Designing, Building, & Maintaining a Science DMZ and Data Architecture

Ken Miller, Jason Zurawski
ken@es.net, zurawski@es.net
ESnet / Lawrence Berkeley National Laboratory

Materials Cyberinfrastructure for Research Data Management Workshop
Princeton, NJ
May 23-24, 2023

National Science Foundation Award #2328479

Outline

- ***Introduction***
- Solution Space
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

Measurable Outcomes

- How to prepare for research use of technology on a campus?
  - "Build CI" – but is this all?
  - ***Improve scientific outcomes in some measurable way***
- Things to consider (pre, during, post):
  - Has it/will it all work as expected? (e.g. more than plugging in wires)
  - Will we ***all*** be satisfied at the end? (researchers are the survey population, not just the IT org …)
  - How do you know when you/we are done? Are you ever done?
- Think of this set of content as a reset – we don't want to build IT for the sake of building IT
  - Tie things back to the user/use cases, and be sensible about the design, installation, and operation

EPOC
Engagement and Performance Operations Center

Network as ~~Infrastructure~~ *Instrument*

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

EPOC
Engagement and Performance Operations Center

Where do we go from here?

1. WHO and WHY are the most important questions
   - Know the users, know the use cases. Those will guide any technical solution
2. WHAT and HOW go together:
   - Figuring out the technology that will help without causing non-productive disruptions
   - Being able to sensibly design, implement, and operate

*We **have** to address item 1, then we will dive into item 2*

EPOC
Engagement and Performance Operations Center

Outline

- Introduction
- ***Solution Space***
  1. ***Understanding the Solution Space (Users, Use Cases, Long Term Impacts)***
  2. Preliminaries (e.g. Network Protocols 101)
  3. Architecture & Design
  4. Monitoring and Measurement
  5. Data Mobility
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

Common Theme / New Mindset

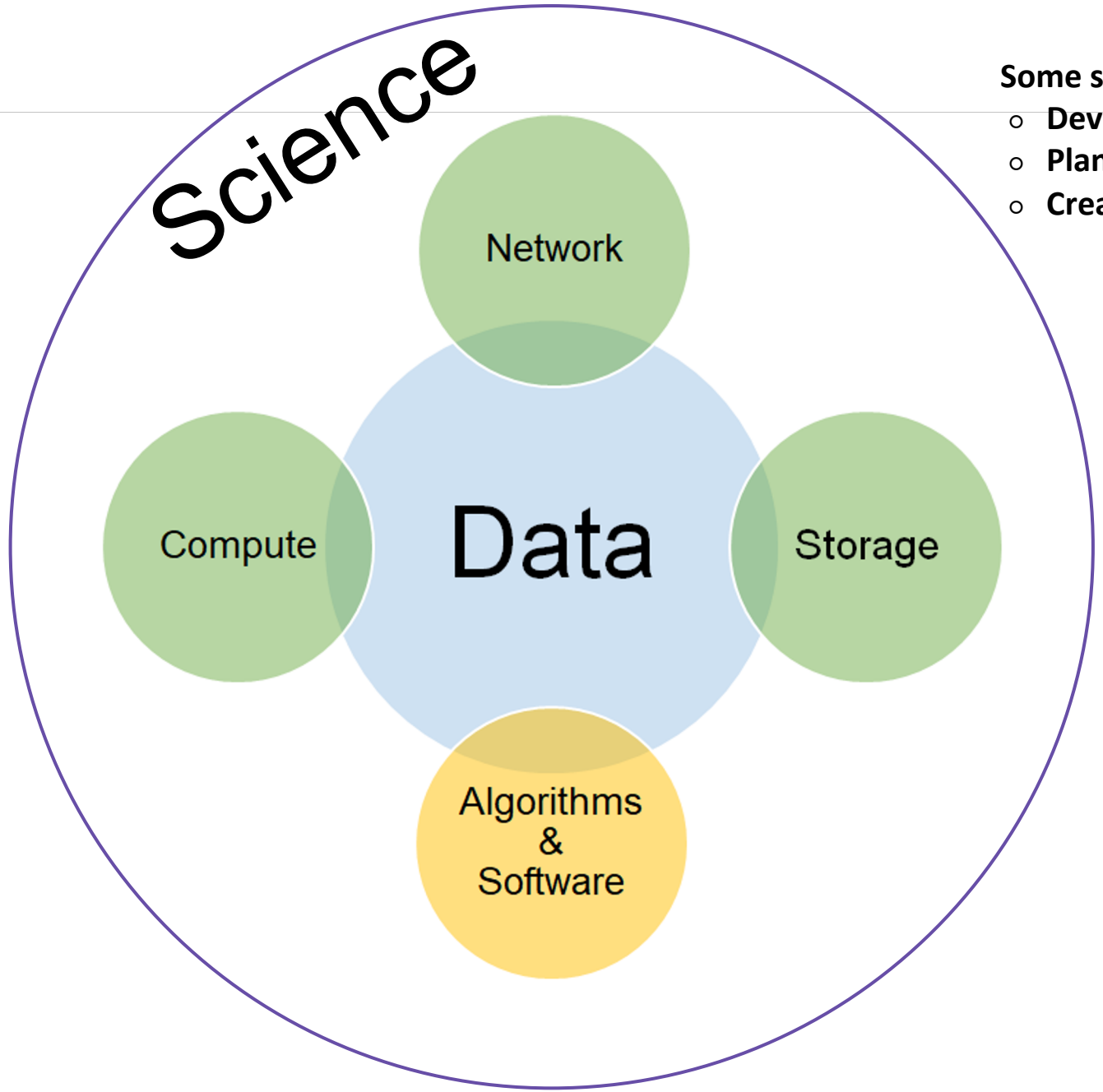
- We aren't building a "Network Architecture", we want a "Data Architecture"
  - A lot of the items that will be thrown at you transcend the traditional network space.
- To get there:
  - Understand the data pipeline for your target user/use case – cradle to retirement home
  - This implies all the things:
    - Creation
    - Usage
    - Transfer/Share
    - Curation

EPOC
Engagement and Performance Operations Center

Common Theme / New Mindset

- What you build must be
  - ***Usable*** – if this becomes a ‘walled garden’, what’s the point? Make it such that people can be easily onboarded and integrated.
  - ***Defensible*** – it is not, nor should it be, the wild west. Control the users and use cases, but don’t impact the usage.
  - ***Scalable*** – as demand grows. Think cornfields and baseball diamonds.
  - ***an institutional capability / source of pride*** – this is something that will draw more users / research dollars if created/marketed/operated correctly. Treat it as such.

EPOC
Engagement and Performance Operations Center

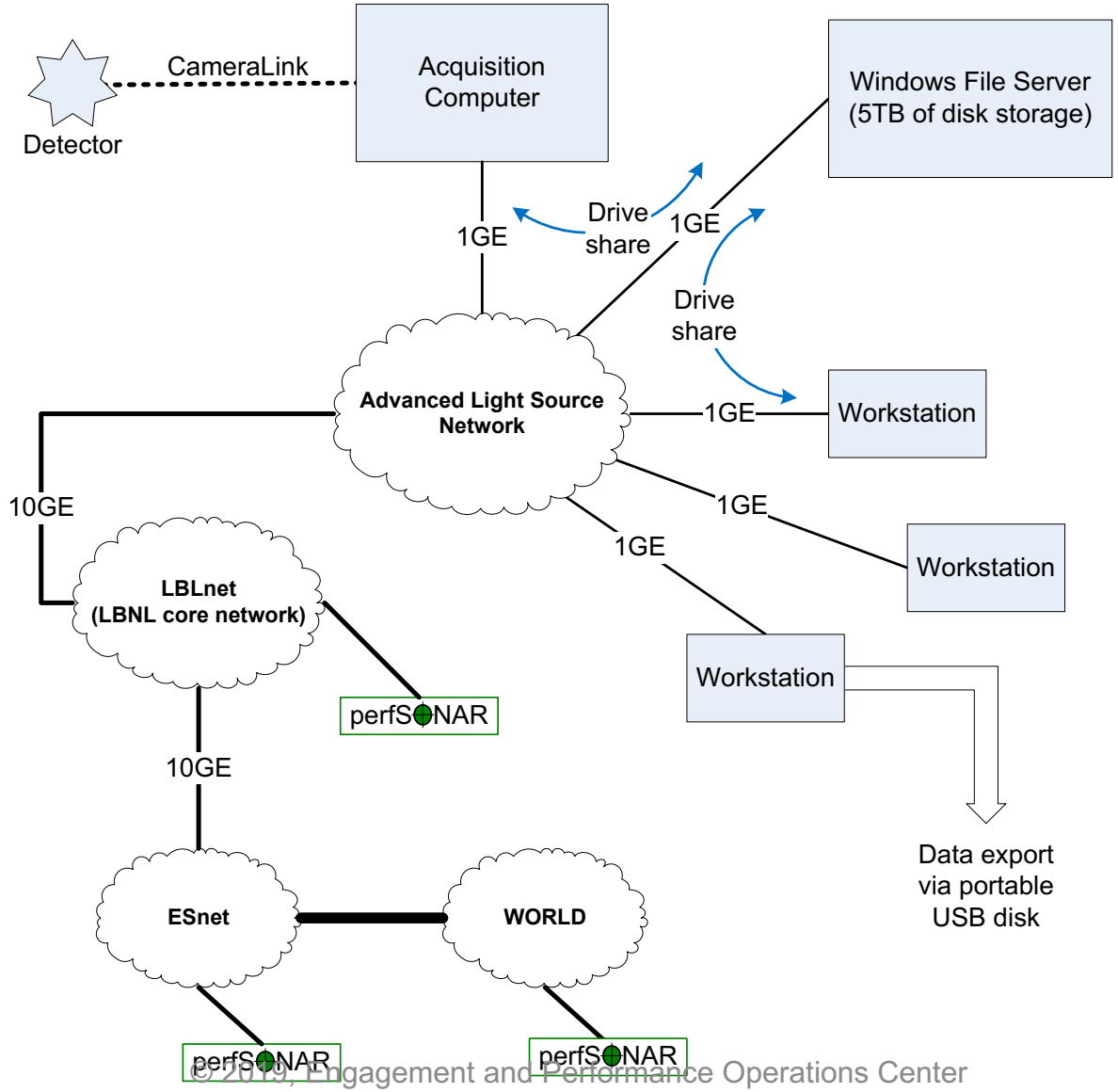


Some specific issues for networks are

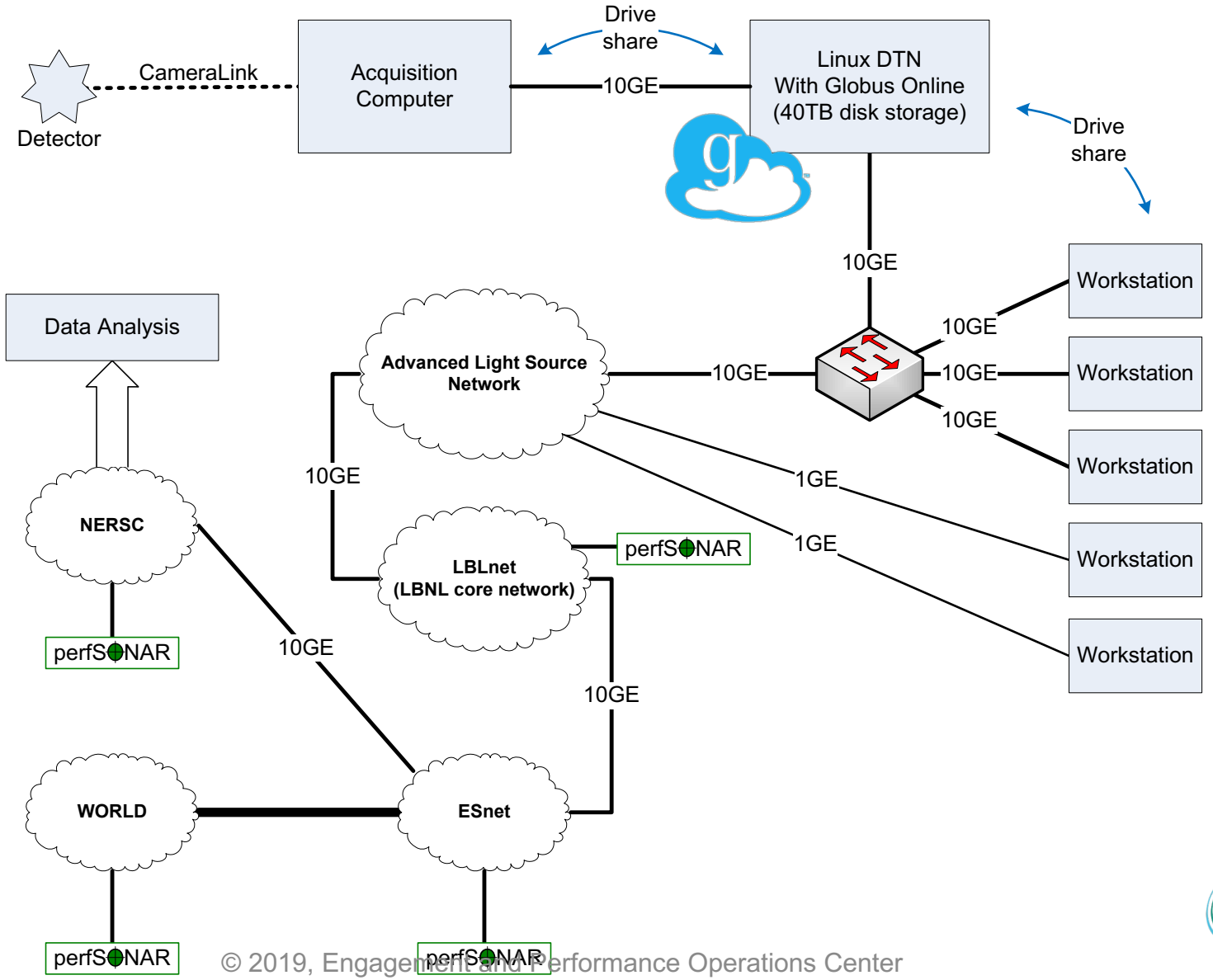
- **Development of services**
- **Planning capacity growth**
- **Creation of collaborations**

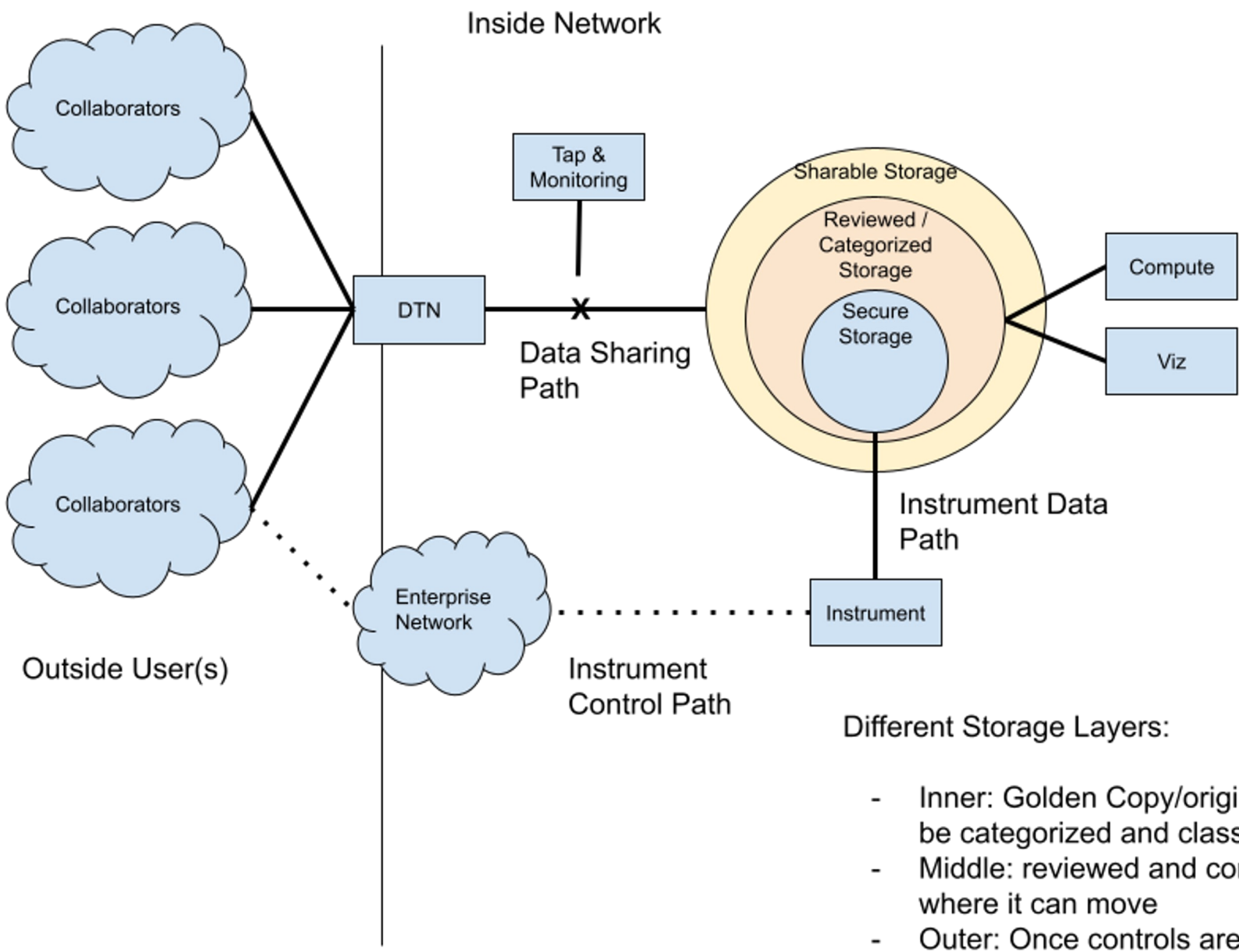
EPOC
Engagement and Performance Operations Center

Science Workflow Consultation



Improved Workflow Infrastructure





Different Storage Layers:

- Inner: Golden Copy/origin until it can be categorized and classified
- Middle: reviewed and controls placed where it can move
- Outer: Once controls are in place, it can be sent to different use cases, maybe several of these (for internal or external use)

EPOC
Engagement and Performance Operations Center

Outline

- Introduction
- ***Solution Space***
  1. Understanding the Solution Space (Users, Use Cases, Long Term Impacts)
  2. ***Preliminaries (e.g. Network Protocols 101)***
  3. Architecture & Design
  4. Monitoring and Measurement
  5. Data Mobility
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

Data Movement / TCP Background

- The data mobility performance requirements for data intensive science are beyond what can typically be achieved using traditional methods
  - Default host configurations (TCP, filesystems, NICs)
  - Converged network architectures designed for commodity traffic
  - Conventional security tools and policies
  - Legacy data transfer tools (e.g. SCP, FTP)
  - Wait-for-trouble-ticket operational models for network performance

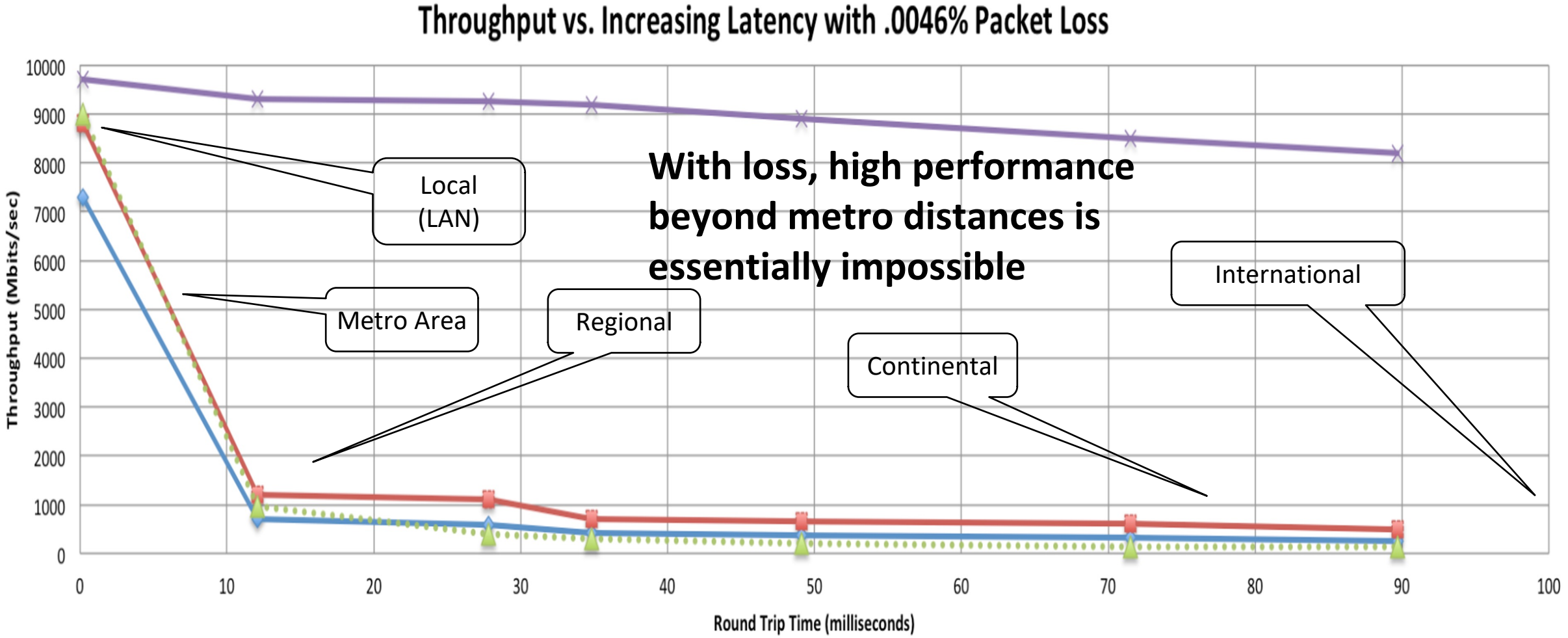
EPOC
Engagement and Performance Operations Center

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
  - ***Congestion*** dictates performance – back off when danger is sensed to preserve/protect resources
- 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



Data Movement / TCP Background

- The Science DMZ model describes a performance-based approach
  - Dedicated infrastructure for wide-area data transfer
    - Well-configured data transfer hosts with modern tools
    - Capable network devices
    - High-performance data path which does not traverse commodity LAN
  - Proactive operational models that enable performance
    - Well-deployed test and measurement tools (perfSONAR)
    - Periodic testing to locate issues instead of waiting for users to complain
  - Security posture well-matched to high-performance science applications

EPOC
Engagement and Performance Operations Center

Th

Consi

- "Fri
  -
  -
  -
  -
- Dec
  -
  -
- Per
  -
- Eng

osθ

AR

z/

Outline

- Introduction
- ***Solution Space***
  1. Understanding the Solution Space (Users, Use Cases, Long Term Impacts)
  2. Preliminaries (e.g. Network Protocols 101)
  3. ***Architecture & Design***
  4. Monitoring and Measurement
  5. Data Mobility
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

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: eliminate sources of packet loss / packet friction

EPOC
Engagement and Performance Operations Center

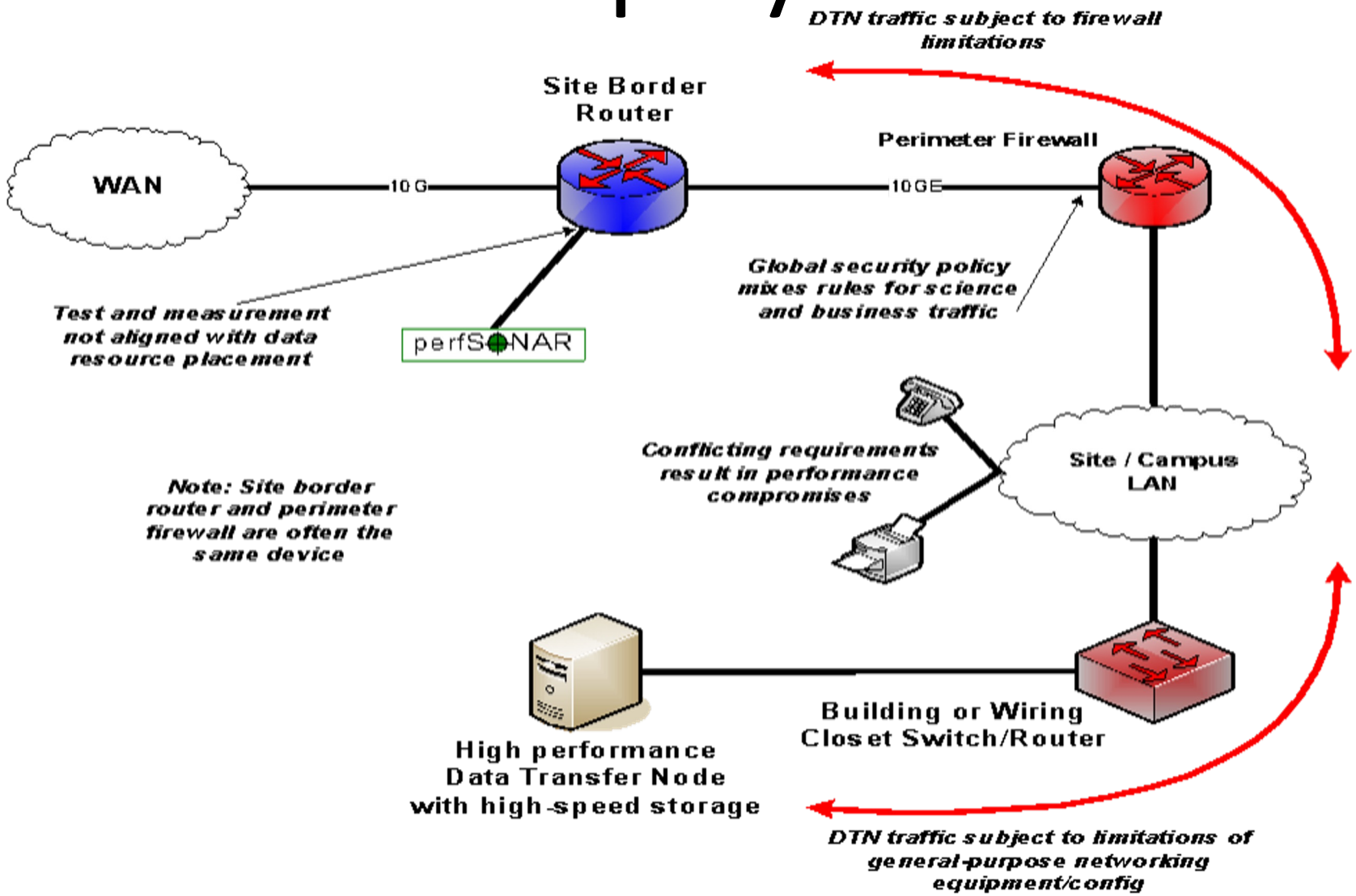
Legacy Method: Ad Hoc DTN Deployment

- 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 lucky, 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.

EPOC
Engagement and Performance Operations Center

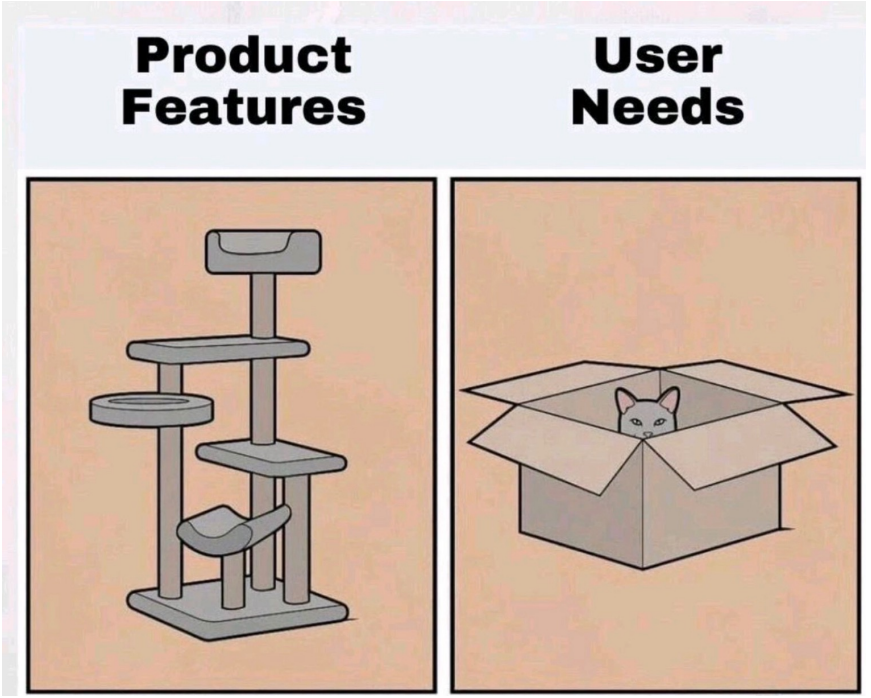
Ad Hoc DTN Deployment



EPOC
Engagement and Performance Operations Center

Design Thoughts – E.g. where does the capability live?

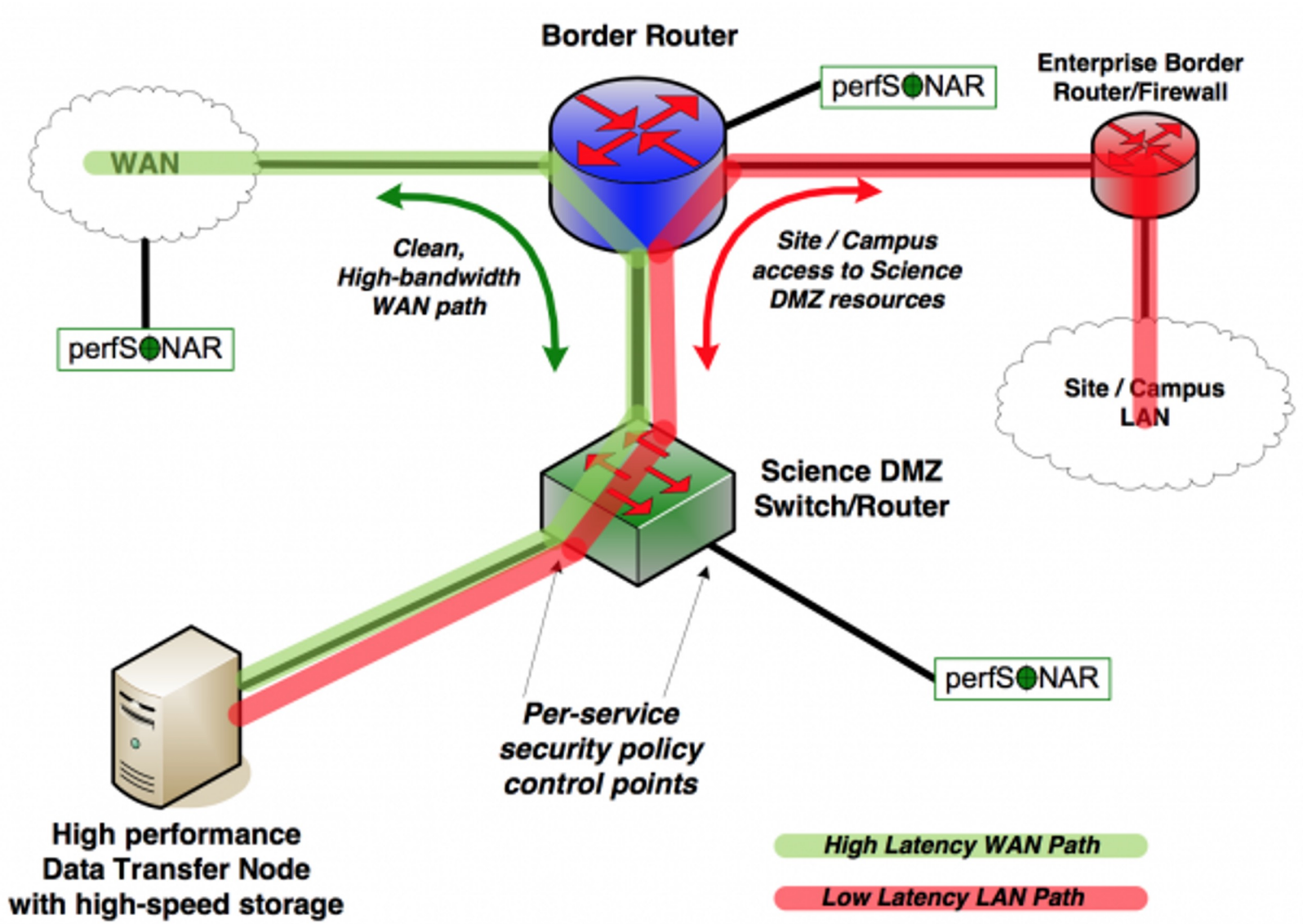
- There are two main schools of thought when going after the CC* design area:
  1. Create an enclave for the science use cases designed specifically to facilitate them. Move things in 1:1
  2. Drop in replacement of a network such that we can capture a bunch of use cases with a big net (*and not dig in too deeply*)



- A strong proposal adopts a stance closer to 1 – its important to know use cases. Its also important to justify expense of equipment to a need
- Number 2 is easier to accomplish (e.g. design, potentially to operate) since its just a 'bigger/better' network that could lift all boats.

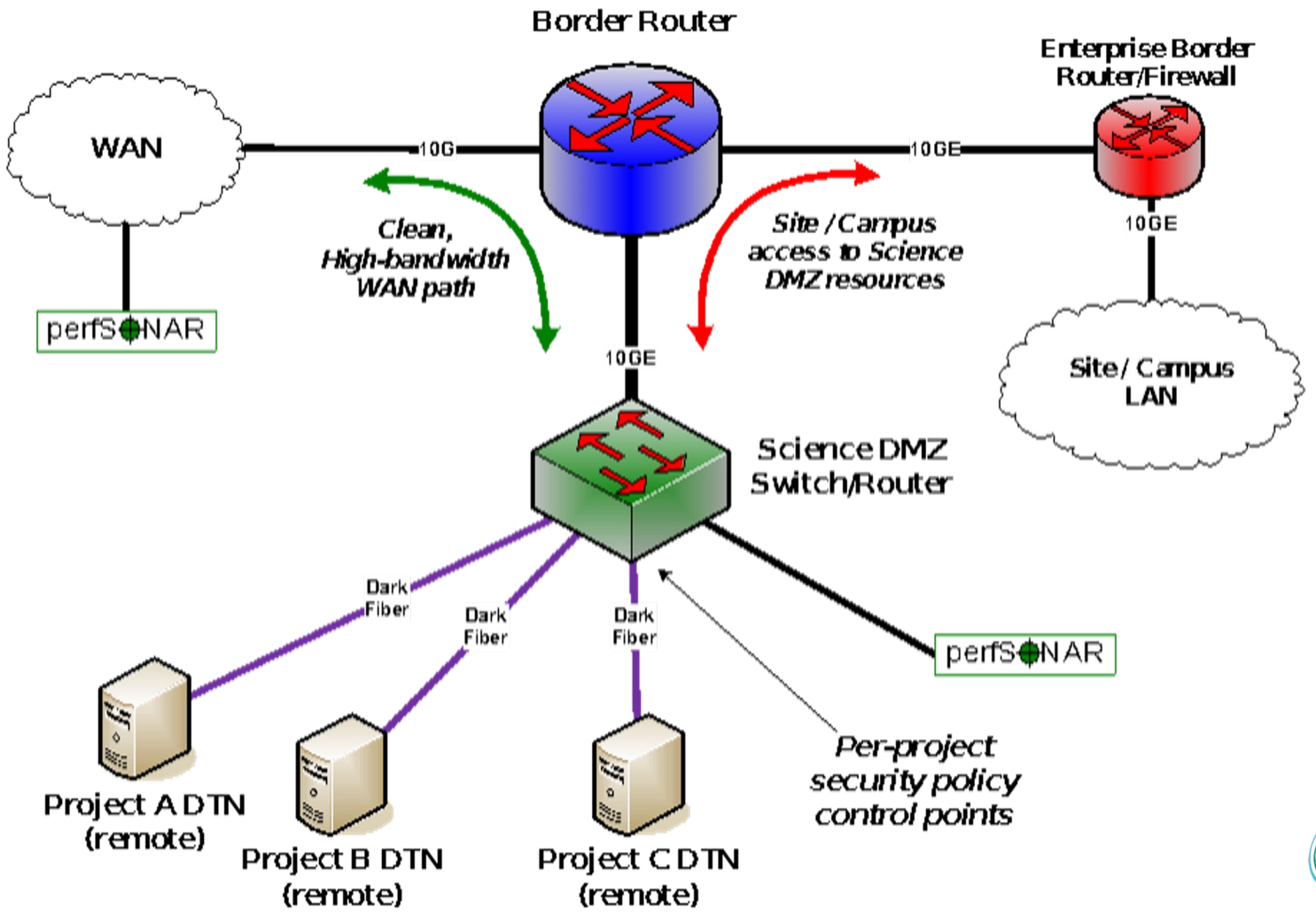
EPOC
Engagement and Performance Operations Center

A better approach: simple Science DMZ



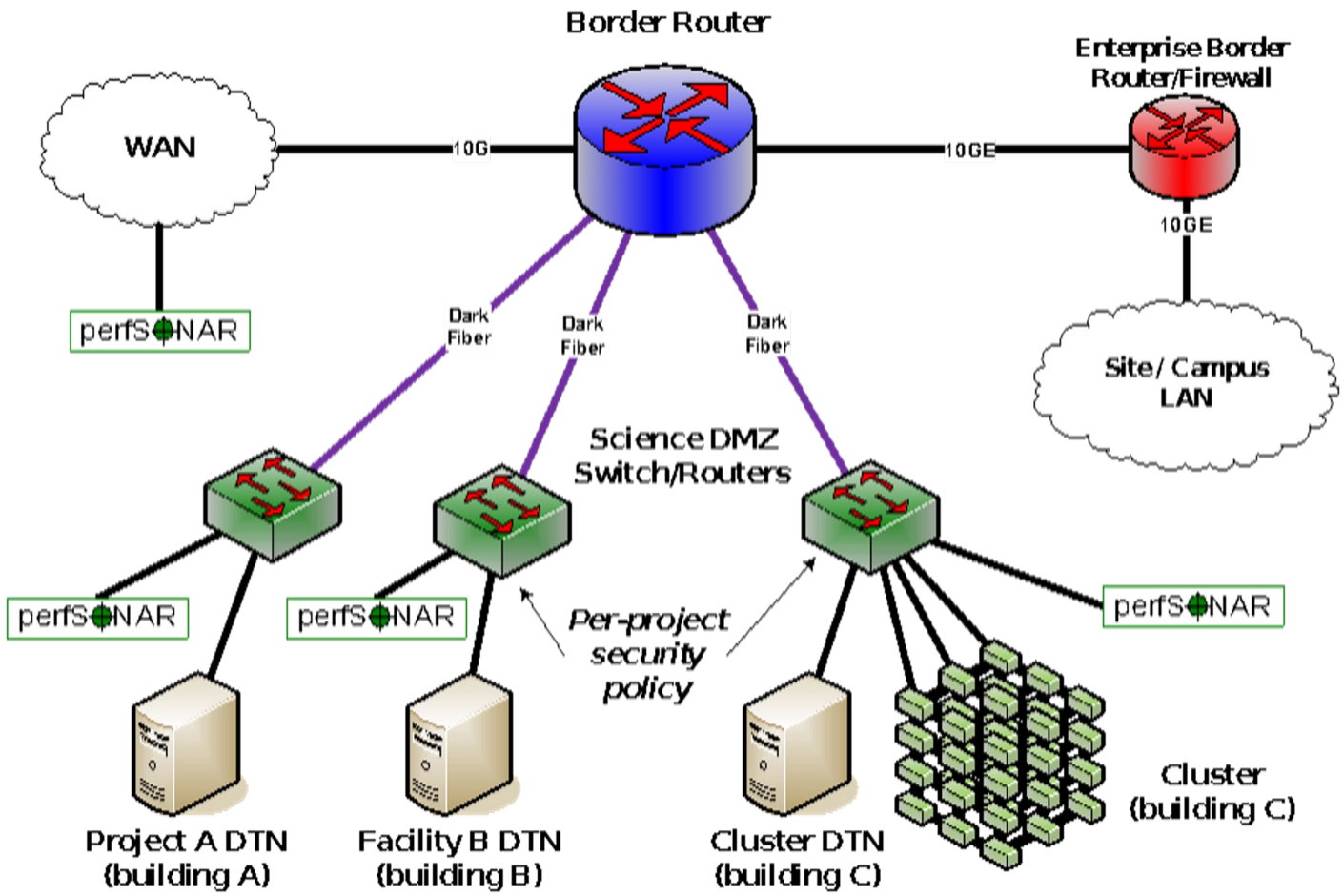
EPOC
Engagement and Performance Operations Center

Distributed Science DMZ – Dark Fiber



EPOC
Engagement and Performance Operations Center

Multiple Science DMZs – Dark Fiber to Dedicated Switches



EPOC
Engagement and Performance Operations Center

Design Thoughts – E.g. where does the capability live?

- There are two main schools of thought when going after the CC* design area:
  1. Create an enclave for the science use cases designed specifically to facilitate them. Move things in 1:1
  2. Drop in replacement of a network such that we can capture a bunch of use cases with a big net (*and not dig in too deeply*)
- A strong proposal adopts a stance closer to 1 – its important to know use cases. Its also important to justify expense of equipment to a need
- Number 2 is easier to accomplish (e.g. design, potentially to operate) since its just a ’bigger/better’ network that could lift all boats.

EPOC
Engagement and Performance Operations Center

Equipment – Routers and Switches

- Requirements for Science DMZ gear are different than the enterprise
  - No need to go for the kitchen sink list of services
  - A Science DMZ box only needs to do a few things, but do them well
  - Support for the latest LAN integration magic with your Windows Active Directory environment is probably not super-important
  - A clean architecture is important
    - How fast can a single flow go?
    - Are there any components that go slower than interface wire speed?
- There is a temptation to go cheap
  - Hey, it only needs to do a few things, right?
  - You typically don't get what you don't pay for
    - (You sometimes don't get what you pay for either)

EPOC
Engagement and Performance Operations Center

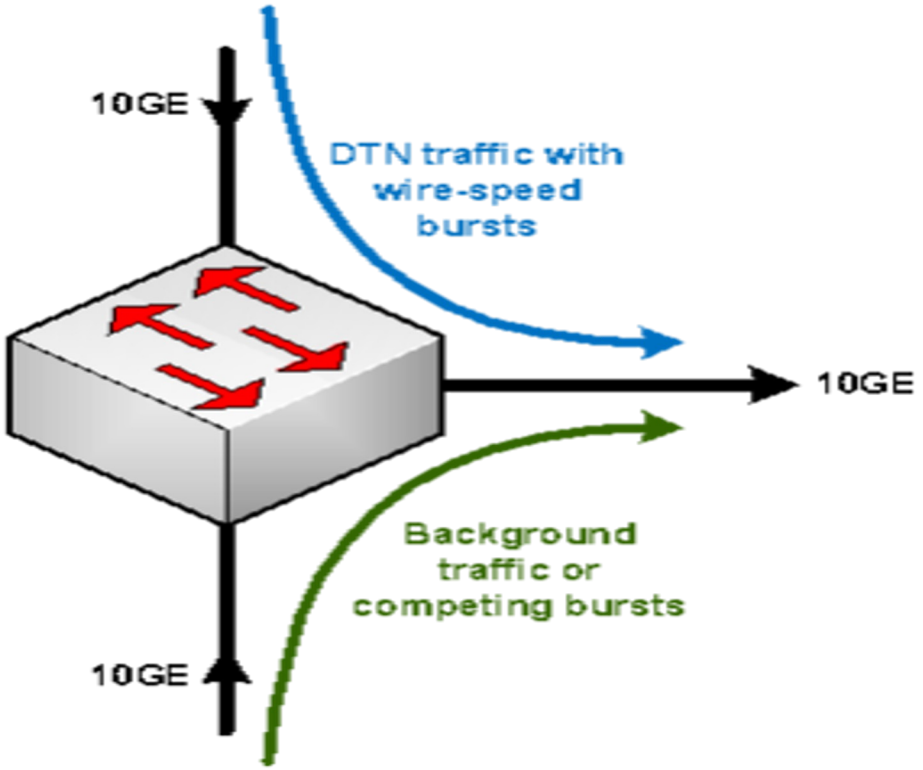
Common Circumstance: Multiple Ingress Data Flows, Common Egress

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



EPOC
Engagement and Performance Operations Center

Some Stuff We Think Is Important

- Deep interface queues (e.g. ***buffer***)
  - Output queue or VOQ – doesn't matter
  - What TCP sees is what matters – fan-in is *not* your friend
  - No, this isn't buffer bloat
- Good counters
  - We like the ability to reliably count *every* packet associated with a particular flow, address pair, etc
    - Very helpful for debugging packet loss
    - Must not affect performance (just count it, don't punt it)
    - sflow support if possible
  - If the box is going to drop a packet, it should increment a counter somewhere indicating that it dropped the packet
    - Magic vendor permissions and hidden commands should not be necessary
    - Some boxes just lie – run away!
- Single-flow performance should be wire-speed

EPOC
Engagement and Performance Operations Center

All About That Buffer (No Cut Through)

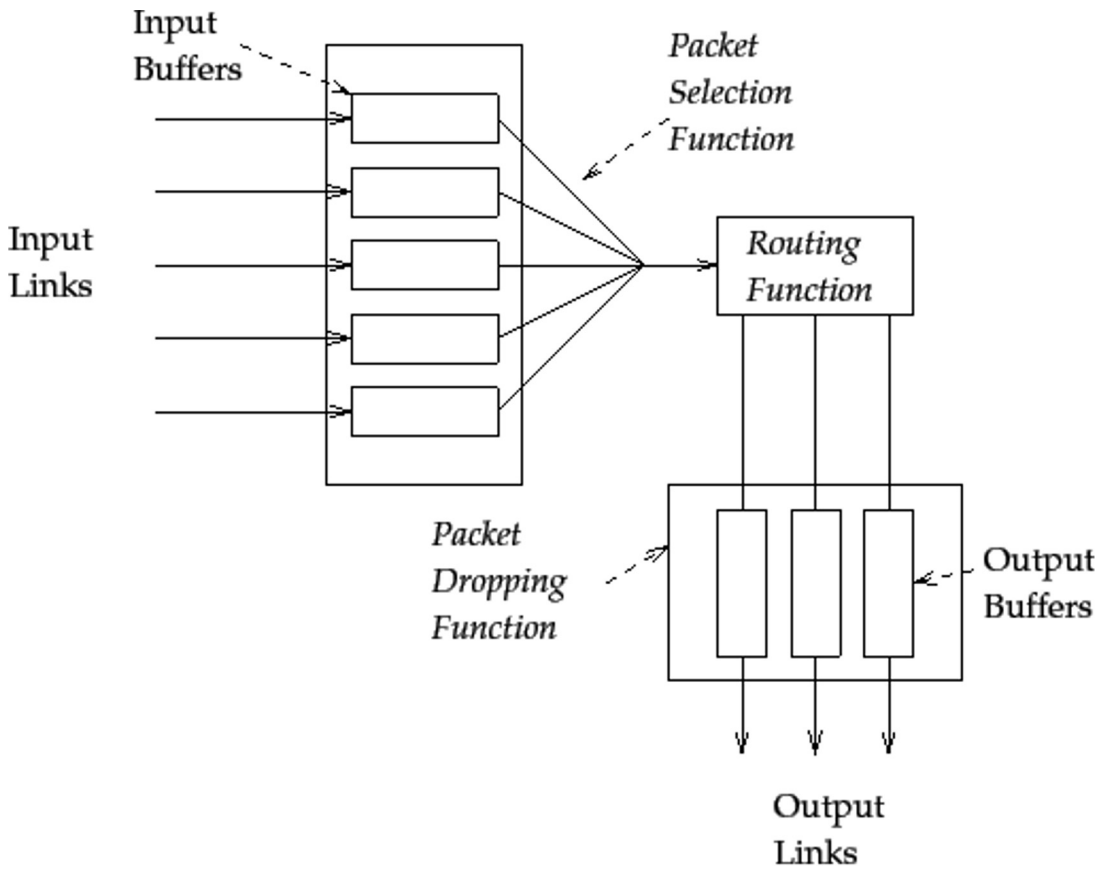


Figure 1: Basic Router Architecture

EPOC
Engagement and Performance Operations Center

All About That Buffer (No Cut Through)

- Data arrives from multiple sources

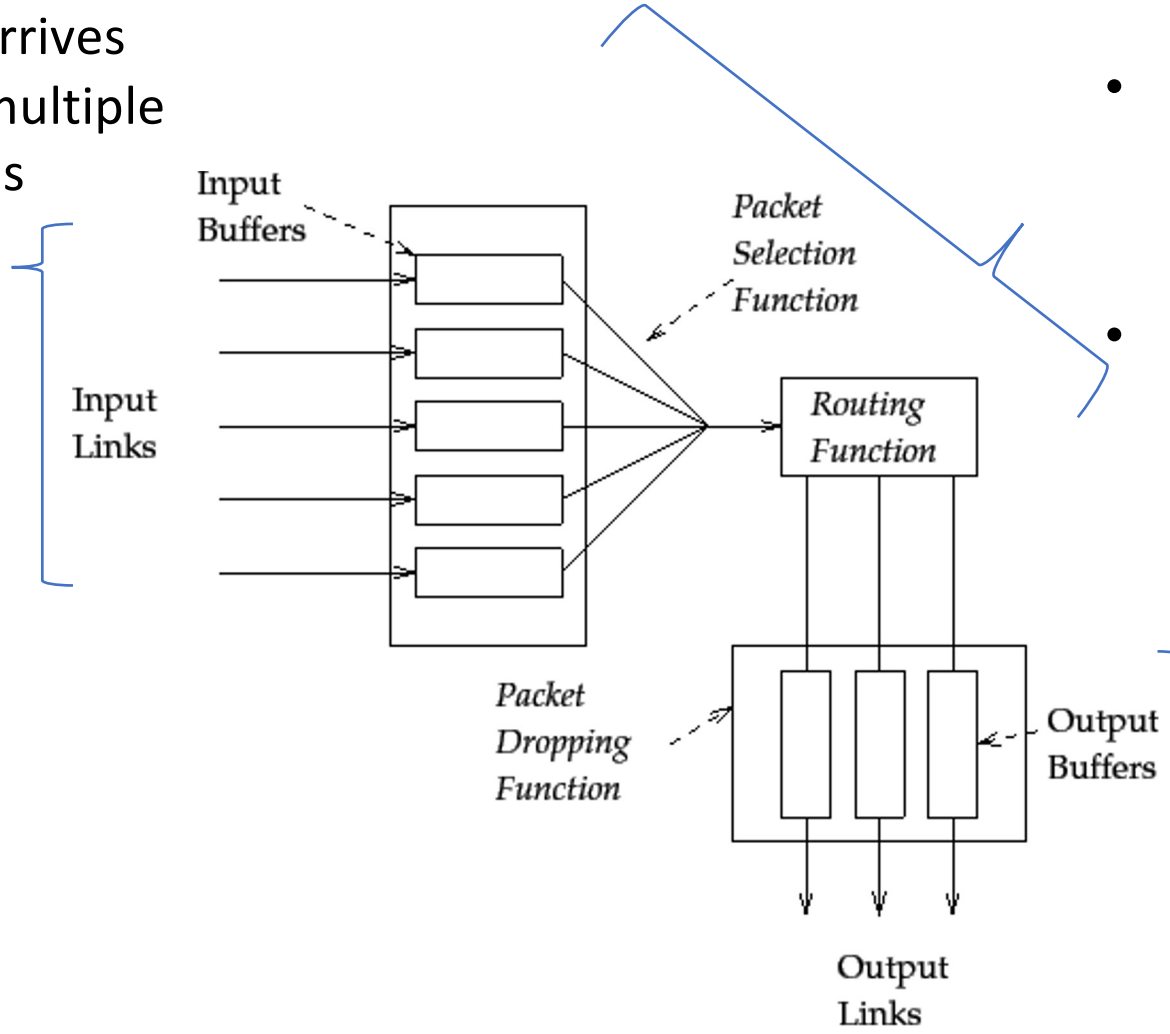


Figure 1: Basic Router Architecture

- Buffers have a finite amount of memory
  - Some have this per interface
  - Others may have access to a shared memory region with other interfaces
- The processing engine will:
  - Extract each packet/frame from the queues
  - Pull off header information to see where the destination should be
  - Move the packet/frame to the correct output queue
- Additional delay is possible as the queues physically write the packet to the transport medium (e.g. optical interface, copper interface)

EPOC
Engagement and Performance Operations Center

All About That Buffer (No Cut Through)

- ***The Bandwidth Delay Product***
  - The amount of “in flight” data for a TCP connection (BDP = bandwidth * round trip time)
- Example: 10Gb/s cross country, ~100ms
  - 10,000,000,000 b/s * .1 s = 1,000,000,000 bits
  - 1,000,000,000 / 8 = 125,000,000 bytes
  - 125,000,000 bytes / (1024*1024) ~ ***125MB***
- Ignore the math aspect: its making sure there is memory to catch and send packets
  - *At **ALL** hops*
  - As the speed increases, there are more packets.
  - If there is not memory, we drop them, and that makes TCP react, and the user sad.

EPOC
Engagement and Performance Operations Center

All About That Buffer (No Cut Through)

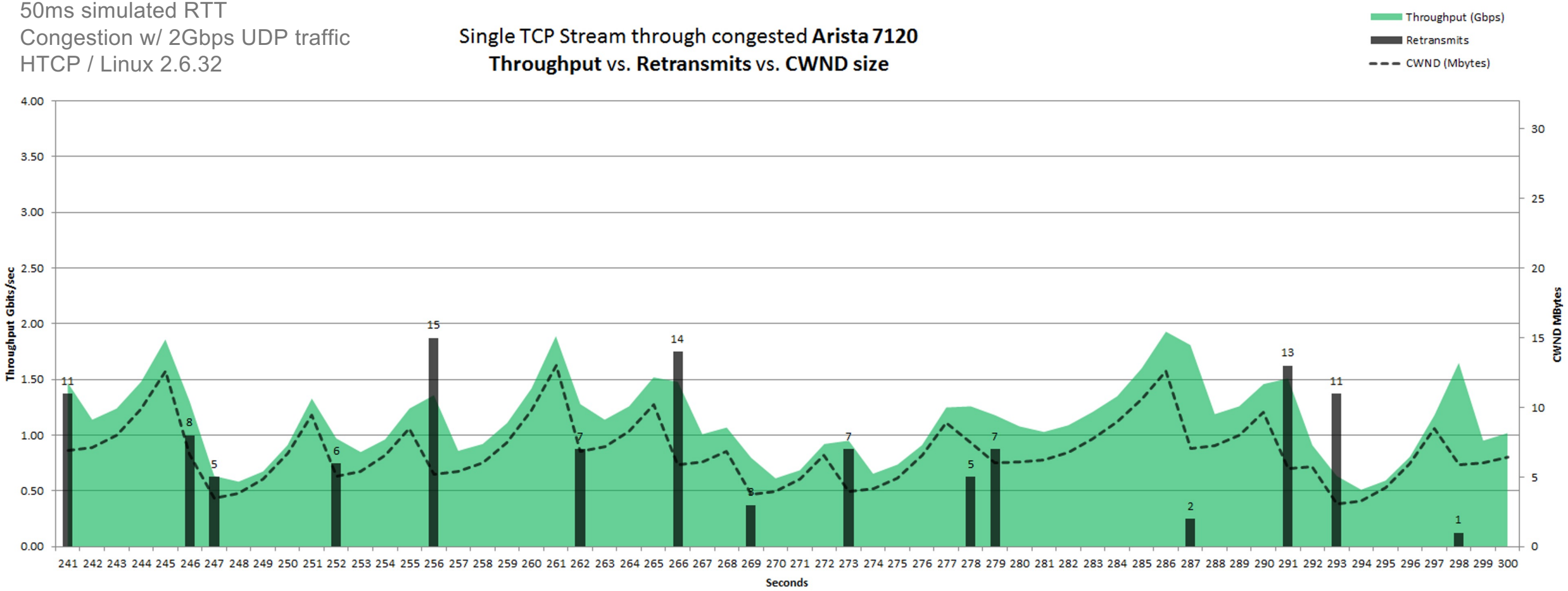
- Buffering isn't as important on the LAN (this is why you are normally pressured to buy 'cut through' devices)
  - Change the math to make the Latency 1ms and the expectation 10Gbps = ***1.25MB***
  - 'Cut through' and low latency switches are designed for the data center, and can handle typical data center loads that don't require buffering (e.g. same to same speeds, destinations within the broadcast domain)

- Buffering **MATTERS** for WAN Transfers
  - Placing something with inadequate buffering in the path reduces the buffer for the entire path. E.g. if you have an expectation of 10Gbps over 100ms – don't place a 12MB buffer anywhere in there – your reality is now ~10x less than it was before (e.g. 10Gbps @ 10ms, or 1Gbps @ 100ms)

EPOC
Engagement and Performance Operations Center

TCP's Congestion Control

50ms simulated RTT
Congestion w/ 2Gbps UDP traffic
HTCP / Linux 2.6.32



Slide from Michael Smitasin, LBLnet

EPOC
Engagement and Performance Operations Center

Outline

- Introduction
- ***Solution Space***
    1. Understanding the Solution Space (Users, Use Cases, Long Term Impacts)
    2. Preliminaries (e.g. Network Protocols 101)
    3. Architecture & Design
    4. ***Monitoring and Measurement***
    5. Data Mobility
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

Test and Measurement – Keeping the Network Clean

- The wide area network, the Science DMZ, and all its systems can be functioning perfectly
- Eventually something is going to break
  - Networks and systems are built with many, many components
  - Sometimes things just break – this is why we buy support contracts
- Other problems arise as well – bugs, mistakes, whatever
- We must be able to find and fix problems when they occur
- Why is this so important? Because we use TCP!

EPOC
Engagement and Performance Operations Center

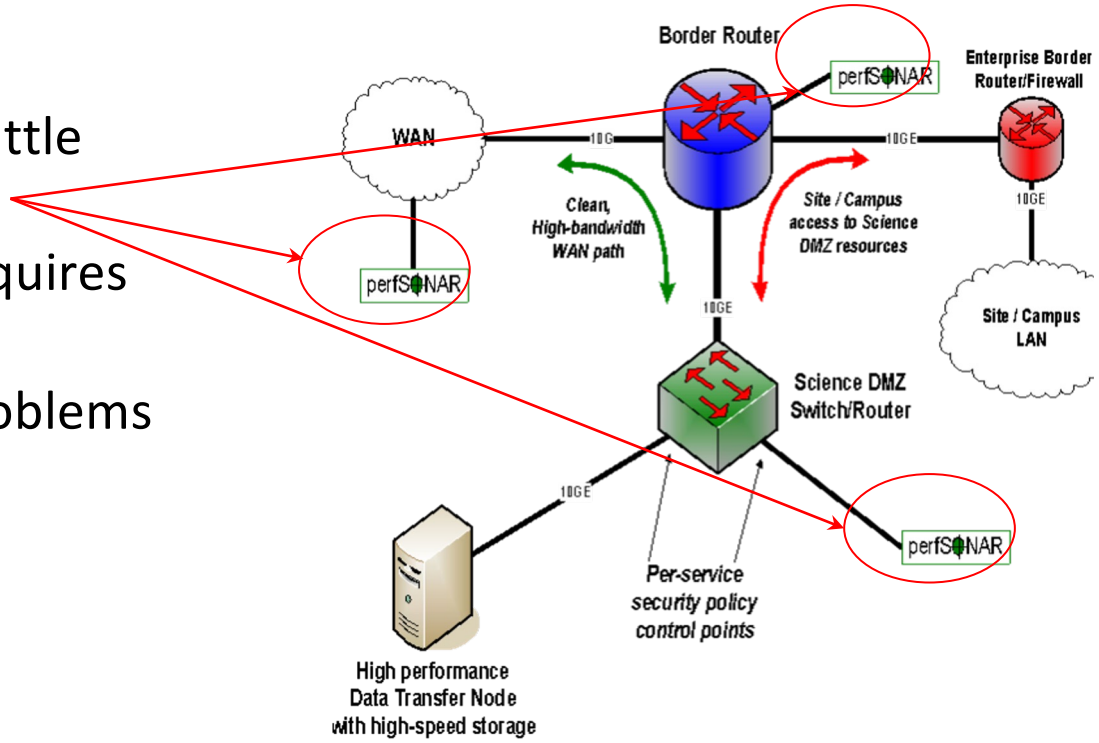
Soft Network Failures

- 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

EPOC
Engagement and Performance Operations Center

perfSONAR

- Network diagrams throughout these materials have little perfSONAR boxes everywhere
  - The reason for this is that consistent behavior requires correctness
  - Correctness requires the ability to find and fix problems
    - ***You can't fix what you can't find***
    - ***You can't find what you can't see***
    - ***perfSONAR lets you see***
- Especially important when deploying high performance services
  - If there is a problem with the infrastructure, need to fix it
  - If the problem is not with your stuff, need to prove it
    - Many players in an end to end path
    - Ability to show correct behavior aids in problem localization



EPOC
Engagement and Performance Operations Center

Outline

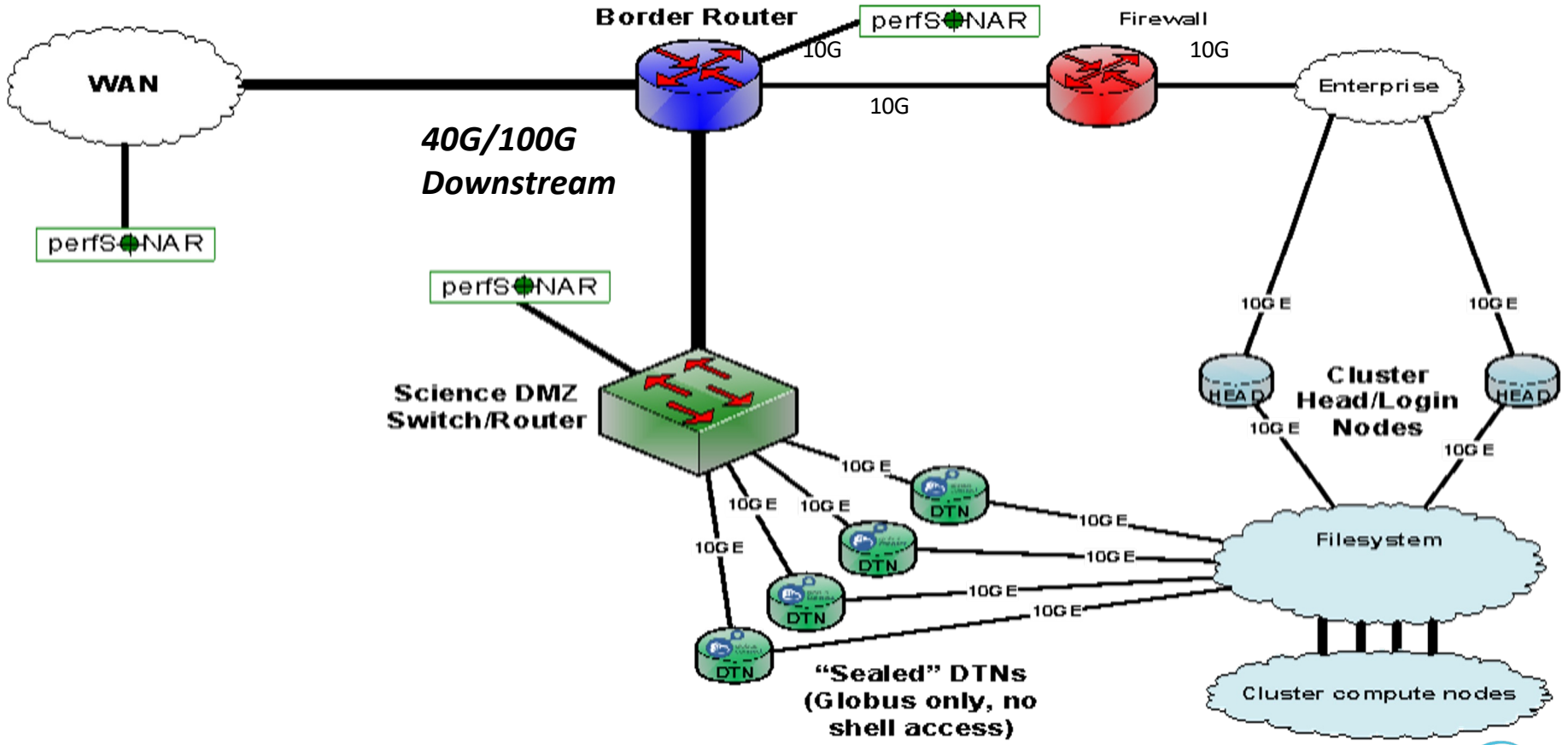
- Introduction
- ***Solution Space***
  1. Understanding the Solution Space (Users, Use Cases, Long Term Impacts)
  2. Preliminaries (e.g. Network Protocols 101)
  3. Architecture & Design
  4. Monitoring and Measurement
  5. ***Data Mobility***
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

Solution Space – Data Mobility

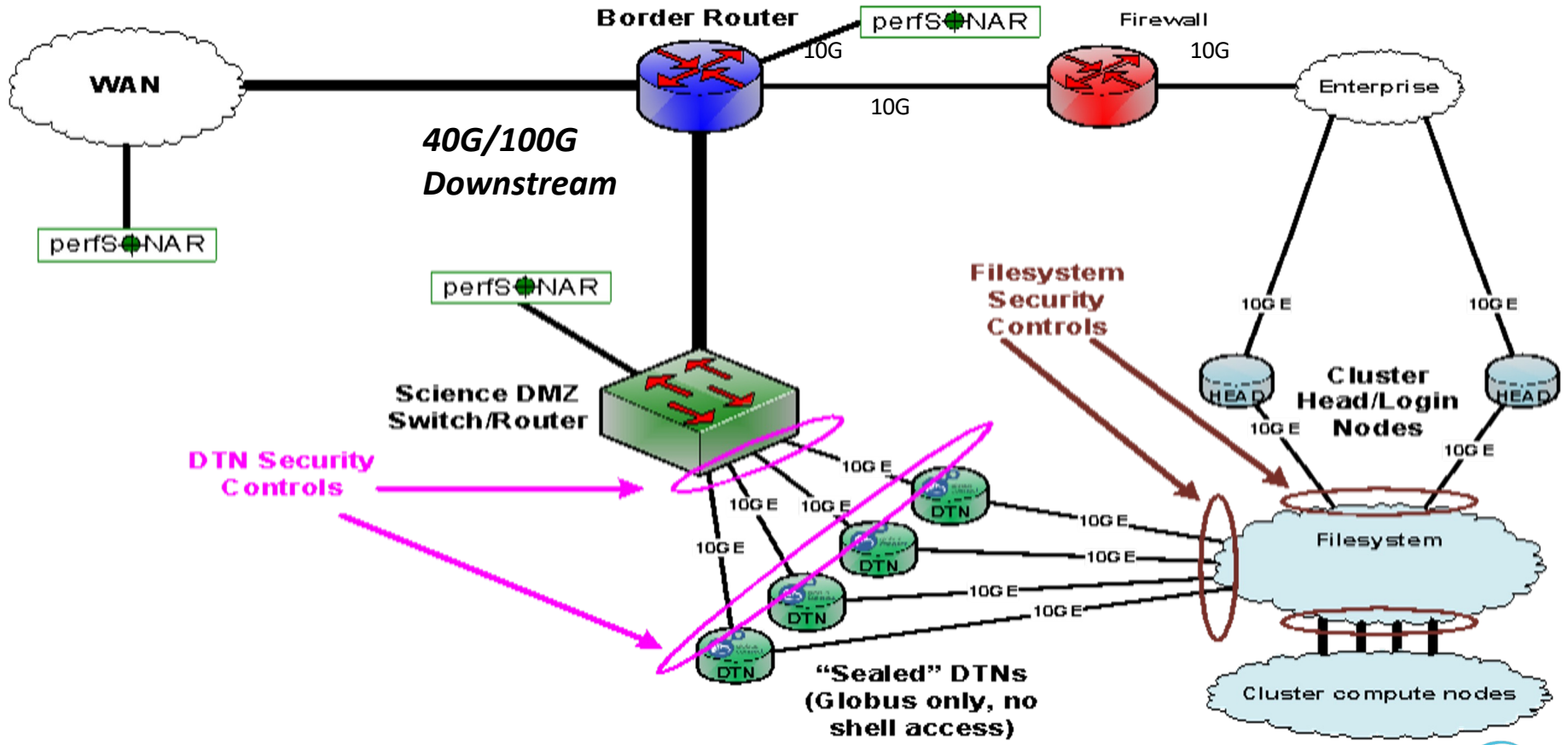
- DTN History & Purpose:
  - Original concept came from initial Science DMZ Design (~2012)
  - Basic idea:
    - Host(s) dedicated to the task of data movement (and only data movement)
    - Limited application set (data movement tools), and users (rarely shell access)
    - Specific security policy 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

Solution Space – Data Mobility



EPOC
Engagement and Performance Operations Center

Solution Space – Data Mobility



EPOC
Engagement and Performance Operations Center

Software – Data Transfer

- Using the right data transfer tool is ***STILL*** very important
- Sample Results: Berkeley, CA to Argonne, IL (near Chicago) RTT = 53 ms, network capacity = 10Gbps.

Tool	Throughput
scp	330 Mbps
wget, GridFTP, FDT, 1 stream	6 Gbps
GridFTP and FDT, 4 streams	8 Gbps (disk limited)

- Notes
  - scp is 24x slower than GridFTP on this path!!
  - to get more than 1 Gbps (125 MB/s) disk to disk requires RAID array.
  - Assume host TCP buffers are set correctly for the RTT

EPOC
Engagement and Performance Operations Center

Data Transfer Performance and Expectations

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	> 100Gbps	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/

EPOC
Engagement and Performance Operations Center

To Reiterate:

- Data movement is hard to get right
- Lots of moving parts
  - Software, Servers, Networks, and People
- Testing will reveal that it may not be ideal
- Testing will also motivate you to make it ideal
- Shared experience around the community – lift all the boats, share all the knowledge, etc.

EPOC
Engagement and Performance Operations Center

Outline

- Introduction
- Solution Space
- ***Conclusions / QA***

EPOC
Engagement and Performance Operations Center

Questions?

- EPOC Helpdesk (send in anything you want):
  - epoc@tacc.utexas.edu

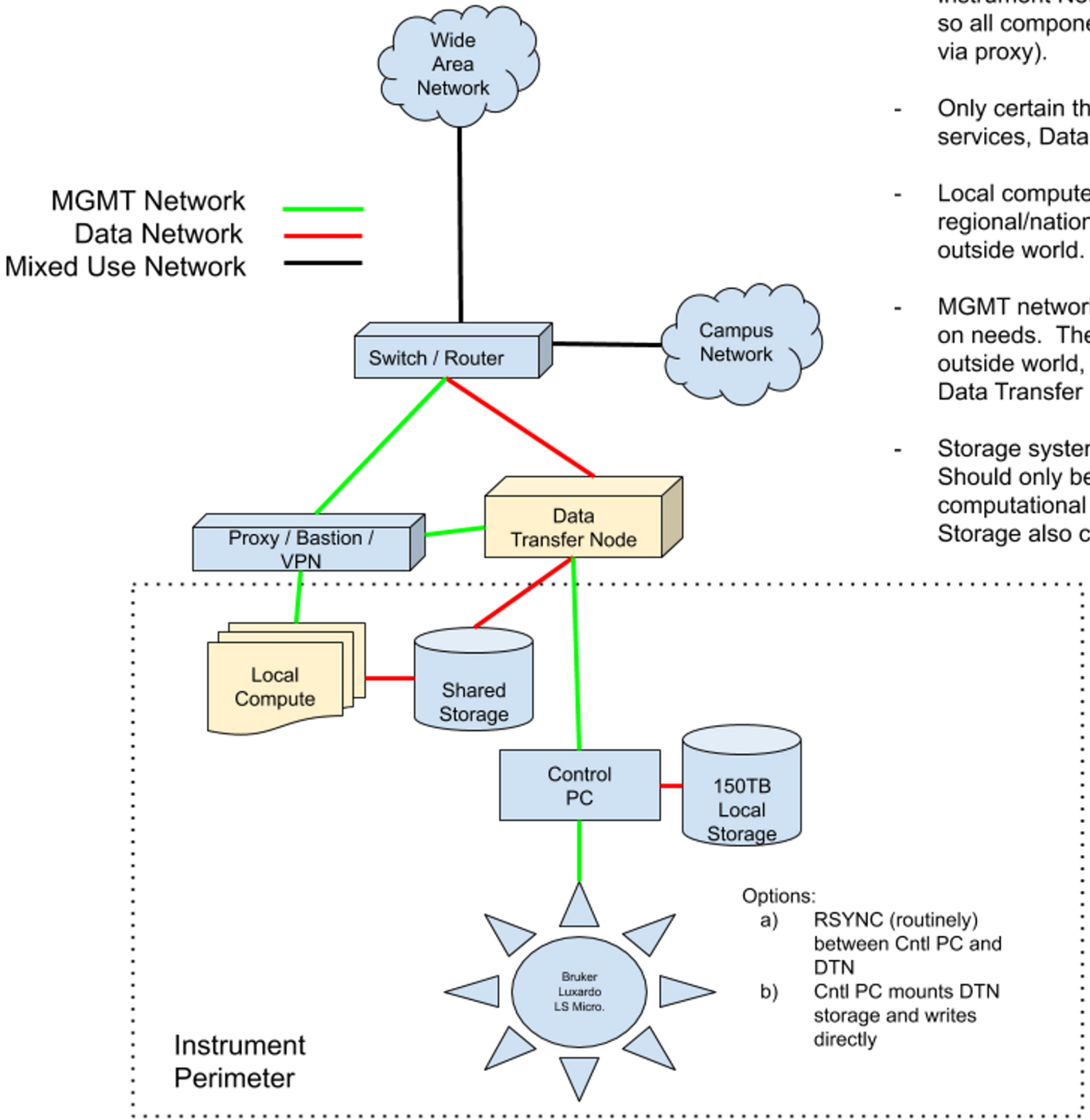
EPOC
Engagement and Performance Operations Center

Designing, Building, & Maintaining a Science DMZ and Data Architecture

Ken Miller, Jason Zurawski
ken@es.net, zurawski@es.net
ESnet / Lawrence Berkeley National Laboratory

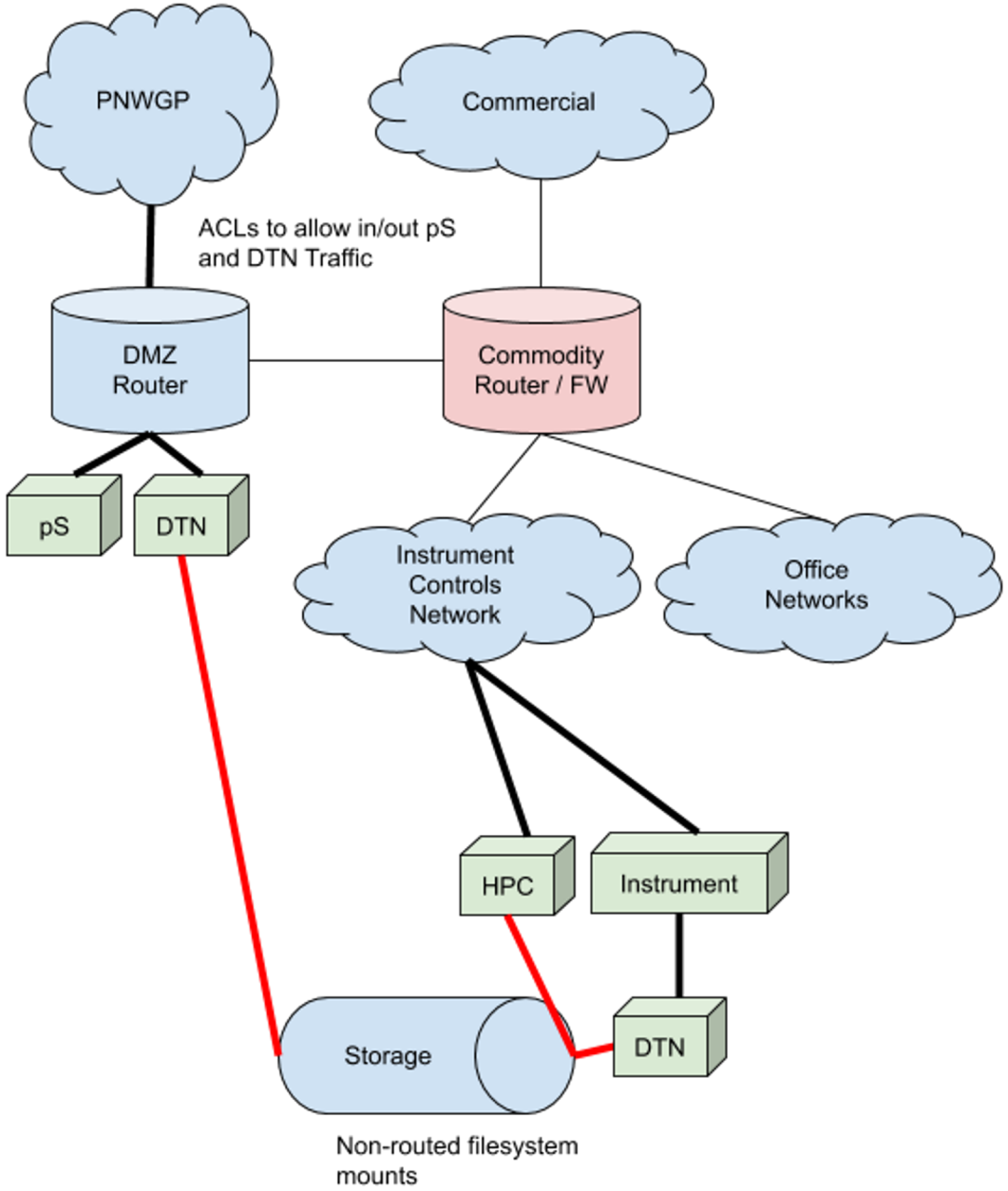
Materials Cyberinfrastructure for Research Data Management Workshop
Princeton, NJ
May 23-24, 2023

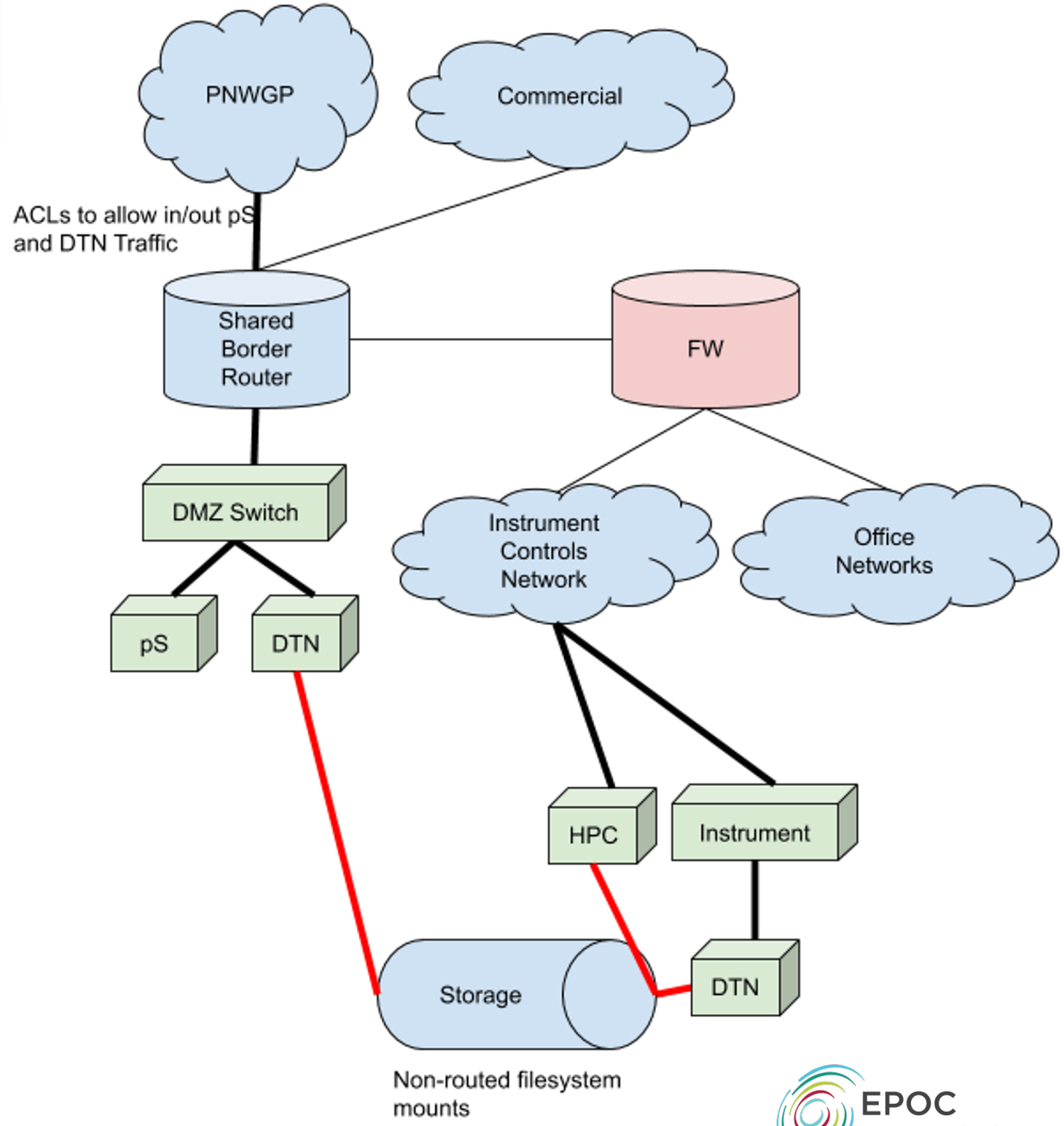
National Science Foundation Award #2328479



- 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.

EPOC
Engagement and Performance Operations Center





EPOC
Engagement and Performance
Operations Center