http://ce.sc.edu/cyberinfra/workshop_2020.html

EPOC
Engagement and Performance Operations Center

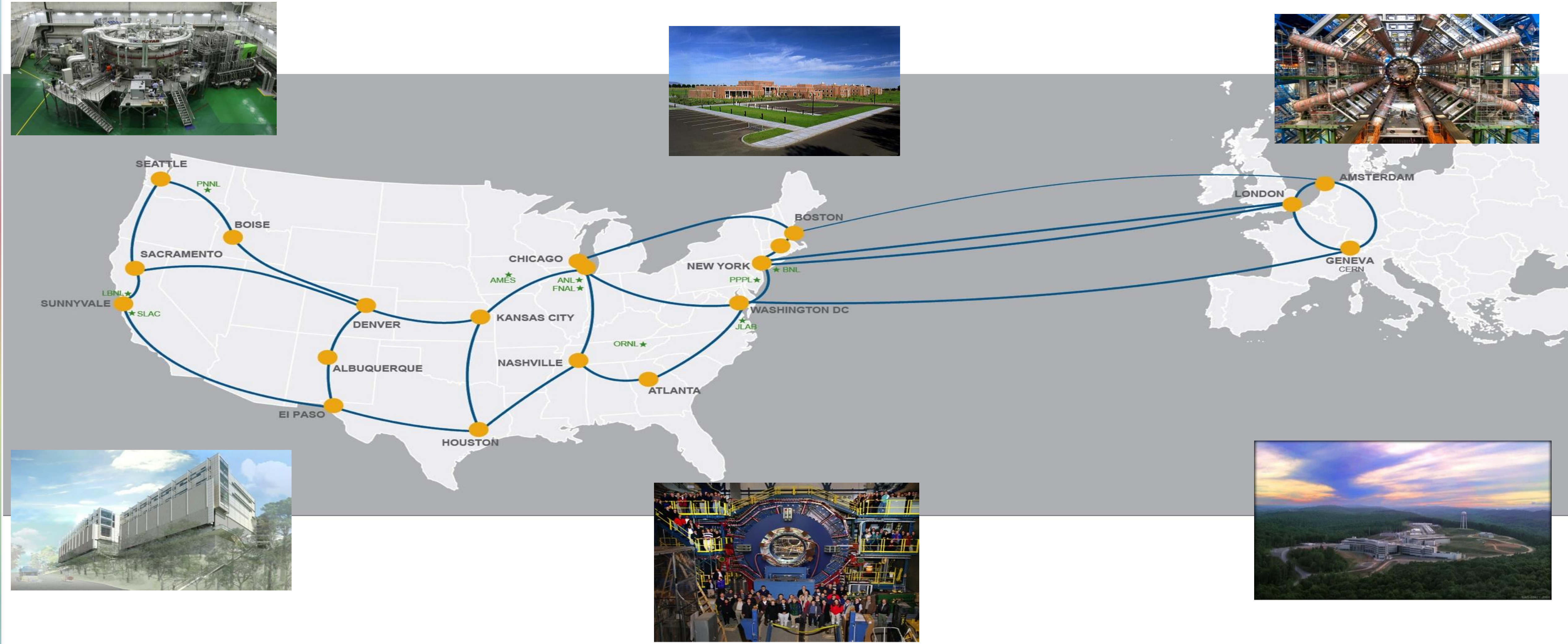
TCP Basics + Science DMZ

Jason Zurawski
zurawski@es.net
ESnet / Lawrence Berkeley National Laboratory

Training Workshop for Educators and Network Engineers on High Speed Network Protocols and Security
May 4th, 2020

ESnet ENERGY SCIENCES NETWORK
INDIANA UNIVERSITY

Network as ~~Infrastructure~~ *Instrument*



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

EPOC
Engagement and Performance Operations Center

The

Consi

- “Fri
  -
  -
  -
  -
- Dec
  -
  -
- Per
  -
- Eng

osθ

NAR

z/

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)

EPOC
Engagement and Performance Operations Center

Packet/Data Loss?!

- ***"Wait a minute! I thought TCP was a reliable protocol? What do you mean 'packet loss', where is the data going!?"***

- We are going to talk about this a lot.
  - The data isn't lost forever, its dropped somewhere on the path.
  - Usually by a device without enough buffer space to accept it, or by someone who thinks the data is corrupted and they won't send it
- Once its dropped, we have a way of knowing its been dropped.
  - TCP is reliable, each end is keeping track of what was sent, and what was received.
  - If something goes missing, its resent.
  - Resending is what takes the time, and causes the slowdown.

EPOC
Engagement and Performance Operations Center

Packet/Data Loss?!

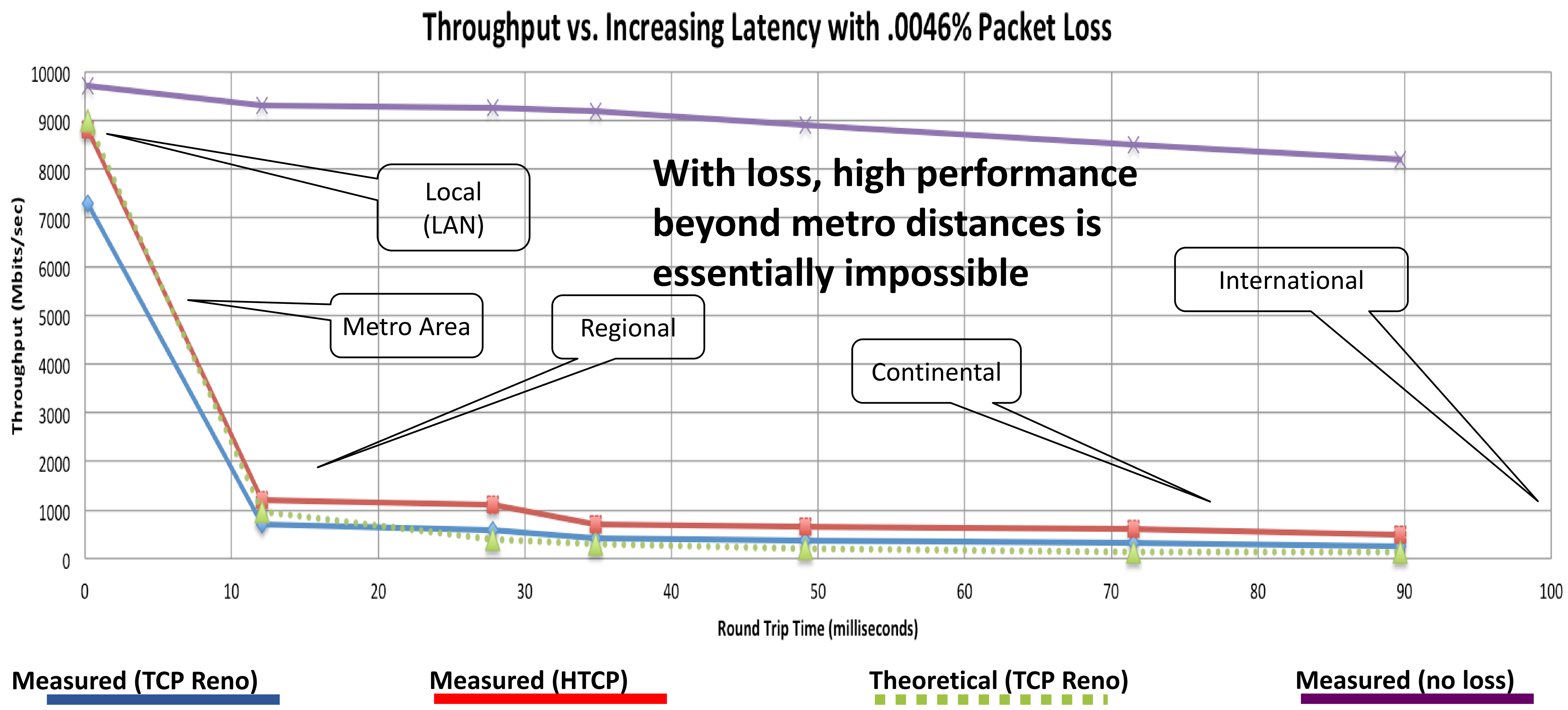
- TCP is able to reliably and transparently recover from packet loss by retransmitting any/all lost packets
  - This is how it provides a reliable data transfer services to the applications which use it, e.g. Web, Email, GridFTP, perfSONAR, etc.
  - The reliability mechanisms dramatically reduce performance when they are exercised

- We want to eliminate the causes of packet loss – so that we don't need to test out the (slow) way that TCP can recover.

- But first … what is the impact of that recovery?

EPOC
Engagement and Performance Operations Center

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



Breaking a Network

- Disk PT Hosts (10G)
  - east-dc-pt1.es.net (New York, NY)
  - lbl-pt1.es.net (Berkeley, CA)
- Path
  - ~70ms RTT

```
[zurawski@lbl-pt1 ~]$ tracepath east-dc-pt1.es.net
 1?: [LOCALHOST]                                             pmtu 9000
 1:  lblmr2-lblpt1.es.net                                      0.266ms asymm  2
 1:  lblmr2-lblpt1.es.net                                      0.210ms asymm  2
 2:  sacrcr5-ip-a-lblmr2.es.net                                2.935ms
 3:  denvcr5-ip-a-sacrcr5.es.net                              24.421ms
 4:  kanscr5-ip-a-denvcr5.es.net                              34.469ms
 5:  chiccr5-ip-a-kanscr5.es.net                              45.426ms
 6:  washcr5-ip-a-chiccr5.es.net                              62.543ms
 7:  eqxashcr5-ip-a-eqxchicr5.es.net                          62.006ms
 8:  washcr5-ip-c-eqxashcr5.es.net                            62.428ms asymm  6
 9:  east-dc-pt1.es.net                                       69.388ms !H
     Resume: pmtu 9000
```

EPOC
Engagement and Performance Operations Center

How Do We Accommodate TCP?

- High-performance wide area TCP flows must get loss-free service
  - Sufficient bandwidth to avoid congestion
  - Deep enough buffers in routers and switches to handle bursts
    - Especially true for long-distance flows due to packet behavior
    - No, this isn't buffer bloat

- Equally important – the infrastructure must be verifiable so that clean service can be provided
  - Stuff breaks
    - Hardware, software, optics, bugs, …
    - How do we deal with it in a production environment?
  - Must be able to prove a network device or path is functioning correctly
    - Regular active tests should be run - perfSONAR
  - Small footprint is a huge win
    - Fewer the number of devices = easier to locate the source of packet loss

EPOC
Engagement and Performance Operations Center

Science DMZ 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)
  - Wait-for-trouble-ticket operational models for network performance

EPOC
Engagement and Performance Operations Center

Science DMZ 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

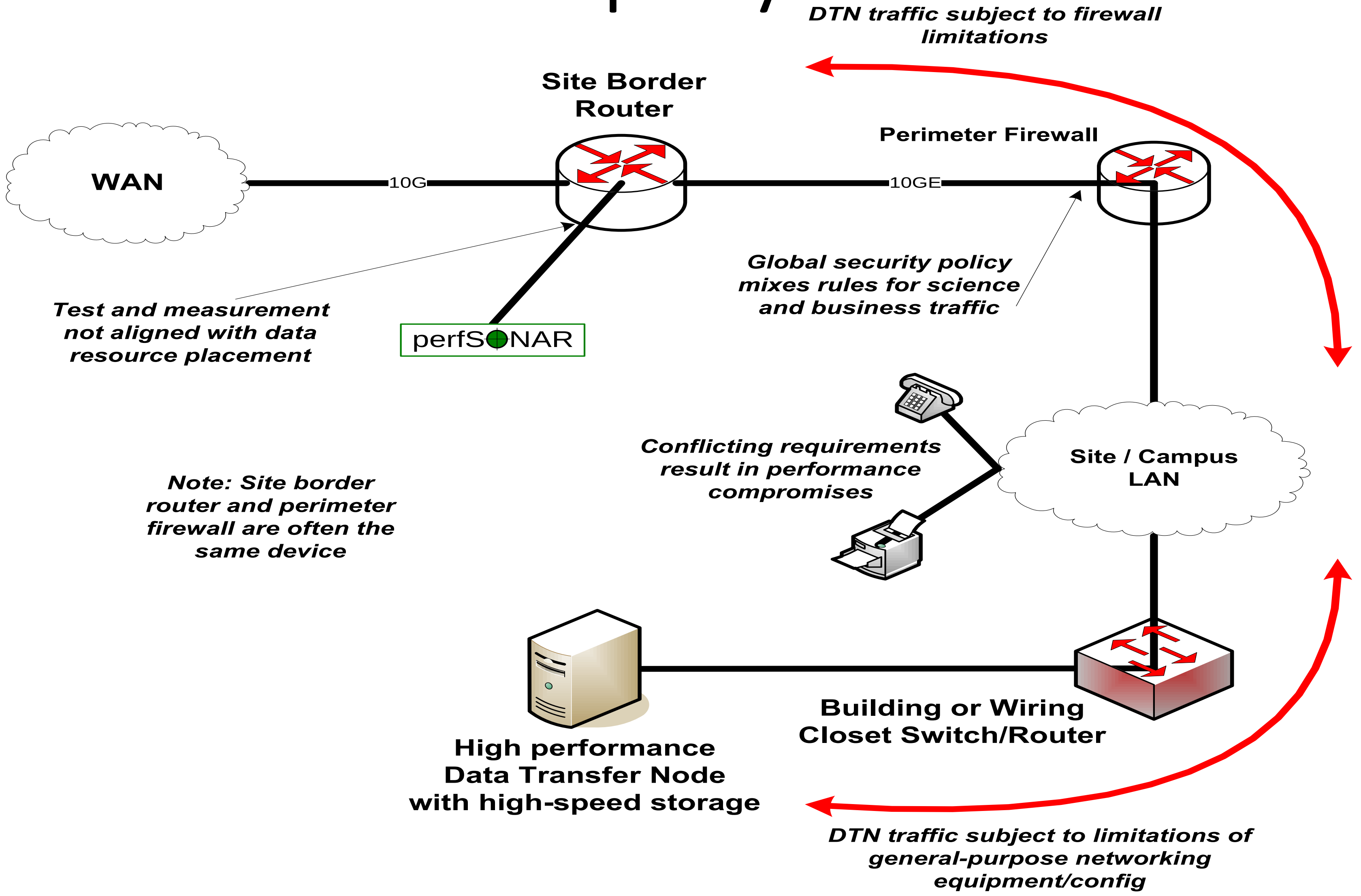
EPOC
Engagement and Performance Operations Center

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

Ad Hoc DTN Deployment



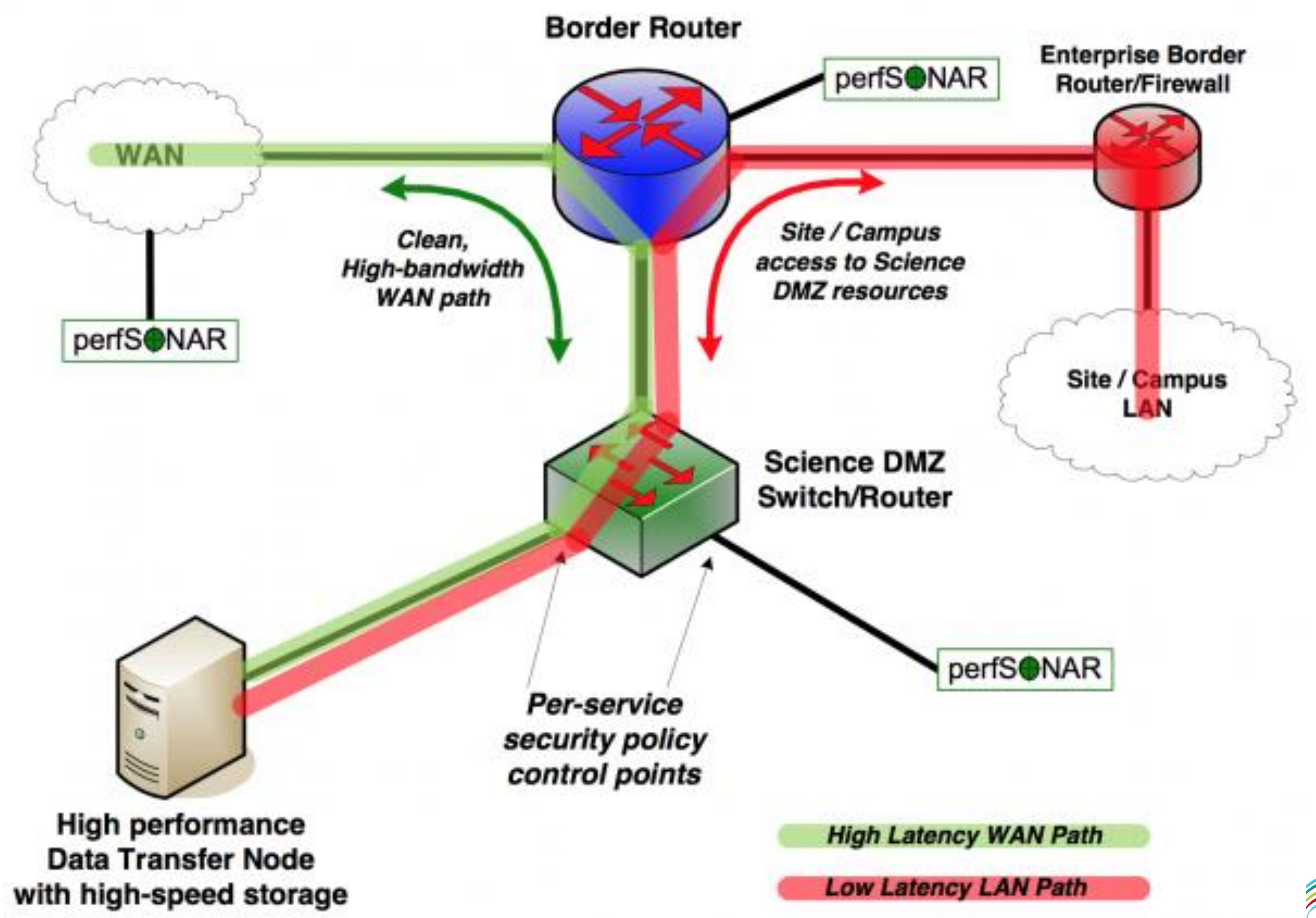
Legacy Method: Ad Hoc DTN Deployment

- 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 performan

- 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

A better approach: simple Science DMZ



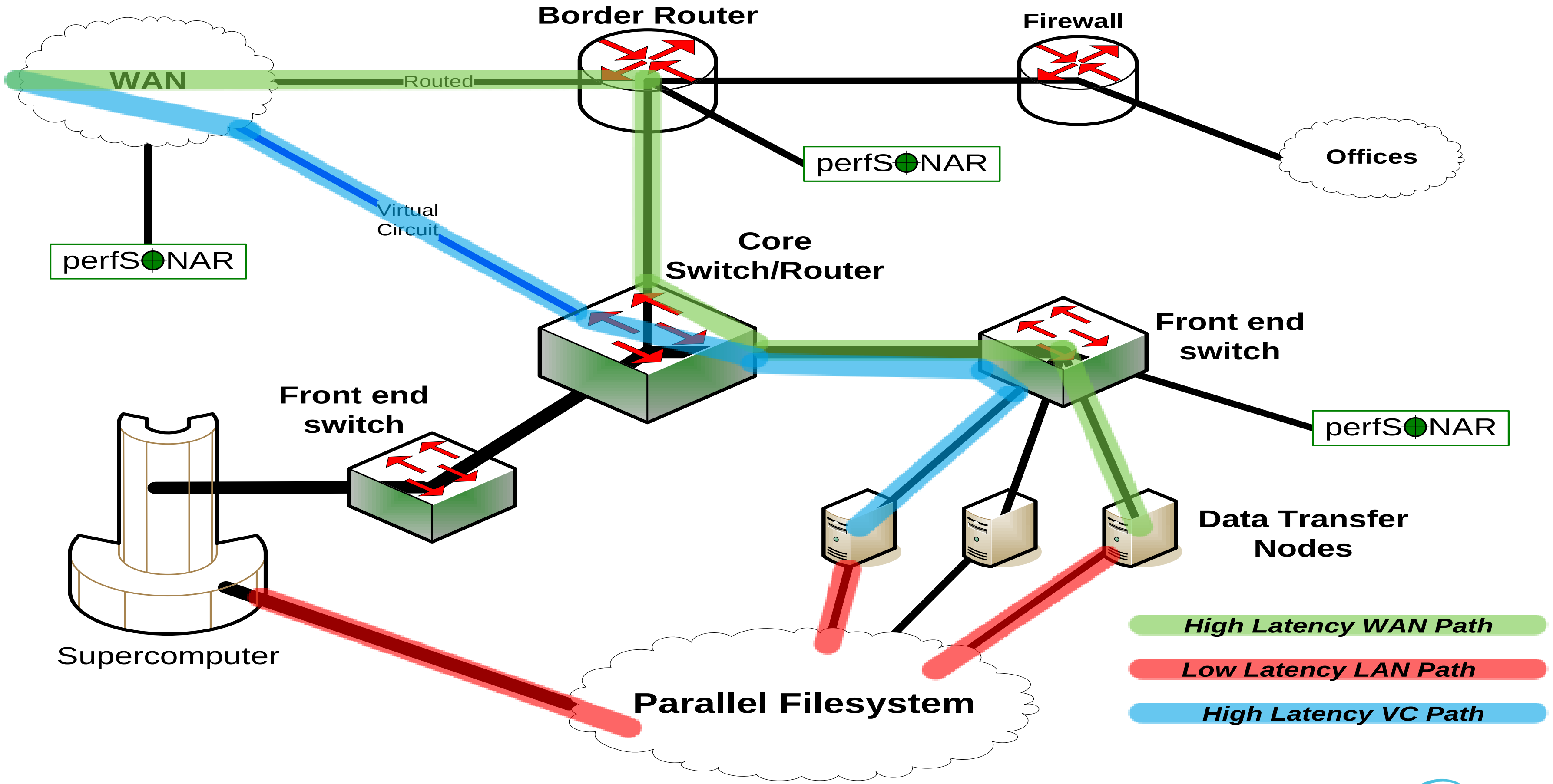
EPOC
Engagement and Performance
Operations Center

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 Data Path



EPOC Engagement and Performance Operations Center

Equipment – Routers and Switches

- Requirements for Science DMZ gear are different
  - 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

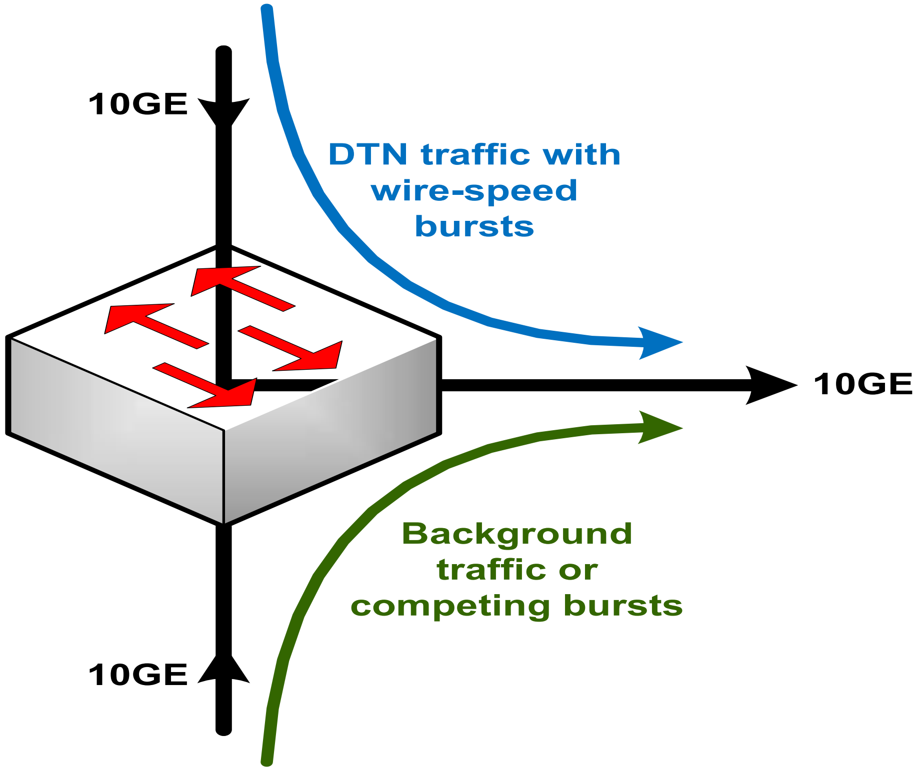
Common Circumstance: Multiple Ingress Data Flows, Common Egress

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

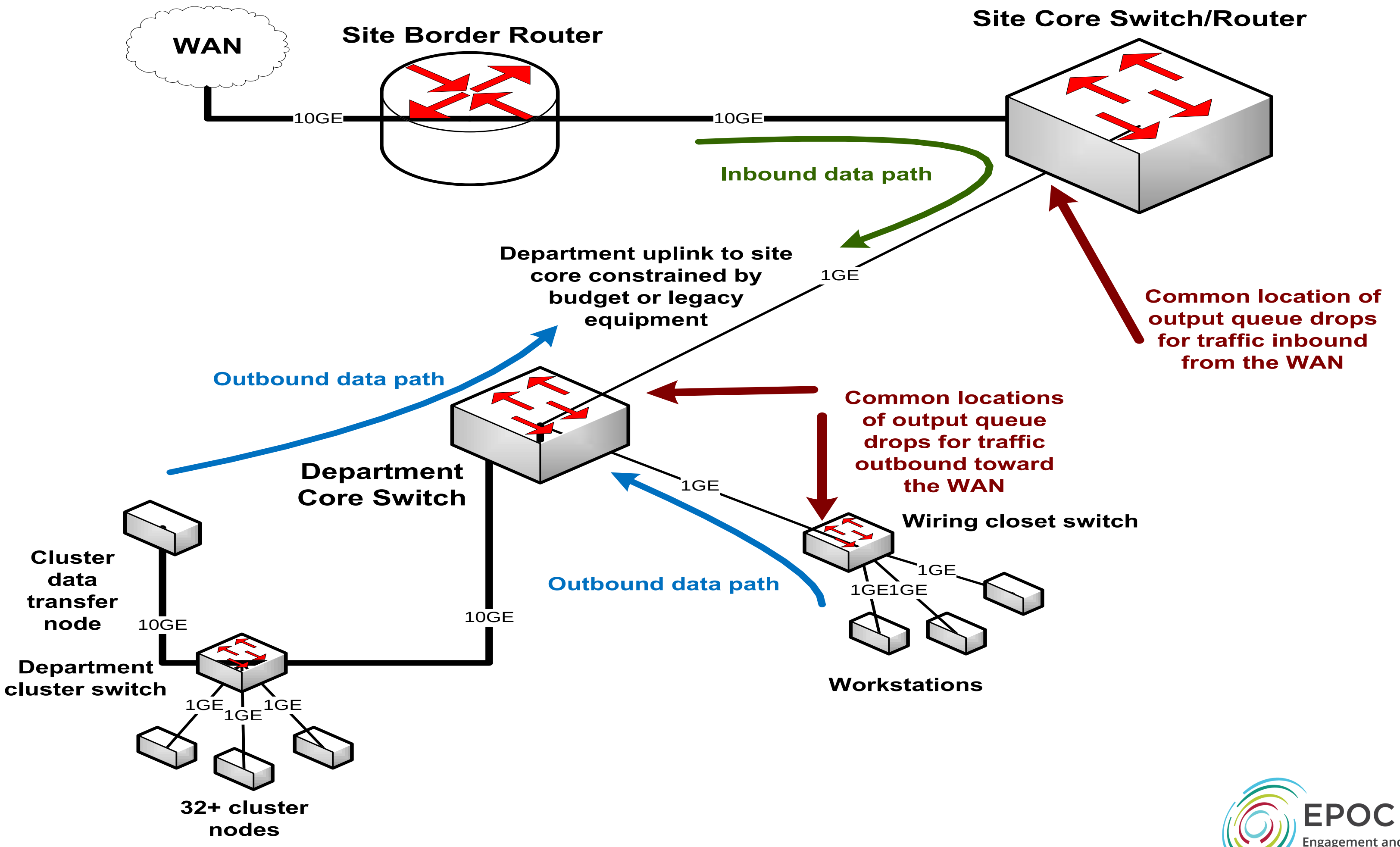
- Instantaneous rate, not average rate
- If TCP has window available and data to send, host sends until there is either no data or no window

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



Output Queue Drops – Common Locations



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

Rant Ahead

N.B. You are entering into rant territory on the matter of switch buffering. If you are going to take away anything from the next section:

1. Under buffered network devices are the ***single greatest threat*** to data intensive use of the network. You can make hosts, operating systems, and application choices perform better for 'free', it will cost $$$ to fix a crappy switch or router
2. You will be steered toward non-optimal choices when you talk with the vendor community because they don't understand simple math (but by the end of this, you will).
3. A 1U/2U data center/racklan network device should never be in the path of your data intensive network use case.
4. Non-passive (e.g. stateful) security devices are the same for buffering, and are actually worse due to the processing overhead.
5. Anytime you jump around the OSI stack – add friction (e.g. routing when you don't need to, application layer inspection, etc.)

EPOC
Engagement and Performance Operations Center

All About That Buffer (No Cut Through)

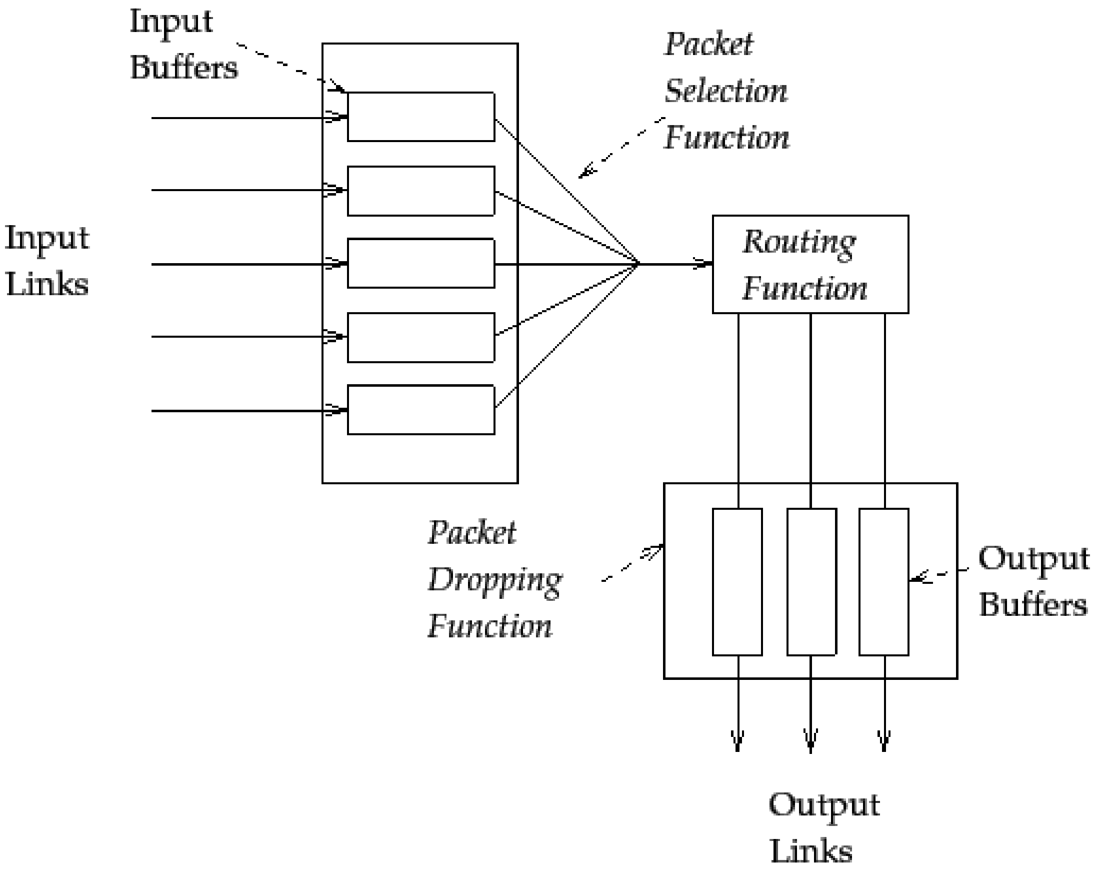


Figure 1: Basic Router Architecture

All About That Buffer (No Cut Through)

- Data arrives from multiple sources

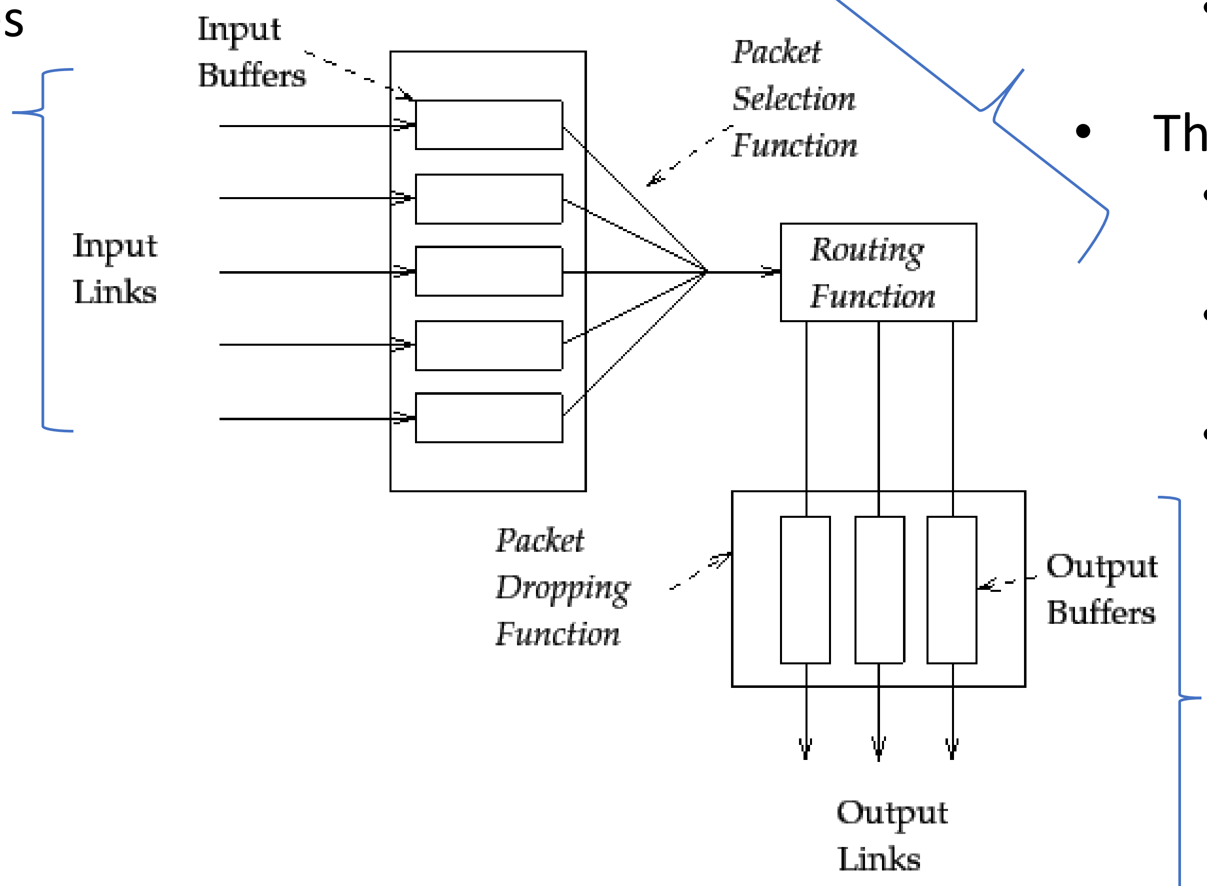


Figure 1: Basic Router Architecture

- 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
  - As the speed increases, there are more packets.
  - If there is not memory, we drop them, and that makes TCP 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 = ***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)

- Ignore the math aspect, its really just about making sure there is memory to catch packets. As the speed increases, there are more packets. If there is not memory, we drop them, and that makes TCP sad.
  - Memory on hosts, and network gear

EPOC
Engagement and Performance Operations Center

All About That Buffer (No Cut Through)

- What does this “look” like to a data transfer? Consider the test of iperf below
  - See TCP ‘ramp up’ and slowly increase the window
  - When something in the path has no more space for packets – a drop occurs. TCP will eventually react to the lost packet, and ‘back off’
  - In the example, this first occurs when we reach a buffer of around 6-8MB. Then after backoff the window is halved a couple of times
  - This happens again later – at a slightly higher buffer limit. This could be because there was cross traffic the first time, etc.

[ID]	Interval		Transfer	Bandwidth	Retr	Cwnd
[14]	0.00-1.00	sec	524 KBytes	4.29 Mbits/sec	0	157 KBytes
[14]	1.00-2.00	sec	3.31 MBytes	27.8 Mbits/sec	0	979 KBytes
[14]	2.00-3.00	sec	17.7 MBytes	148 Mbits/sec	0	***5.36 MBytes***
[14]	***3.00-4.00***	***sec***	***18.8 MBytes***	***157 Mbits/sec***	***214***	***1.77 MBytes***
[14]	4.00-5.00	sec	11.2 MBytes	94.4 Mbits/sec	0	***1.88 MBytes***
[14]	5.00-6.00	sec	10.0 MBytes	83.9 Mbits/sec	0	***2.39 MBytes***
[14]	6.00-7.00	sec	16.2 MBytes	136 Mbits/sec	0	3.63 MBytes
[14]	7.00-8.00	sec	23.8 MBytes	199 Mbits/sec	0	5.50 MBytes
[14]	8.00-9.00	sec	38.8 MBytes	325 Mbits/sec	0	8.23 MBytes
[14]	9.00-10.00	sec	57.5 MBytes	482 Mbits/sec	0	11.8 MBytes
[14]	10.00-11.00	sec	81.2 MBytes	682 Mbits/sec	0	***16.2 MBytes***
[14]	***11.00-12.00***	***sec***	***50.0 MBytes***	***419 Mbits/sec***	***35***	***3.93 MBytes***
[14]	12.00-13.00	sec	15.0 MBytes	126 Mbits/sec	0	***2.20 MBytes***
[14]	13.00-14.00	sec	11.2 MBytes	94.4 Mbits/sec	0	***2.53 MBytes***
[14]	14.00-15.00	sec	13.8 MBytes	115 Mbits/sec	1	1.50 MBytes
[14]	15.00-16.00	sec	6.25 MBytes	52.4 Mbits/sec	5	813 KBytes
[14]	16.00-17.00	sec	5.00 MBytes	41.9 Mbits/sec	0	909 KBytes
[14]	17.00-18.00	sec	5.00 MBytes	41.9 Mbits/sec	0	1.37 MBytes
[14]	18.00-19.00	sec	10.0 MBytes	83.9 Mbits/sec	0	2.43 MBytes
[14]	19.00-20.00	sec	17.5 MBytes	147 Mbits/sec	0	4.22 MBytes

EPOC
Engagement and Performance Operations Center

Decoding Specifications

- *"The buffering behaviors of the switches and their operating system, such as behavior under memory stress, are typically proprietary information and not well documented"* http://www.measurementlab.net/blog/traffic-microbursts-and-their-effect-on-internet-measurement/
- "Even if you know ***how much*** packet buffer is in the switch, assumptions on ***how it is deployed*** that are not backed up by testing can lead to unhappiness. What we like to say is that is ***the job of the network engineers to move bottlenecks around***."
  - *Jim Warner*
- *http://people.ucsc.edu/~warner/buffer.html*

EPOC
Engagement and Performance Operations Center

Decoding Specifications

- So lets say the spec sheet says this:

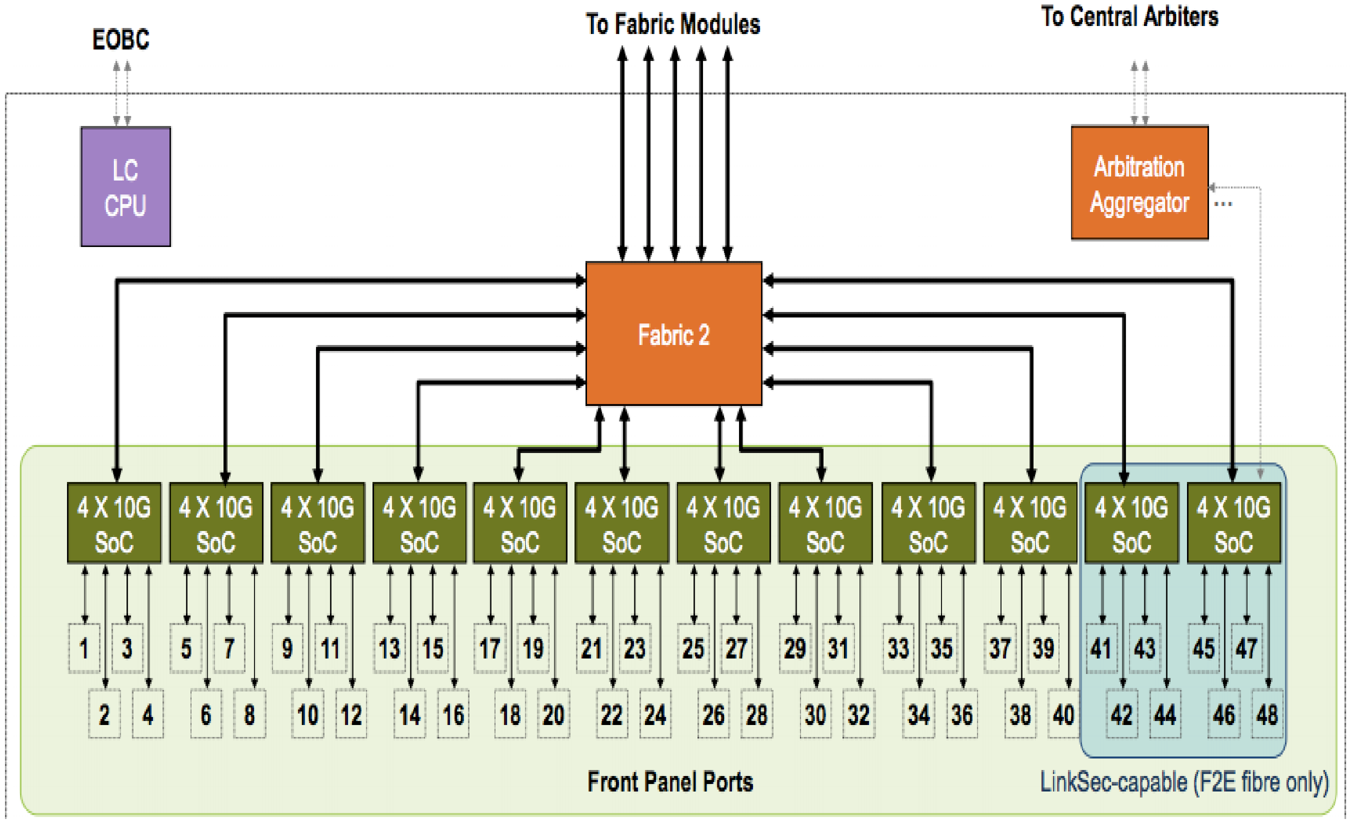
VOQ buffer	72 MB per module

- What does ‘module’ mean?
  - Typically this means the amount of memory for the entire switch (if it’s a single unit) or a blade (if the chassis supports more than one.
  - BUT … this memory can be allocated in a number of different ways:
    - ***Shared between all ports***
    - ***Dedicated (smaller) amounts per-port***
    - ***Shared between ASICS, which control a bank of ports***

EPOC
Engagement and Performance Operations Center

Decoding Specifications

- Consider this architecture
  - 48 Ports
    - 12 ASICS
    - 4 Ports per ASIC
  - ***72MB*** total
    - ***6MB per ASIC***
    - If all ports are in use – expect that each port has access to ***1.5MB***. If only one in use, it can use 6MB
  - Additional memory is often available in a 'burst buffer' in the fabric



ASIC = application-specific integrated circuit, think 'small routing engine'

EPOC
Engagement and Performance Operations Center

Decoding Specifications

- Recall: https://www.switch.ch/network/tools/tcp_throughput/
- What does 6MB get you?
  - 1Gbps @ <= 48ms (e.g. ½ needed for coast-to-coast)
  - 10Gbps @ <= 4.8ms (e.g. metro area)
- What does 1.5MB get you?
  - 1Gbps @ <= 12ms (e.g. regional area)
  - 10Gbps @ <= 1.2ms (e.g. data center [or more accurately, rack or row)]
- In either case – remember this assumes you are the only thing using that memory … congestion is a more likely reality

EPOC
Engagement and Performance Operations Center

Takeaways

- Try before you buy
  - Request a demo unit (or two)
  - Learn all the ins and outs
- Develop tests for worst case scenario
  - Plug in all the ports, and create traffic with a hardware tester (IXIA, SPIRENT) or a perfSONAR resource
  - Cross traffic within the switch
  - Testing to far away resources (latency is your friend and enemy)
- If you can’t get single stream TCP to work well, buffers are often the core of the problem
- Its worth spending the extra $ on buffer, really

EPOC
Engagement and Performance Operations Center

http://ce.sc.edu/cyberinfra/workshop_2020.html

EPOC
Engagement and Performance Operations Center

TCP Basics + Science DMZ

Jason Zurawski
zurawski@es.net
ESnet / Lawrence Berkeley National Laboratory

Training Workshop for Educators and Network Engineers on High Speed Network Protocols and Security
May 4th, 2020

ESnet
ENERGY SCIENCES NETWORK

INDIANA UNIVERSITY