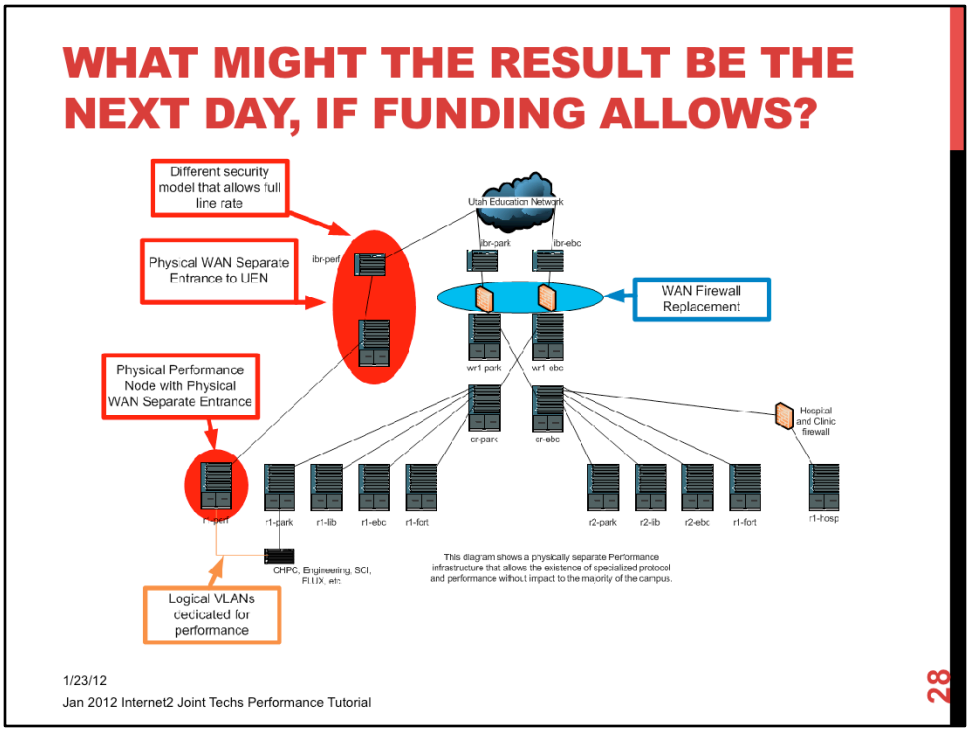
Exploration of additional mechanisms for protecting but simultaneously keeping out of the way.



- Partial snapshot of campus backbone with a MPLS tunnel providing a backbone path that goes from a distribution node, through the core to the WAN router, around the firewall and terminates traffic on the Internet Border Router. The traffic ingresses/egresses directly on the IBR and on the distribution router. The end customer provides own routing or routes on the distribution router.



- New physical IBR in order to separate the performance research/science DMZ network traffic from the rest of the U WAN traffic in order to mitigate risk. At first, the idea was to implement a performance distribution node first, but, the WAN is the higher risk, i.e. filling pipe or different security rule gone awry.



- New physical IBR in order to separate the performance research/science DMZ network traffic from the rest of the U WAN traffic in order to mitigate risk
- Complete separate infrastructure – NOT 5 nines, no dual-homing (under discussion), possibly different network vendor infrastructure.

ISSUES ALONG THE WAY

- **7yr firewall hardware**
  - Operational graphs did not reflect the packet drops and did not show the limited throughput.
  - Graphs of firewall throughput looked like existing firewalls were within expected parameters, though, some anomalies had arisen.

ISSUES ALONG THE WAY

Are you testing IPv4 or IPv6 with your active measurement infrastructure?

- Dual-stack is nice for servers but problematic for measurement infrastructure. What are you really testing?

Graphs not showing? What really is the path MTU?

- MPLS overhead causing mismatch in MTU, etc.
- New firewalls have different MTU max than previous firewalls.

ISSUES ALONG THE WAY

Ability to release the bottlenecks at University can potentially flood upstream provider – Make sure you are collaborating tightly!

- UEN has temporary single 10Gb/s feeding Level 3 PoP which houses Internet2 connectivity and multiple Commodity Internet connections.
- Waiting on metro optical network to relieve bottleneck.
- Filling research pipes causes commodity to slow down dramatically leading to some concern.

Resources available

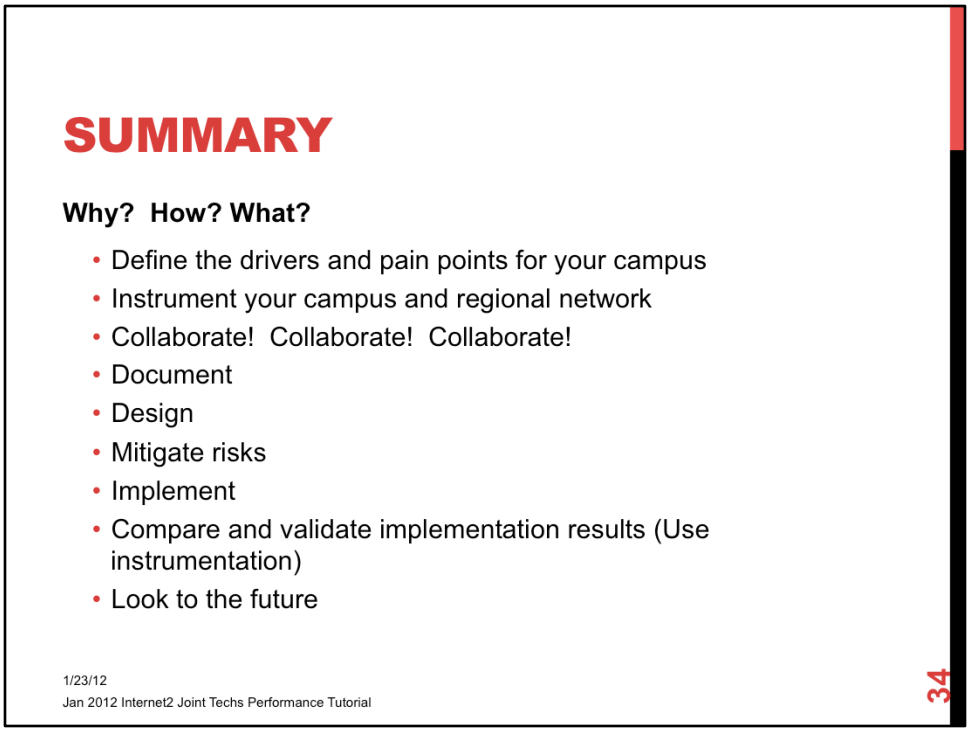
- Timing with major data center project
- Timing with other major projects

PANACEA? NOPE, AT LEAST NOT YET.

- Still working with MTU issues with new firewalls, MPLS tunnels and router settings
- Still educating and trying to get researchers to use high performance transfer tools
- Trying to finish policy
- Trying to obtain funding
- Still seeing changes in the world affect transfers

NEED DEVELOPMENT NETWORK TOO

- **Research/Science DMZ Network *NOT* a Network Development Sandbox**
  - Need pseudo-production focused on performance and unique flows
- **Need development sandbox testbed too**
  - Need to play with technologies such as OpenFlow in a network sandbox and then roll to the Research DMZ



- Make a list of issues that are affecting your campus
- Instrument your campus and regional network with perfSONAR
- Collaborate with your research community, your security group, your NOC, your Compliance group, your IT leadership, your regional NOC, your national backbone provider (I2/ESnet/etc.), your colleagues at peer institutions, …