http://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/

CC* Data Movement Workshop and Exhibition – Science DMZ & Security

Jason Zurawski, Eli Dart
zurawski@es.net, dart@es.net
ESnet / Lawrence Berkeley National Laboratory

Dr. Jennifer M. Schopf
jmschopf@indiana.edu
Indiana University International Networks

CC/CICI PI Meeting Pre-Workshop*
September 22nd 2019

National Science Foundation Award #1826994

https://epoc.global

Motivation

- Networks are an essential part of data-intensive science
  - Connect data sources to data analysis
  - Connect collaborators to each other
  - Enable machine-consumable interfaces to data and analysis resources (e.g. portals), automation, scale
- Performance is critical
  - Exponential data growth
  - Constant human factors
  - Data movement and data analysis must keep up
- Effective use of wide area (long-haul) networks by scientists has historically been difficult

EPOC
Engagement and Performance Operations Center

The Central Role of the Network

- The very structure of modern science assumes science networks exist: high performance, feature rich, global scope
- What is “The Network” anyway?
  - “The Network” is the set of devices and applications involved in the use of a remote resource
    - This is not about supercomputer interconnects
    - This is about data flow from experiment to analysis, between facilities, etc.
  - User interfaces for “The Network” – portal, data transfer tool, workflow engine
  - Therefore, servers and applications must also be considered
- What is important? Ordered list:
  1. Correctness
  2. Consistency
  3. Performance

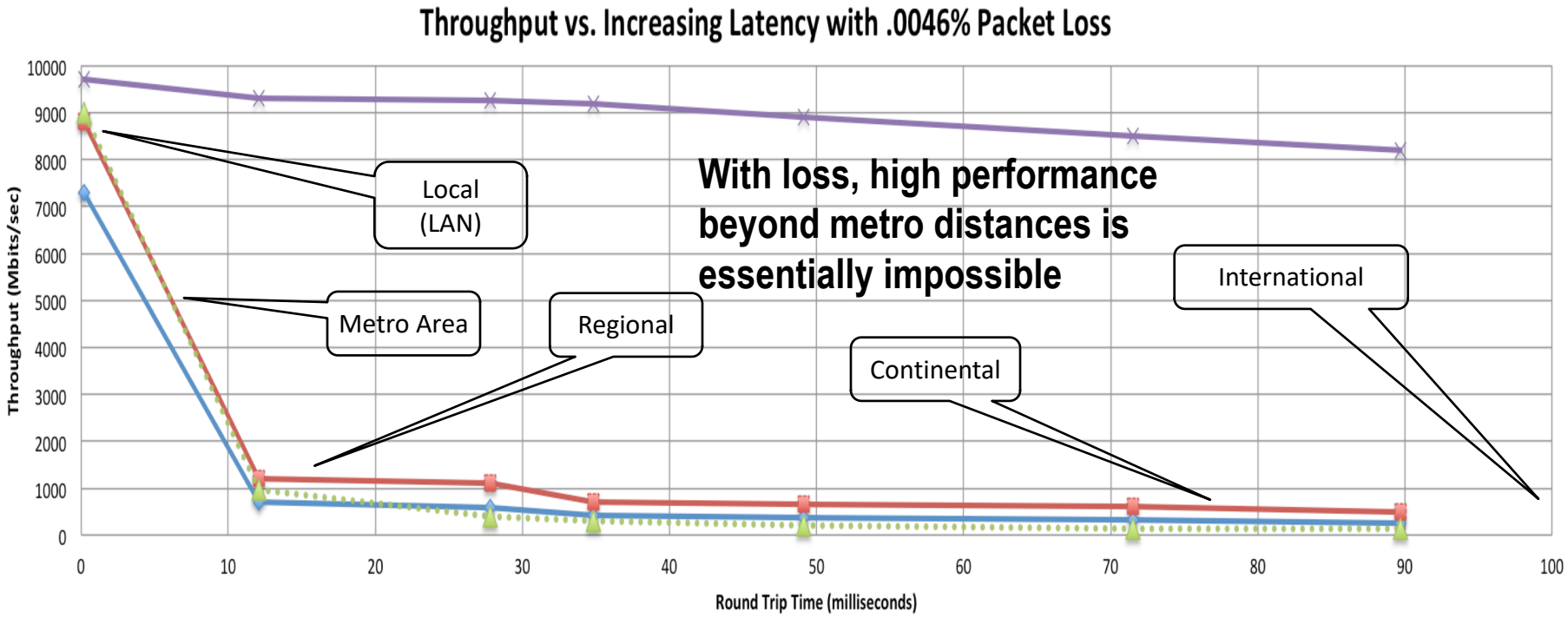
EPOC
Engagement and Performance Operations Center

TCP – Ubiquitous and Fragile

- Networks provide connectivity between applications on hosts – how do they see the network?
  - From an application's perspective, the interface to "the other end" is a socket
  - 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 (more than 95% of ESnet traffic is TCP)

EPOC
Engagement and Performance Operations Center

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



Working With TCP In Practice

- Far easier to support TCP than to fix TCP
  - People have been trying to fix TCP for years – limited success
  - Like it or not we're stuck with TCP in the general case
- Pragmatically speaking, we must accommodate TCP
  - Sufficient bandwidth to avoid congestion
  - Zero packet loss
  - Verifiable infrastructure
    - Networks are complex
    - Must be able to locate problems quickly
    - Small footprint is a huge win – small number of devices so that problem isolation is tractable

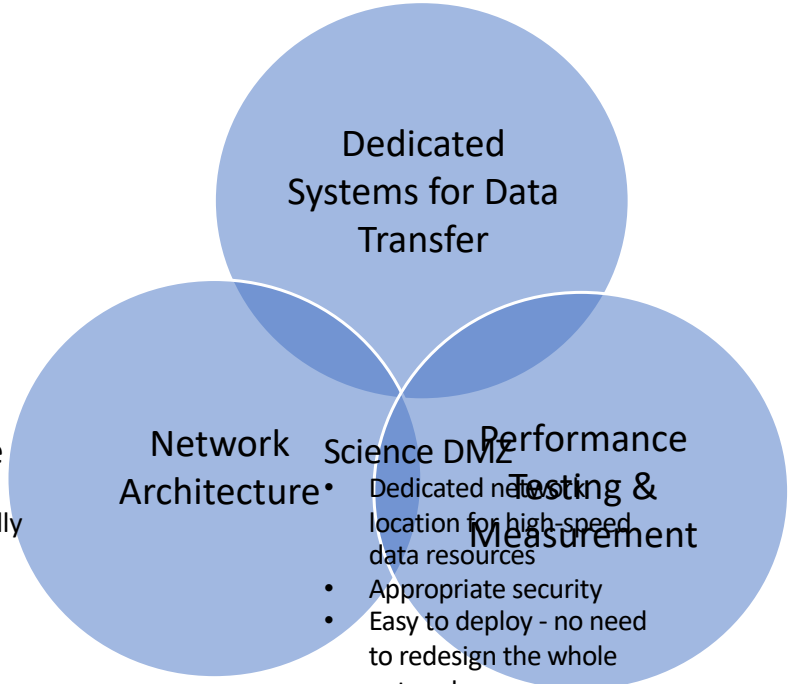
EPOC
Engagement and Performance Operations Center

Putting A Solution Together

- Effective support for TCP-based data transfer
  - Design for correct, consistent, high-performance operation
  - Design for ease of troubleshooting
- Easy adoption is critical
  - Large laboratories and universities have extensive IT deployments
  - Drastic change is prohibitively difficult
- Cybersecurity – defensible without compromising performance
- Borrow ideas from traditional network security
  - Traditional DMZ
    - Separate enclave at network perimeter ("Demilitarized Zone")
    - Specific location for external-facing services
    - Clean separation from internal network
  - Do the same thing for science – ***Science DMZ***

EPOC
Engagement and Performance Operations Center

The Science DMZ Design Pattern



Data Transfer Node
- High performance
- Configured specifically for data transfer
- Proper tools

Science DMZ
- Dedicated network location for high-speed data resources
- Appropriate security
- Easy to deploy - no need to redesign the whole network

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

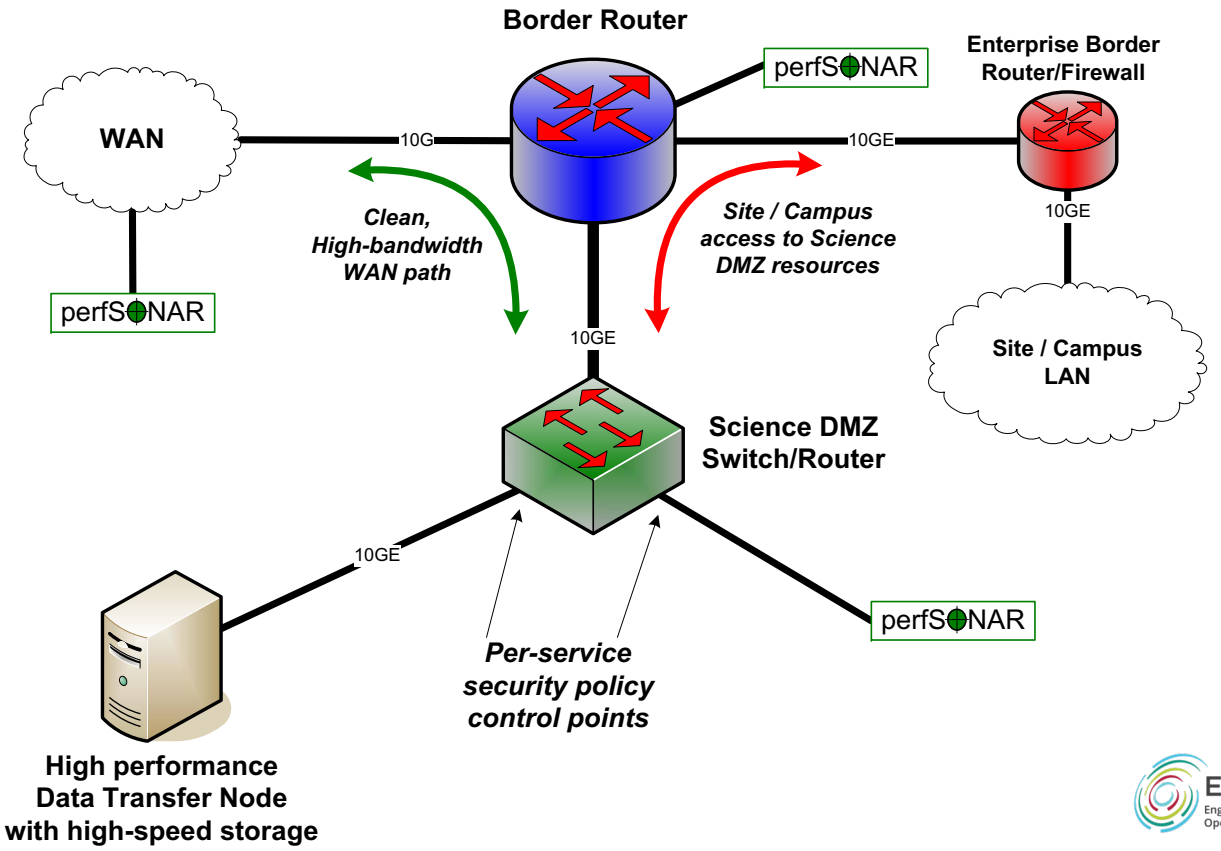
EPOC
Engagement and Performance Operations Center

Abstract 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

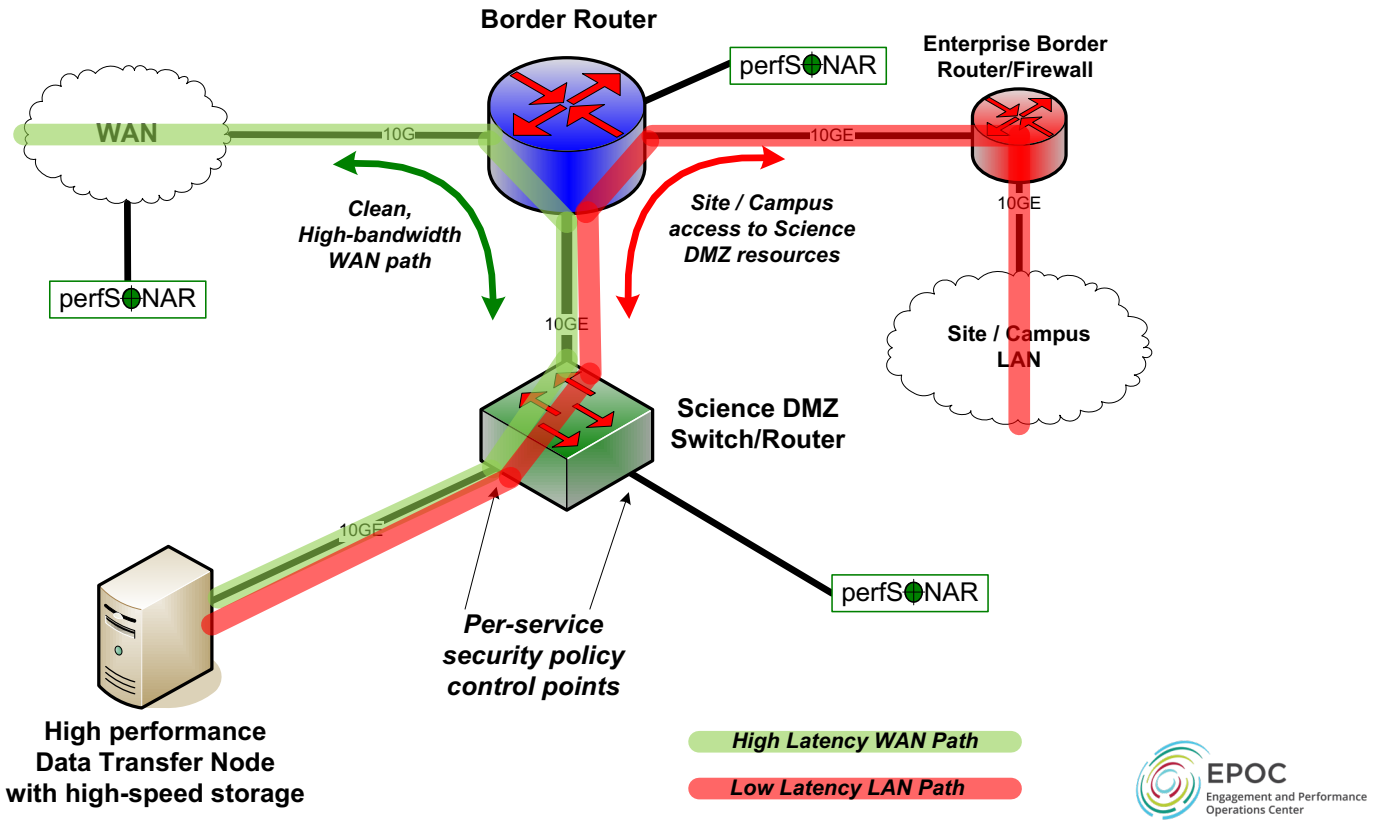
EPOC
Engagement and Performance Operations Center

Science DMZ Design Pattern (Abstract)



EPOC
Engagement and Performance Operations Center

Local And Wide Area Data Flows

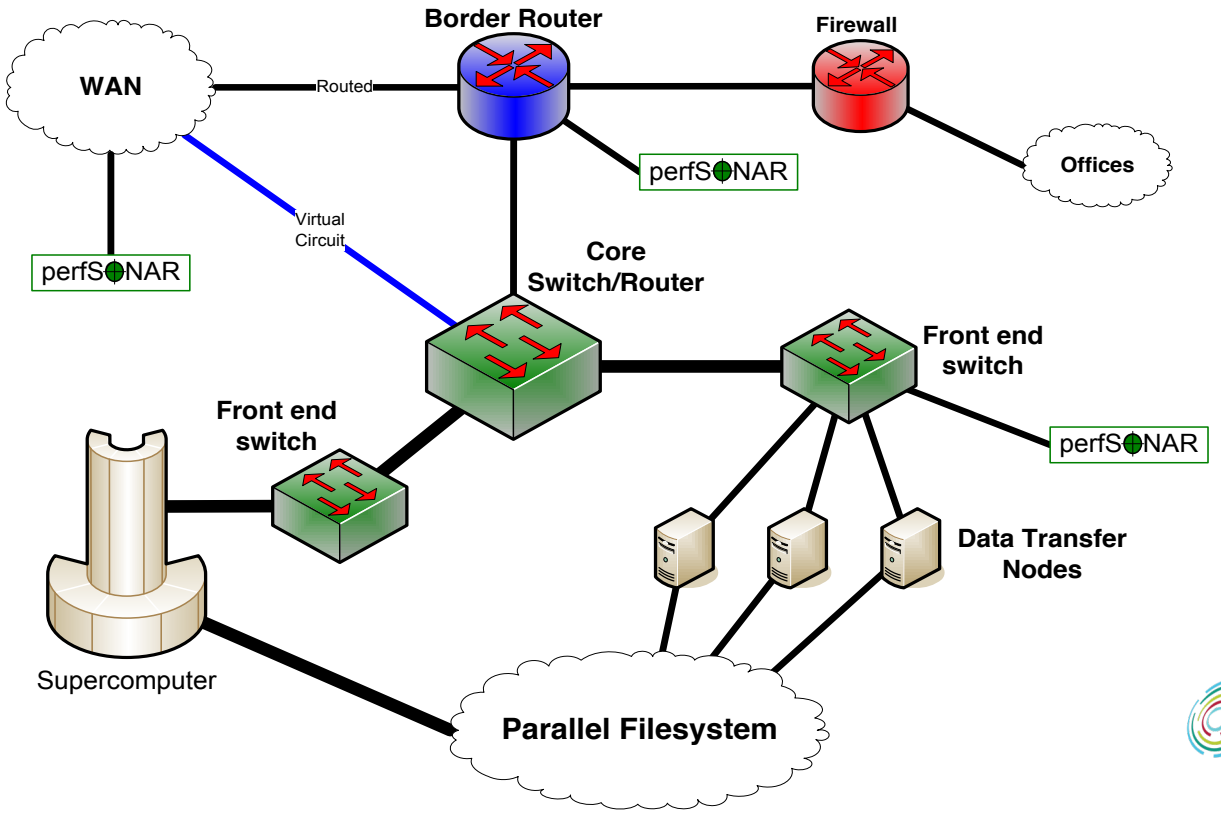


Supercomputer Center Deployment

- 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

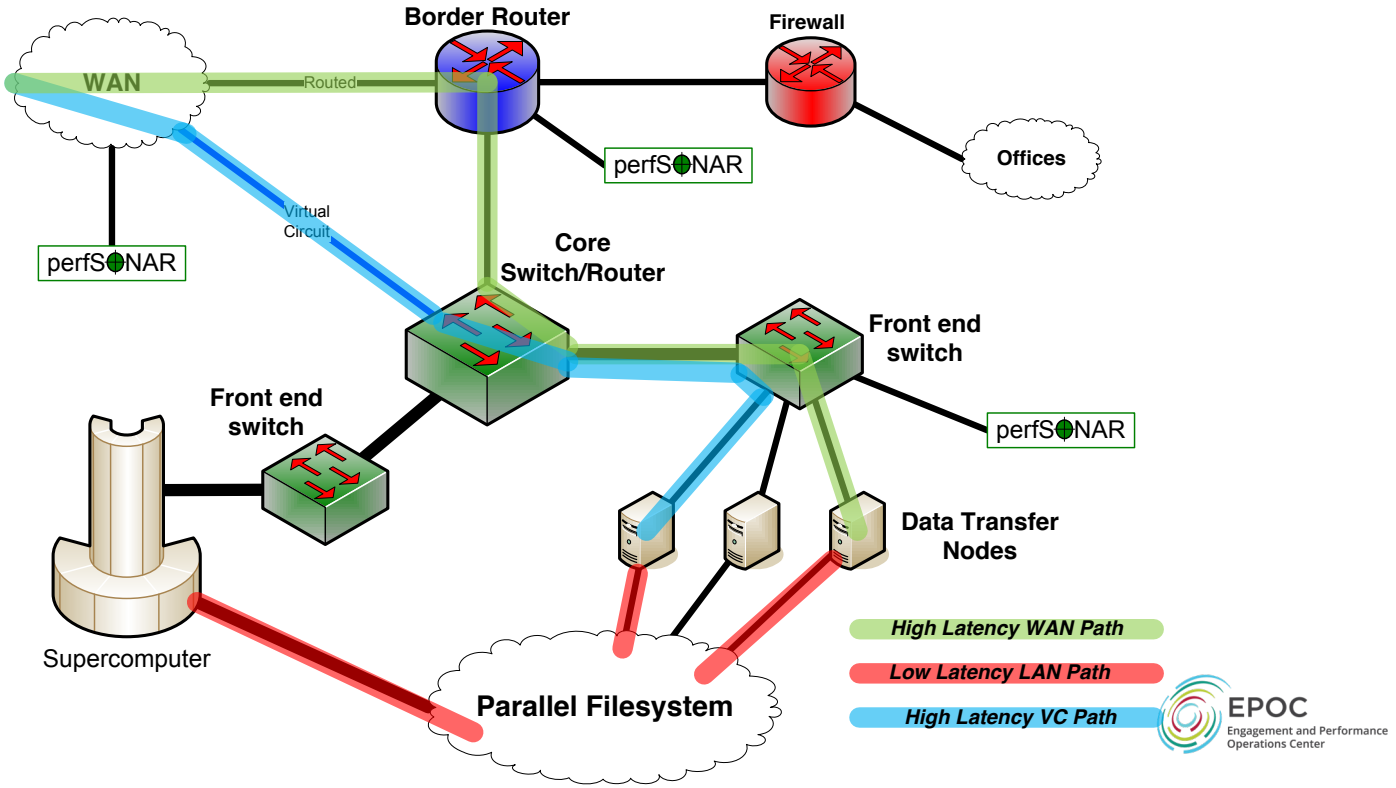
EPOC
Engagement and Performance Operations Center

Supercomputer Center



EPOC
Engagement and Performance Operations Center

Supercomputer Center Data Path

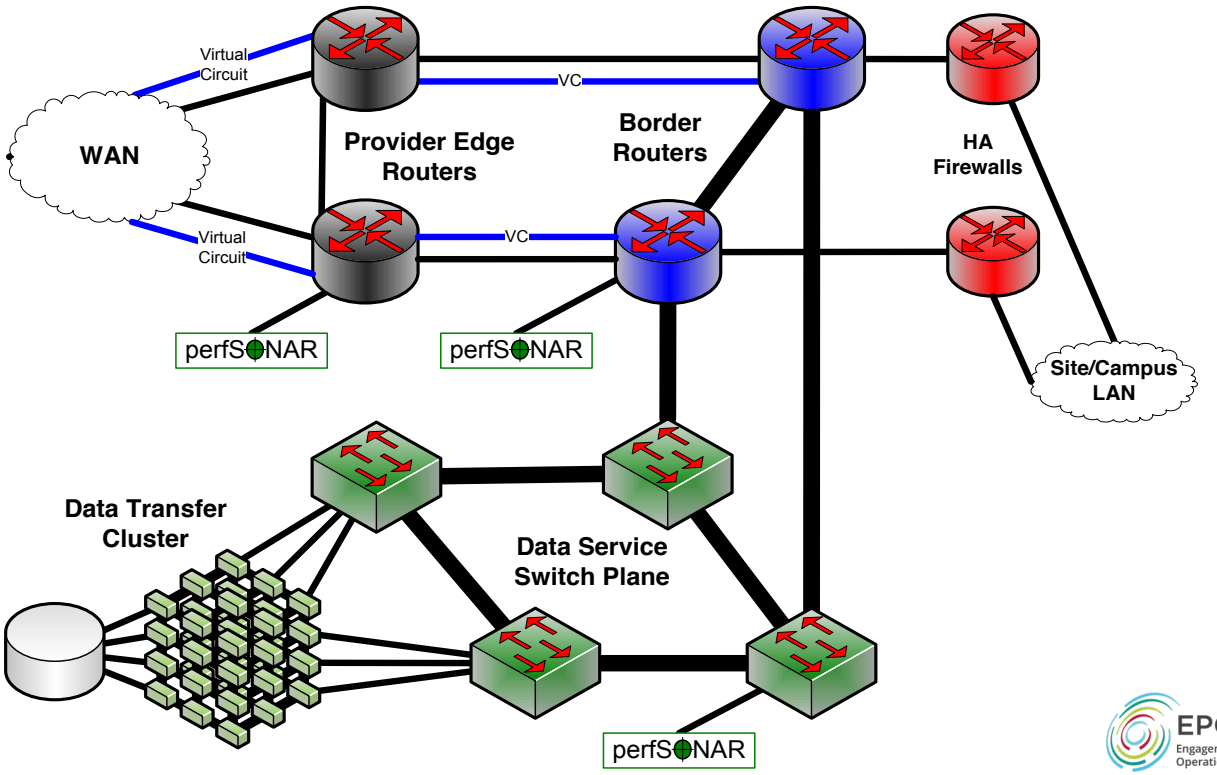


Major Data Site Deployment

- In some cases, large scale data service is the major driver
  - Huge volumes of data (Petabytes or more) – ingest, export
  - Large number of external hosts accessing/submitting data
- Single-pipe deployments don't work
  - Everything is parallel
    - Networks (Nx100G LAGs, soon to be Nx200G or Nx400G)
    - Hosts – data transfer clusters, no individual DTNs
    - WAN connections – multiple entry, redundant equipment
  - Choke points (e.g. firewalls) just cause problems

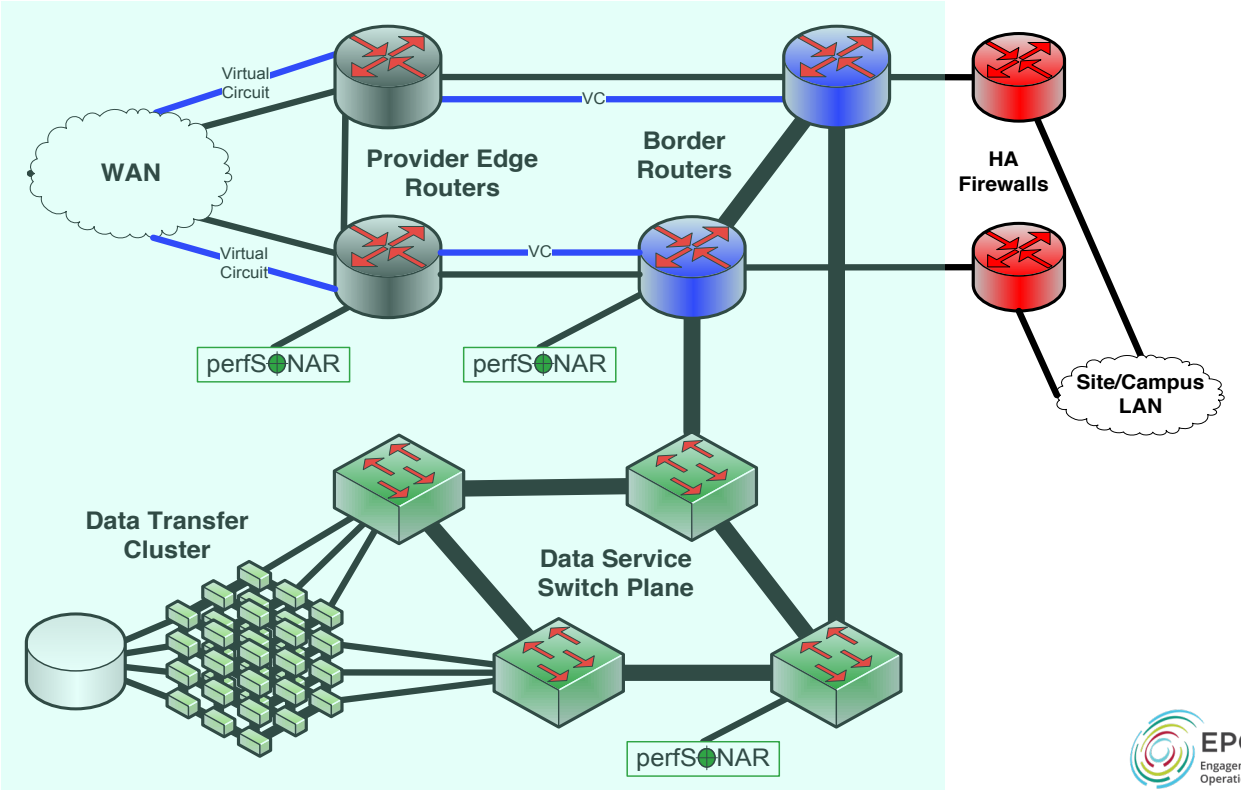
EPOC
Engagement and Performance Operations Center

Data Site – Architecture



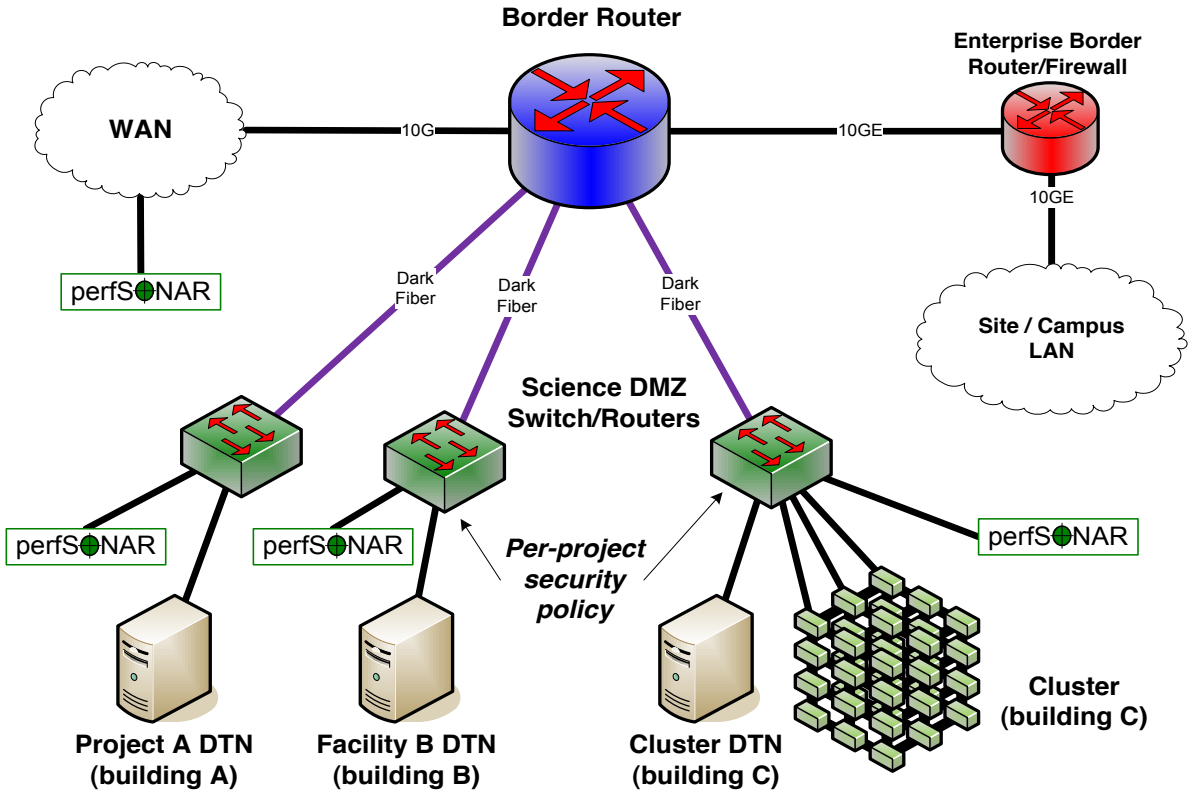
EPOC
Engagement and Performance Operations Center

Data Site – Data Path



EPOC
Engagement and Performance Operations Center

Multiple Science DMZs – Dark Fiber



EPOC
Engagement and Performance Operations Center

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

EPOC
Engagement and Performance Operations Center

Components in Detail

- Performance monitoring – perfSONAR
- Data Transfer Nodes (DTNs)
- Security

EPOC
Engagement and Performance Operations Center

Science DMZ Security

- **Goal**: Disentangle security policy and enforcement for science flows from security for business systems
- **Rationale**
  - Science data traffic is simple from a security perspective
  - Narrow application set on Science DMZ
    - Data transfer, data streaming packages
    - No printers, document readers, web browsers, building control systems, financial databases, staff desktops, etc.
  - Security controls that are typically implemented to protect business resources often cause performance problems
- **Separation allows each to be optimized**

EPOC
Engagement and Performance Operations Center

Science DMZ as Security Architecture

- Allows for better segmentation of risks, more granular application of controls to those segmented risks.
  - Limit risk profile for high-performance data transfer applications
  - Apply specific controls to data transfer hosts
  - Avoid including unnecessary risks, unnecessary controls
- Remove degrees of freedom – focus only on what is necessary
  - Easier to secure
  - Easier to achieve performance
  - Easier to troubleshoot

EPOC
Engagement and Performance Operations Center

Science DMZ Security Myth

- **The big myth:** The main goal of the Science DMZ is to avoid firewalls and other security controls.
  - Leads to all sorts of odd (and wrong) claims like:
  - "Our whole backbone is a Science DMZ because there is no firewall in front of the backbone."
  - "The Science DMZ doesn't allow for **any** security controls."
  - "The Science DMZ requires a default-permit policy."
- **The reality:**
  - The Science DMZ is about being performant ***and*** being secure
  - Reduce degrees-of-freedom, reduce complexity → reduce risk
  - Ensure that the devices in the data path are high performance

EPOC
Engagement and Performance Operations Center

From Myth To Reality

- Contrary to myth, the Science DMZ *is a security architecture.*
- The Science DMZ is a form of security *control,* not something that needs to be controlled.
- At the same time, the Science DMZ enables us to do a better job of risk-based security through segmentation.

EPOC
Engagement and Performance Operations Center

Network Segmentation

- Think about residence hall networks, business application networks, and the networks that are primarily in research areas:
  - The risk profiles are clearly different
  - It makes sense to segment along these lines
- Your institution may already be doing this for things like HIPAA and PCI-DSS. Why? *Because of the controls!*
- The Science DMZ follows the same concept, from a security perspective.
- Using a Science DMZ to segment research traffic (especially traffic from specialized research instruments) can actually *improve* campus security posture.

EPOC
Engagement and Performance Operations Center

Placement Outside the Firewall

- The Science DMZ resources are placed outside the enterprise firewall for performance reasons
  - The meaning of this is specific – ***Science DMZ traffic does not traverse the firewall data plane***
  - Packet filtering is fine – just don't do it with a firewall
- Lots of heartburn over this, especially from the perspective of a conventional firewall manager
  - Lots of organizational policy directives mandating firewalls
  - Firewalls are designed to protect converged enterprise networks
  - Why would you put critical assets outside the firewall???
- The answer is that firewalls are typically a poor fit for high-performance science applications

EPOC
Engagement and Performance Operations Center

Firewall Capabilities and Science Traffic

- Firewalls have a lot of sophistication in an enterprise setting
  - Application layer protocol analysis (HTTP, POP, MSRPC, etc.)
  - Built-in VPN servers
  - User awareness
- Data-intensive science flows typically don't match this profile
  - Common case – data on filesystem A needs to be on filesystem Z
    - Data transfer tool verifies credentials over an encrypted channel
    - Then open a socket or set of sockets, and send data until done (1TB, 10TB, 100TB, …)
  - One workflow can fill 50% of a 100G network link
- Do we have to use a firewall?

EPOC
Engagement and Performance Operations Center

Firewalls As Access Lists

- When you ask a firewall administrator to allow data transfers through the firewall, what do they ask for?
  - IP address of your host
  - IP address of the remote host
  - Port range
  - ***That looks like an ACL to me!***
- No special config for advanced protocol analysis – just address/port
- Router ACLs are better than firewalls at address/port filtering
  - ACL capabilities are typically built into the router
  - Router ACLs typically do not drop traffic permitted by policy

EPOC
Engagement and Performance Operations Center

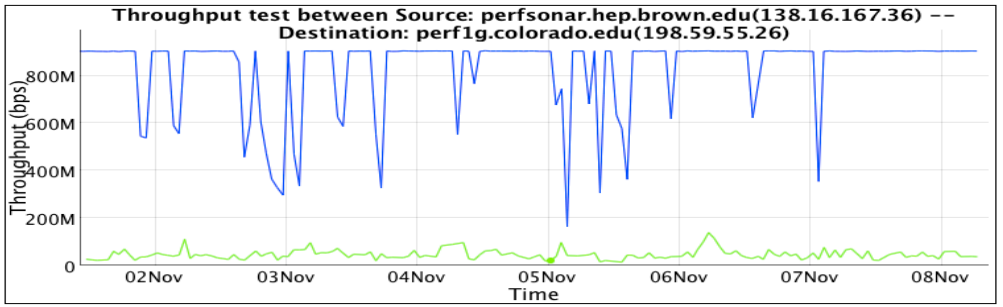
Security Without Firewalls

- Data intensive science traffic interacts poorly with firewalls
- Does this mean we ignore security? ***NO!***
  - We **must** protect our systems
  - We just need to find a way to do security that does not prevent us from getting the science done
- ***Key point – security policies and mechanisms that protect the Science DMZ should be implemented so that they do not compromise performance***
- Traffic permitted by policy should not experience performance impact as a result of the application of policy

EPOC
Engagement and Performance Operations Center

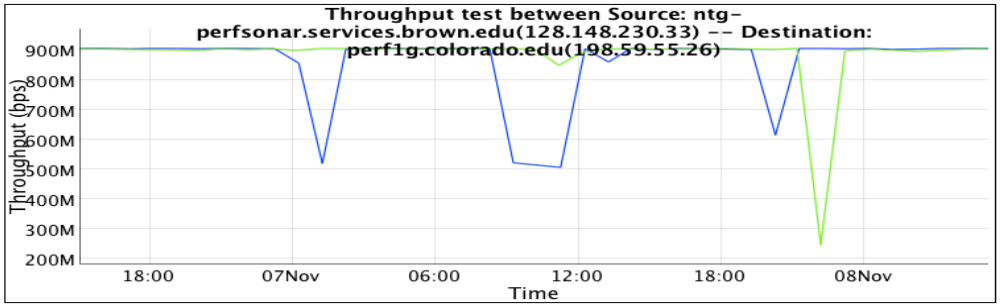
Firewall Performance Example

- Observed performance, via perfSONAR, through a firewall:



Almost 20 times slower through the firewall

- Observed performan



Huge improvement without the firewall

EPOC
Engagement and Performance Operations Center

If Not Firewalls, Then What?

- Remember – the goal is to protect systems in a way that allows the science mission to succeed
- I like something I heard at NERSC – paraphrasing: “Security controls should enhance the utility of science infrastructure.”
- There are multiple ways to solve this – some are technical, and some are organizational/sociological
- I’m not going to lie to you – this is harder than just putting up a firewall and closing your eyes

EPOC
Engagement and Performance Operations Center

Collaboration Within The Organization

- All stakeholders should collaborate on Science DMZ design, policy, and enforcement
- The security people have to be on board
  - Remember: security people already have political cover – it's called the firewall
  - If a host gets compromised, the security officer can say they did their due diligence because there was a firewall in place
  - If the deployment of a Science DMZ is going to jeopardize the job of the security officer, expect pushback
- The Science DMZ is a strategic asset, and should be understood by the strategic thinkers in the organization
  - Changes in security models
  - Changes in operational models
  - Enhanced ability to compete for funding
  - Increased institutional capability – greater science output

EPOC
Engagement and Performance Operations Center

Wrapup

- The Science DMZ design pattern provides a flexible model for supporting high-performance data transfers and workflows
- Key elements:
  - Accommodation of TCP
    - Sufficient bandwidth to avoid congestion
    - Loss-free IP service
  - Location – near the site perimeter if possible
  - Test and measurement
  - Dedicated systems
  - Appropriate security
- Support for advanced capabilities is much easier with a Science DMZ

EPOC
Engagement and Performance Operations Center

Links

- ESnet fasterdata knowledge base
  - http://fasterdata.es.net/
- Science DMZ paper
  - http://www.es.net/assets/pubs_presos/sc13sciDMZ-final.pdf
- Science DMZ email list
  - https://gab.es.net/mailman/listinfo/sciencedmz
- perfSONAR
  - http://fasterdata.es.net/performance-testing/perfsonar/
  - http://www.perfsonar.net

EPOC
Engagement and Performance Operations Center

http://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/

CC* Data Movement Workshop and Exhibition – Science DMZ & Security

Jason Zurawski, Eli Dart
zurawski@es.net, dart@es.net
ESnet / Lawrence Berkeley National Laboratory

Dr. Jennifer M. Schopf
jmschopf@indiana.edu
Indiana University International Networks

CC/CICI PI Meeting Pre-Workshop*
September 22nd 2019

National Science Foundation Award #1826994

https://epoc.global

Extra Slides – Community Science DMZ Deck

EPOC
Engagement and Performance Operations Center

Overview

- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- Science DMZ Security
- User Engagement
- Wrap Up

Motivation

- Networks are an essential part of data-intensive science
  - Connect data sources to data analysis
  - Connect collaborators to each other
  - Enable machine-consumable interfaces to data and analysis resources (e.g. portals), automation, scale
- Performance is critical
  - Exponential data growth
  - Constant human factors
  - Data movement and data analysis must keep up
- Effective use of wide area (long-haul) networks by scientists has historically been difficult

Data Mobility in a Given Time Interval

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

This table available at:

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

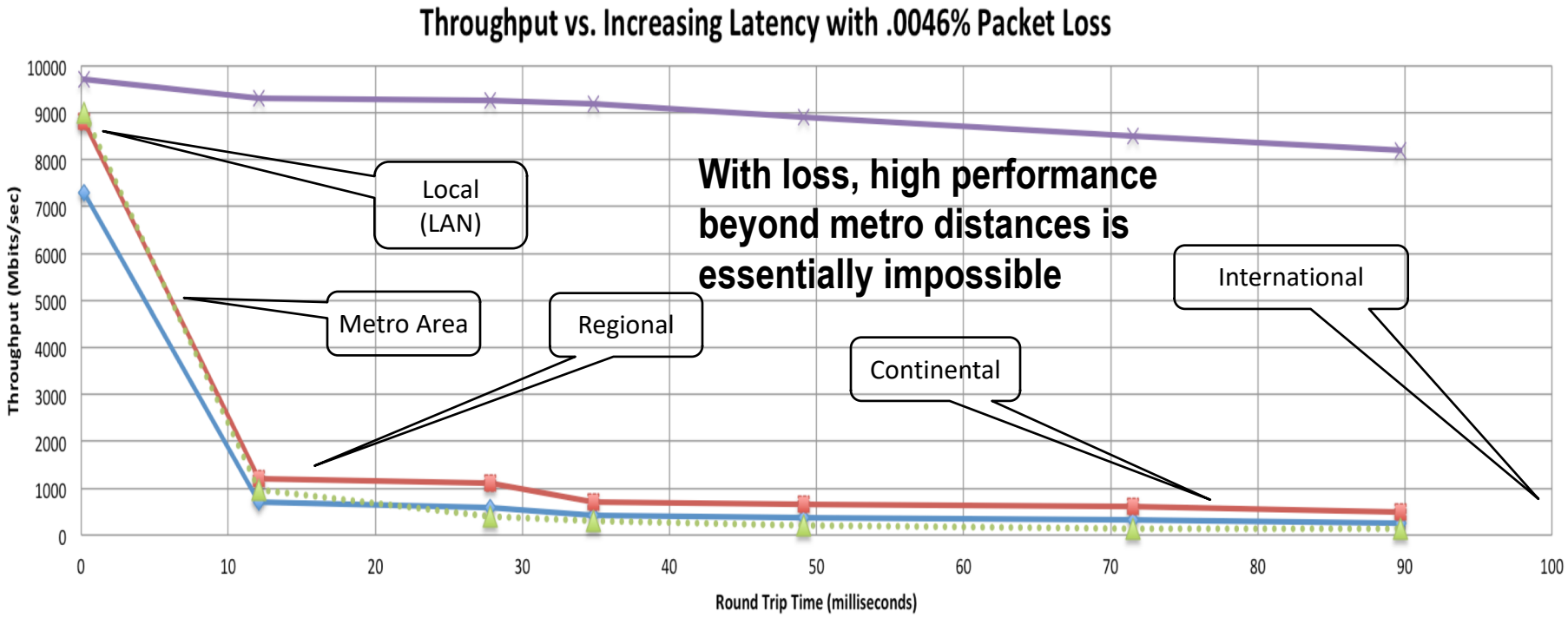
The Central Role of the Network

- The very structure of modern science assumes science networks exist: high performance, feature rich, global scope
- What is "The Network" anyway?
  - "The Network" is the set of devices and applications involved in the use of a remote resource
    - This is not about supercomputer interconnects
    - This is about data flow from experiment to analysis, between facilities, etc.
  - User interfaces for "The Network" – portal, data transfer tool, workflow engine
  - Therefore, servers and applications must also be considered
- What is important? Ordered list:
  1. Correctness
  2. Consistency
  3. Performance

TCP – Ubiquitous and Fragile

- 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
  - 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 (more than 95% of ESnet traffic is TCP)

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



Working With TCP In Practice

- Far easier to support TCP than to fix TCP
  - People have been trying to fix TCP for years – limited success
  - Like it or not we're stuck with TCP in the general case
- Pragmatically speaking, we must accommodate TCP
  - Sufficient bandwidth to avoid congestion
  - Zero packet loss
  - Verifiable infrastructure
    - Networks are complex
    - Must be able to locate problems quickly
    - Small footprint is a huge win – small number of devices so that problem isolation is tractable

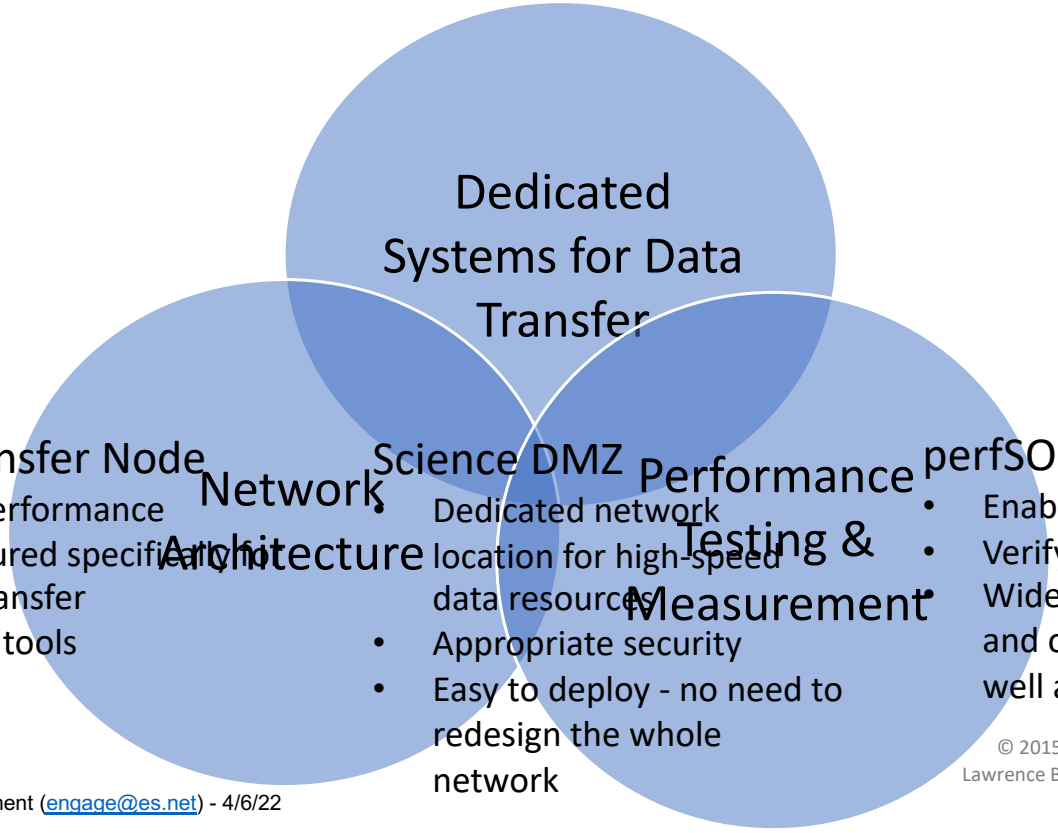
Putting A Solution Together

- Effective support for TCP-based data transfer
  - Design for correct, consistent, high-performance operation
  - Design for ease of troubleshooting
- Easy adoption is critical
  - Large laboratories and universities have extensive IT deployments
  - Drastic change is prohibitively difficult
- Cybersecurity – defensible without compromising performance
- Borrow ideas from traditional network security
  - Traditional DMZ
    - Separate enclave at network perimeter ("Demilitarized Zone")
    - Specific location for external-facing services
    - Clean separation from internal network
  - Do the same thing for science – ***Science DMZ***

Overview

- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- Science DMZ Security
- User Engagement
- Wrap Up

The Science DMZ Design Pattern



Data Transfer Node
- High performance
- Configured specifically for data transfer
- Proper tools

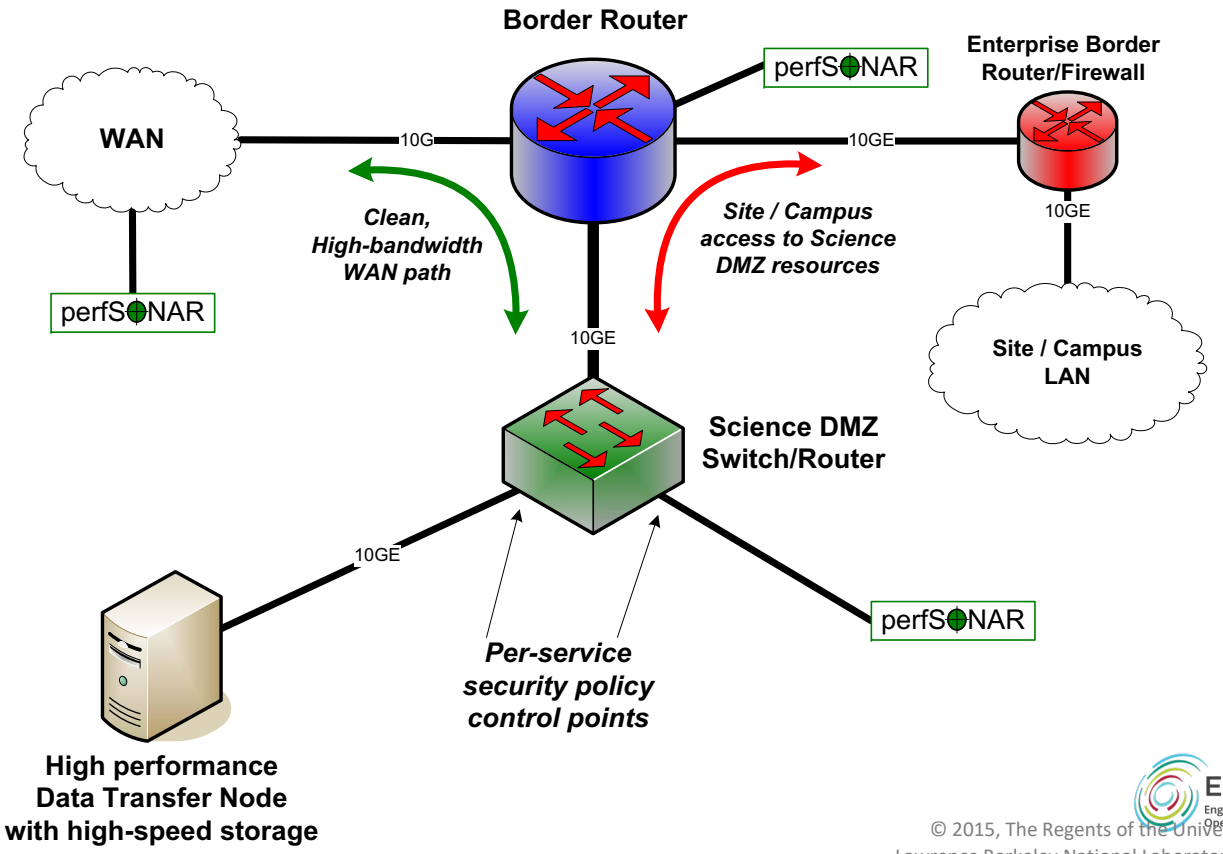
Science DMZ
- Dedicated network location for high-speed data resources
- Appropriate security
- Easy to deploy - no need to redesign the whole network

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

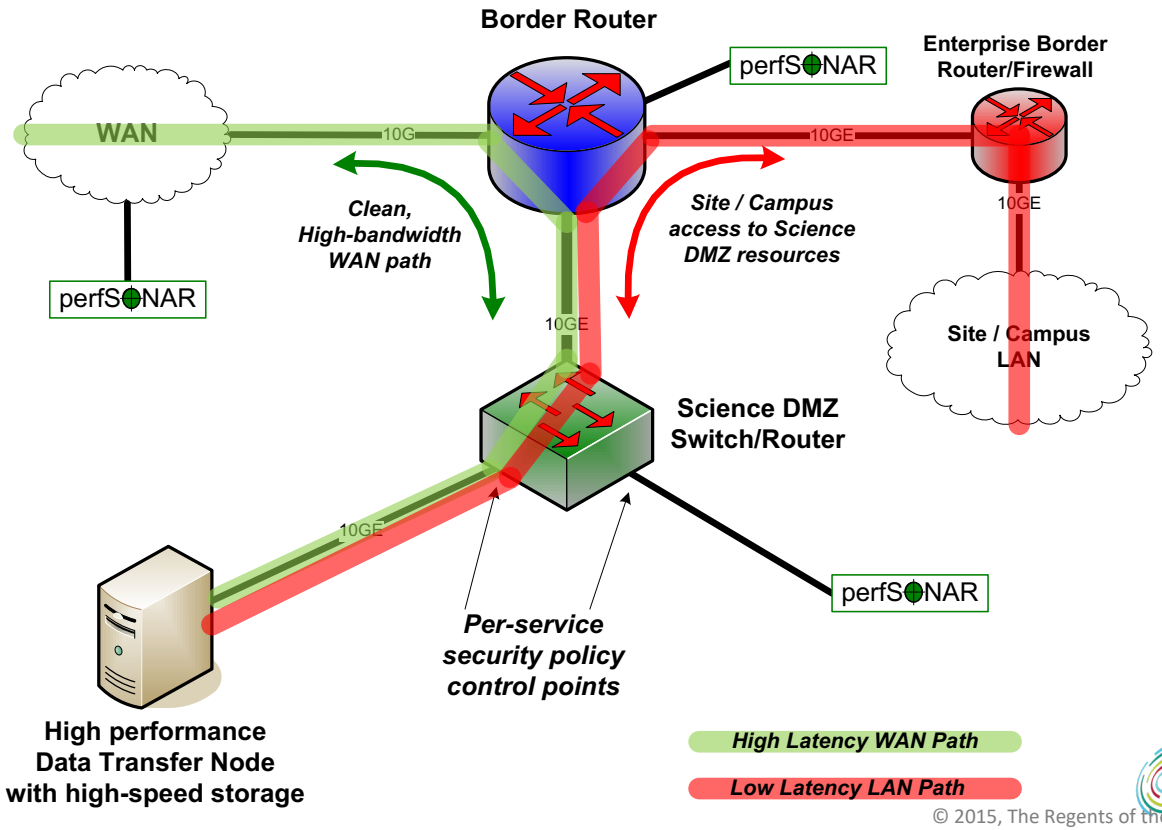
Abstract or Prototype Deployment

- 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

Science DMZ Design Pattern (Abstract)



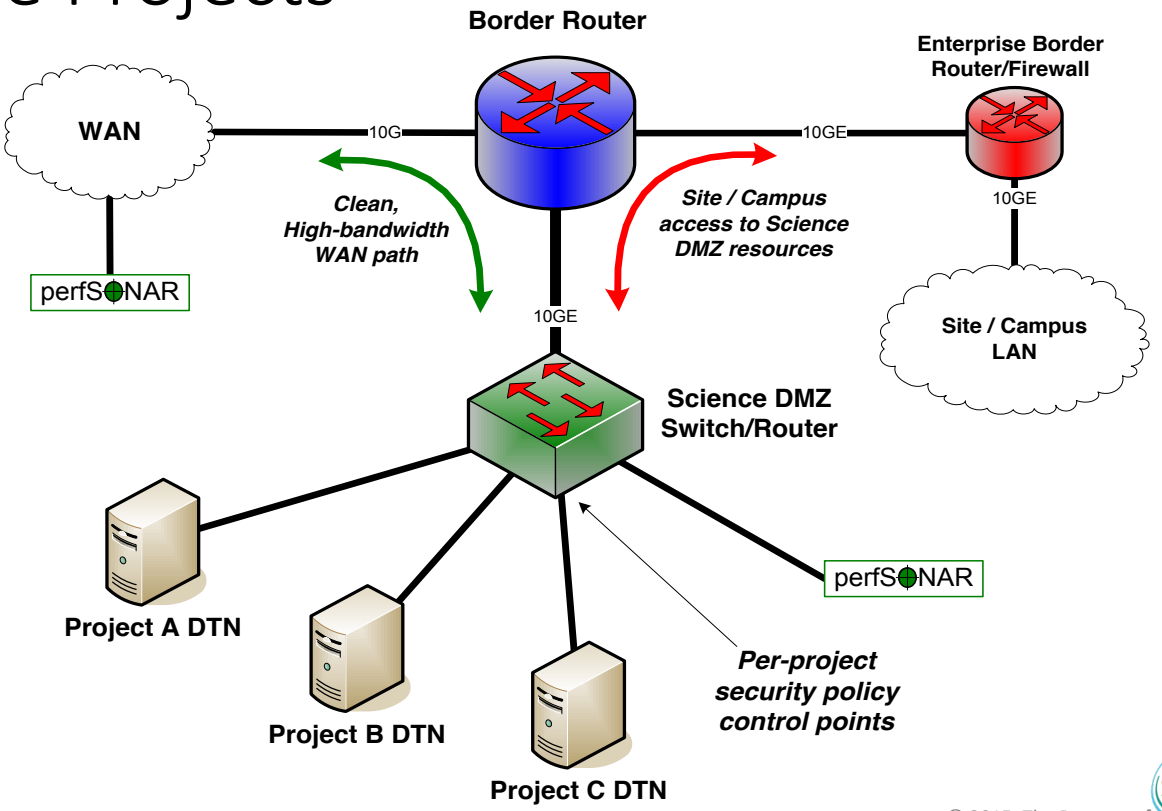
Local And Wide Area Data Flows



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?

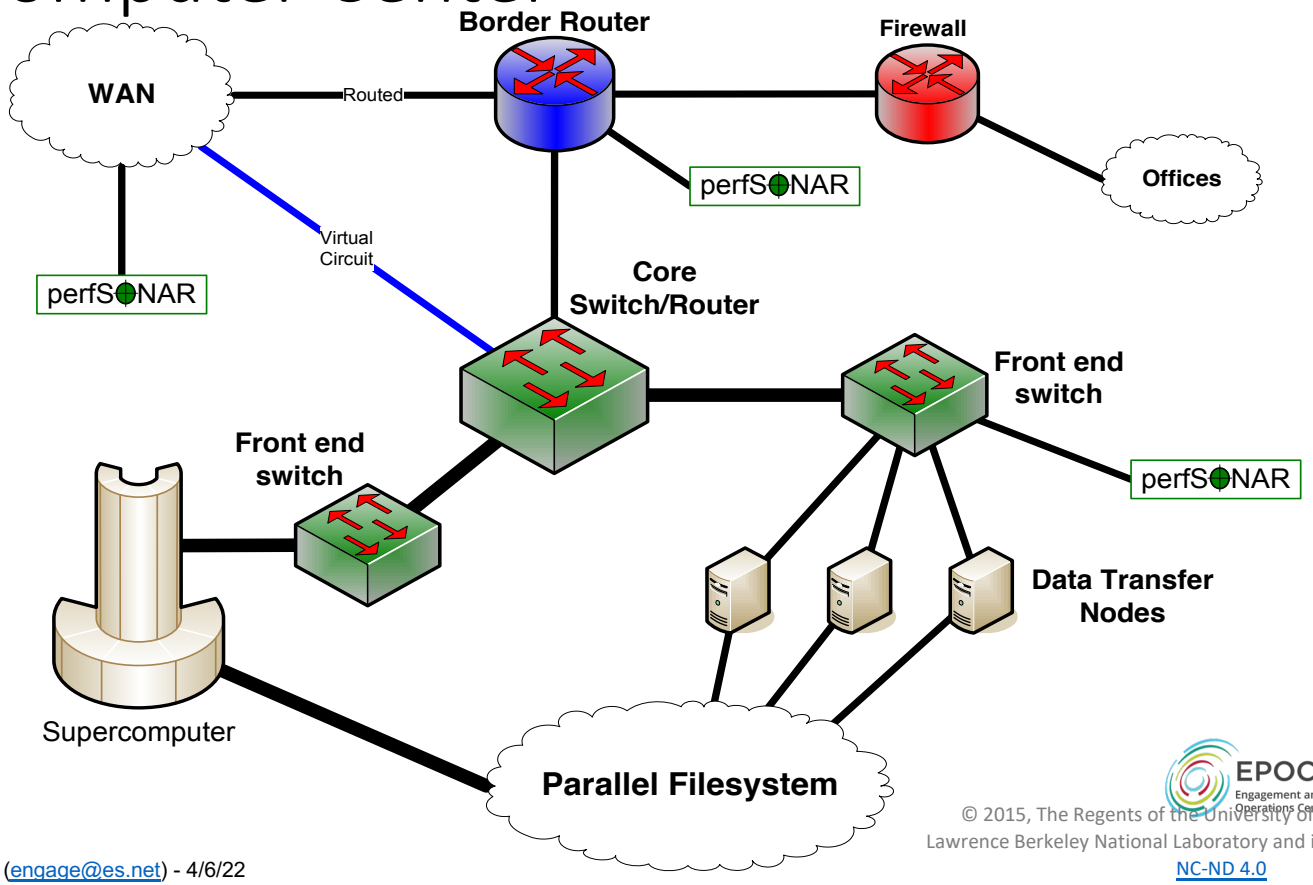
Multiple Projects



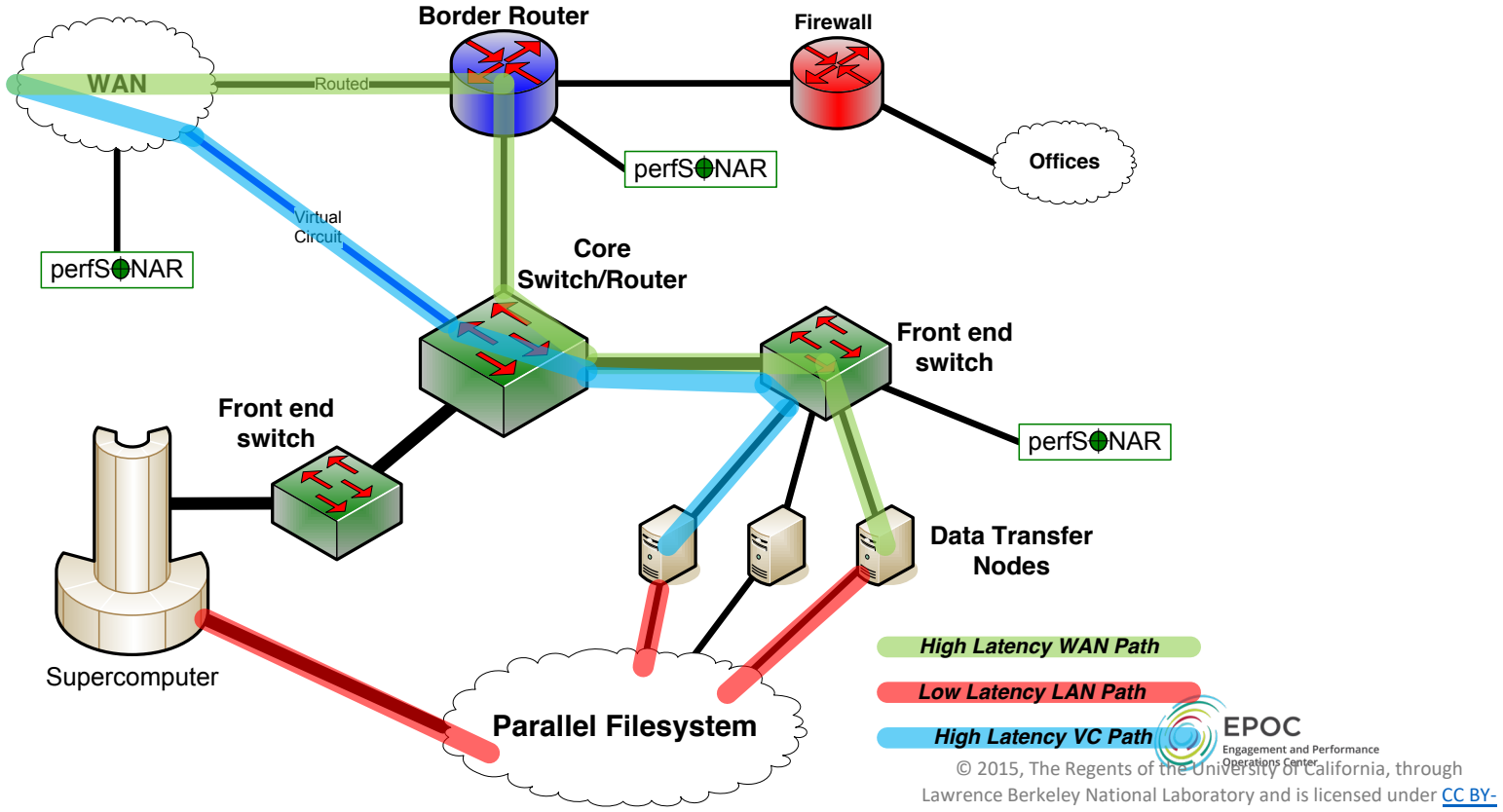
Supercomputer Center Deployment

- 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



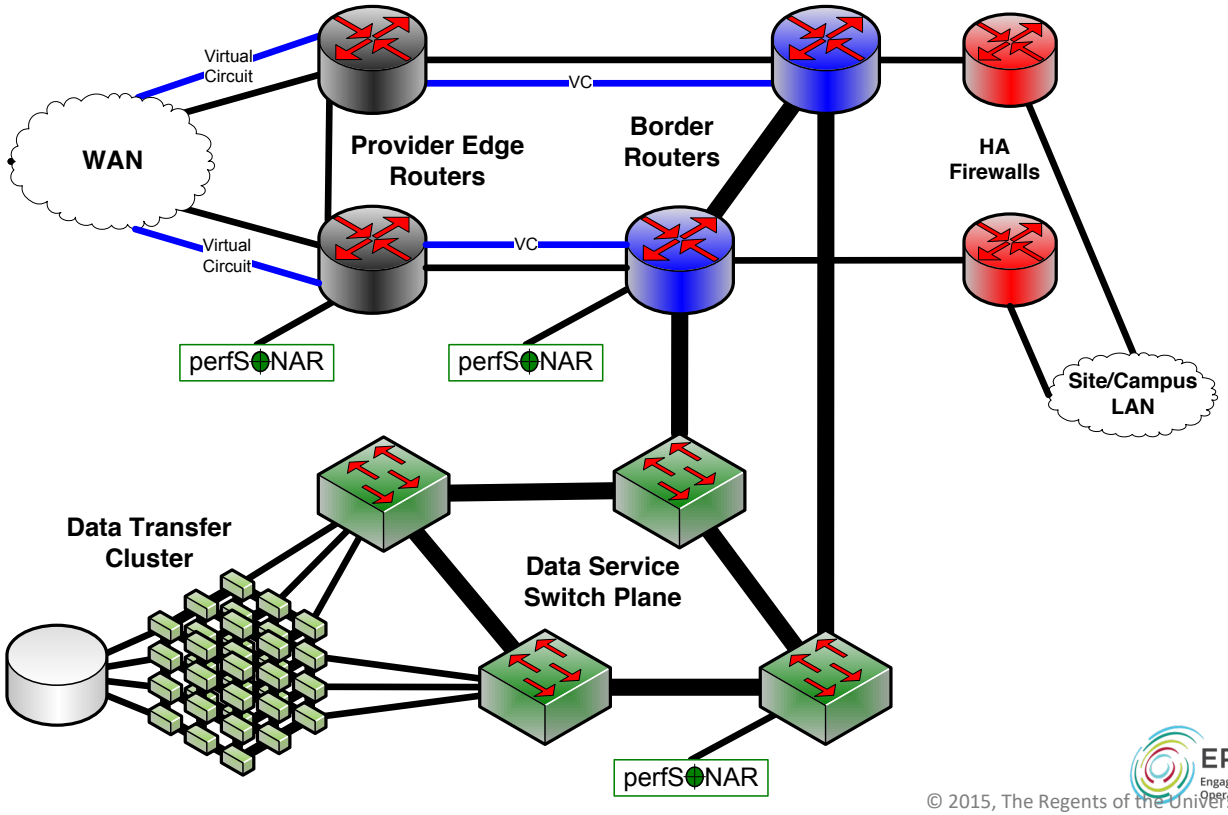
Supercomputer Center Data Path



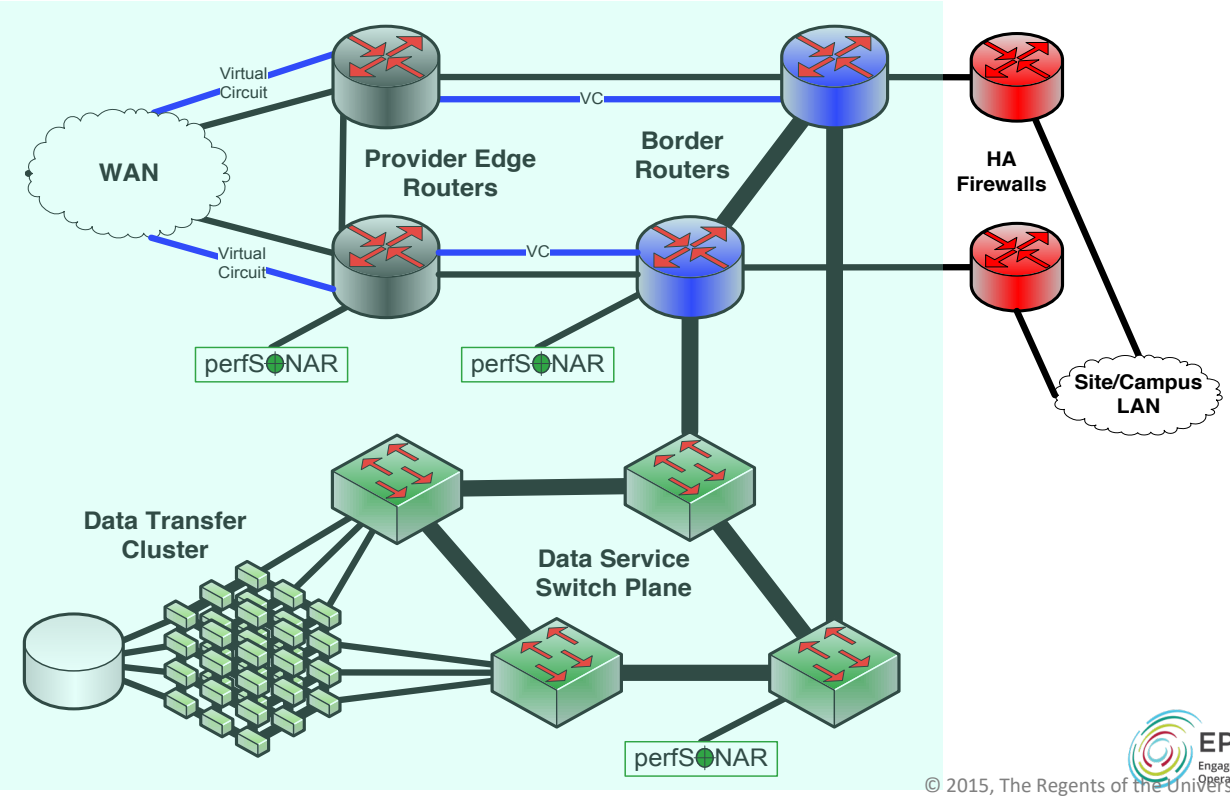
Major Data Site Deployment

- In some cases, large scale data service is the major driver
  - Huge volumes of data (Petabytes or more) – ingest, export
  - Large number of external hosts accessing/submitting data
- 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) just cause problems

Data Site – Architecture



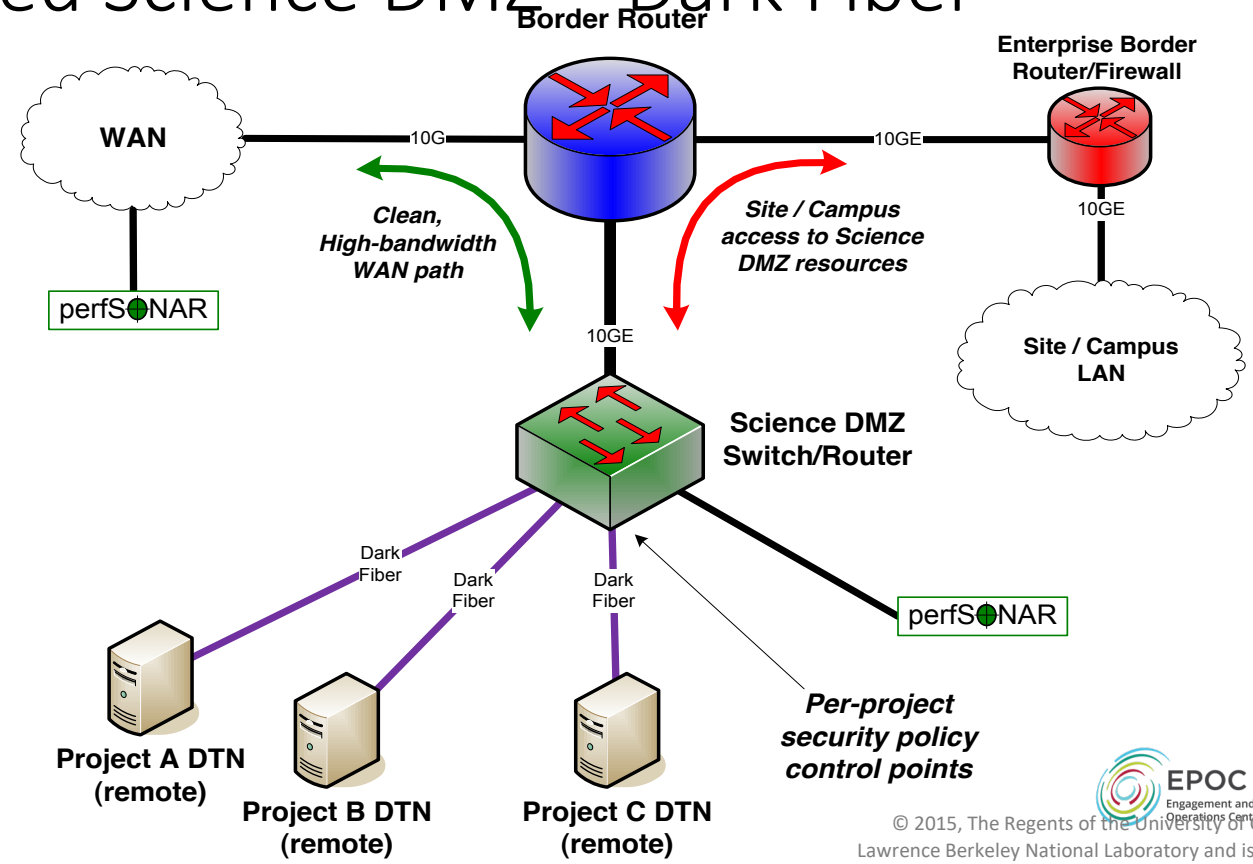
Data Site – Data Path



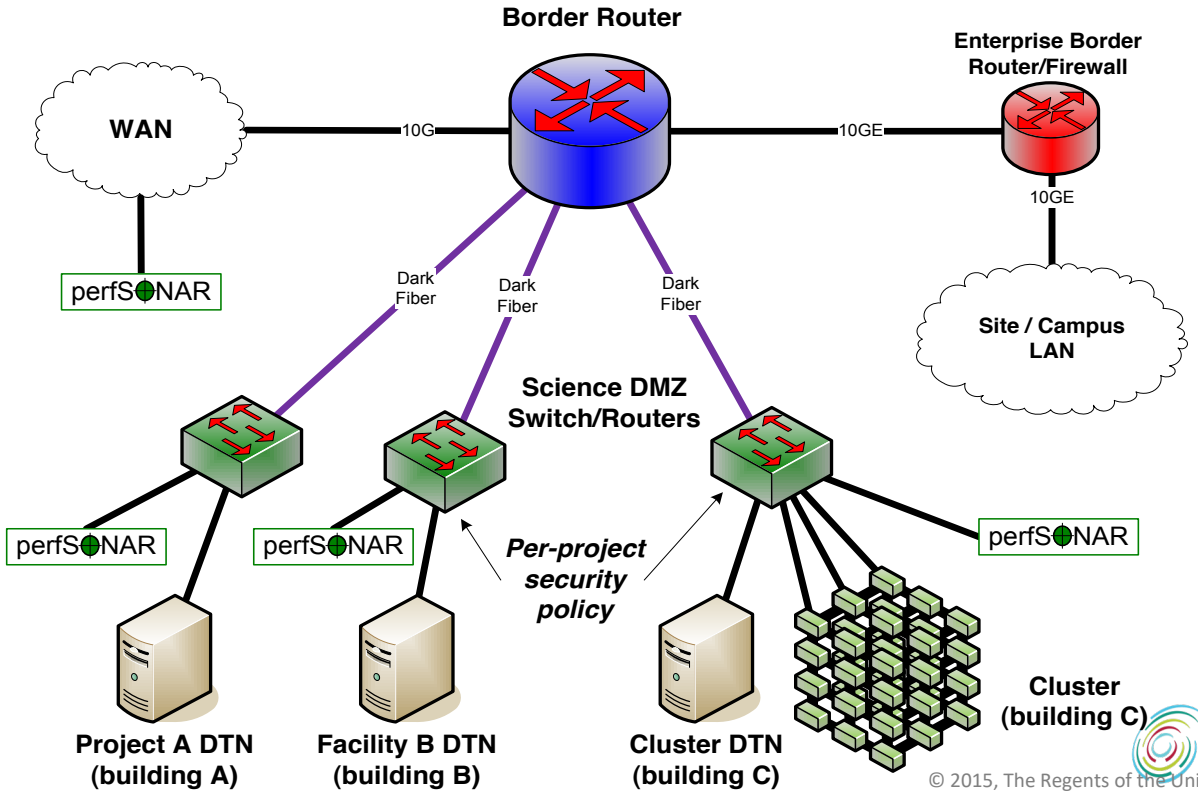
Distributed Science DMZ

- Fiber-rich environment enables a 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 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



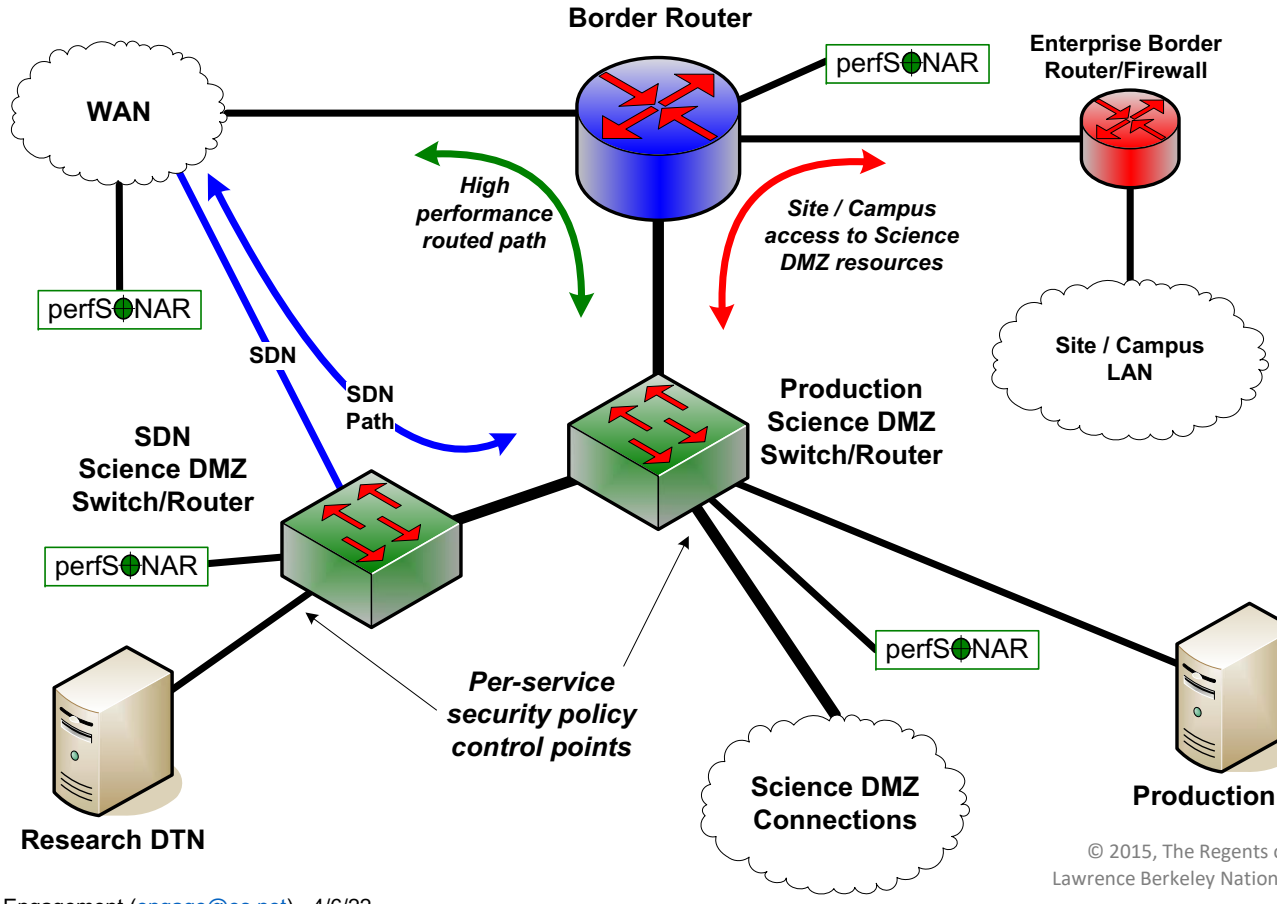
Development Environment

- One thing that often happens is that an early power user of the Science DMZ is the network engineering group that builds it
  - Service prototyping
  - Deployment of test applications for other user groups to demonstrate value
- The production Science DMZ is just that – production
  - Once users are on it, you can't take it down to try something new
  - Stuff that works tends to attract workload
- ***Take-home message: plan for multiple Science DMZs from the beginning – at the very least you're going to need one for yourself***
- The Science DMZ model easily accommodates this

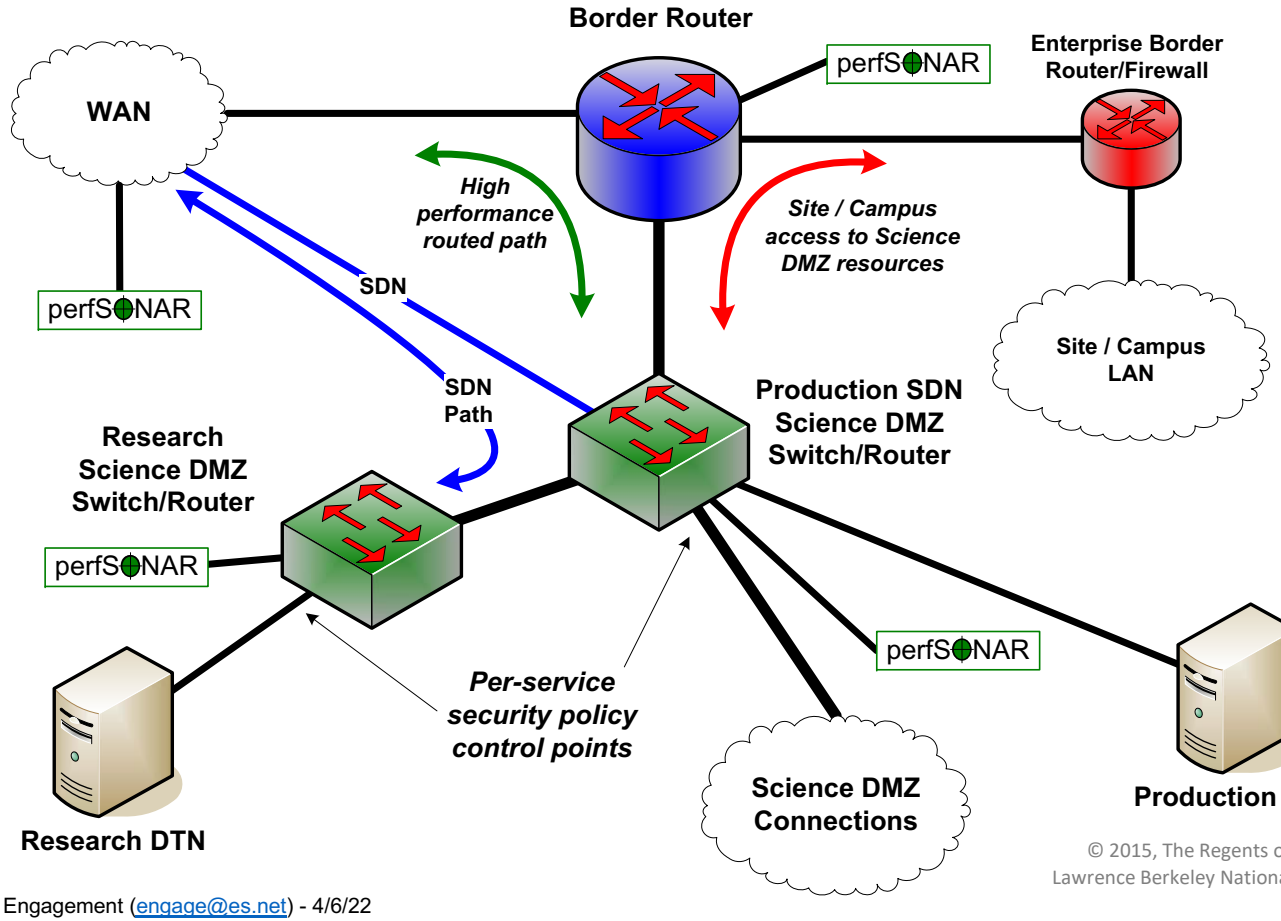
Science DMZ – Flexible Design Pattern

- The Science DMZ design pattern is highly adaptable to research
- Deploying a research Science DMZ is straightforward
  - The basic elements are the same
    - Capable infrastructure designed for the task
    - Test and measurement to verify correct operation
    - Security policy well-matched to the environment, application set is strictly limited to reduce risk
  - Connect the research DMZ to other resources as appropriate
- The same ideas apply to supporting an SDN effort
  - Test/research areas for development
  - Transition to production as technology matures and need dictates
  - One possible trajectory follows…

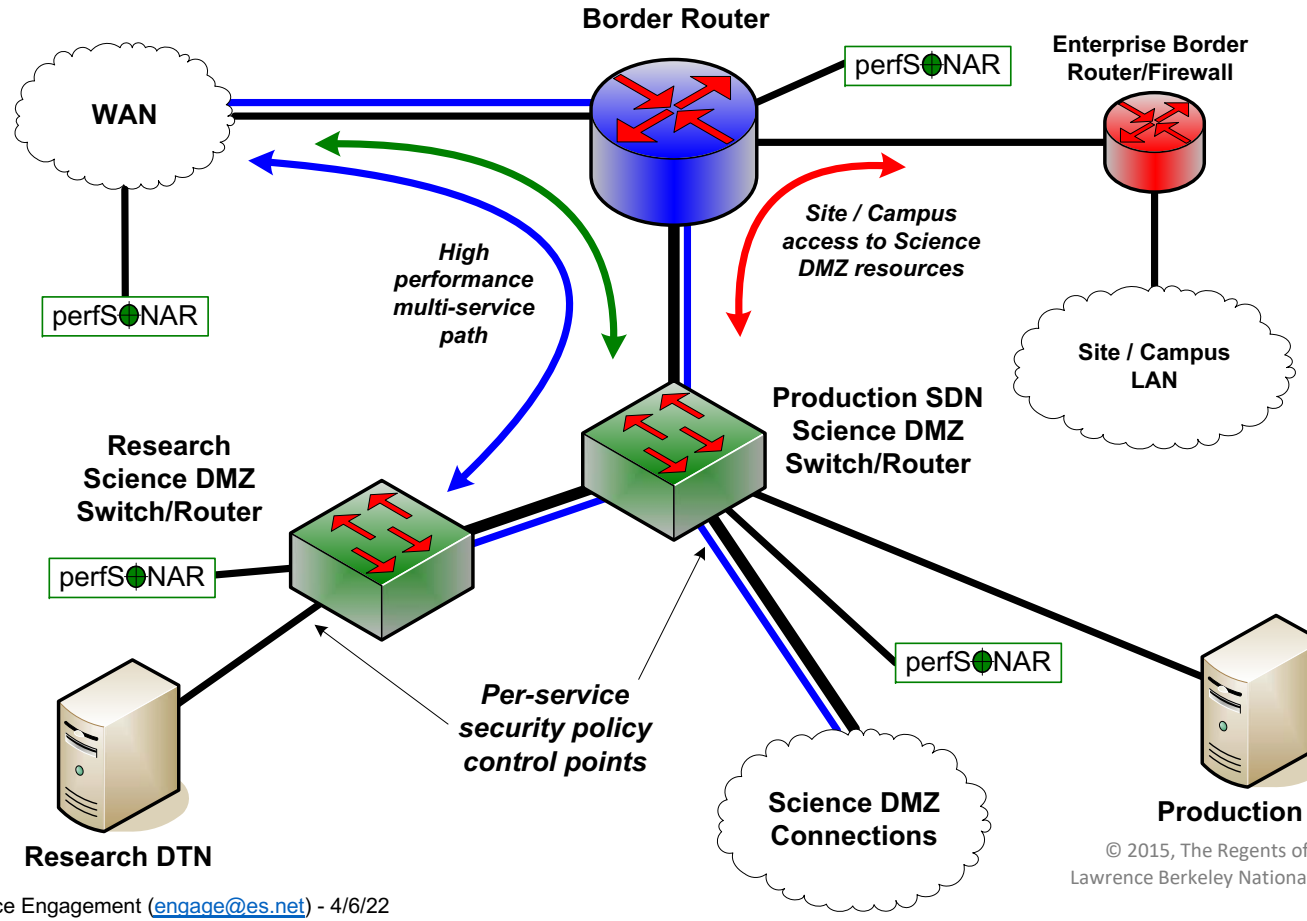
Science DMZ – Separate SDN Connection



Science DMZ – Production SDN Connection



Science DMZ – SDN Campus Border



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

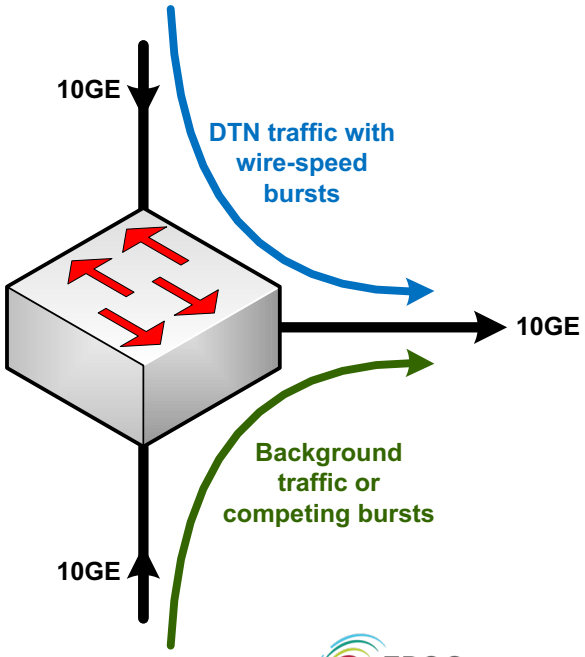
Multiple Ingress Flows, Common Egress

Hosts will typically send packets at the speed of their interface (1G, 10G, etc.)

- Instantaneous rate, not average rate
- If TCP has window available and data to send, host sends until there is either no data or no window

Hosts moving big data (e.g. DTNs) can send large bursts of back-to-back packets

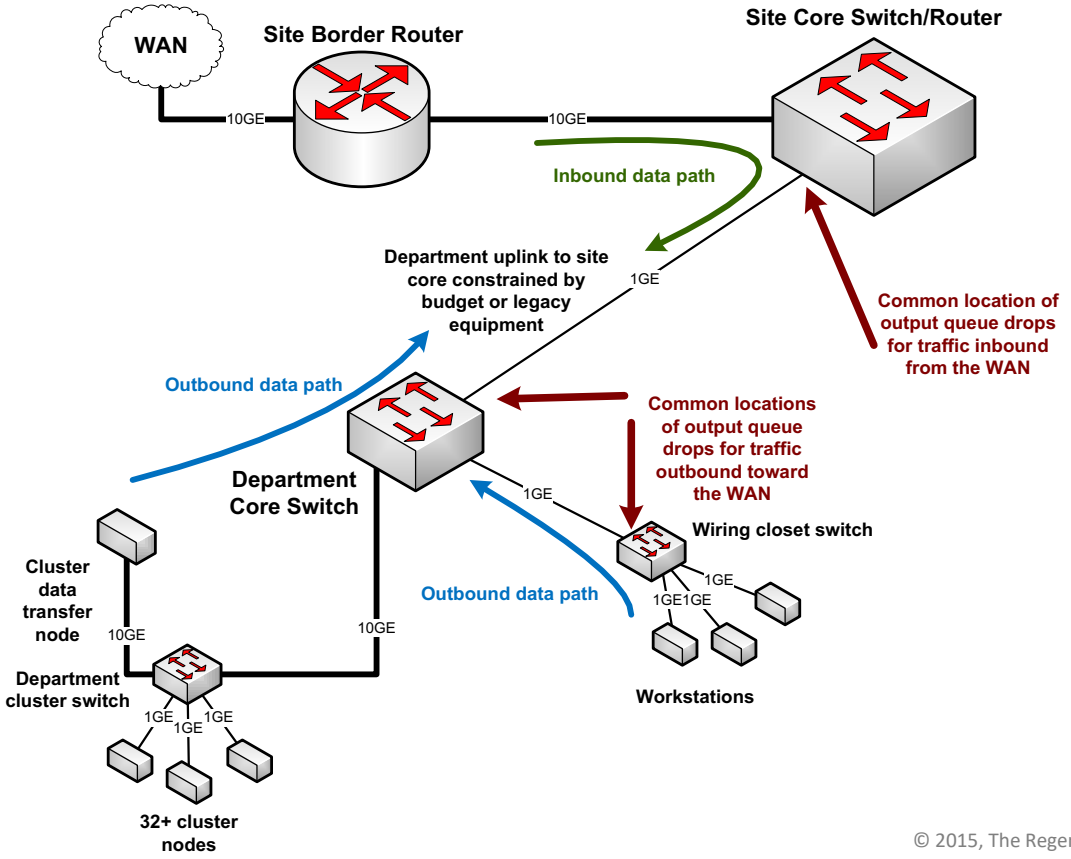
- This is true even if the average rate as measured over seconds is slower (e.g. 4Gbps)
- On microsecond time scales, there is often congestion
- Router or switch must queue packets or drop them



Router and Switch Output Queues

- Interface output queue allows the router or switch to avoid causing packet loss in cases of momentary congestion
- In network devices, queue depth (or 'buffer') is often a function of cost
  - Cheap, fixed-config LAN switches (especially in the 10G space) have inadequate buffering. Imagine a 10G 'data center' switch as the guilty party
  - Cut-through or low-latency Ethernet switches typically have inadequate buffering (the whole point is to avoid queuing!)
- Expensive, chassis-based devices are more likely to have deep enough queues
  - Juniper MX and Alcatel-Lucent 7750 used in ESnet backbone
  - Other vendors make such devices as well - details are important
  - Thx to Jim: http://people.ucsc.edu/~warner/buffer.html
- This expense is one driver for the Science DMZ architecture – only deploy the expensive features where necessary

Output Queue Drops – Common Locations



Overview

- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- Science DMZ Security
- User Engagement
- Wrap Up

perfSONAR

Performance Monitoring

- Everything may function perfectly when it is deployed
- Eventually something is going to break
  - Networks and systems are complex
  - Bugs, mistakes, …
  - Sometimes things just break – this is why we buy support contracts
- Must be able to find and fix problems when they occur
- Must be able to find problems in other networks (your network may be fine, but someone else's problem can impact your users)

- 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 RFC793, 1981)

perfSONAR

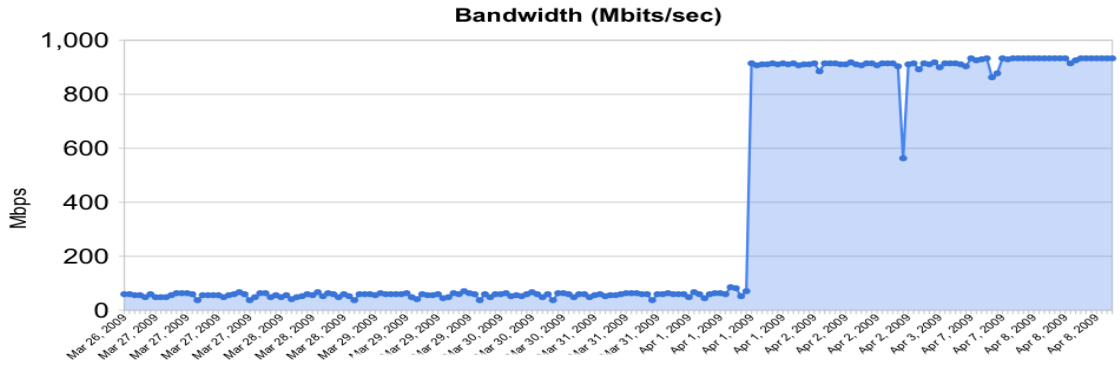
Soft Network Failures – Hidden Problems

- Hard failures are well-understood
  - Link down, system crash, software crash
  - Traditional network/system monitoring tools designed to quickly find hard failures
- Soft failures result in degraded capability
  - Connectivity exists
  - Performance impacted
  - Typically something in the path is functioning, but not well
- Soft failures are hard to detect with traditional methods
  - No obvious single event
  - Sometimes no indication at all of any errors
- Independent testing is the only way to reliably find soft failures

EPOC Engagement and Performance Operations Center

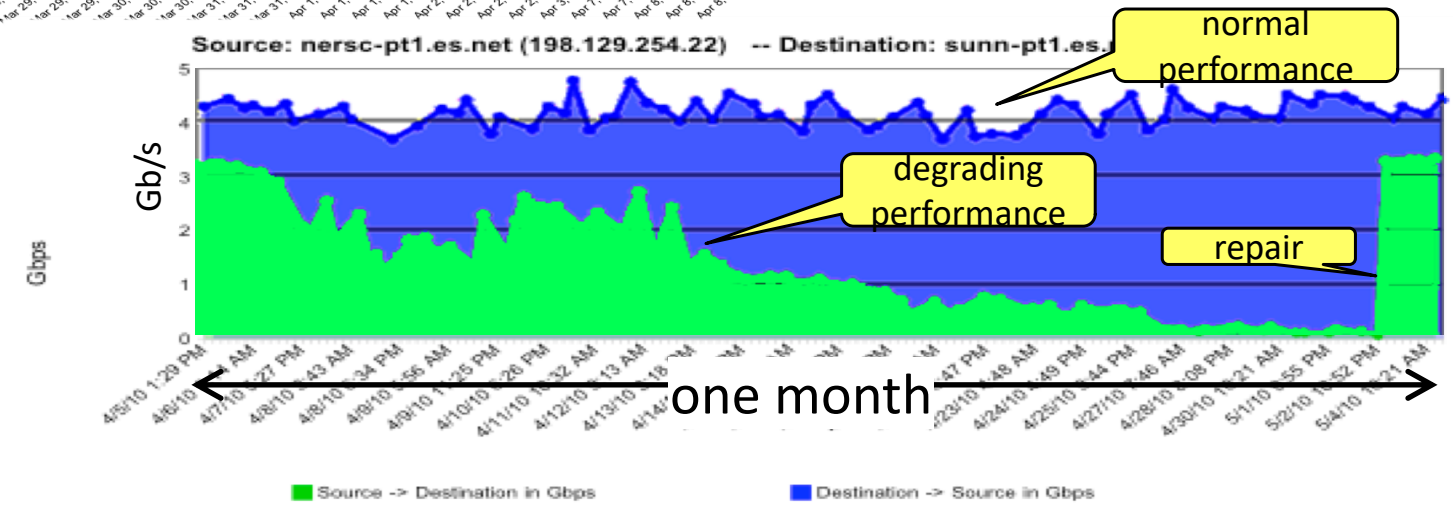
Sample Soft Failures

perfSONAR



Rebooted router with full route table

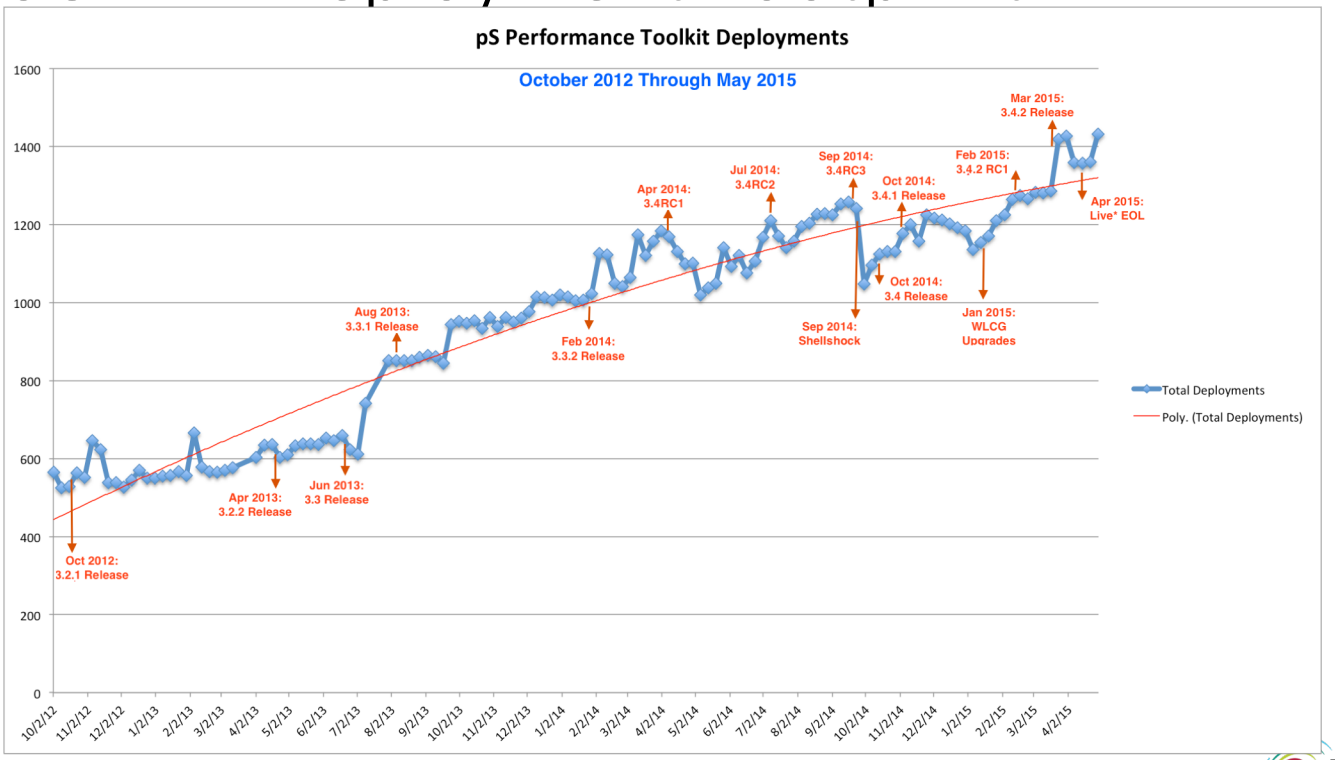
Gradual failure of optical line card



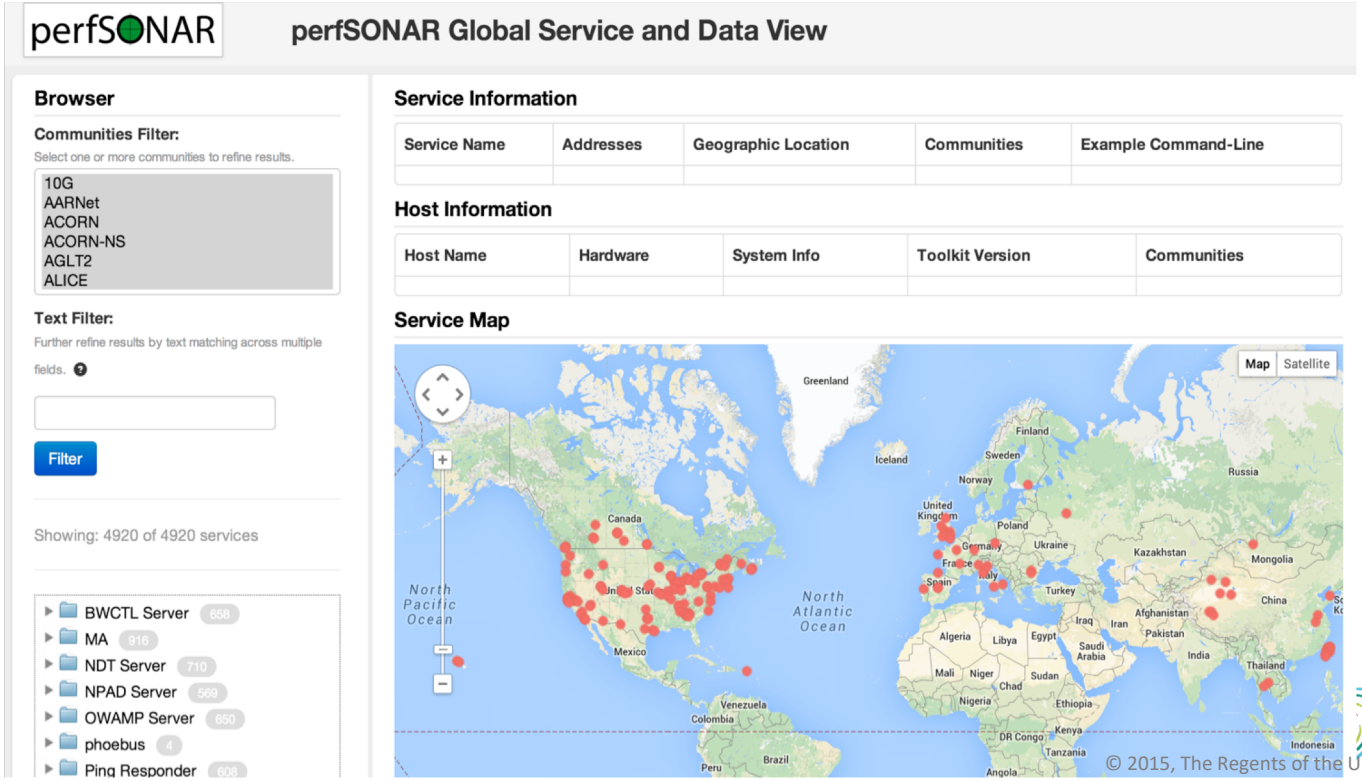
Testing Infrastructure – perfSONAR

- perfSONAR is:
  - A widely-deployed test and measurement infrastructure
    - ESnet, Internet2, US regional networks, international networks
    - Laboratories, supercomputer centers, universities
  - A suite of test and measurement tools
  - A collaboration that builds and maintains the toolkit
- By installing perfSONAR, a site can leverage over 1100 test servers deployed around the world
- perfSONAR is ideal for finding soft failures
  - Alert to existence of problems
  - Fault isolation
  - Verification of correct operation

perfSONAR Deployment Footprint



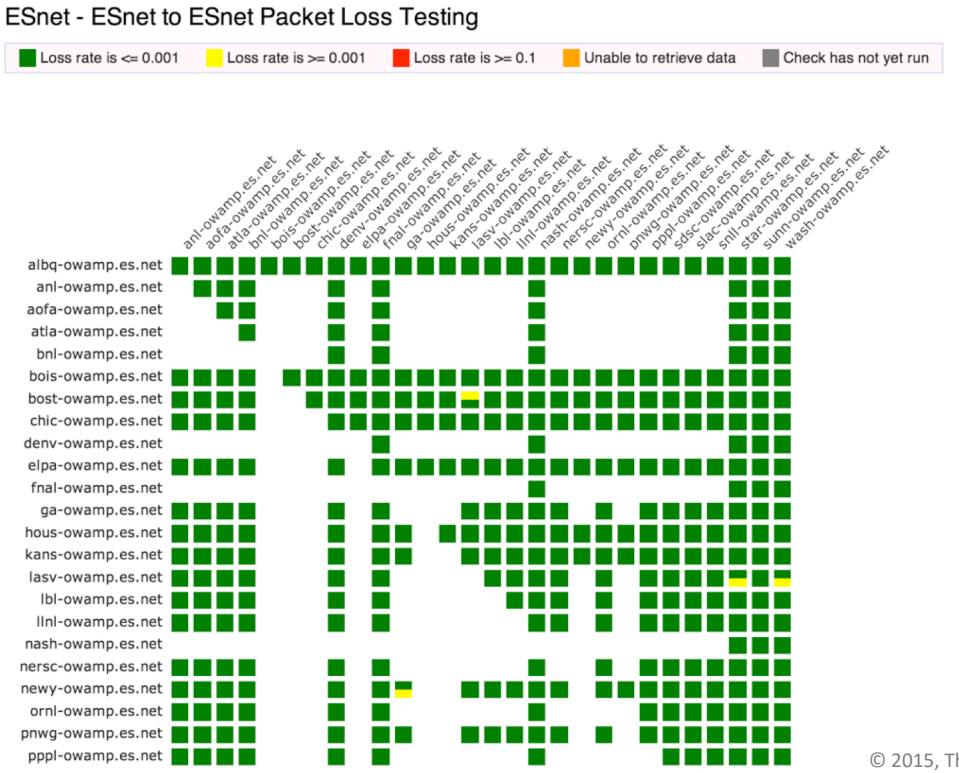
Lookup Service Directory Search: http://stats.es.net/ServicesDirectory/



EPOC
Engagement and Performance Operations Center

perfSONAR Dashboard: http://ps-dashboard.es.net

perfSONAR



EPOC Engagement and Performance Operations Center

Overview

- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- Science DMZ Security
- User Engagement
- Wrap Up

Dedicated Systems – Data Transfer Node

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

Data Transfer Tools For DTNs

- Parallelism is important
  - It is often easier to achieve a given performance level with four parallel connections than one connection
  - Several tools offer parallel transfers, including Globus/GridFTP
- Latency interaction is critical
  - Wide area data transfers have much higher latency than LAN transfers
  - Many tools and protocols assume a LAN
- Workflow integration is important
- Key tools: Globus Online, HPN-SSH

Data Transfer Tool Comparison

- In addition to the network, using the right data transfer tool is critical

•Data transfer test from 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 properly engineered storage (RAID, parallel filesystem, etc.)

Overview

- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- Science DMZ Security
- User Engagement
- Wrap Up

Science DMZ Security

- Goal – disentangle security policy and enforcement for science flows from security for business systems
- Rationale
  - Science data traffic is simple from a security perspective
  - Narrow application set on Science DMZ
    - Data transfer, data streaming packages
    - No printers, document readers, web browsers, building control systems, financial databases, staff desktops, etc.
  - Security controls that are typically implemented to protect business resources often cause performance problems
- Separation allows each to be optimized

Performance Is A Core Requirement

- Core information security principles
  - Confidentiality, Integrity, Availability (CIA)
  - Often, CIA and risk mitigation result in poor performance
- In data-intensive science, performance is an additional core mission requirement: CIA → PICA
  - CIA principles are important, but ***if performance is compromised the science mission fails***
  - Not about "how much" security you have, but how the security is implemented
  - Need a way to appropriately secure systems without performance compromises

Placement Outside the Firewall

- The Science DMZ resources are placed outside the enterprise firewall for performance reasons
  - The meaning of this is specific – ***Science DMZ traffic does not traverse the firewall data plane***
  - Packet filtering is fine – just don't do it with a firewall
- Lots of heartburn over this, especially from the perspective of a conventional firewall manager
  - Lots of organizational policy directives mandating firewalls
  - Firewalls are designed to protect converged enterprise networks
  - Why would you put critical assets outside the firewall???
- The answer is that firewalls are typically a poor fit for high-performance science applications

Firewall Internals

- Typical firewalls are composed of a set of processors which inspect traffic in parallel
  - Traffic distributed among processors such that all traffic for a particular connection goes to the same processor
  - Simplifies state management
  - Parallelization scales deep analysis
- Excellent fit for enterprise traffic profile
  - High connection count, low per-connection data rate
  - Complex protocols with embedded threats
- Each processor is a fraction of firewall link speed
  - Significant limitation for data-intensive science applications
  - Overload causes packet loss – performance crashes

What's Inside Your Firewall?

- Vendor: "but wait – we don't do this anymore!"
  - It is true that vendors are working toward line-rate 10G firewalls, and some may even have them now
  - 10GE has been deployed in science environments for over 10 years
  - Firewall internals have only recently started to catch up with the 10G world
  - 100GE is being deployed now, 40Gbps host interfaces are available now
  - Firewalls are behind again
- In general, IT shops want to get 5+ years out of a firewall purchase
  - This often means that the firewall is years behind the technology curve
  - Whatever you deploy now, that's the hardware feature set you get
  - When a new science project tries to deploy data-intensive resources, they get whatever feature set was purchased several years ago

The Firewall State Table

- Many firewalls use a state table to improve performance
  - State table lookup is fast
  - No need to process entire ruleset for every packet
  - Also allows session tracking (e.g. TCP sequence numbers)
- State table built dynamically
  - Incoming packets are matched against the state table
  - If no state table entry, go to the ruleset
  - If permitted by ruleset, create state table entry
  - Remove state table entry after observing connection teardown
- Semantically similar to punt-and-switch model of traffic forwarding used on many older routers

State Table Issues

- If the state table is not pruned, it will overflow
  - Not all connections close cleanly
    - I shut my laptop and go to a meeting
    - Software crashes happen
  - Some attacks try to fill state tables
- Solution: put a timer on state table entries
  - When a packet matches the state table entry, update the timer
  - If the timer expires, delete the state table entry
- What if I just pause for a few minutes?
  - This turns out to be a problem – state table timers are typically in the 5-15 minute range, while host keepalive timers are 2 hours
  - If a connection pauses (e.g. control channel waits for a large transfer), the firewall will delete the state table entry from under it – the control connection now hangs
  - We have seen this in production environments

Firewall Capabilities and Science Traffic

- Firewalls have a lot of sophistication in an enterprise setting
  - Application layer protocol analysis (HTTP, POP, MSRPC, etc.)
  - Built-in VPN servers
  - User awareness
- Data-intensive science flows typically don't match this profile
  - Common case – data on filesystem A needs to be on filesystem Z
    - Data transfer tool verifies credentials over an encrypted channel
    - Then open a socket or set of sockets, and send data until done (1TB, 10TB, 100TB, …)
  - One workflow can use 10% to 50% or more of a 10G network link
- Do we have to use a firewall?

Firewalls As Access Lists

- When you ask a firewall administrator to allow data transfers through the firewall, what do they ask for?
  - IP address of your host
  - IP address of the remote host
  - Port range
  - ***That looks like an ACL to me!***
- No special config for advanced protocol analysis – just address/port
- Router ACLs are better than firewalls at address/port filtering
  - ACL capabilities are typically built into the router
  - Router ACLs typically do not drop traffic permitted by policy

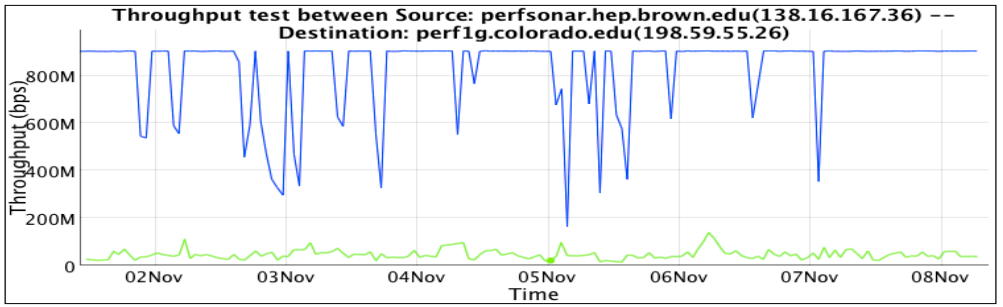
Security Without Firewalls

- Data intensive science traffic interacts poorly with firewalls
- Does this mean we ignore security? ***NO!***
  - We **must** protect our systems
  - We just need to find a way to do security that does not prevent us from getting the science done
- ***Key point – security policies and mechanisms that protect the Science DMZ should be implemented so that they do not compromise performance***
- Traffic permitted by policy should not experience performance impact as a result of the application of policy

Firewall Performance Example

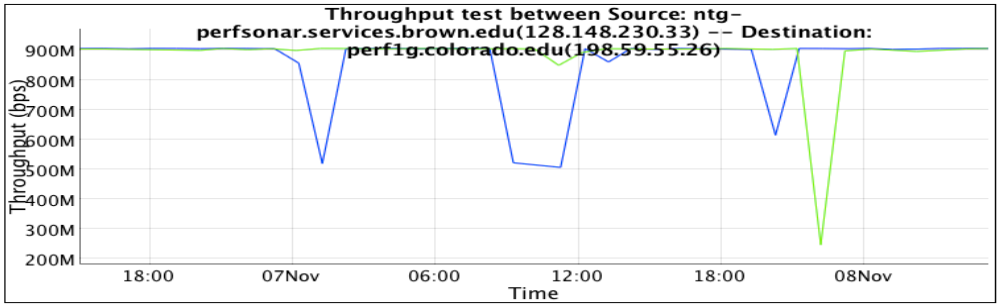
- Observed performance, via perfSONAR, through a firewall:

Almost 20 times slower through the firewall



- Observed performan

Huge improvement without the firewall



If Not Firewalls, Then What?

- Remember – the goal is to protect systems in a way that allows the science mission to succeed
- I like something I heard at NERSC – paraphrasing: "Security controls should enhance the utility of science infrastructure."
- There are multiple ways to solve this – some are technical, and some are organizational/sociological
- I'm not going to lie to you – this is harder than just putting up a firewall and closing your eyes

Other Technical Capabilities

- Intrusion Detection Systems (IDS)
  - One example is Bro – http://bro-ids.org/
  - Bro is high-performance and battle-tested
    - Bro protects several high-performance national assets
    - Bro can be scaled with clustering: http://www.bro-ids.org/documentation/cluster.html
  - Other IDS solutions are available also
- Netflow and IPFIX can provide intelligence, but not filtering
- Openflow and SDN
  - Using Openflow to control access to a network-based service seems pretty obvious
  - This could significantly reduce the attack surface for any authenticated network service
  - This would only work if the Openflow device had a robust data plane

Other Technical Capabilities (2)

- Aggressive access lists
  - More useful with project-specific DTNs
  - If the purpose of the DTN is to exchange data with a small set of remote collaborators, the ACL is pretty easy to write
  - Large-scale data distribution servers are hard to handle this way (but then, the firewall ruleset for such a service would be pretty open too)
- Limitation of the application set
  - One of the reasons to limit the application set in the Science DMZ is to make it easier to protect
  - Keep desktop applications off the DTN (and watch for them anyway using logging, netflow, etc – take violations seriously)
  - This requires collaboration between people – networking, security, systems, and scientists

Collaboration Within The Organization

- All stakeholders should collaborate on Science DMZ design, policy, and enforcement
- The security people have to be on board
  - Remember: security people already have political cover – it's called the firewall
  - If a host gets compromised, the security officer can say they did their due diligence because there was a firewall in place
  - If the deployment of a Science DMZ is going to jeopardize the job of the security officer, expect pushback
- The Science DMZ is a strategic asset, and should be understood by the strategic thinkers in the organization
  - Changes in security models
  - Changes in operational models
  - Enhanced ability to compete for funding
  - Increased institutional capability – greater science output

Overview

- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- Science DMZ Security
- User Engagement
- Wrap Up

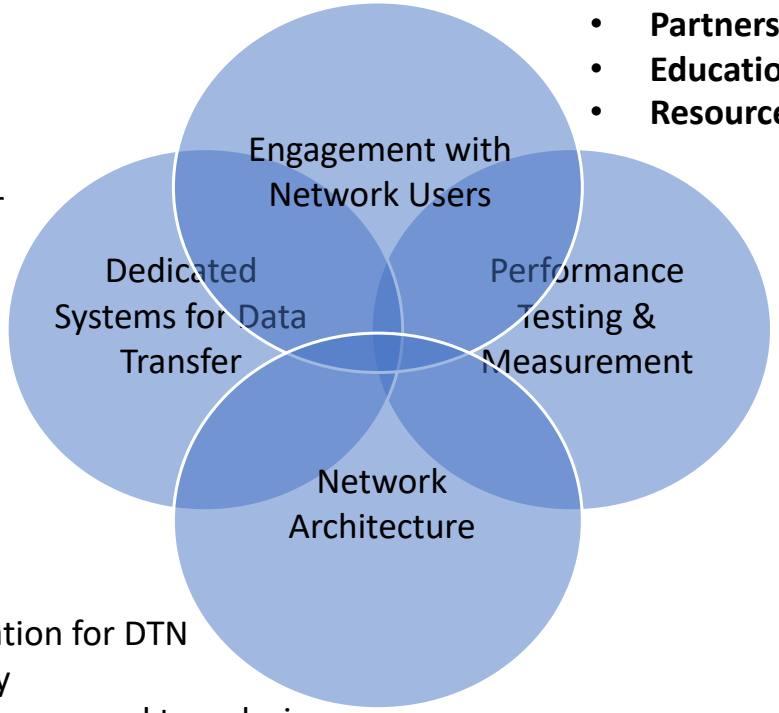
Science DMZ Superfecta: Engagement

Engagement

- **Partnerships**
- **Education & Consulting**
- **Resources & Knowledgebase**

Data Transfer Node

- High performance
- Configured for data transfer
- Proper tools



perfSONAR

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

Science DMZ

- Dedicated location for DTN
- Proper security
- Easy to deploy - no need to redesign the whole network

Challenges to Network Adoption

- Causes of performance issues are complicated for users.
- Lack of communication and collaboration between the CIO's office and researchers on campus.
- Lack of IT exp
  collaboration
- User's perfor
  ("The networ
  it didn't work").
- Cultural change is hard ("we've always shipped disks!").
- Scientists want to do science not IT support

The Capability Gap

EPOC
Engagement and Performance Operations Center

Overview

- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- On the Topic of Security
- User Engagement
- Wrap Up

Futures

- The Science DMZ design pattern is highly adaptable to new technologies
  - Software Defined Networking (SDN)
  - Non-IP protocols (RDMA over Ethernet)
- Deploying new technologies in a Science DMZ is straightforward
  - The basic elements are the same
    - Capable infrastructure designed for the task
    - Test and measurement to verify correct operation
    - Security policy well-matched to the environment
    - Application set strictly limited to reduce security risk
  - Change footprint is small – often just a single router or switch
  - The rest of the infrastructure need not change

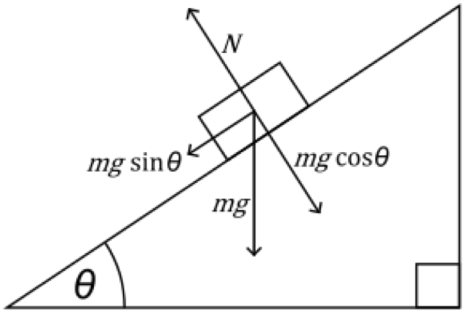
Wrapup

- The Science DMZ design pattern provides a flexible model for supporting high-performance data transfers and workflows
- Key elements:
  - Accommodation of TCP
    - Sufficient bandwidth to avoid congestion
    - Loss-free IP service
  - Location – near the site perimeter if possible
  - Test and measurement
  - Dedicated systems
  - Appropriate security
- Support for advanced capabilities (e.g. SDN) is much easier with a Science DMZ

The Science DMZ in 1 Slide

Consists of **three key components**, all required:

- "Friction free" network path
  - Highly capable network devices (wire-speed, deep queues)
  - Virtual circuit connectivity option
  - Security policy and enforcement specific to science workflows
  - Located at or near site perimeter if possible
- Dedicated, high-performance Data Transfer Nodes (DTNs)
  - Hardware, operating system, libraries all optimized for transfer
  - Includes optimized data transfer tools such as Globus Online and GridFTP
- Performance measurement/test node
  - perfSONAR
- Engagement with end users



globus

perfSONAR

Details at http://fasterdata.es.net/science-dmz/

EPOC
Engagement and Performance Operations Center

Links

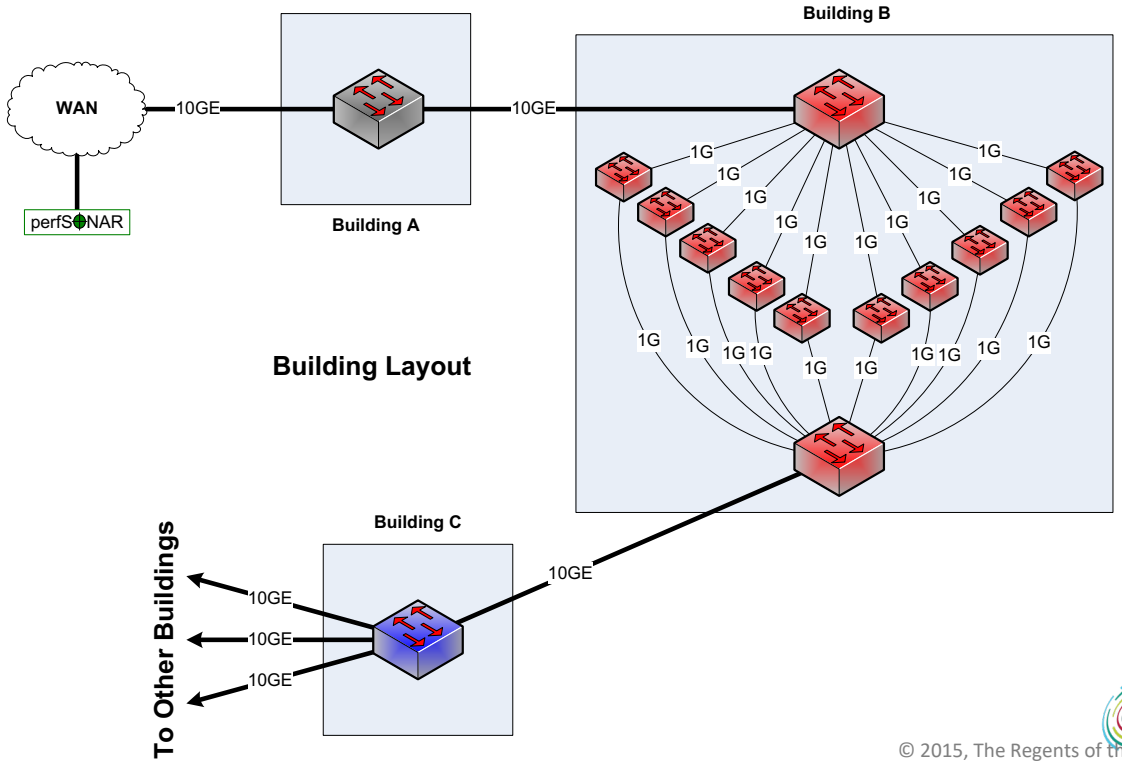
- ESnet fasterdata knowledge base
  - http://fasterdata.es.net/
- Science DMZ paper
  - http://www.es.net/assets/pubs_presos/sc13sciDMZ-final.pdf
- Science DMZ email list
  - https://gab.es.net/mailman/listinfo/sciencedmz
- perfSONAR
  - http://fasterdata.es.net/performance-testing/perfsonar/
  - http://www.perfsonar.net

Extra Slides – Firewall Internals

Thought Experiment

- We're going to do a thought experiment
- Consider a network between three buildings – A, B, and C
- This is supposedly a 10Gbps network end to end (look at the links on the buildings)
- Building A houses the border router – not much goes on there except the external connectivity
- Lots of work happens in building B – so much that the processing is done with multiple processors to spread the load in an affordable way, and results are aggregated after
- Building C is where we branch out to other buildings
- Every link between buildings is 10Gbps – this is a 10Gbps network, right???

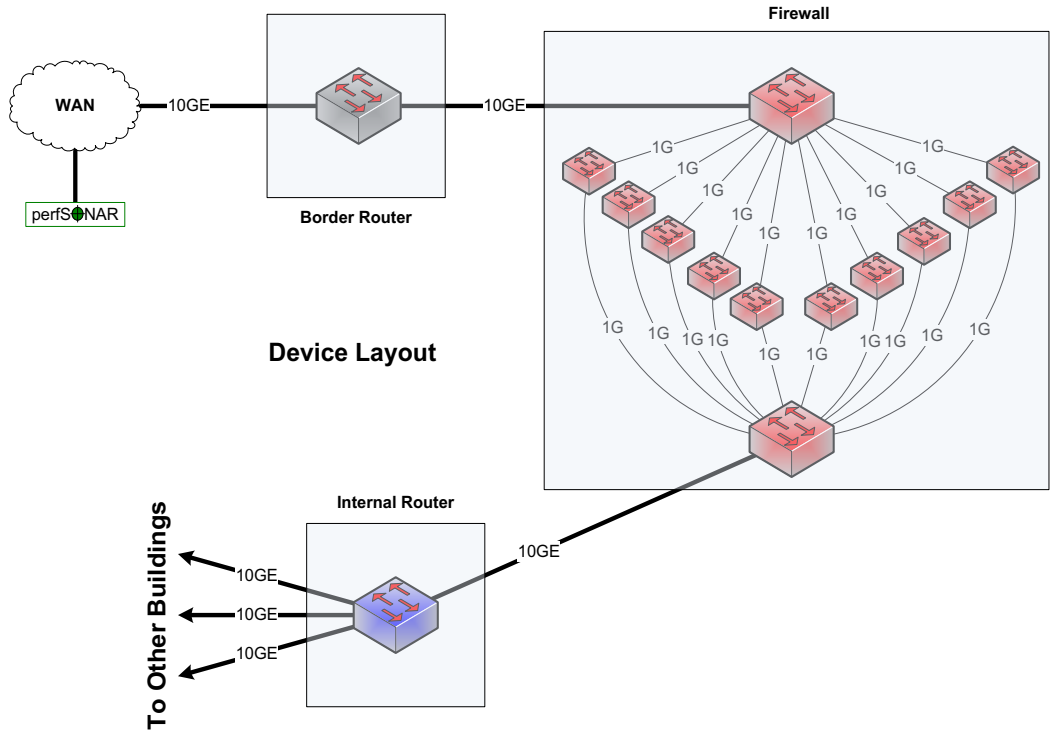
Notional 10G Network Between Buildings



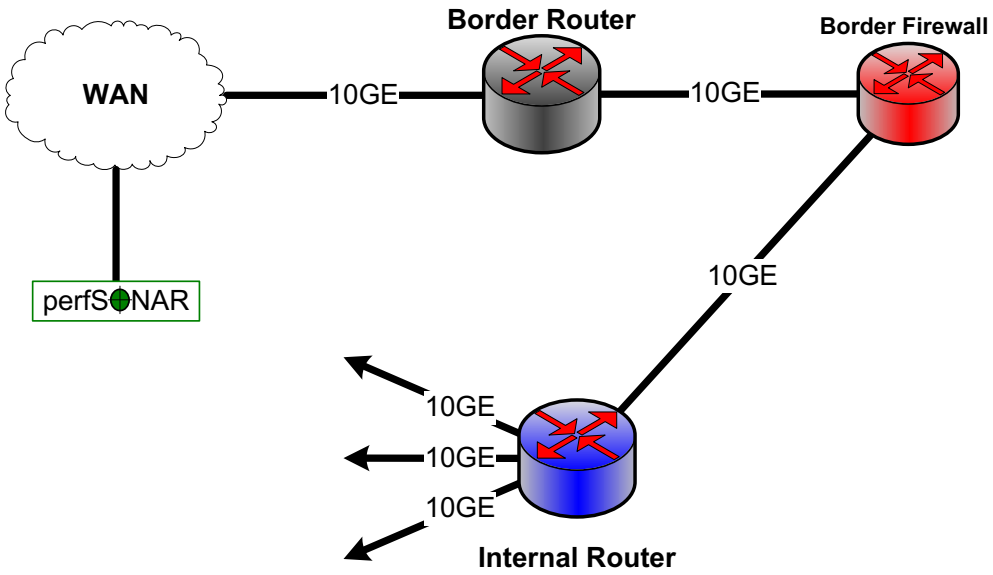
Clearly Not A 10Gbps Network

- If you look at the inside of Building B, it is obvious from a network engineering perspective that this is not a 10Gbps network
  - Clearly the maximum per-flow data rate is 1Gbps, not 10Gbps
  - However, if you convert the buildings into network elements while keeping their internals intact, you get routers and firewalls
  - What firewall did the organization buy? What's inside it?
  - Those little 1G "switches" are firewall processors
- This parallel firewall architecture has been in use for years
  - Slower processors are cheaper
  - Typically fine for a commodity traffic load
  - Therefore, this design is cost competitive and common

Notional 10G Network Between Devices



Notional Network Logical Diagram



Extra Slides – Initial Data Intensive Science Network

Ad Hoc DTN 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

- If present, perfSONAR is 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
- This usually fails to provide the necessary performance.