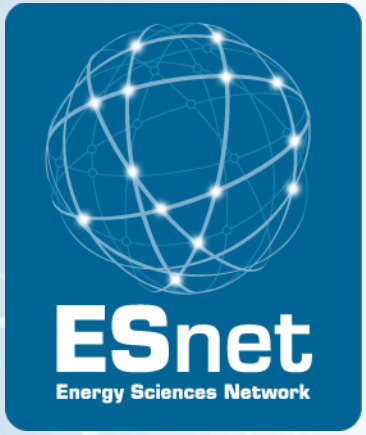


Achieving the Science DMZ

Eli Dart, Network Engineer

ESnet Network Engineering Group

Joint Techs, Winter 2012

Baton Rouge, LA

January 22, 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

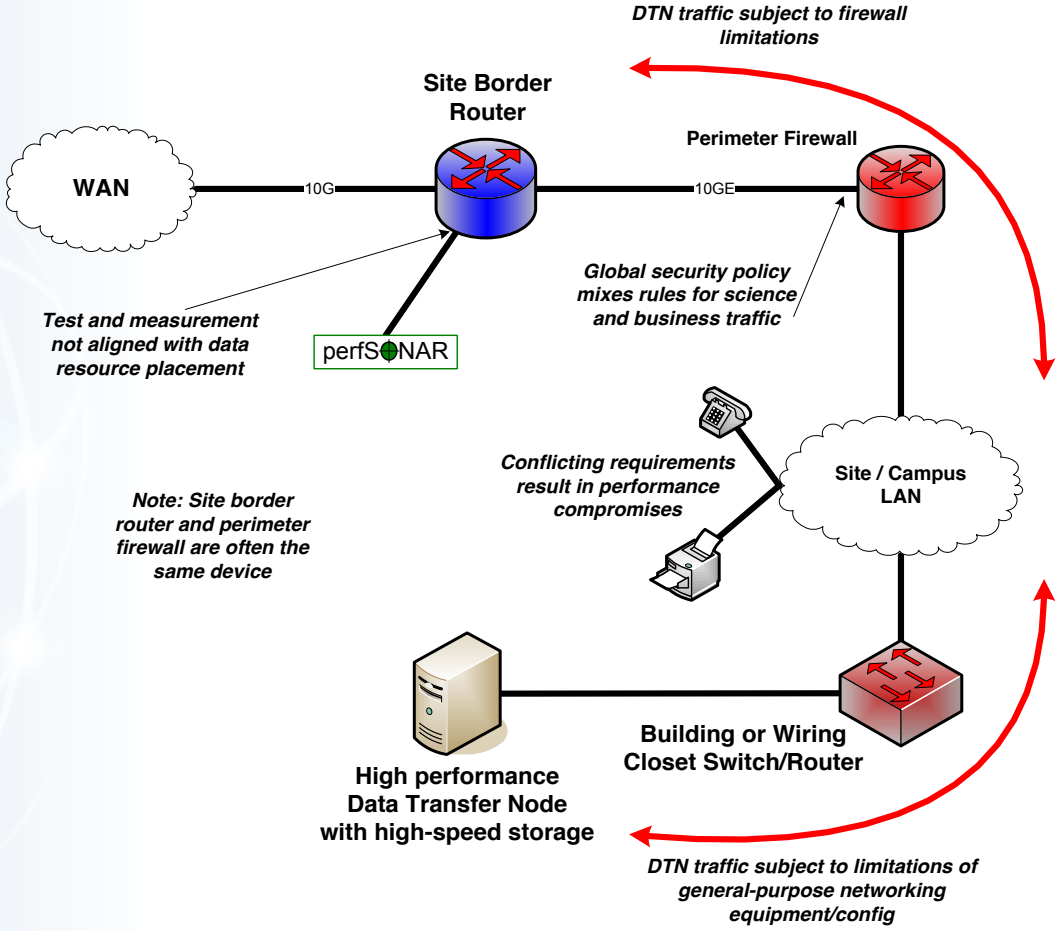
perfSONAR at the border

- 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

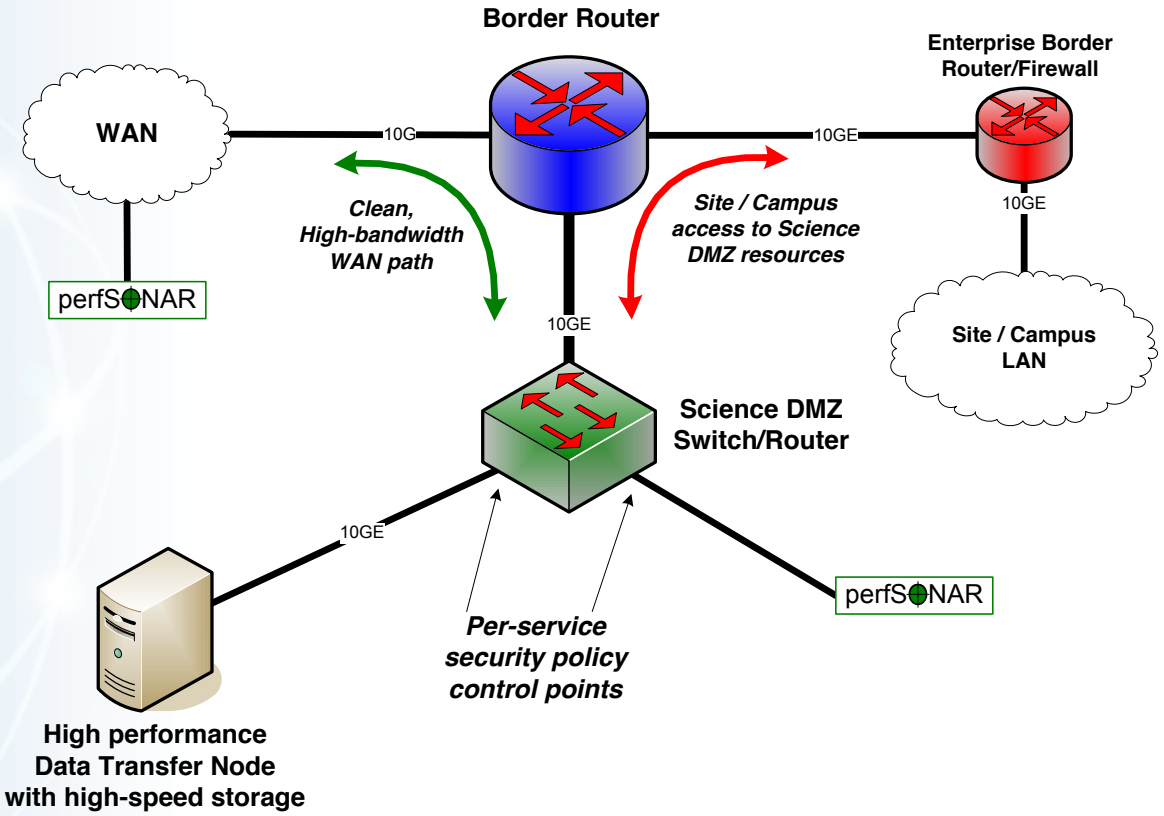
Easy to experiment with components and technologies

- DTN prototyping
- perfSONAR testing

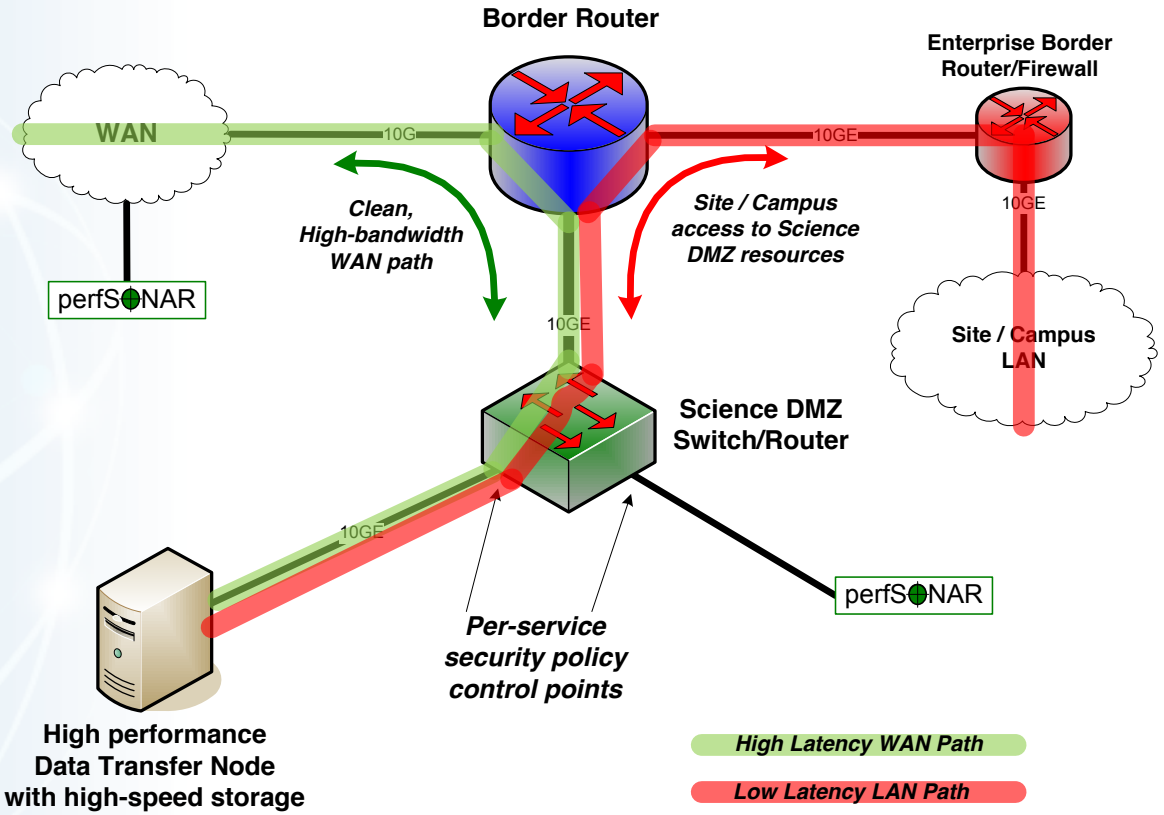
Limited scope makes security policy exceptions easy

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

Prototype Science DMZ



Prototype Science DMZ Data Path



Prototype With Virtual Circuits

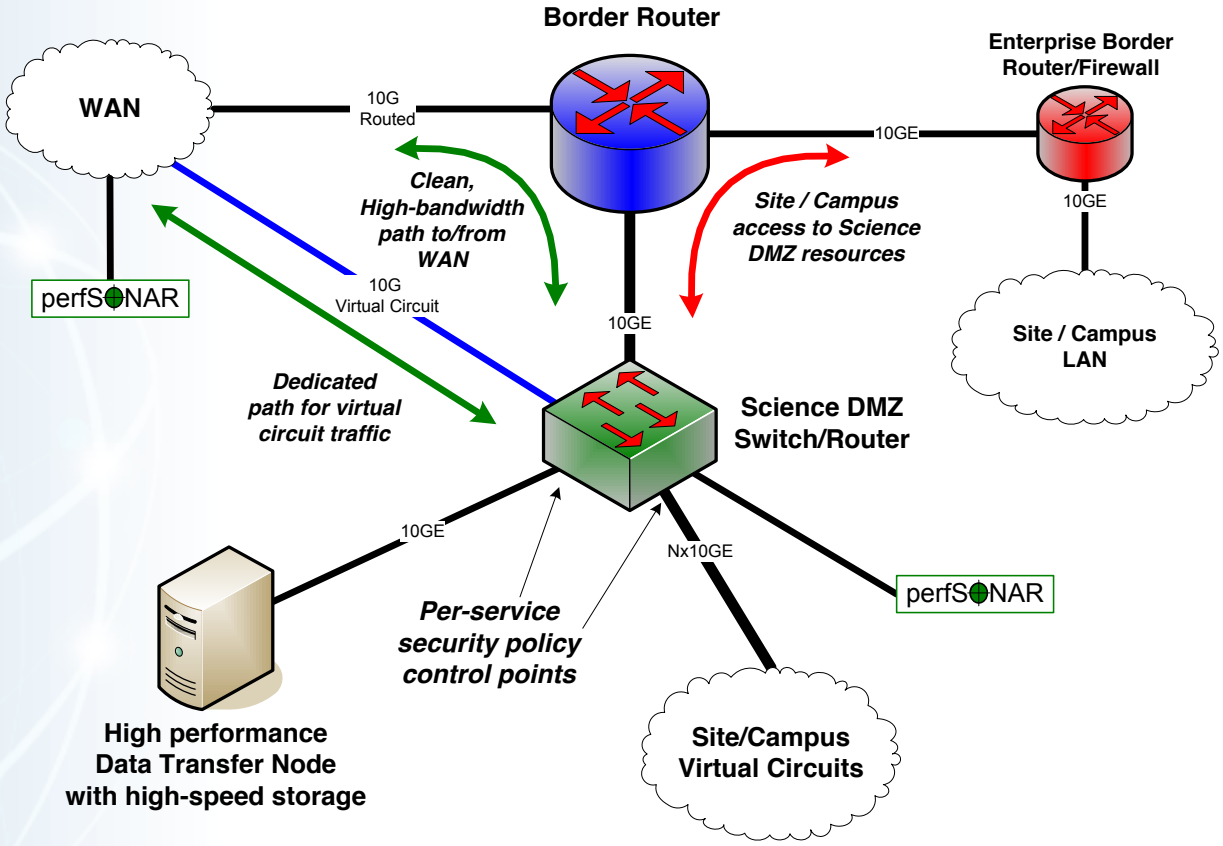
Small virtual circuit prototype can be done in a small Science DMZ

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

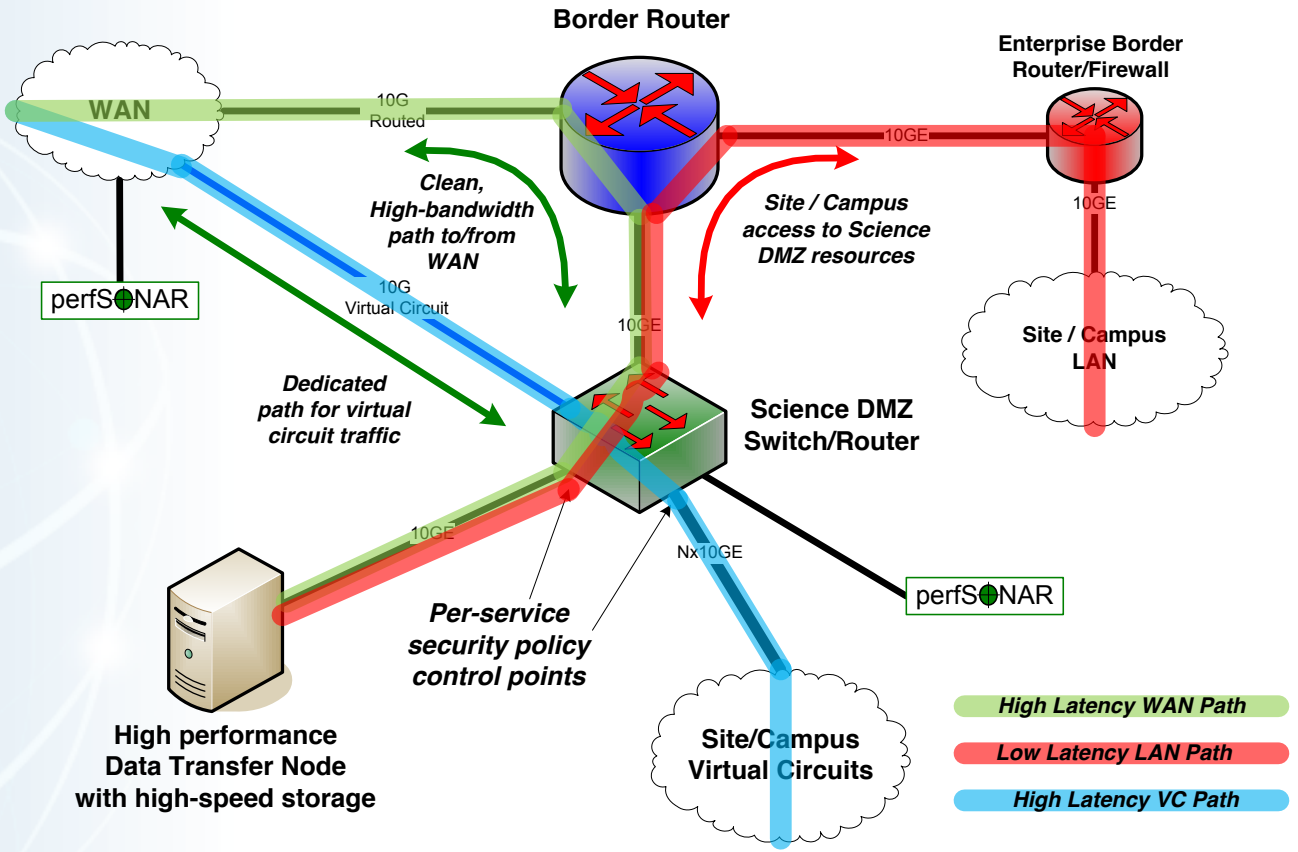
As with any Science DMZ deployment, this can be expanded as need grows

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

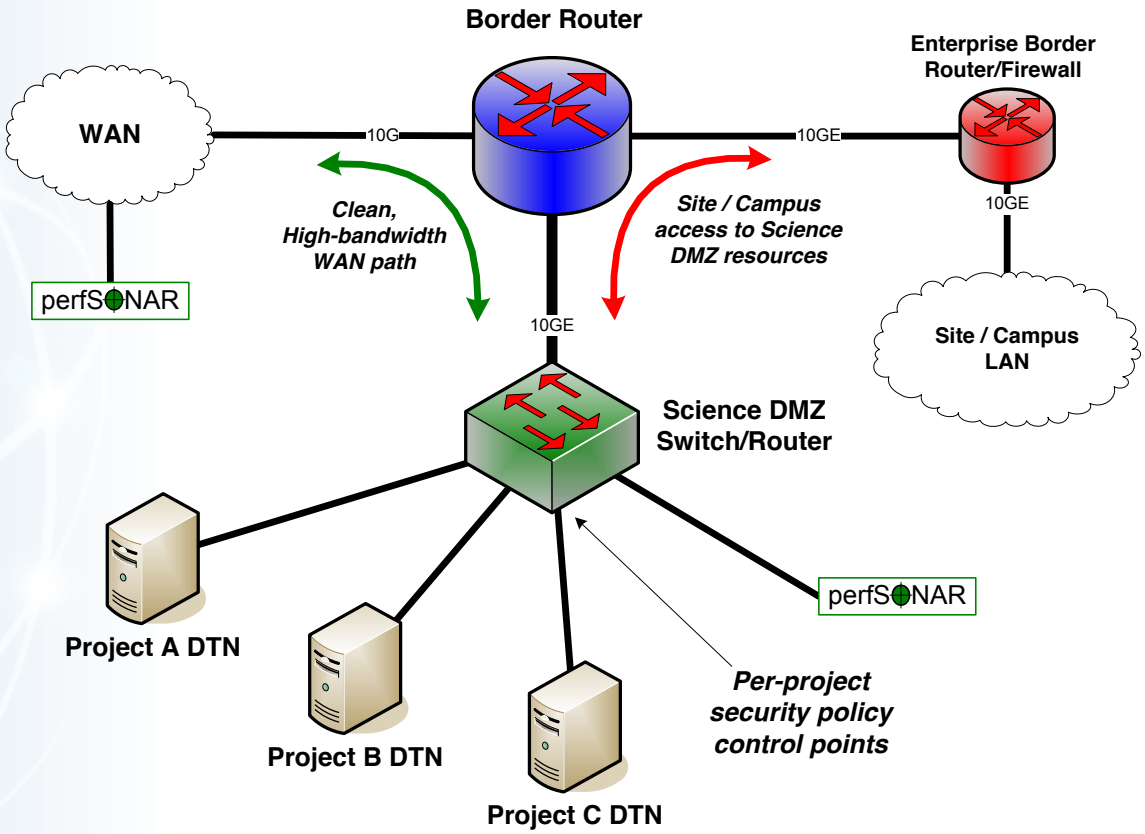
This may or may not work well depending on institutional politics

- 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

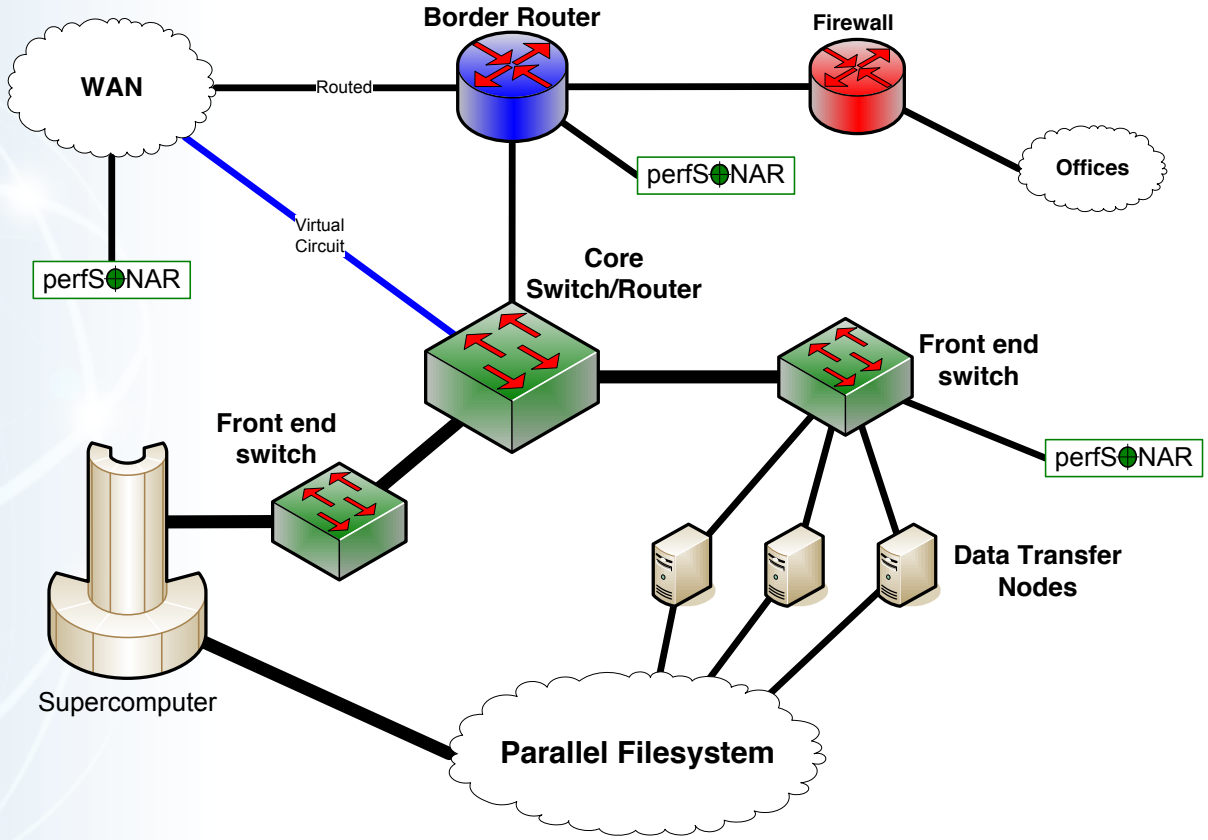
"Science DMZ" may not look like a discrete entity here

- 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.

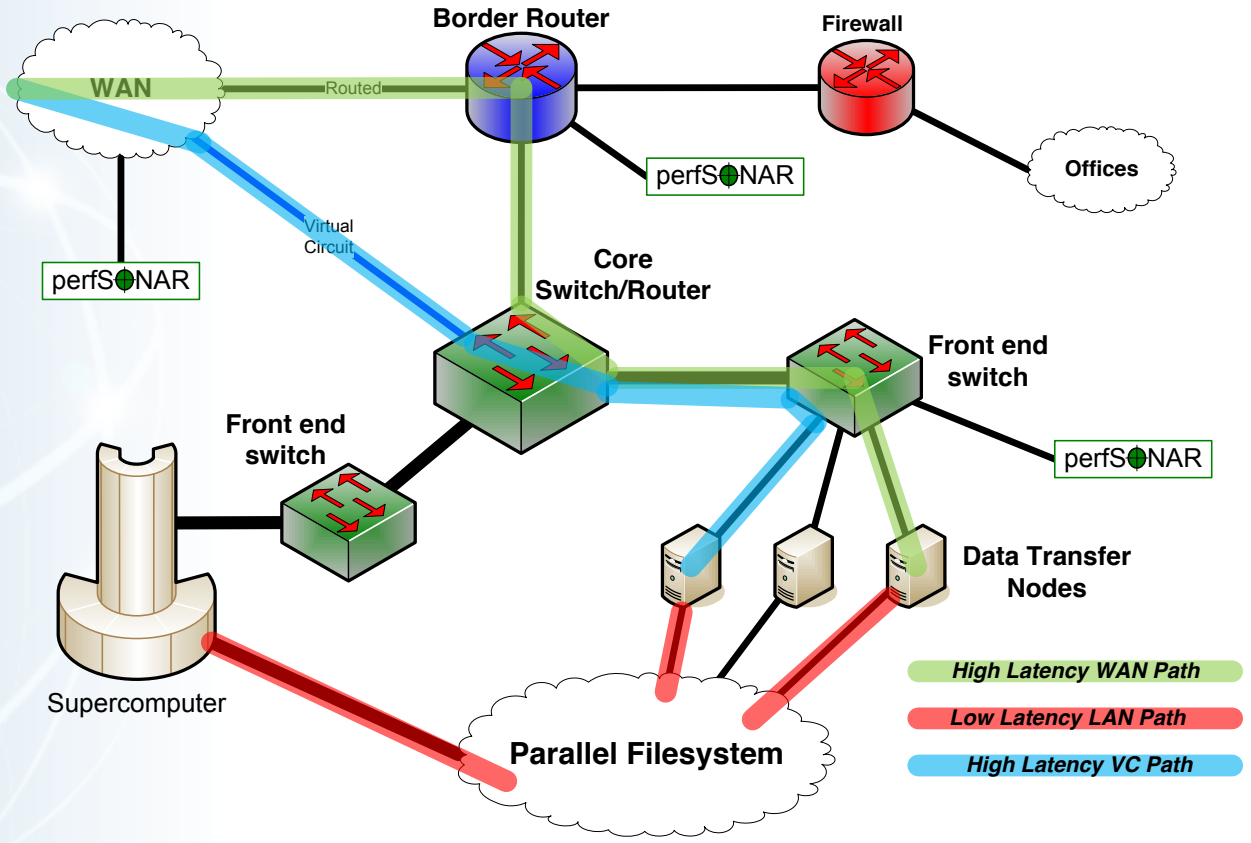
Office networks can look like an afterthought, but they aren't

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

Supercomputer Center



Supercomputer Center Data Path



Major Data Site Deployment

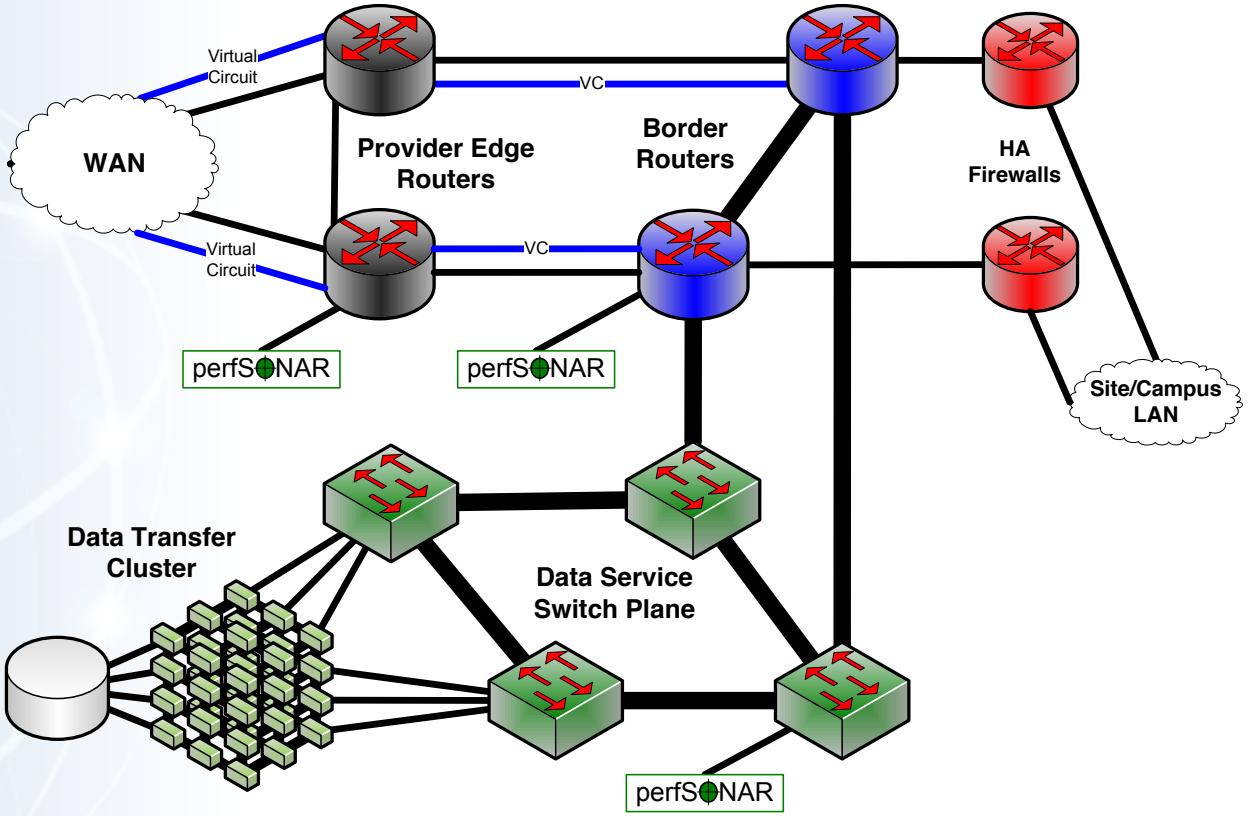
In some cases, large scale data service is the major driver

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

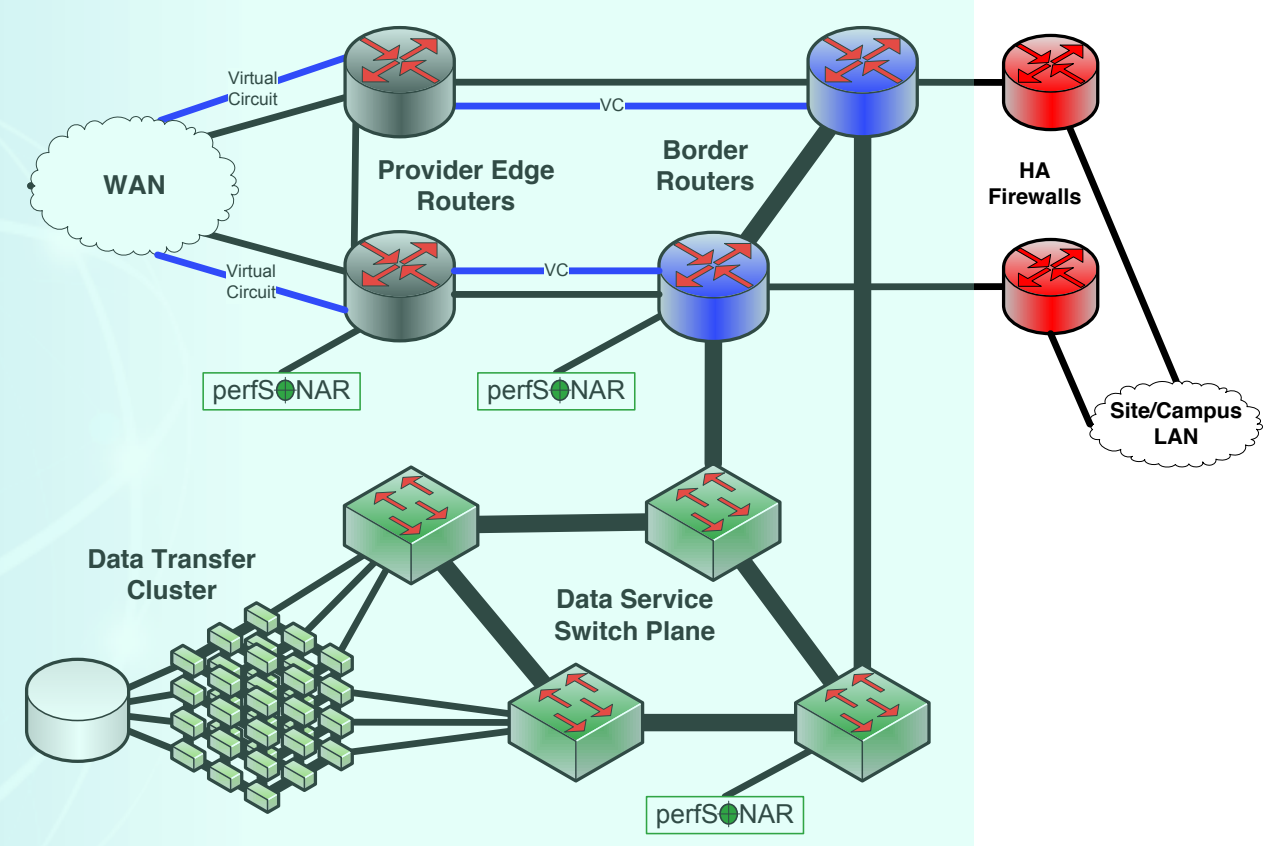
Single-pipe deployments don't work

- 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

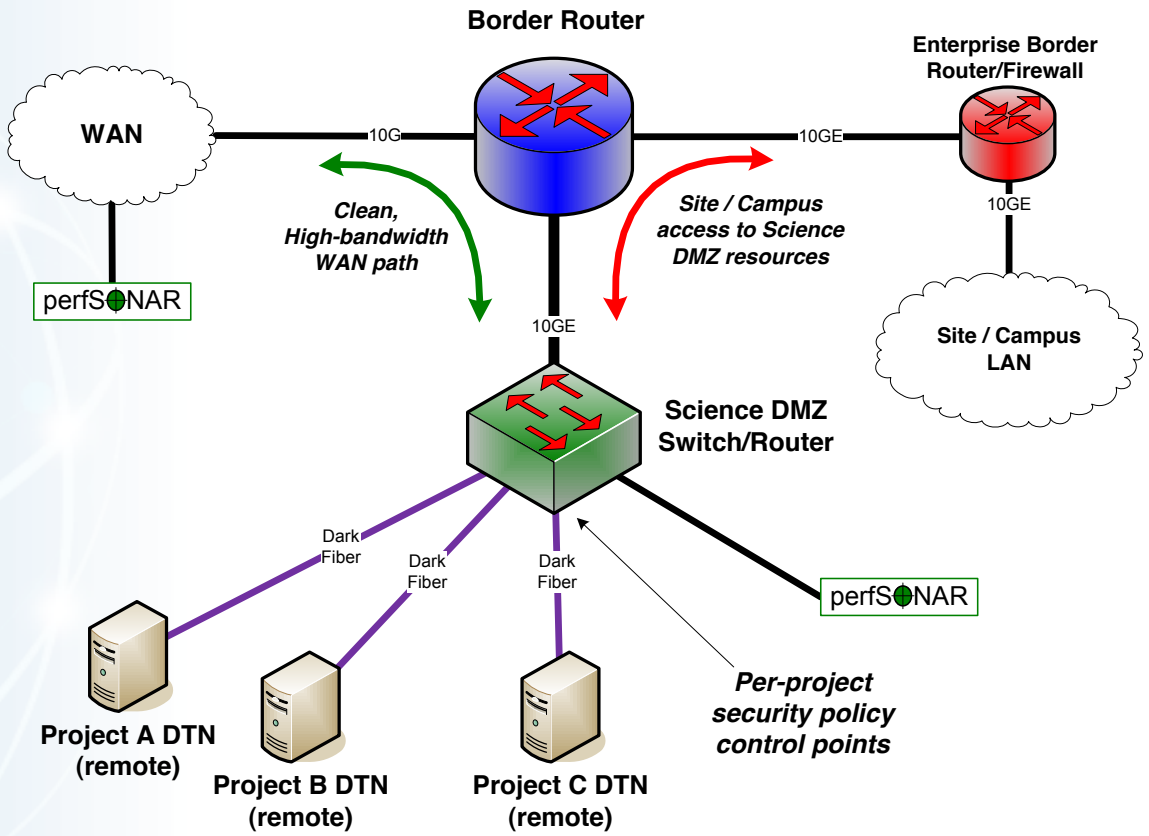
Dark fiber distributes connectivity to Science DMZ services throughout the site

- 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

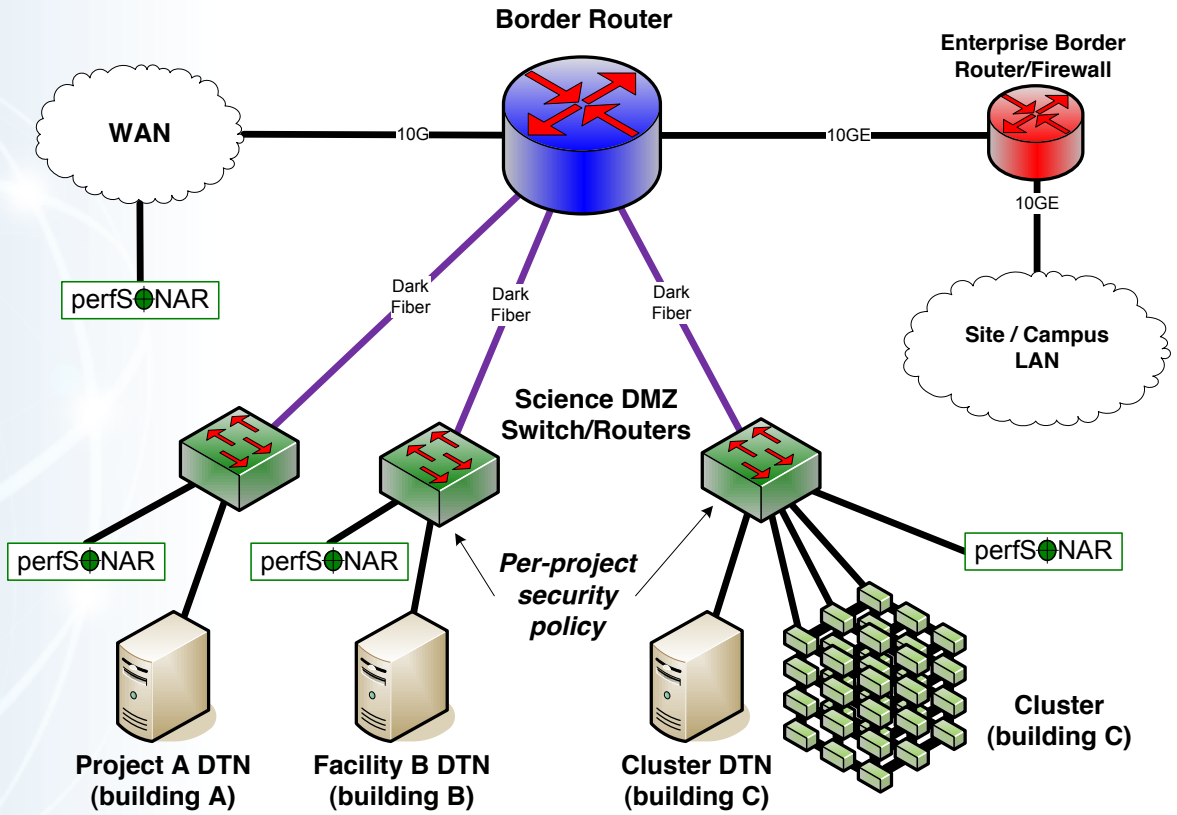
Security is made more complex

- 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

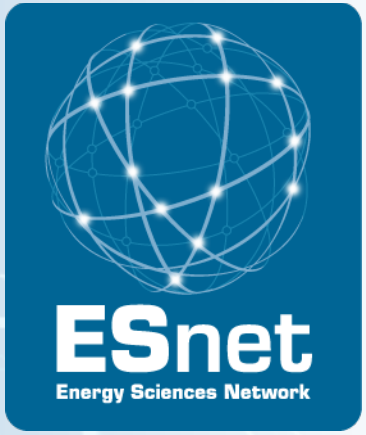
Ability to support science that might otherwise be impossible

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



Questions?

Thanks!

Eli Dart - dart@es.net

http://www.es.net/

http://fasterdata.es.net/