ESnet
ENERGY SCIENCES NETWORK

An overview of Science DMZs and Data Transfer Nodes (DTNs)

ESnet Science Engagement

Lawrence Berkeley National Laboratory

Engagement and Performance Operations Center (EPOC)

Science DMZ and DTN Overview
Hands-On Workship Networking Topics-EPOC, NYSERNET, UoSC

April 5, 2022

Ken Miller ken@es.net

Outline

- **Introduction**
- Science DMZ Architecture
- Science DMZ Designs
- Data Transfer Nodes

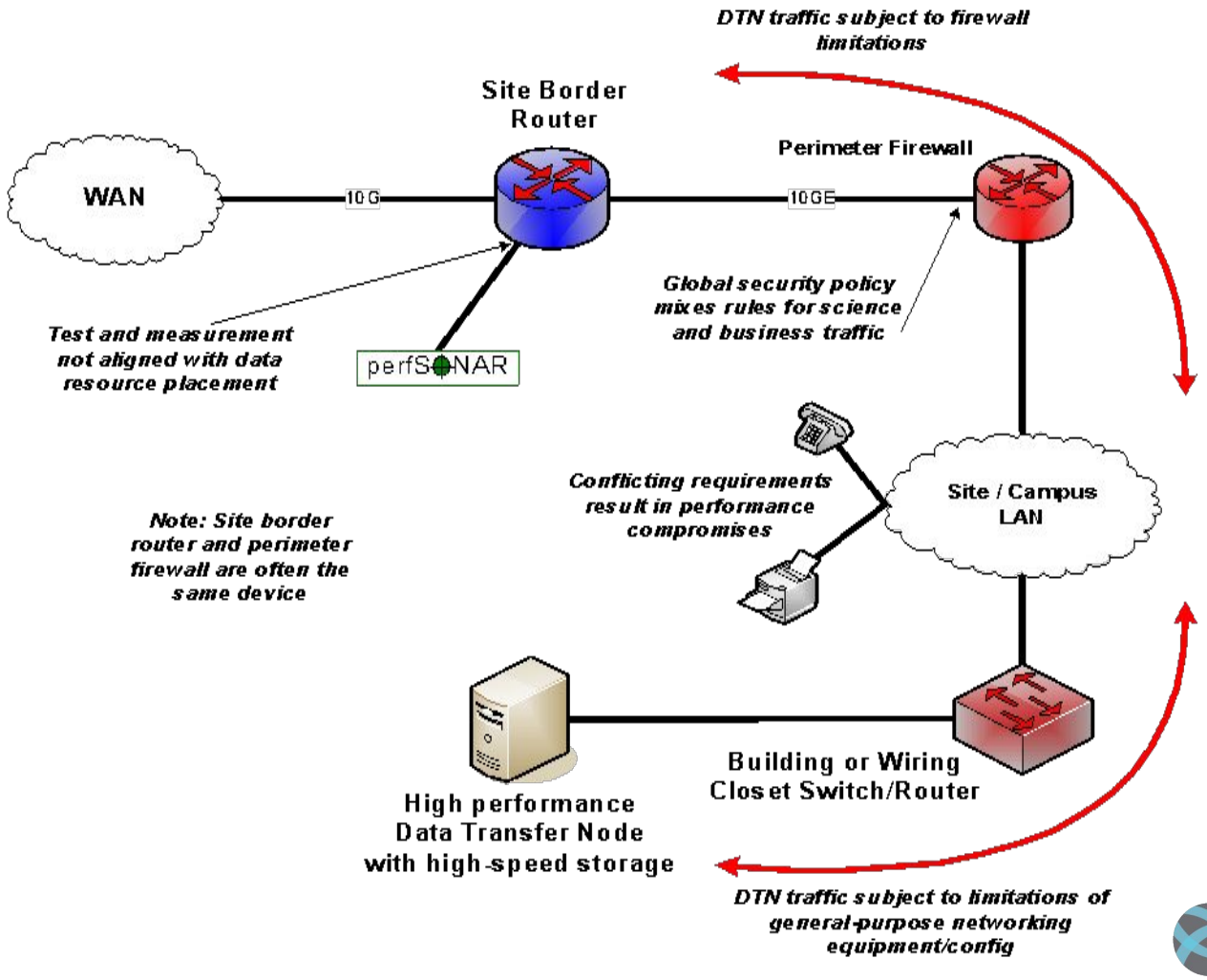
ESnet

Campus Utilization …

- Can campus infrastructure handle research compute, storage, applications, and networking efficiently?
- One researcher's workflow can
  - Use more compute that a campus VM infrastructure
  - More storage than an entire department
  - More network capacity than the entire campus wifi
- Campus infrastructure resources are typically designed around the following:
  - "Speeds and Feeds" of networking access
  - Number of Servers and VMs to run services
  - Storage to facilitate services
- How is this IT infrastructure measured? Uptime/Availability/SLAs
- When is user or application performance measured? After trouble tickets?

ESnet

Legacy Data Transfer Server or Data Portal Deployment



ESnet

Performance At Different Data Scales

Data set size	1 Minute	5 Minutes	20 Minutes	1 Hour
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
Time to transfer				

This table available at:

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

ESnet

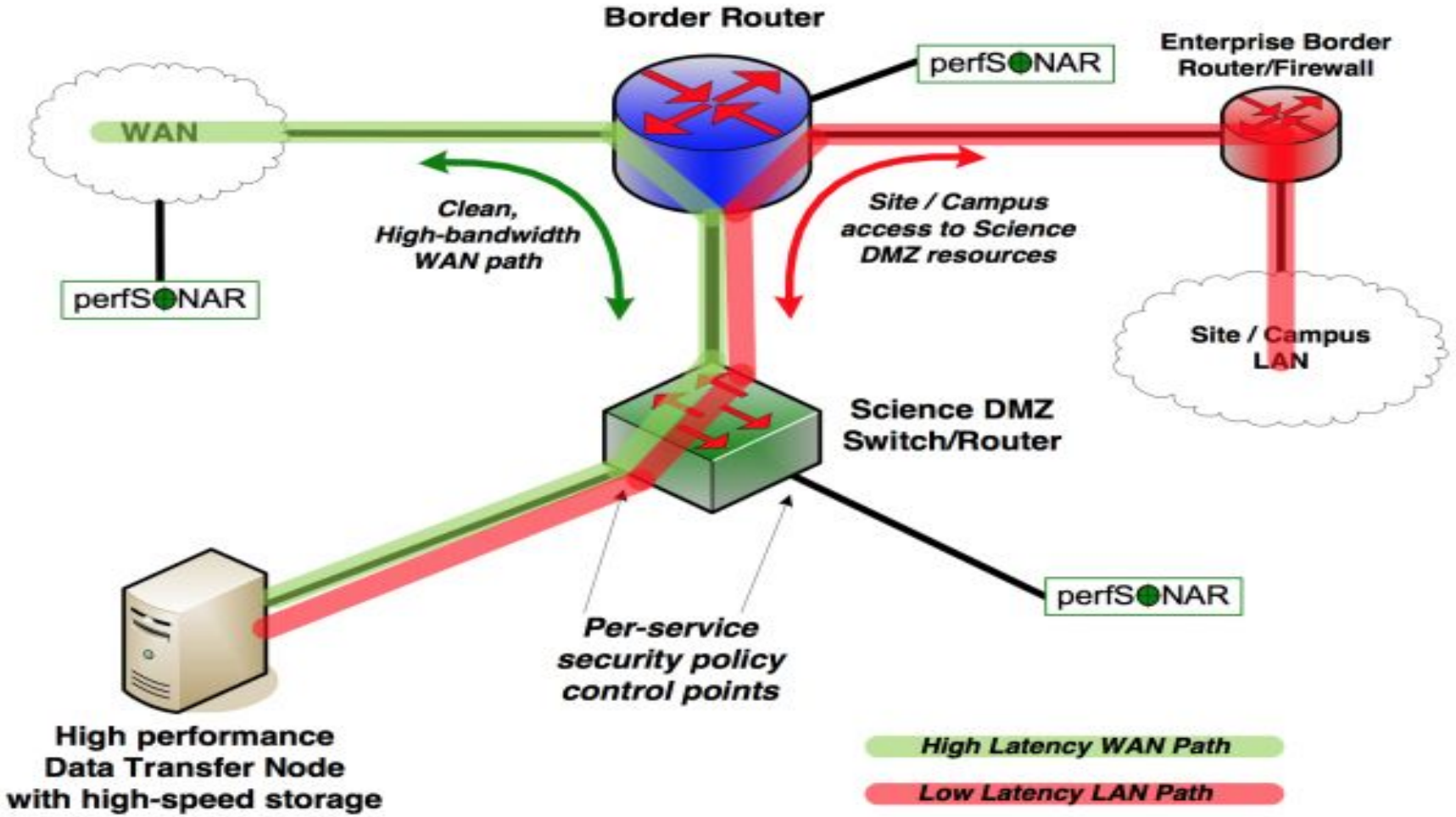
Campus Utilization or Performance? Why not Both!

- What if performance was included in the campus design?
- How well does IT know your researchers requirements and expectations?
- How well do researchers know what to expectat from IT and infrastructure performance?

- Think of this set of content as a reset – we don't want to build IT for the sake of building IT
- Tie infrastructure back to the user/use cases, and be sensible about the design, installation, security, performance, and operation.
- One reference of this type of implementation is the Science DMZ design pattern

ESnet

A Better Approach: Science DMZ Design



ESnet

Outline

- Introduction
- **Science DMZ Architecture**
- Science DMZ Designs
- Data Transfer Nodes

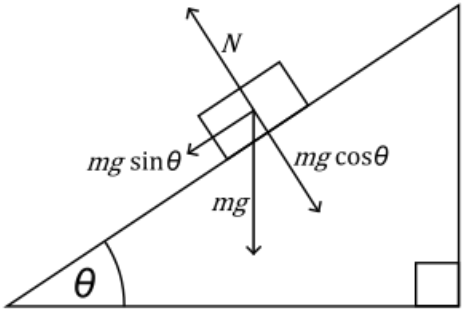
ESnet

The Science DMZ in 1 Slide

Consists of **four 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

globus

Performance measurement/test node

- perfSONAR

perfSONAR

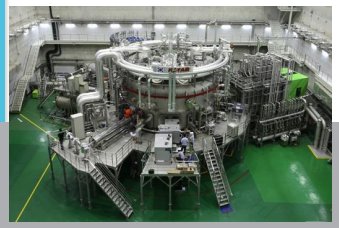
Engagement with users and use cases.

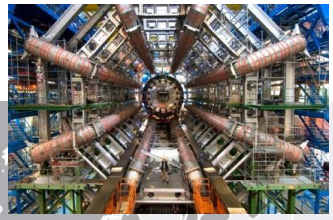
- Education, Partnership, Knowledgebase

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

ESnet

Network as ~~Infrastructure~~ *Instrument*



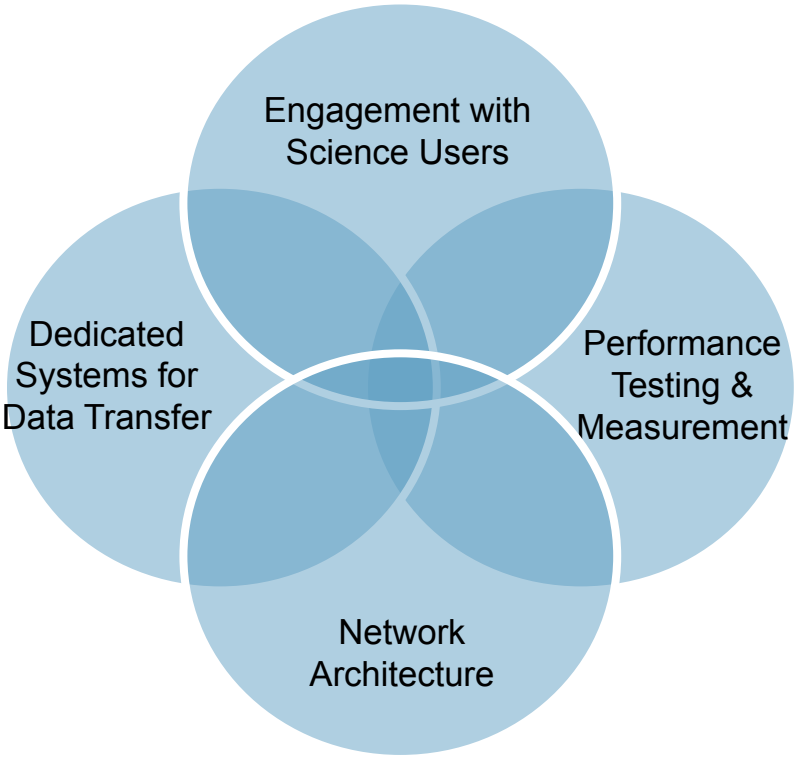


Connectivity is the first step – ***usability*** must follow

The Data Transfer Architecture: Science DMZ Model

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

ESnet

Outline

- Introduction
- Science DMZ Architecture
- **Science DMZ Designs**
- Data Transfer Nodes

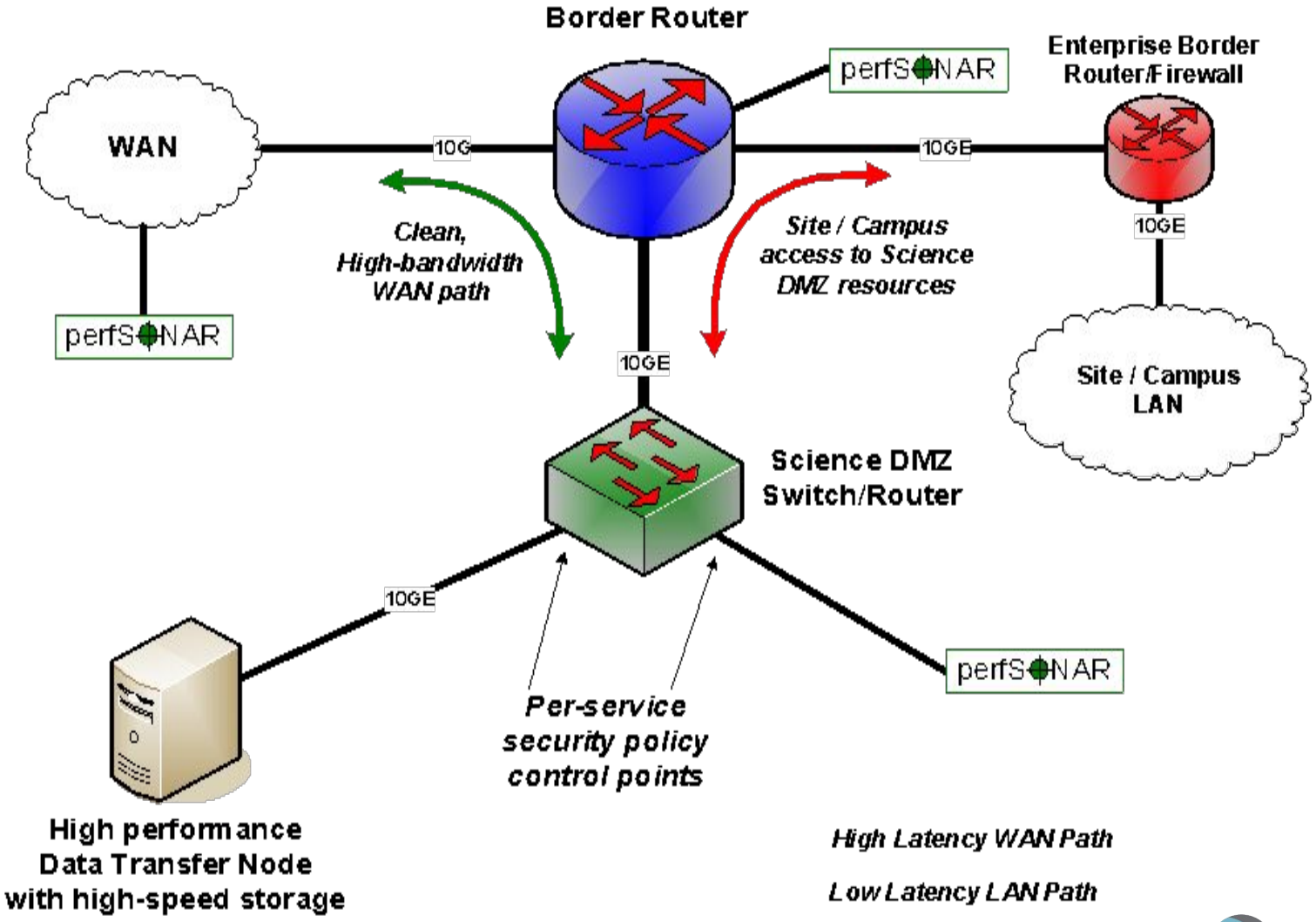
ESnet

Science Data Architecture Examples

- Data Transfer Expectations
- Science DMZ - Data Transfer Node with Local Storage
- Distributed Science DMZ - Data Transfer Node per project
- Multiple Science DMZs – Data Transfer Node per project
- Supercomputer Center Data Architectures Paths -
  - Data Transfer Node with Connected Storage
  - Clustered DTNs with Connected Storage
- Next-Generation Data Portal: Science DMZ, DTN pool, Central Data Store
- Instrument Data Architecture

ESnet

Science DMZ Design Pattern with Local And Wide Area Data Flows

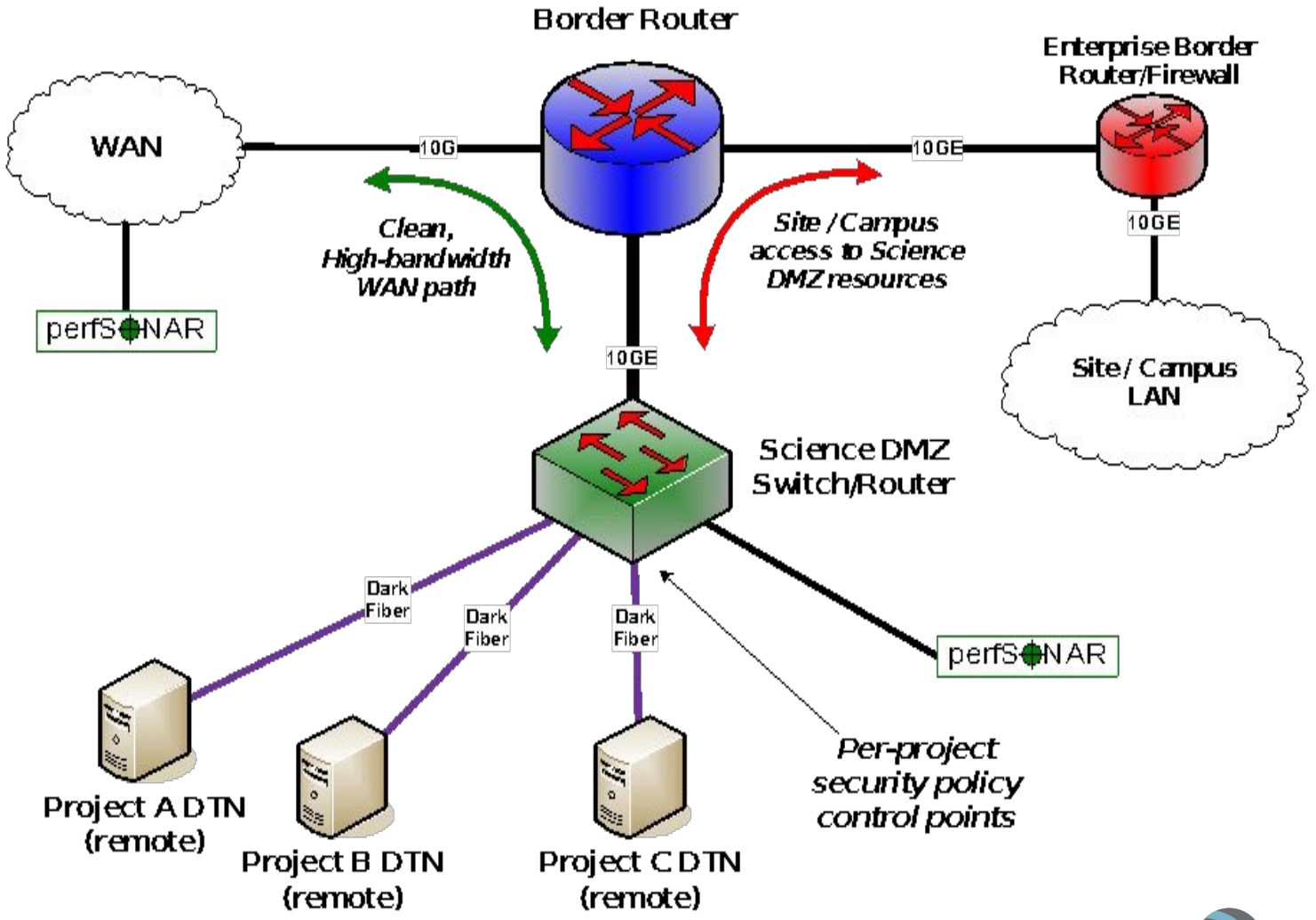


Science DMZ Design Pattern with Local And Wide Area Data Flows

- A network architecture explicitly designed for high-performance applications, where the science network is distinct from the general-purpose network
- The use of dedicated systems for data transfer
- Performance measurement and network testing systems that are regularly used to characterize the network and are available for troubleshooting
- Security policies and enforcement mechanisms that are tailored for high performance science environments

ESnet

Distributed Science DMZ – Dark Fiber



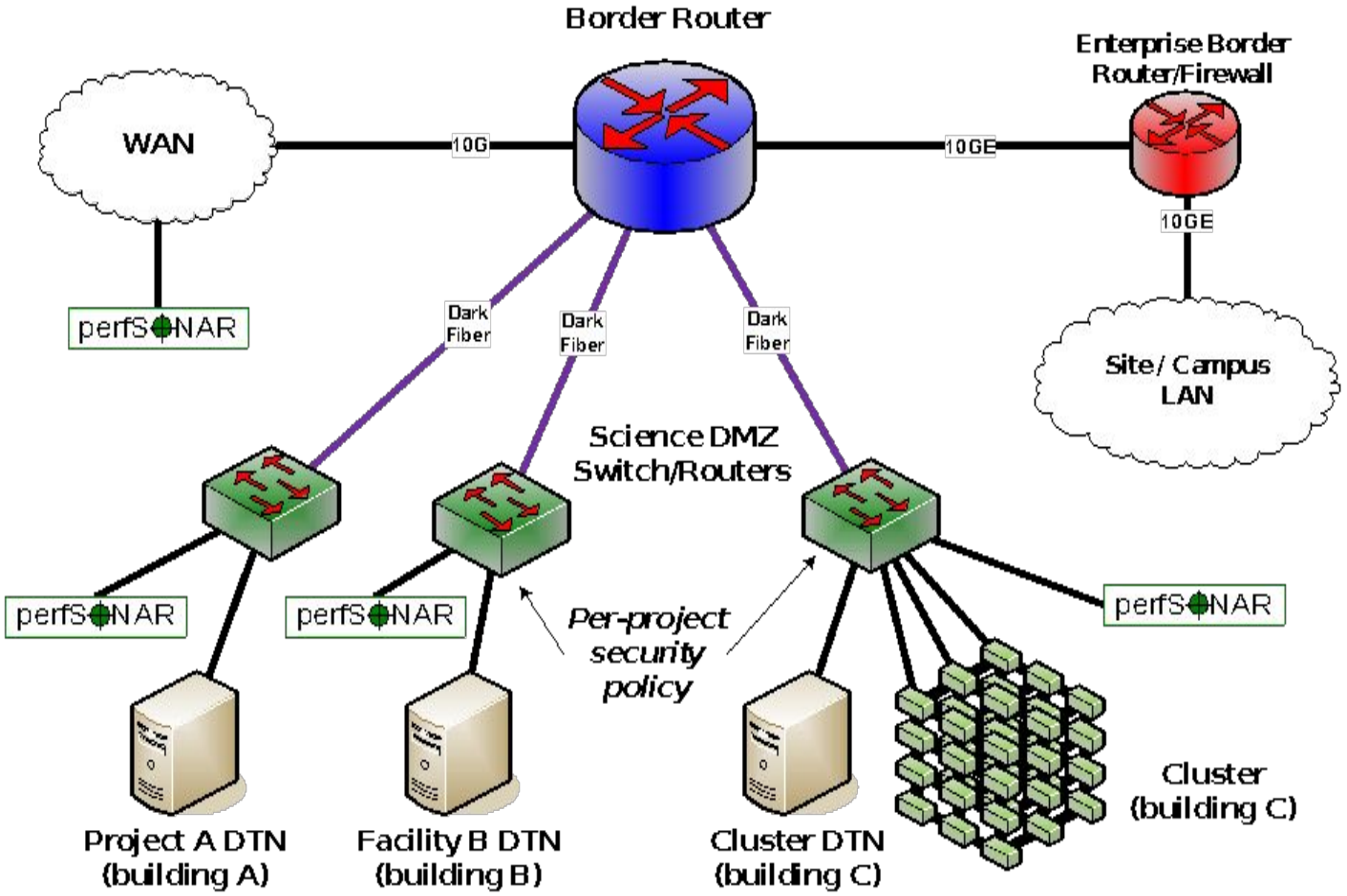
ESnet

Distributed Science DMZ – Dark Fiber

- Pros:
  - Extends Science DMZ and DTN further into the campus with direct fiber instead of traversing over campus network equipment
  - This extended DTN can attach to instrument DTN
  - This design keeps the DTN attached to a deep buffer switch instead of a local access layer switch.
  - Keeps Elephant Science Flows off of campus network
  - Reduced need for deep buffer switch at Science Edge
- Cons:
  - Switching optics from SR to LR
  - Requires use of a dedicated pair of fiber across campus fiber plant
  - Ok for a single DTN host. If more devices, like storage, need connectivity, requires use of a deep buffer switch at Science/Instrument Edge to support the DTN. It's better to connect the DTN directly to storage instead dual homing the DTN.

ESnet

Multiple Science DMZs – Dark Fiber to Dedicated Switches

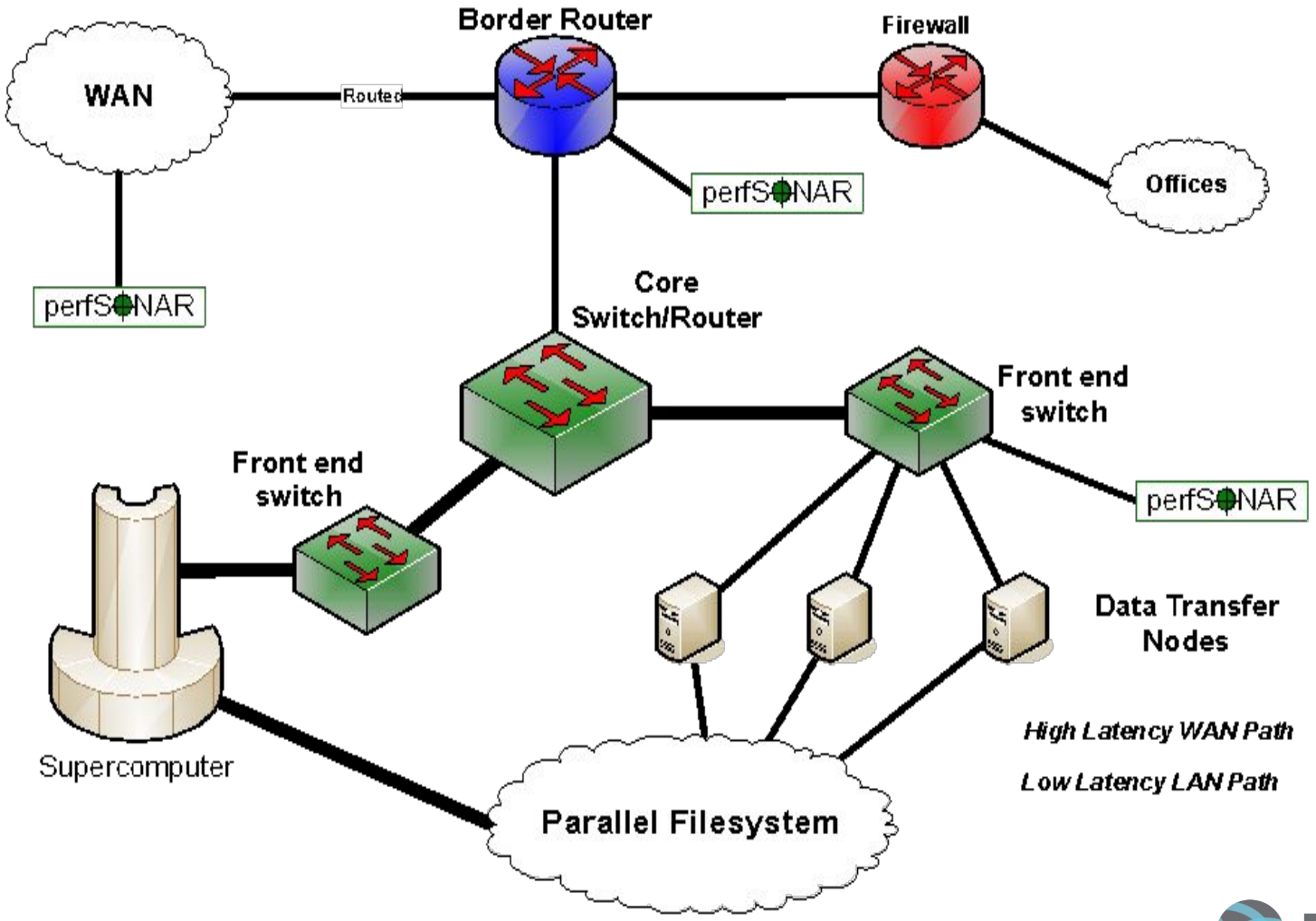


Distributed Science DMZ – Dark Fiber to Dedicated Switches

-
- Pros:
  - Extends Science DMZ and DTN further into the campus research area with direct fiber instead of traversing over campus network equipment
  - This extended switch can attach to multiple instrument DTNs, storage, instrument remote control, sensors, or other instrumentation.
  - This design provides the DTNs, storage, or other instrumentation, a deep buffer switch to connect to instead of a local access layer switch.
  - Keeps Elephant Science Flows off of campus network
  - Provides per-project security controls for multiple projects
- Cons:
  - Requires use of a dedicated pair of fiber across campus fiber plant
  - Requires additional equipment

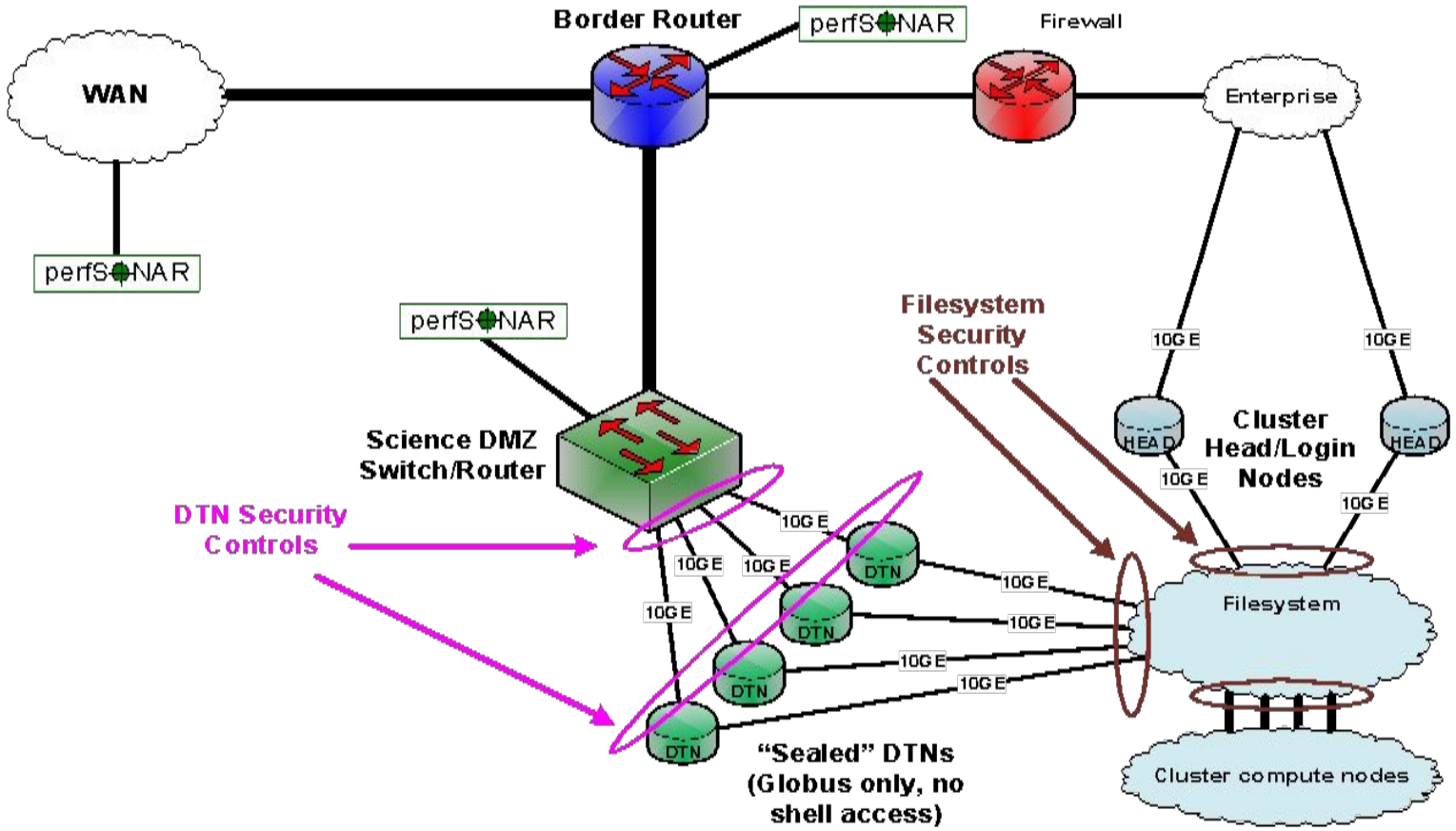
ESnet

SC/HPC Center Data Architectures Paths



ESnet

Clustered DTNs with Connected Storage



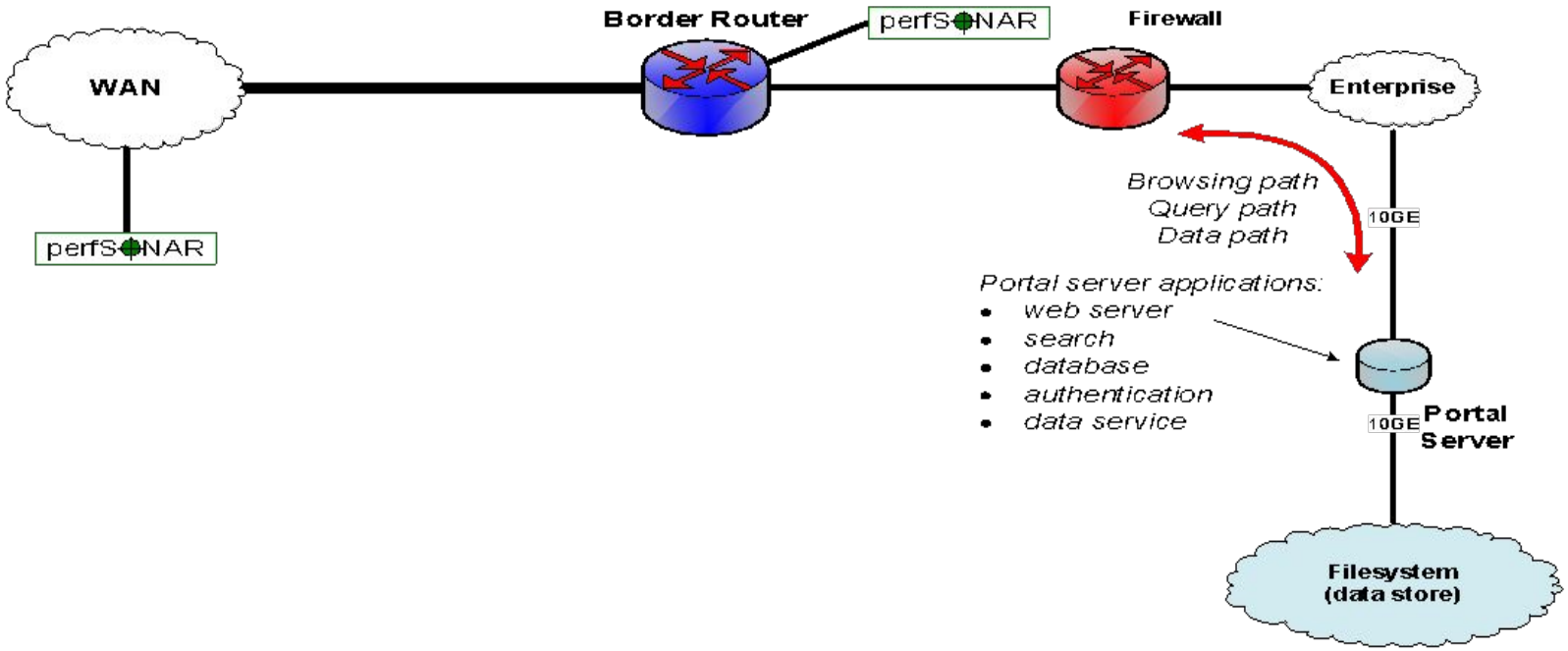
ESnet

Supercomputer Center Data Architectures Paths

- Pros:
  - Clustered DTNs copy data directly to and from SC/HPC storage
  - No Double Copy
  - Multiple DTNs to handle load
  - Multiple DTNs to provide redundancy
  - Follows Science DMZ
  - Splits the Data Transfer Portal Application to just the API access to the transfer tools
- Cons:
  - More equipment to manage
  - More systems and services to secure
  - Multiple devices and layers to troubleshoot

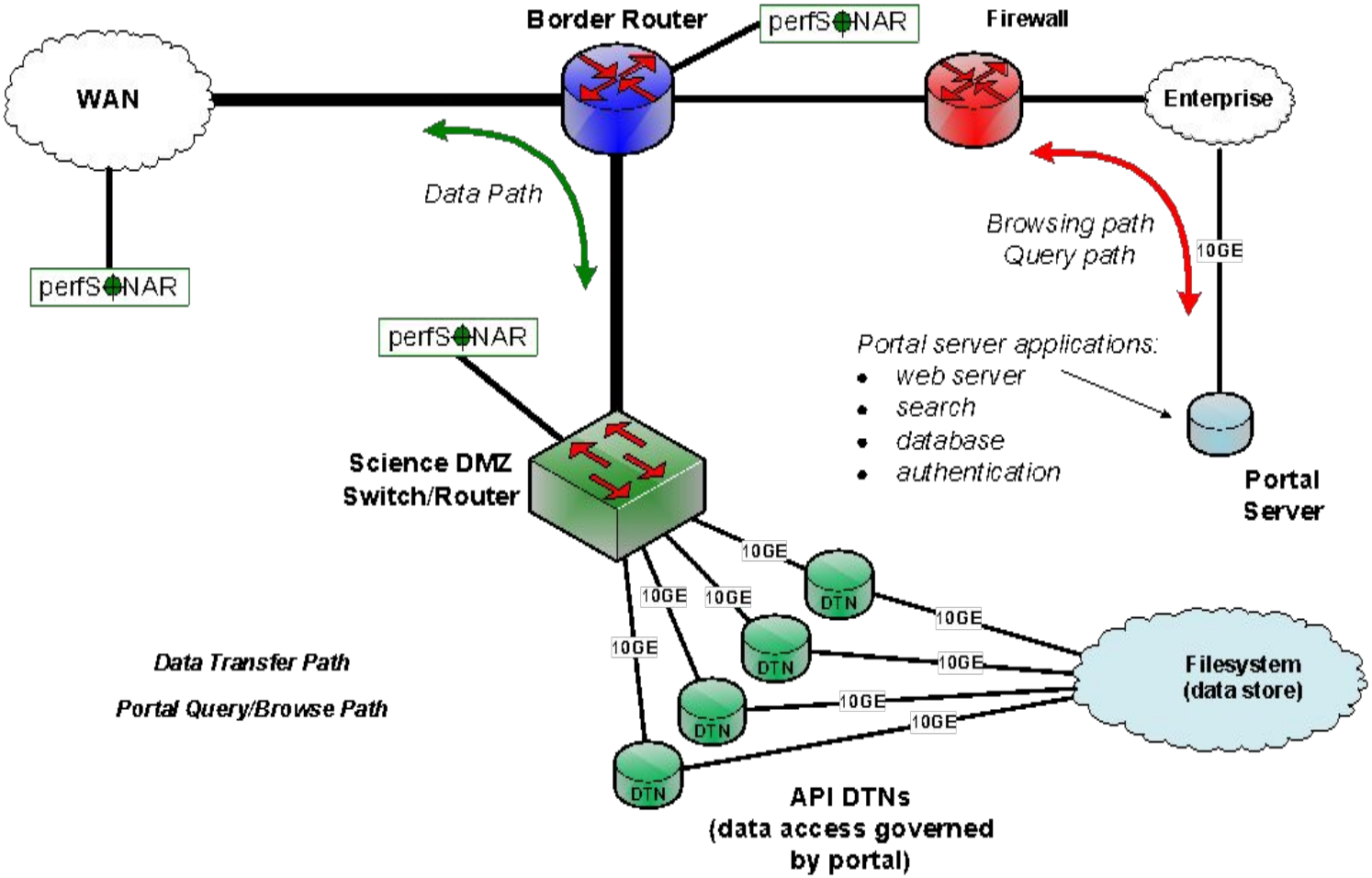
ESnet

Legacy Data Portal Design



ESnet

Next-Generation Portal Leverages Science DMZ, DTN pool, Central Data Store



https://peerj.com/articles/cs-144/

ESnet

Next-Generation Portal Leverages Science DMZ, DTN pool, Central Data Store

Pros

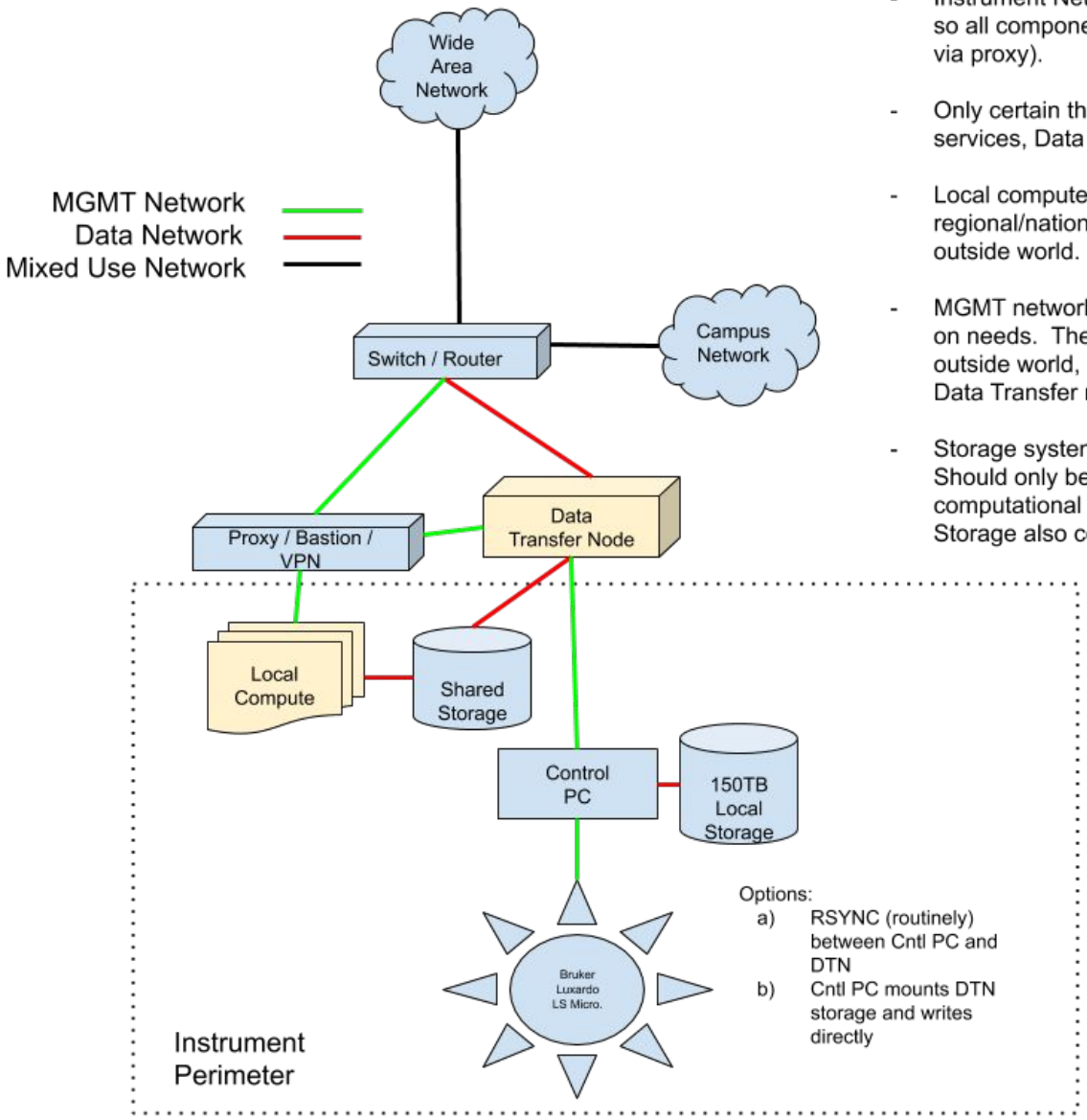
- Data handling is separate, and scalable
  - High-performance DTNs in the Science DMZ
  - Scale as much as you need to without modifying the portal software
- Portal is still its normal self, but enhanced
  - Portal GUI, database, search, etc. all function as they did before
  - Query returns pointers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

Cons

- Separating the data handling from the portal logic is more complex, but performant
- Two or more security profiles to manage
- Multiple points of infrastructure to manage and troubleshoot

ESnet

Instrument Data Architecture



- Instrument Network can features static internal addressing scheme, so all components can function without external networking (except via proxy).
- Only certain things exposed with external address: Proxy/internet services, Data Transfer Node, Bastion/VPN.
- Local compute can be bolted on to complete analysis. Can also use regional/national compute, and use Data Transfer node to send to outside world.
- MGMT network could have connections to multiple things - depends on needs. The idea here is that the control PC is isolated from the outside world, and has to Proxy through either the VPN/Bastion or Data Transfer node.
- Storage system is meant to be protected from external access. Should only be accessible by instrument, data transfer, and computational resources (e.g. establish a 'data VLAN' for access). Storage also could just be inside of the data transfer node.

ESnet

EPOC stories: Other designs causing challenges

- Air-gapped/Isolated research network as a Science DMZ and no firewalls
  - Completely separated from enterprise network
  - Research and Education Routes only limits access to cloud and pub services like github
  - Desktops and workstations on this network
  - Typically unusable
  - **Recommendations: Integrate Research Network as a segment with Enterprise network to make it usable and build Science DMZ for transfers**
- Enterprise network like an ISP has border firewalls, data center firewalls, campus/department firewalls
  - Limits cross campus transfers from instrument to storage in data center
  - Researcher is stating their Cryo-EM transfers off-campus are faster the across campus to campus storage.
  - Numerous MTU mismatches found within DC and HPC network
  - **Recommendations: Tune for jumbo frames and the Science DMZ design pattern can be applied on campus to bypass DC firewall and setup a dedicated DTN for Cryo-EM**

ESnet

Outline

- Introduction
- Science DMZ Architecture
- Science DMZ Designs
- **Data Transfer Nodes**

ESnet

DTN Design, History & Purpose

- Original concept came from initial Science DMZ Design (~2012)
- Basic idea:
  - Host(s) dedicated to the task of data movement (and only data movement)
  - Limited application set (data movement tools), and users (rarely shell access)
  - Specific security policy enforced on the switch/router ACLs
    - Ports for data movement tools, most in a ‘closed wait’ state
    - Nothing to impact the data channel
  - Typically 2 footed:
    - Limited reach into local network (e.g. ‘control channel’: shared filesystem, instruments)
    - WAN piece that the data tools use (e.g. ‘data channel’)
- Position this, and the pS node, in the DMZ enclave near the border

ESnet

Why a DTN?

- DTN = Data Transfer Node
- Purpose built server to transfer data
- Tuned to the performance as necessary
- A tuned 10G is better than 25,40,100G untuned servers
- DTNs can have local storage, connected storage, or both
- Multiple DTNs can be setup for various projects
- Also, Multiple DTNs can be clustered together

- **Match the DTN to the capabilities of the site and wide area network infrastructure**

ESnet

DTN Design Considerations

- Single Resource for moving data
- Largest possible NIC to match needs and cover data transfer speed Typically 10G or 25/40 for campus.
- At a minimum, connect directly to border router with R&E connectivity or in a Science DMZ environment.
- Fast CPU of 3.3 Ghz or greater to support higher speeds
- Multiple CPU cores of 8+ to scale with parallel data transfers
- Sufficient local storage and options to connect external for growth if necessary

Reference Architecture or Use cases:

https://fasterdata.es.net/science-dmz/DTN/hardware-selection/

ESnet

DTN Reference Architecture - wash-dtn1.es.net

- CPU
  - 2 x Intel Cascade lake Xeon Gold 6246
  - 12 cores each
  - 3.3GHz 165W TDP processor
- Memory
  - 12 x 16G DDR4 2933 ECC RDIMM (192G total)
- Disk
  - 10 x Intel P4610 1.6TB U.2/2.5” PCIe NVMe 3.0 x4 Drives
  - 2 x Enterprise 960G 2.5“ SATA SSD (OS, onboard Intel SATA Raid 1)
- Network
  - Mellanox ConnectX-5 EN MCX516A-CCAT 40/50/100GbE dual-port QSFP28 NIC
- Application
  - Globus
  - https://app.globus.org/file-manager?origin_id=2a6a759c-5cfe-4402-ac5e-a06d9d7f7c37&origin_path=%2F

ESnet

Data Mobility Benchmark

- Try to benchmark your DTNs and Data Architectures monthly or after any changes.
- Download ESnet data Climate Data Sets from Wash-DTN1.es.net or another ESnet server to test your write speeds
  - https://app.globus.org/file-manager?origin_id=2a6a759c-5cfe-4402-ac5e-a06d9d7f7c37&origin_path=%2F
  - Climate-Small, ~245GB, 1496 files, 305 folders
  - Climate-Medium, ~245GB, 117 files, 1 folder
  - Climate-Large, ~245GB, 11 files, 1 folder
  - Climate-Huge, ~245GB, 2 files, 1 folder
- For larger systems, try the DME datasets:
  - https://app.globus.org/file-manager?origin_id=5837354e-7087-4d0d-b7bc-e3655f883899&origin_path=%2F
  - ds08, ~1TB, 30076 files, 1 folder
  - ds10, ~1TB, 100 files, 1 folder
  - ds16, ~1TB, 4 files, 1 folder
- Once downloaded, you can re-upload to test your read speeds.

ESnet

Data Transfer Scorecard with Rates by Audience

Host Transfer Rates	⅛ PetaScale (Minimum)	⅓ PetaScale	½ PetaScale		PetaScale: 1 PB/wk	PetaScale: 1 PB/day
	10G Capable DTN				10xG, 25G, 40G, 100G DTNs	
Data Transfer Rate/Volume (Researcher)	1 TB/hr	2 TB/hr	3 TB/hr		5.95 TB/hr	41.67 TB/hr
Network Transfer Rate (Network Admin)	2.22 Gb/s	4.44 Gb/s	6.67 Gb/s		13.23 Gb/s	92.59 Gb/s
Storage Transfer Rate (Sys/Storage Admin)	277.78 MB/s	555.54 MB/s	833.33 MB/s		1.65 GB/s	11.57 GB/s

A benchmark table is provided to gauge data architecture performance, which can vary depending on number of files, folders, size of files, distance between sites, CI performance (network, server, disk/filesystem), as well as data transfer tool.

ESnet

Science DMZ Tuning Recommendations

- Verify your MTUs match for your Science workflow as well as data transfer paths, storage, and nodes: EPOC paper on MTUs
- Adjust BGP to prefer Research and Education Networks
- Setup perfSONAR at your border router as well as near the edge of your science resources (near an instrument)
- Research the Modern Research Data Portal for data distribution
- Setup Weekly Top 10 Source and Destination reports to keep an eye on large transfers. This also help find researchers to engage with. If tools are limited, check out NetSage with your regional network provider.
- Well tuned 10G DTNs go a long way vs a single 100G DTN.
- If large transfer must go through border firewall for Restricted or Controlled data, research if your firewall has an ASIC path (FastPath/ExpressPath)
- DTN Tuning: https://fasterdata.es.net/DTN/tuning/

ESnet

ESnet
ENERGY SCIENCES NETWORK

Questions?

EPOC contact
epoc@iu.edu

Ken Miller
ken@es.net

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB