Science DMZ, Congestion Control and Buffer Sizing

Doug Southworth
dsouthworth@tacc.utexas.edu
Texas Advanced Computing Center

Workshop on P4 Programmable Switches
August 29, 2023

National Science Foundation Award #1826994

Outline

- ***Introduction***
- Solution Space
- Putting it Together
- 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

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
  6. Security and Policy
- Putting it Together
- 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

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
  6. Security and Policy
- Putting it Together
- 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

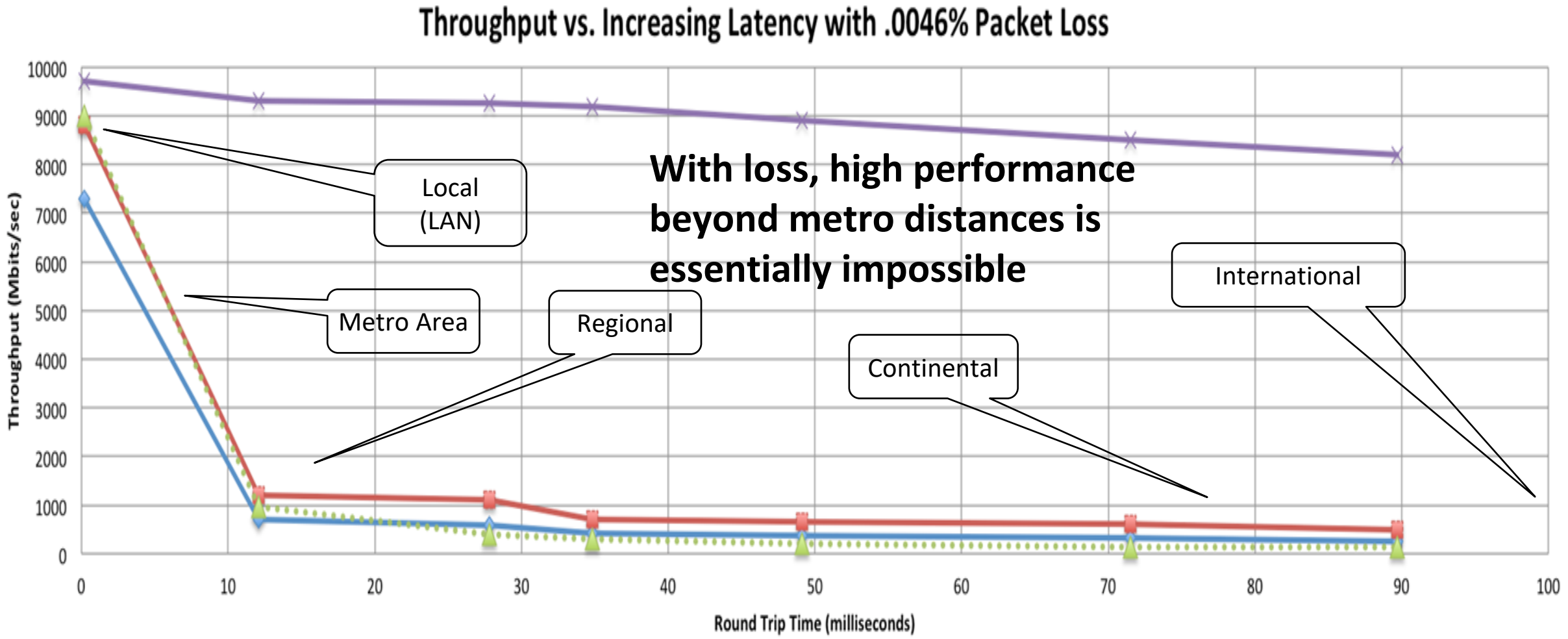
EPOC
Engagement and Performance Operations Center

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

The

Consi

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

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
    6. Security and Policy
- Putting it Together
- 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

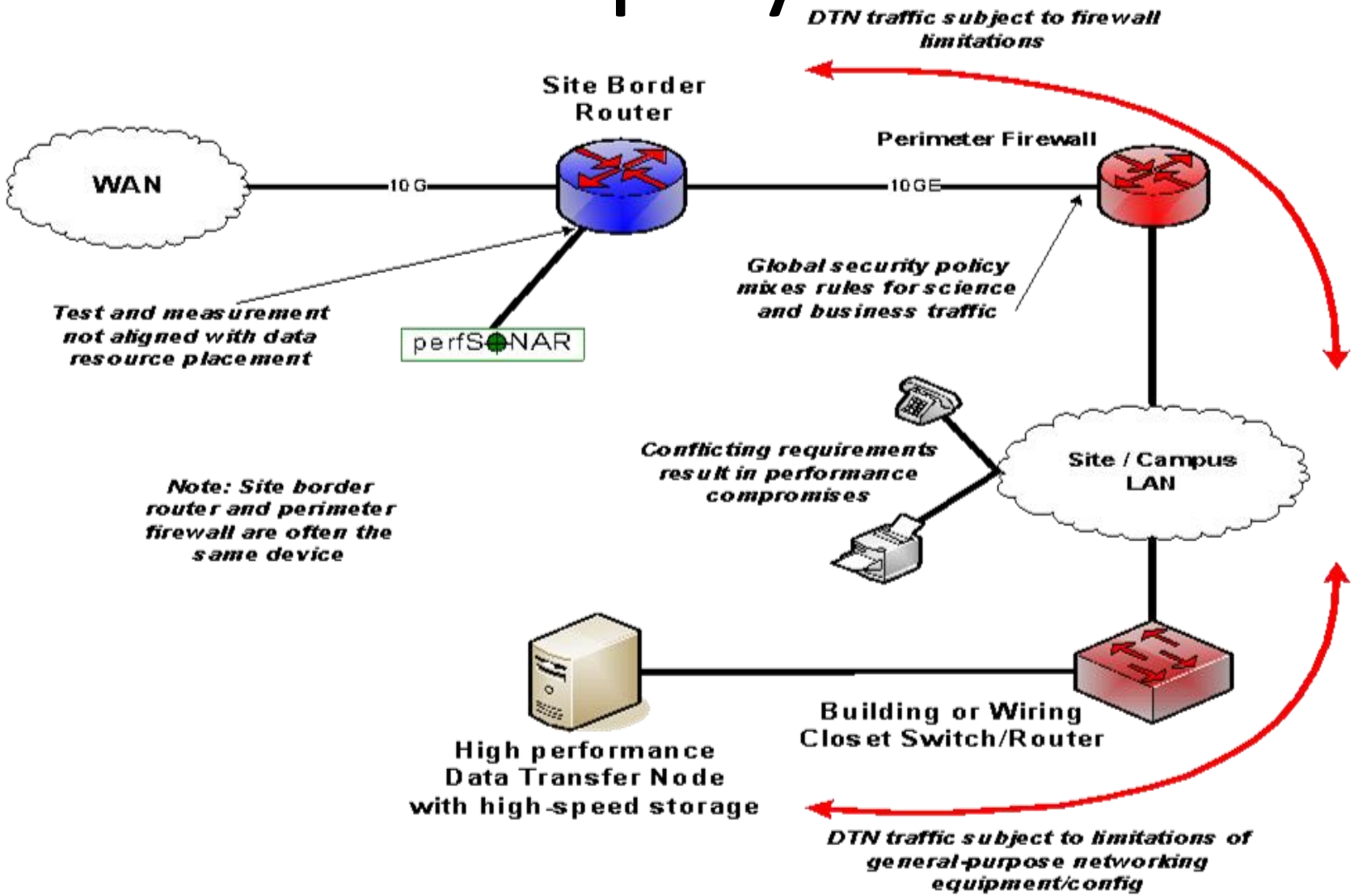
Legacy Method: Ad Hoc DTN Deployment

- 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.

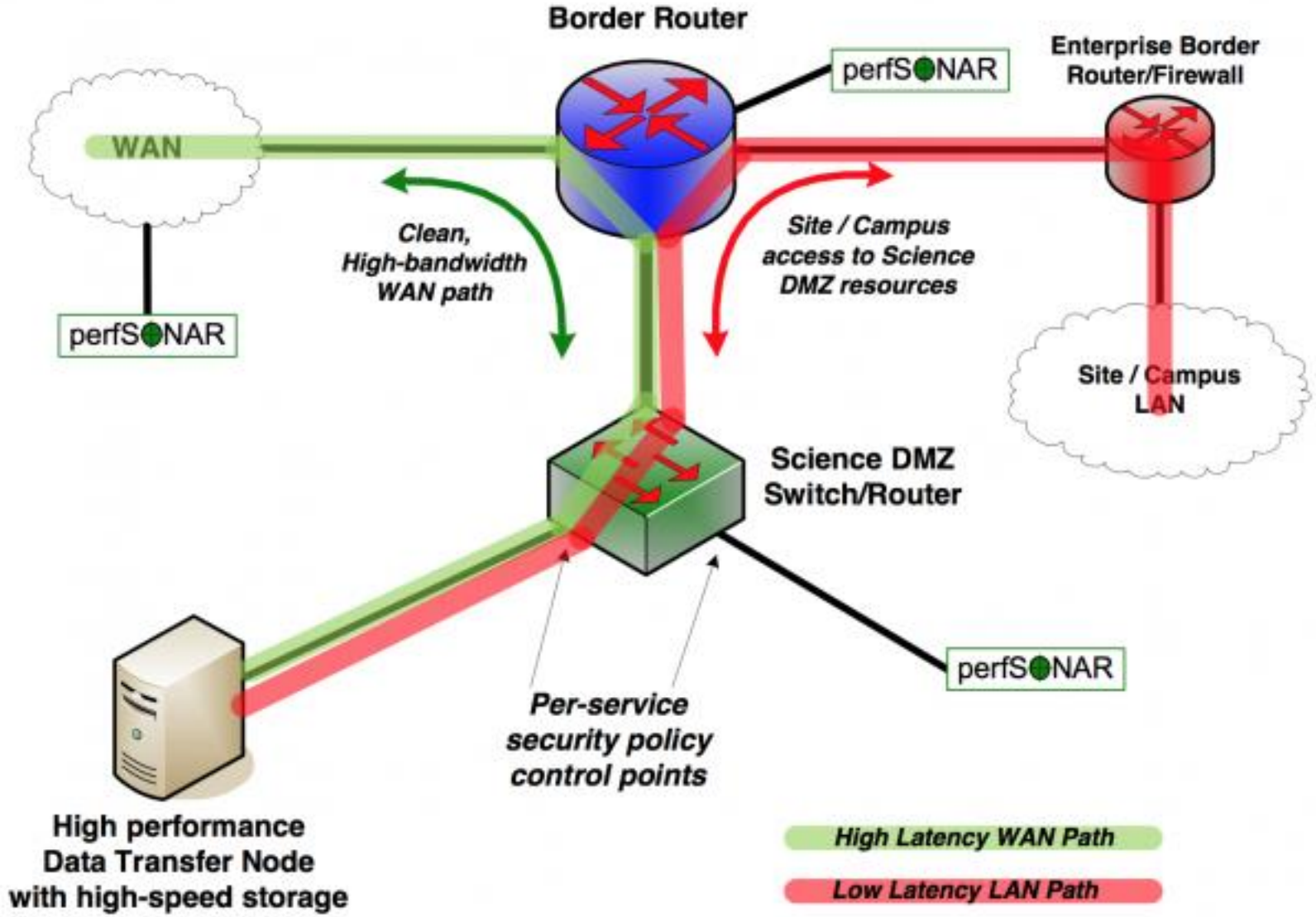
EPOC
Engagement and Performance Operations Center

Ad Hoc DTN Deployment



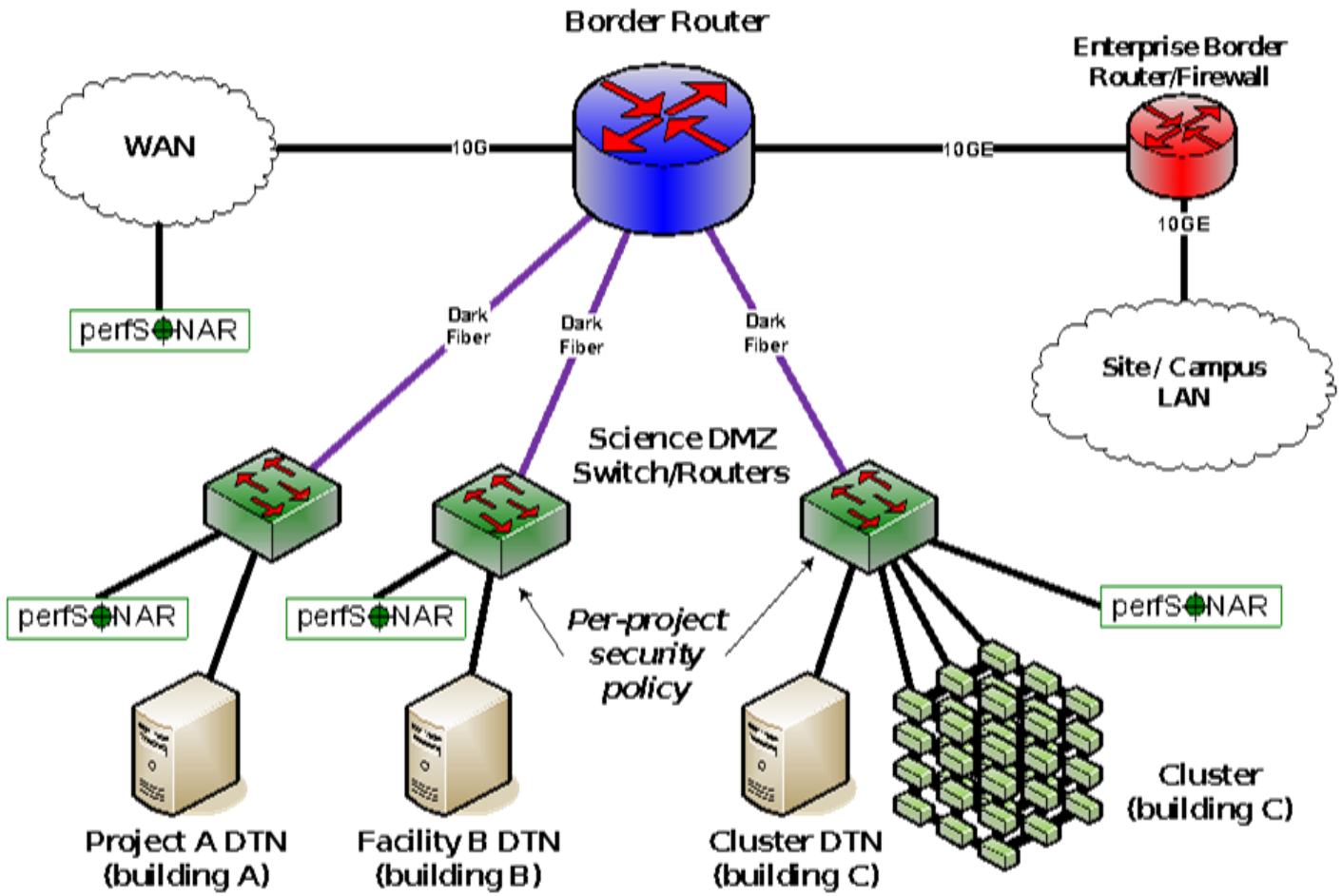
EPOC
Engagement and Performance Operations Center

A better approach: simple Science DMZ



EPOC
Engagement and Performance Operations Center

Multiple Science DMZs – Dark Fiber to Dedicated Switches



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

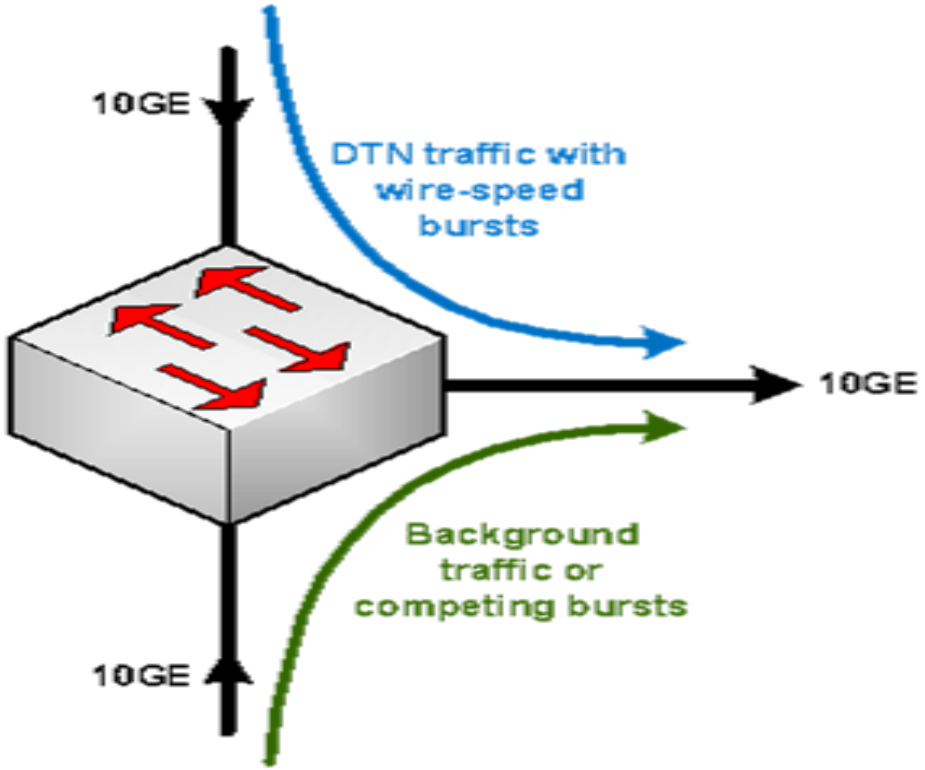
Common Circumstance: Multiple Ingress Data Flows, Common Egress

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

All About That Buffer (No Cut Through)

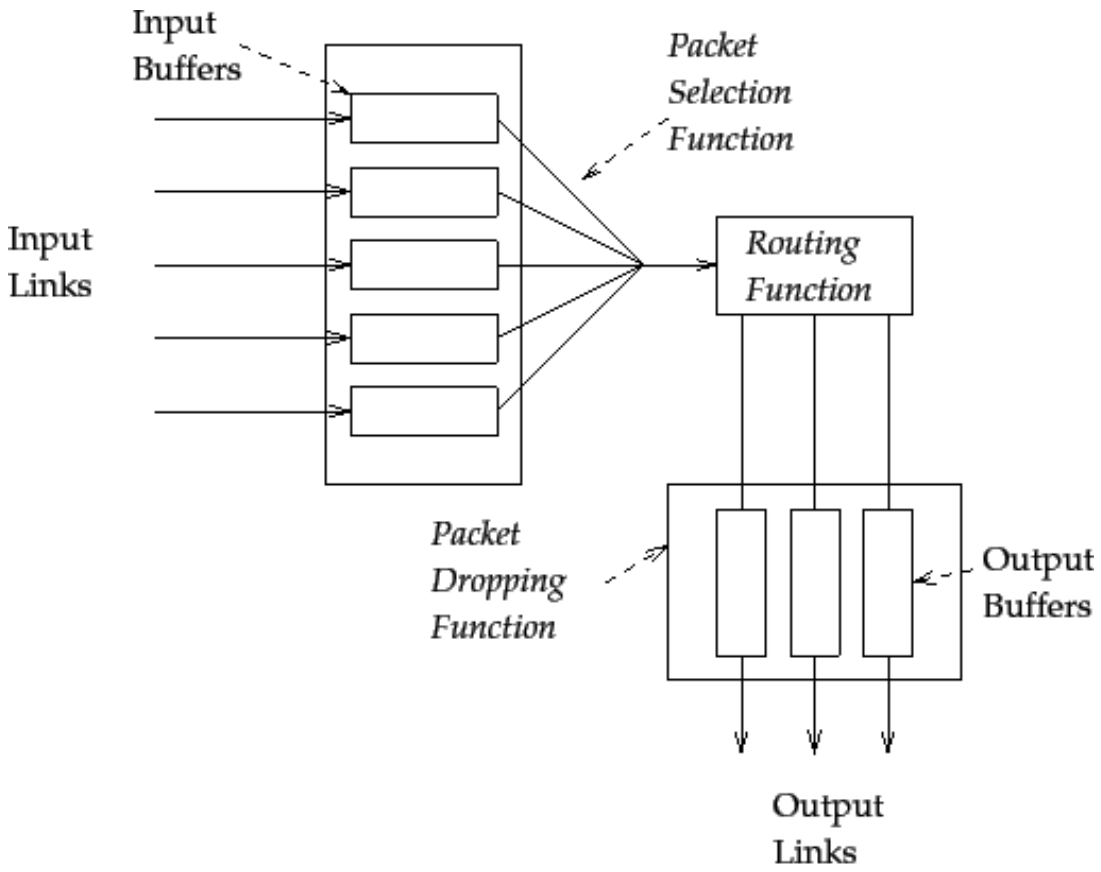


Figure 1: Basic Router Architecture

EPOC
Engagement and Performance Operations Center

All About That Buffer (No Cut Through)

- Data arrives from multiple sources

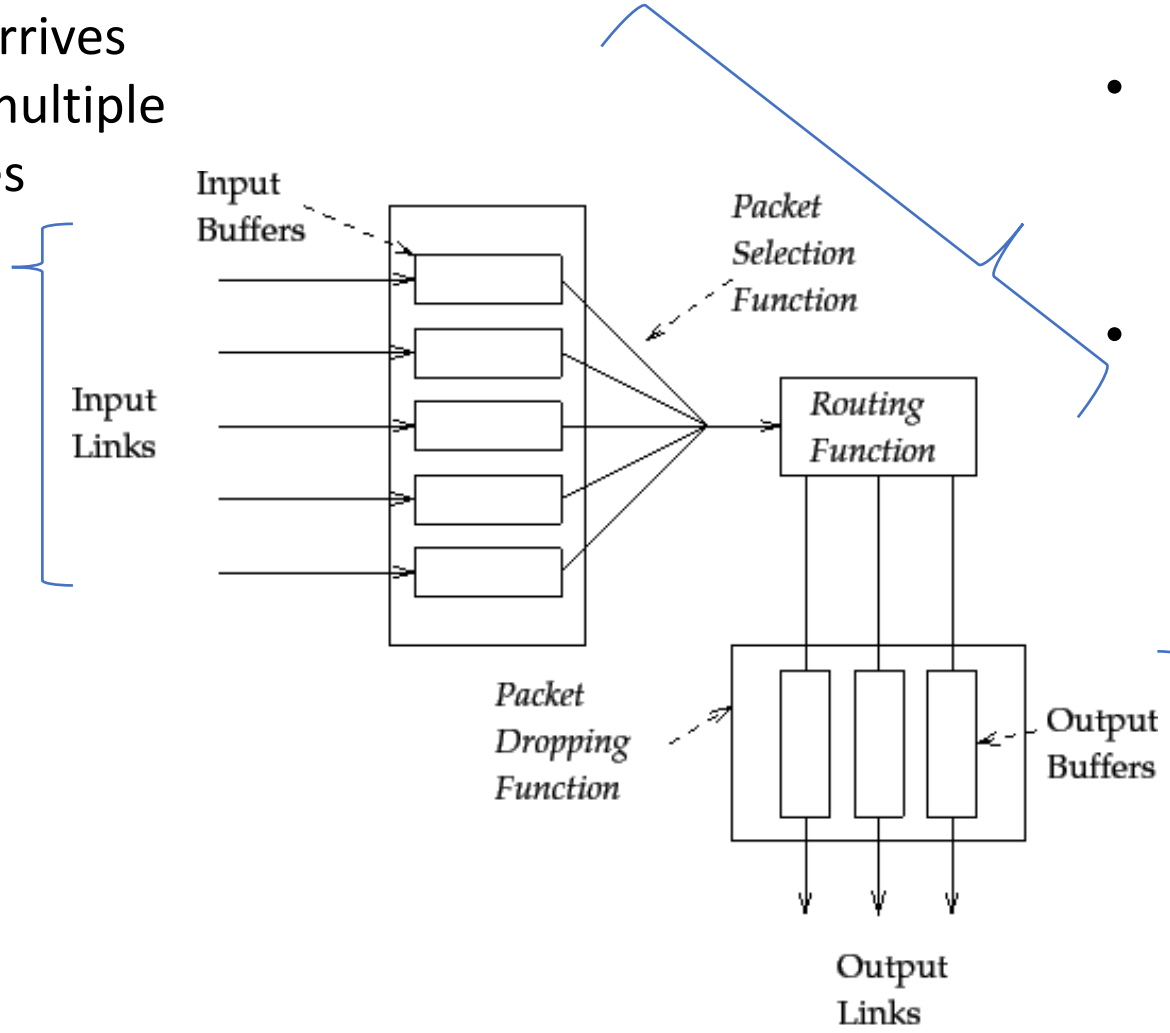


Figure 1: Basic Router Architecture

- 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)

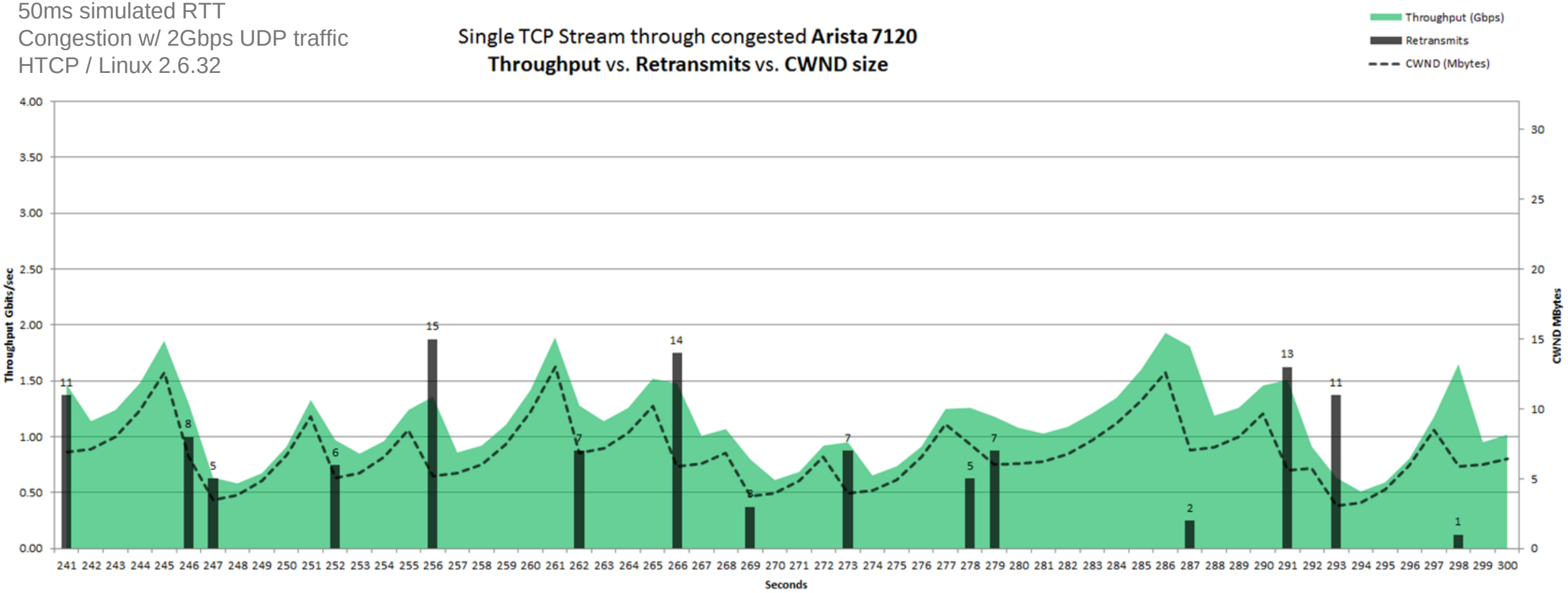
- Buffering isn't as important on the LAN (this is why you are normally pressured to buy 'cut through' devices)
  - Change the math to make the Latency 1ms and the expectation 10Gbps = ***1.25MB***
  - 'Cut through' and low latency switches are designed for the data center, and can handle typical data 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
    6. Security and Policy
- Putting it Together
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

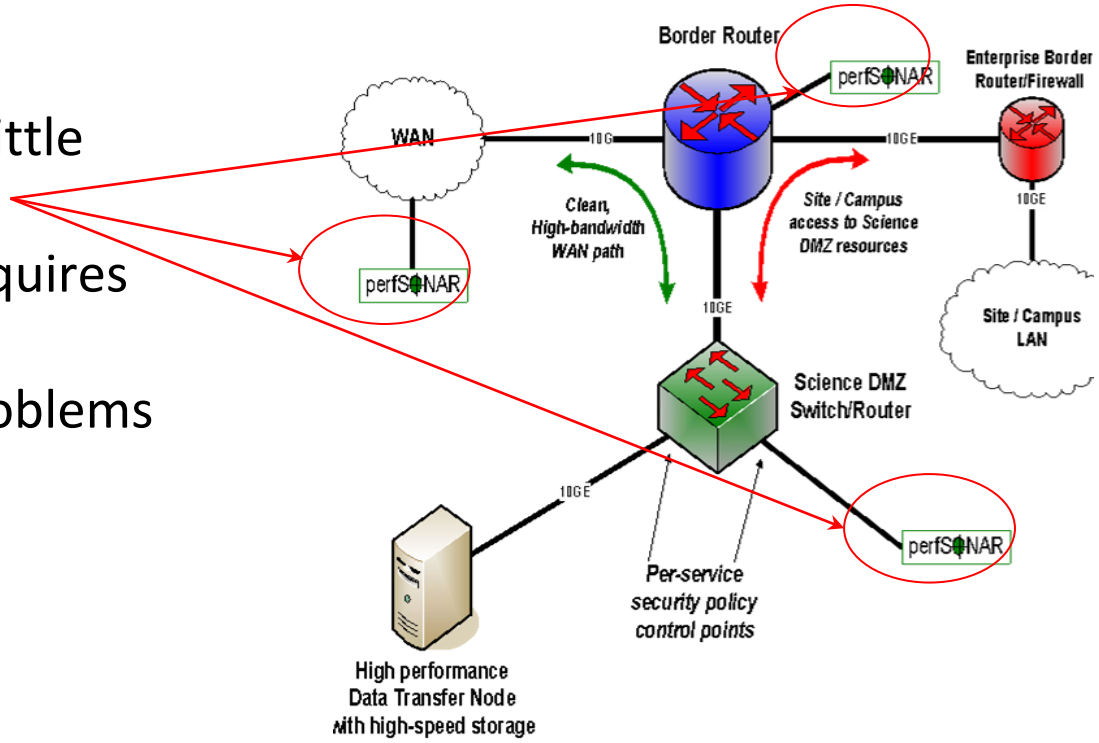
Test and Measurement – Keeping the Network Clean

- 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

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

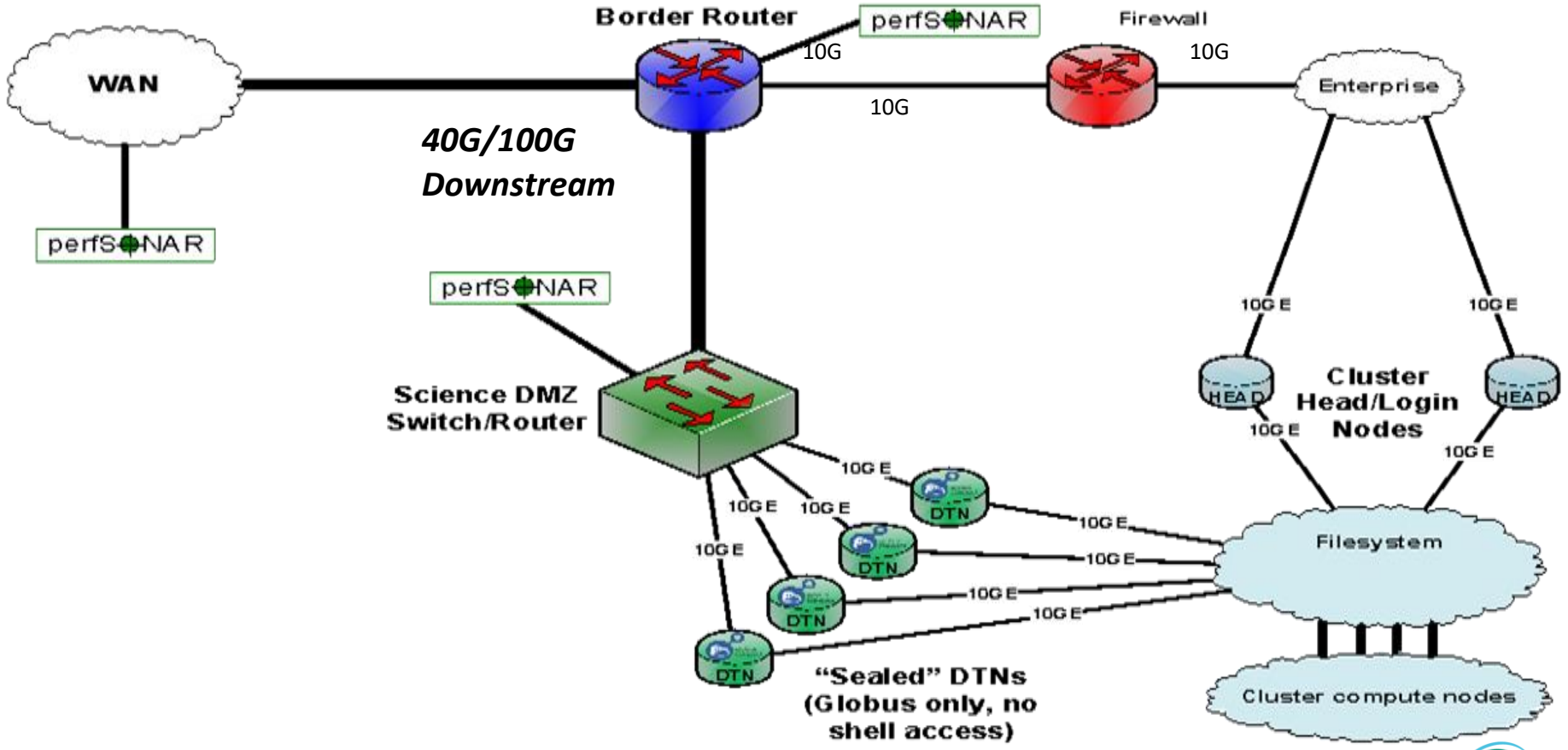
- Introduction
- ***Solution Space***
    1. Understanding the Solution Space (Users, Use Cases, Long Term Impacts)
    2. Preliminaries (e.g. Network Protocols 101)
    3. Architecture & Design
    4. Monitoring and Measurement
    5. ***Data Mobility***
    6. Security and Policy
- Putting it Together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

EPOC
Engagement and Performance Operations Center

Solution Space – Data Mobility

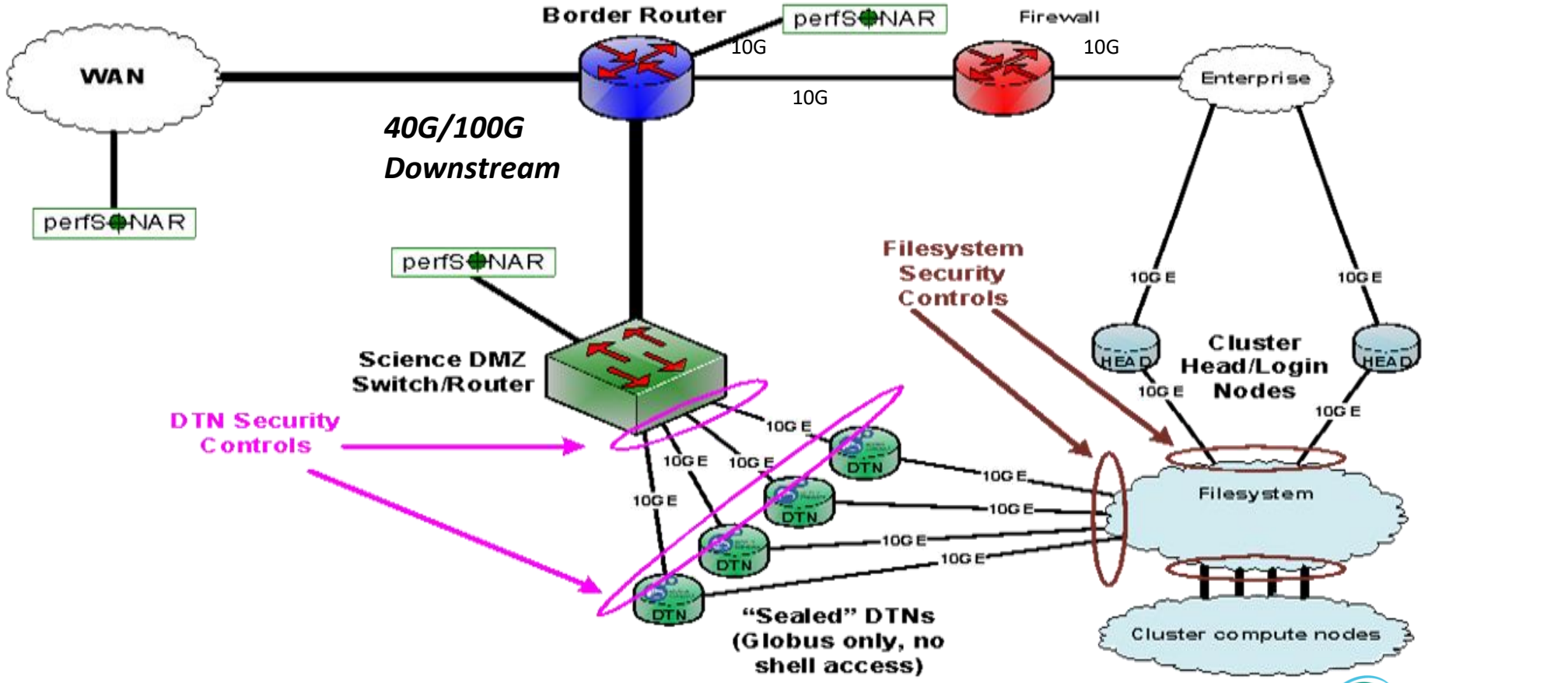
- DTN History & Purpose:
  - Original concept came from initial Science DMZ Design (~2012)
  - Basic idea:
    - Host(s) dedicated to the task of data movement (and only data movement)
    - Limited application set (data movement tools), and users (rarely shell access)
    - Specific security policy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

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***
    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
    6. ***Security and Policy***
- Putting it Together
- Conclusions / QA

EPOC
Engagement and Performance Operations Center

Science DMZ Security

- **Goal**: Disentangle security policy and enforcement for science flows from enterprise / 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 ***routinely*** 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

Outline

- Introduction
- Solution Space
- ***Putting it Together***
- Conclusions / QA

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

Sensible Usage Policies

- Define access methods
  - E.g. shared DTN that plugs into storage vs. someone's laptop
  - Tools that can be used
  - People who get accounts
  - Consider: http://fasterdata.es.net/science-dmz/science-dmz-users/
- Define AUP
  - What can/should/will be sent across the infrastructure
  - What happens when something bad occurs
  - How often the AUP is reviewed
  - Consider: http://fasterdata.es.net/science-dmz/usage-policy/
- Define monitoring/measurement/security expectations
  - Let the pros monitor/keep things up to date
  - Let the users just use
- ***BE TRANSPARENT***

EPOC
Engagement and Performance Operations Center

Putting it All Together

- Know the users, know the use cases
- Data Architecture – e.g. support the ingress and egress of information at all layers
  - Network Gear
  - Data Transfer Software / Hardware
  - Measurement / Monitoring
  - Security Infrastructure
- Facilitate Usage
- Have a good story for long term usage/onboarding/expansion/maintenance

EPOC
Engagement and Performance Operations Center

Outline

- Introduction
- Solution Space
- Putting it Together
- ***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

Science DMZ, Congestion Control and Buffer Sizing

Doug Southworth
dsouthworth@tacc.utexas.edu
Texas Advanced Computing Center

Workshop on P4 Programmable Switches
August 29, 2023

National Science Foundation Award #1826994