Science DMZ Architecture

Jason Zurawski - zurawski@es.net
Kate Petersen Mace – kate@es.net

ESnet Science Engagement – engage@es.net

http://fasterdata.es.net

Science DMZ Overview

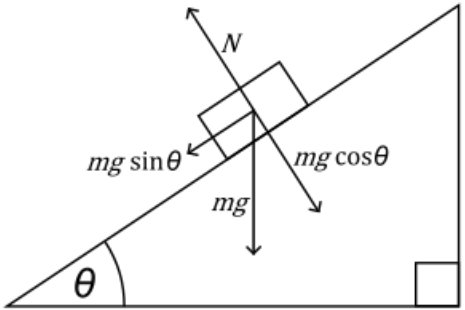
OIN Operating Innovative Networks

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 and GridFTP
- ***Performance measurement/test node***
  - perfSONAR
- ***Once it's up, users often need training – Engagement is key***



perfSONAR

OIN Operating Innovative Networks

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

ESnet

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

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

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)

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 loss 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.

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

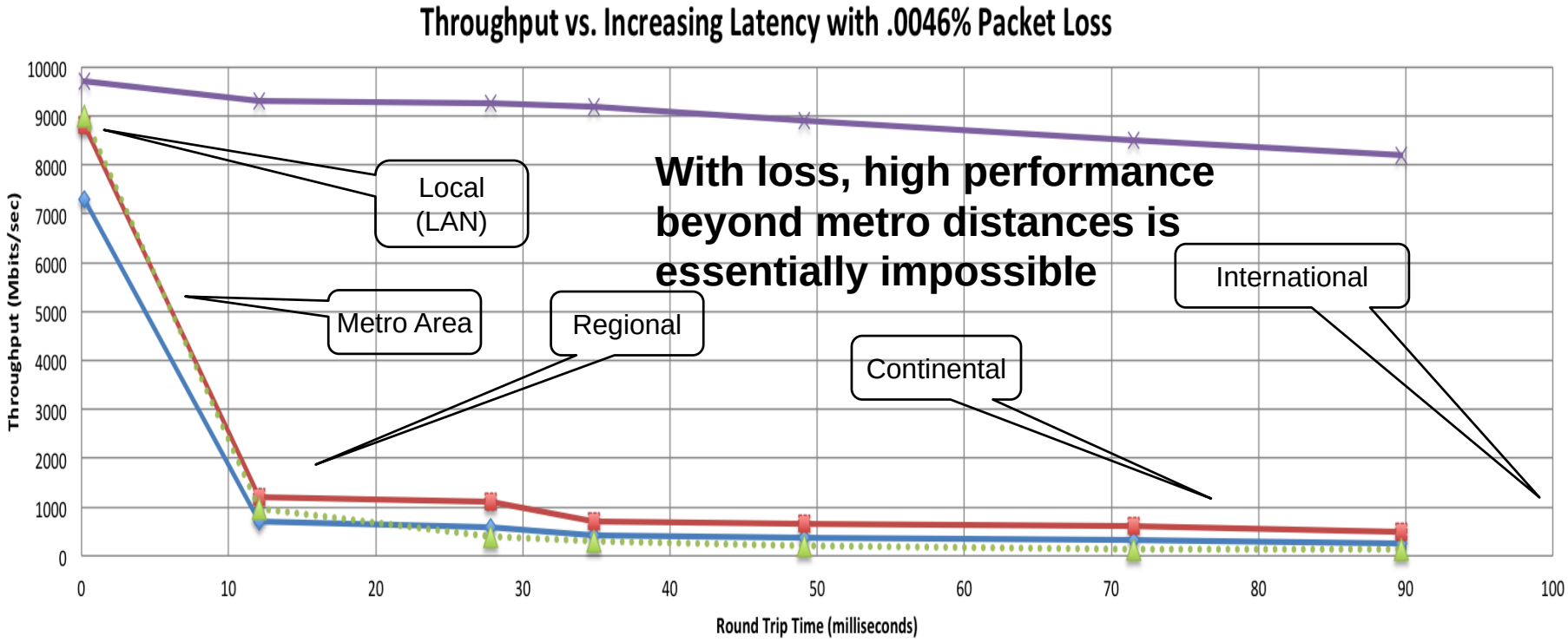
- We want to eliminate the causes of packet loss – so that we don't need to test out the (slow) way that TCP can recover.

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

OIN Operating Innovative Networks

ESnet

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



Firewall Limits

Breaking a Network

- Disk PT Hosts (10G)
  - bnl-diskpt1.es.net (Upton, NY)
  - lbl-diskpt1.es.net (Berkeley, CA)
- Path
  - ~75ms RTT

```
traceroute to 198.124.238.150 (198.124.238.150), 30 hops max, 60 byte packets
 1  lblmr2-lbl-diskpt1.es.net (198.129.77.101)  0.252 ms  0.226 ms  0.187 ms
 2  sunncr5-ip-a-lblmr2.es.net (134.55.49.1)  2.299 ms  2.609 ms  2.992 ms
 3  sacrcr5-ip-a-sunncr5.es.net (134.55.40.5)  4.729 ms  4.947 ms  5.352 ms
 4  denvcr5-ip-a-sacrcr5.es.net (134.55.50.202)  25.709 ms  25.912 ms  26.223 ms
 5  kanscr5-ip-a-denvcr5.es.net (134.55.49.58)  37.057 ms  37.093 ms  37.452 ms
 6  chiccr5-ip-a-kanscr5.es.net (134.55.43.81)  47.076 ms  47.773 ms  47.817 ms
 7  starcr5-ip-a-chiccr5.es.net (134.55.42.42)  47.533 ms  47.340 ms  47.644 ms
 8  bostcr5-ip-a-starcr5.es.net (134.55.218.189)  68.946 ms  68.844 ms  69.092 ms
 9  newycr5-ip-a-bostcr5.es.net (134.55.209.34)  73.474 ms  73.415 ms  73.682 ms
10  bnlmr2-ip-a-newycr5.es.net (134.55.221.133)  74.806 ms  74.829 ms  74.960 ms
11  bnl-diskpt1.es.net (198.124.238.150)  74.928 ms  74.933 ms  74.912 ms
```

OIN Operating Innovative Networks

ESnet

Forcing Bad Performance (to illustrate behavior)

- Add 10% Loss to a specific host

```
sudo /sbin/tc qdisc delete dev eth0 root
sudo /sbin/tc qdisc add dev eth0 root handle 1: prio
sudo /sbin/tc qdisc add dev eth0 parent 1:1 handle 10: netem loss 10%
sudo /sbin/tc filter add dev eth0 protocol ip parent 1:0 prio 3 u32 match ip dst 198.129.254.78/32 flowid 1:1
```

- Add 10% Duplication to a specific host

```
sudo /sbin/tc qdisc delete dev eth0 root
sudo /sbin/tc qdisc add dev eth0 root handle 1: prio
sudo /sbin/tc qdisc add dev eth0 parent 1:1 handle 10: netem duplicate 10%
sudo /sbin/tc filter add dev eth0 protocol ip parent 1:0 prio 3 u32 match ip dst 198.129.254.78/32 flowid 1:1
```

- Add 10% Corruption to a specific host

```
sudo /sbin/tc qdisc delete dev eth0 root
sudo /sbin/tc qdisc add dev eth0 root handle 1: prio
sudo /sbin/tc qdisc add dev eth0 parent 1:1 handle 10: netem corrupt 10%
sudo /sbin/tc filter add dev eth0 protocol ip parent 1:0 prio 3 u32 match ip dst 198.129.254.78/32 flowid 1:1
```

- Reorder packets: 50% of packets (with a correlation of 75%) will get sent immediately, others will be delayed by 75ms.

```
sudo /sbin/tc qdisc delete dev eth0 root
sudo /sbin/tc qdisc add dev eth0 root handle 1: prio
sudo /sbin/tc qdisc add dev eth0 parent 1:1 handle 10: netem delay 10ms reorder 25% 50%
sudo /sbin/tc filter add dev eth0 protocol ip parent 1:0 prio 3 u32 match ip dst 198.129.254.78/32 flowid 1:1
```

- Reset things

```
sudo /sbin/tc qdisc delete dev eth0 root
```

OIN Operating Innovative Networks

ESnet

How Do We Accommodate TCP?

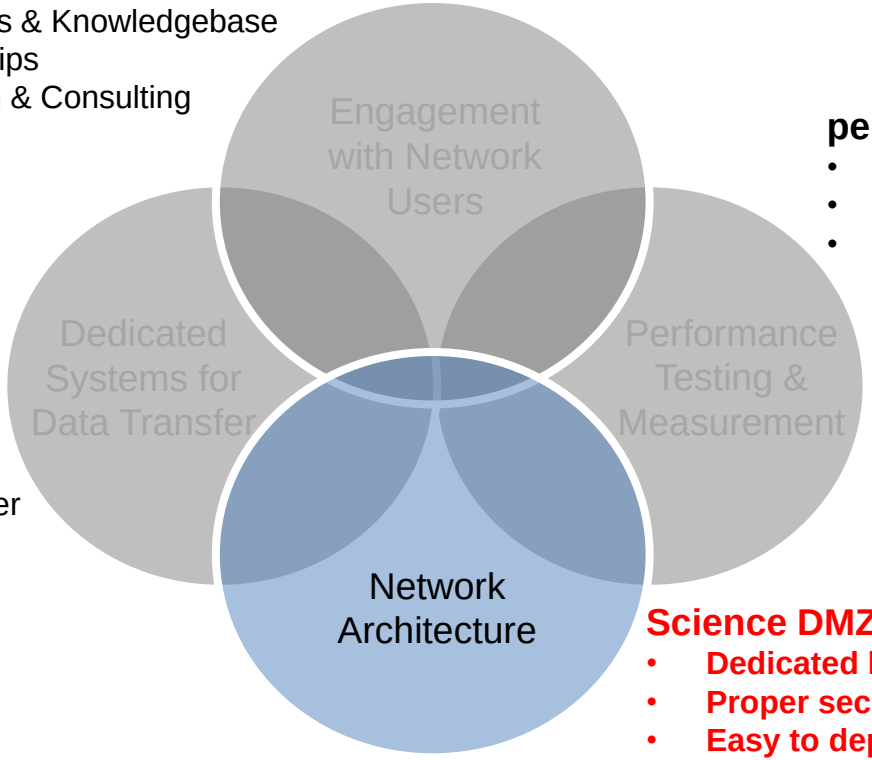
- High-performance wide area TCP flows must get loss-free service
  - Sufficient bandwidth to avoid congestion
  - Deep enough buffers in routers and switches to handle bursts
    - Especially true for long-distance flows due to packet behavior
    - No, this isn't buffer bloat
- Equally important –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

OIN Operating Innovative Networks

ESnet

The Data Transfer Superfecta: Science DMZ Model

Engagement
- Resources & Knowledgebase
- Partnerships
- Education & Consulting



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

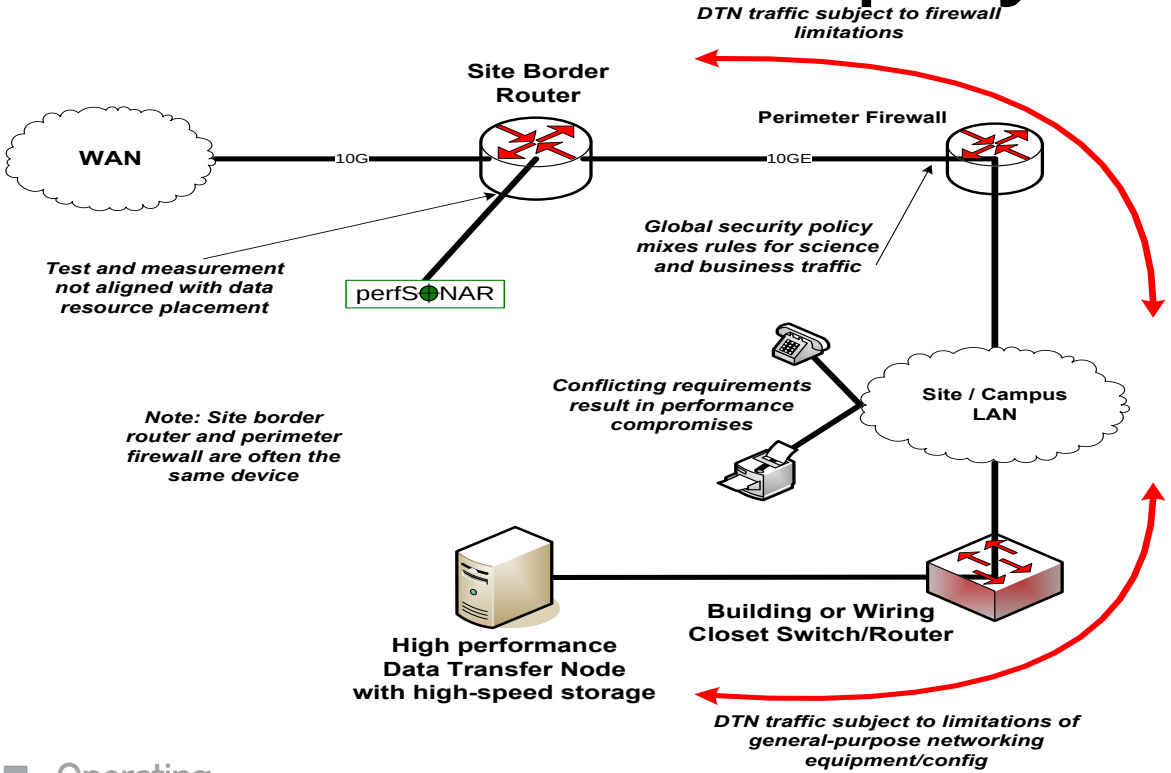
Data Transfer Node
- Configured for data transfer
- High performance
- Proper tools

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

OIN Operating Innovative Networks

ESnet

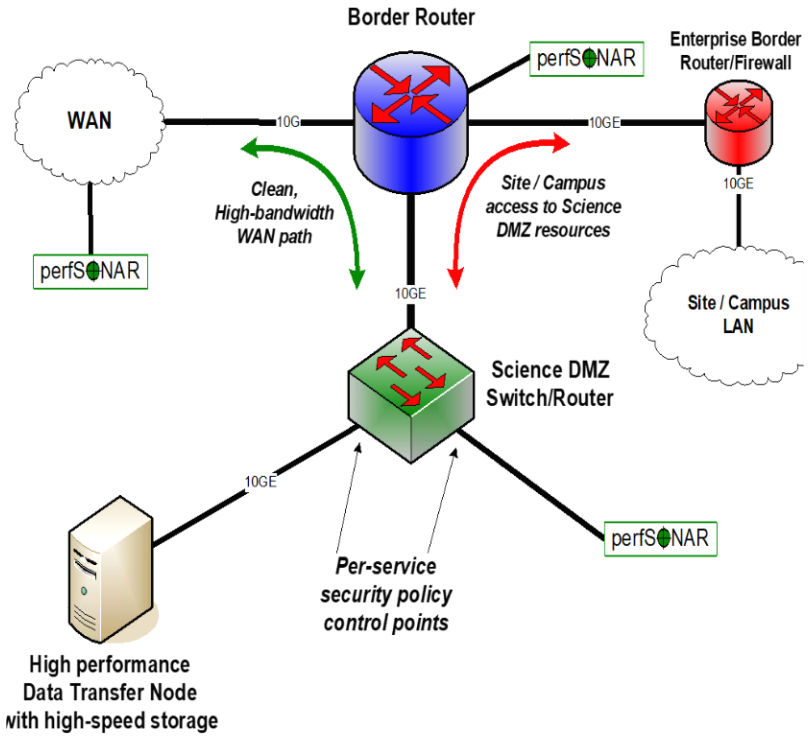
Ad Hoc DTN Deployment



OIN Operating Innovative Networks

ESnet

A better approach: simple Science DMZ



OIN Operating Innovative Networks

ESnet

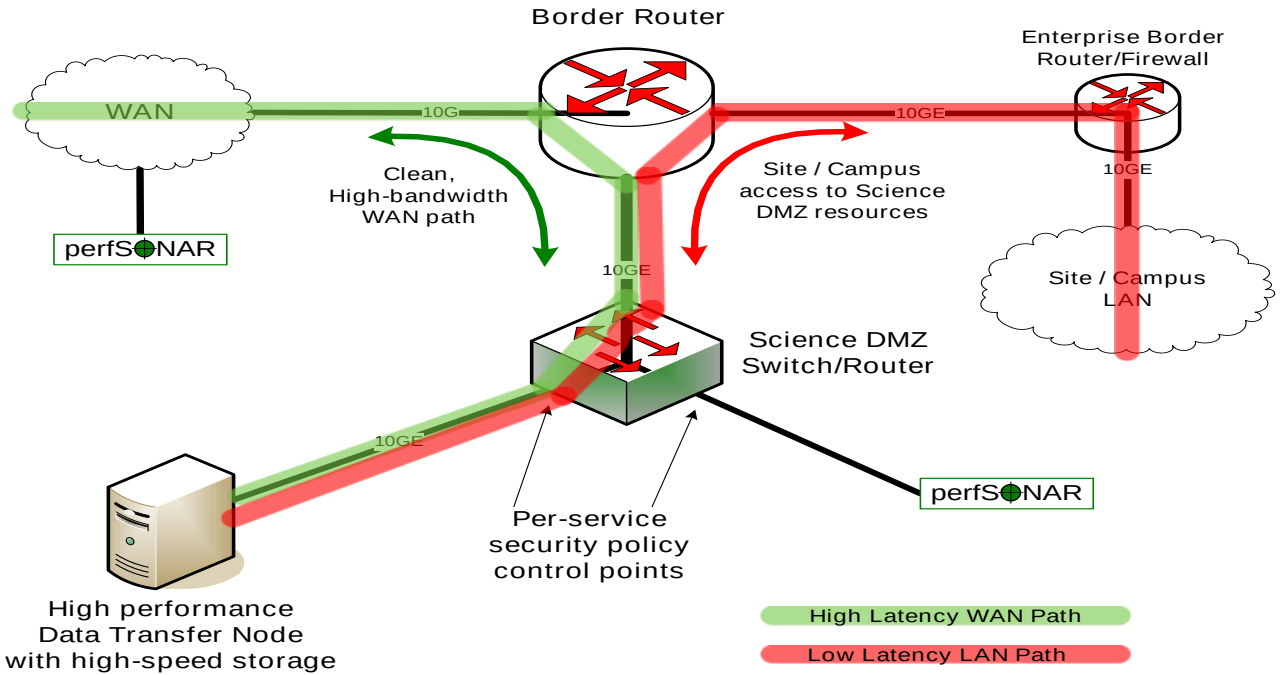
Small-scale Science DMZ Deployment

- 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

OIN Operating Innovative Networks

ESnet

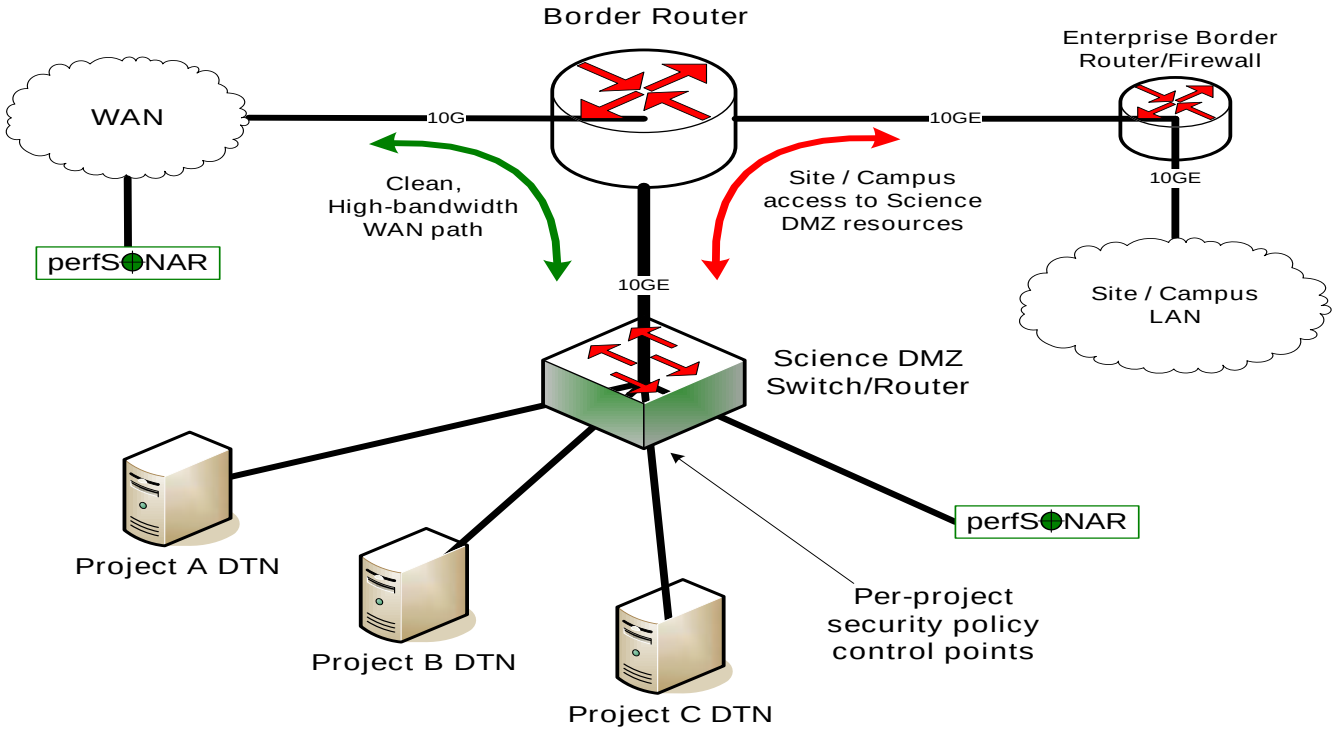
Prototype Science DMZ Data Path



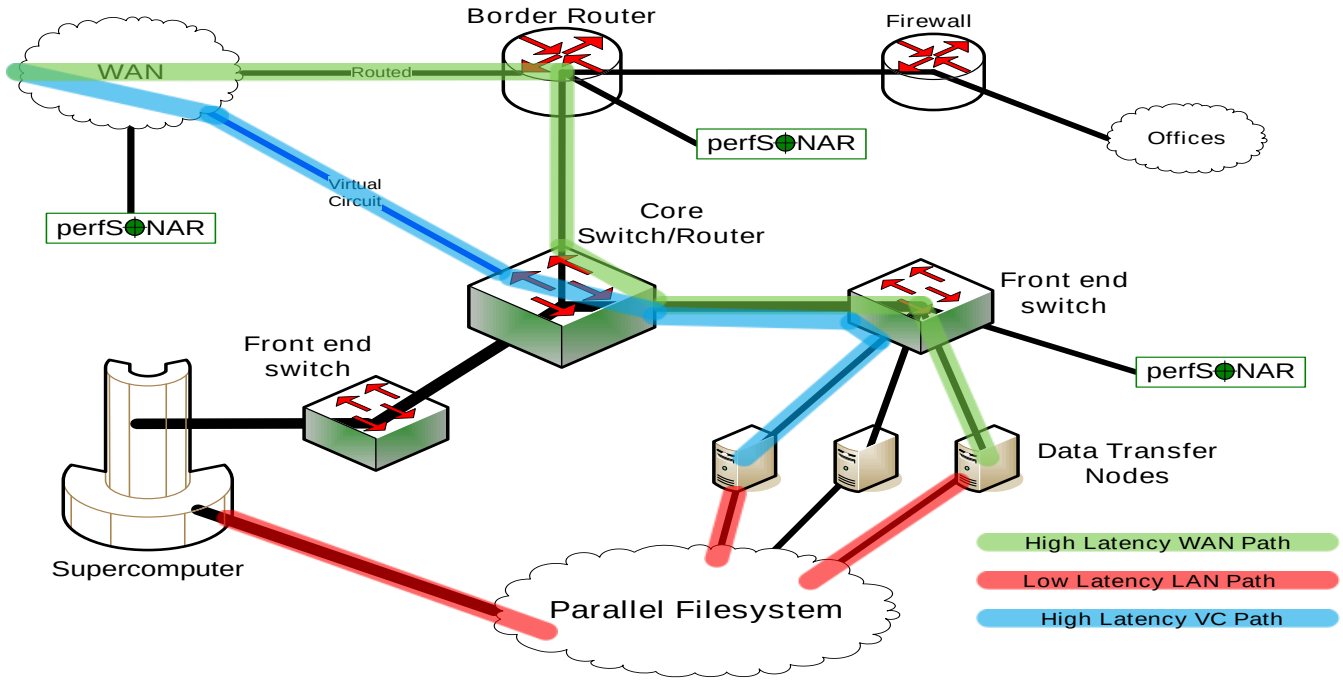
OIN Operating Innovative Networks

ESnet

Multiple Projects



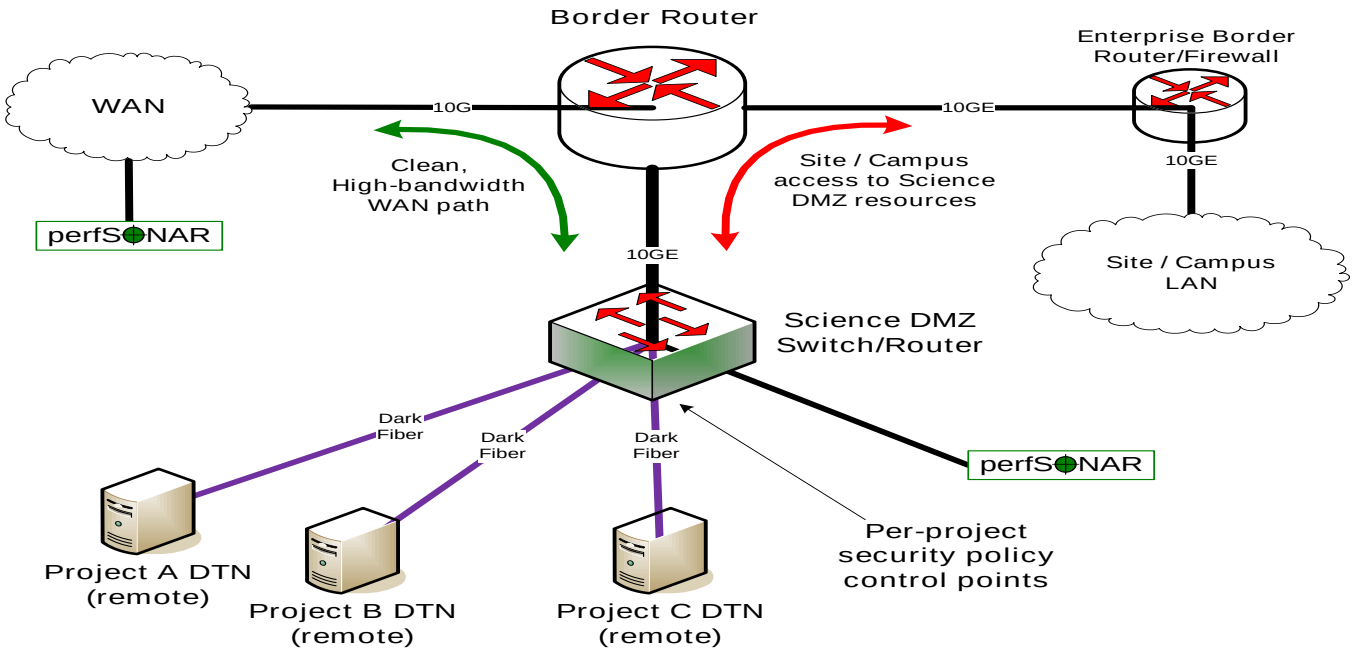
Supercomputer Center Data Path



OIN Operating Innovative Networks

ESnet

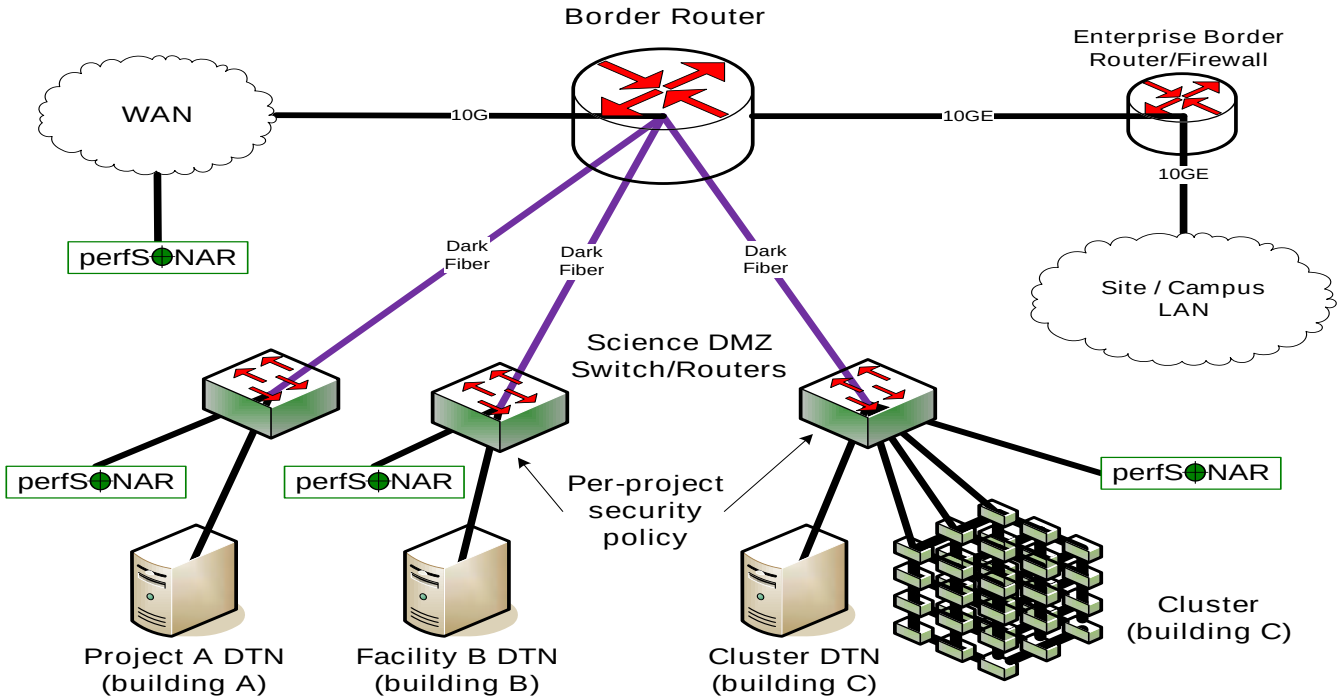
Distributed Science DMZ – Dark Fiber



OIN Operating Innovative Networks

ESnet

Multiple Science DMZs – Dark Fiber



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

OIN Open Innovative Networks

ESnet

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)

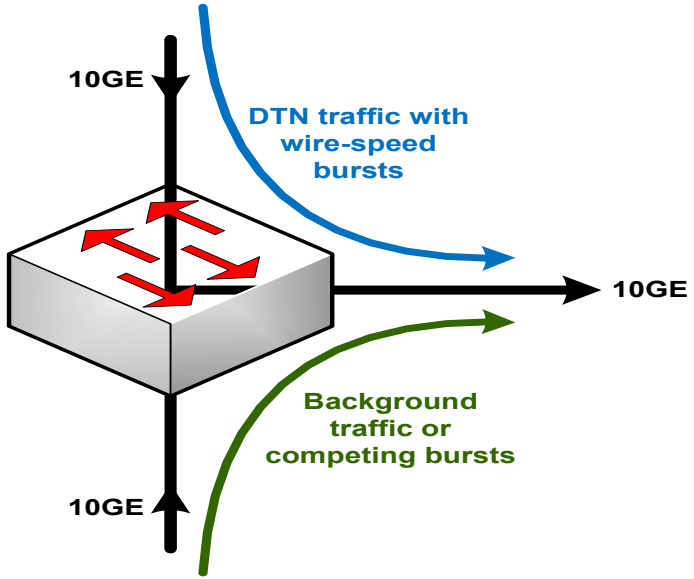
Common Circumstance: Multiple Ingress Data Flows, Common Egress

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

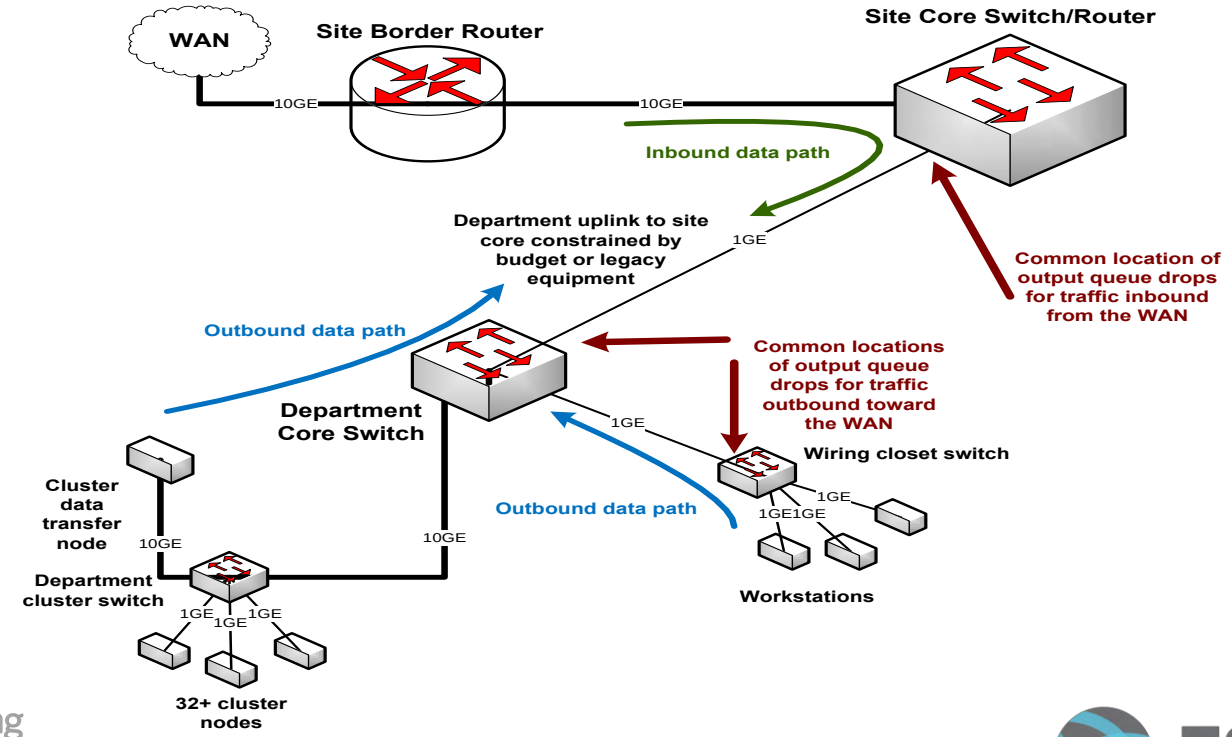
- This is true even if the average rate as measured over seconds is slower (e.g. 4Gbps)
- On microsecond time scales, there is often congestion
- Router or switch must queue packets or drop them



OIN Operating Innovative Networks

ESnet

Output Queue Drops – Common Locations



OIN Operating Innovative Networks

ESnet

ScienceDMZ Security

OIN Operating Innovative Networks

ESnet

“Please tell me – what should I buy??”

- We get this question a lot
  - Hard to answer for several reasons
  - Some have to do with us, and some have to do with you
- We can’t endorse anybody
  - We can’t recommend one business over another
  - No, really – we’ll get slapped
  - We are also unable to accept any liability
- We have no idea what’s right for your environment
  - Different networks are different – specific vendor, multi-vendor, single vendor, management style, home-grown tools, commercial tools, …
  - Our goal is to describe our understanding of what works and why
  - You are free to use that information in whatever way you see fit
- **Know Your Network**

OIN Operating Innovative Networks

ESnet

You are not alone

- Lots of community resources
  - Ask folks who have already done it
  - Ask the Science DMZ mailing list: sciencedmz@es.net
- Vendors can be very helpful – just ask the right questions
  - Request an eval box (or preferably two)
  - Ask for config examples to implement a particular feature
    - E.g. “Please give me the QoS config for the following:”
      - 1 queue for network control (highest priority) – 5% of interface buffer memory
      - 1 queue configured for tail-drop (lower priority) – 95% of interface buffer memory
      - With that config, how many milliseconds of buffer are in the tail-drop queue when measured at interface wire speed?

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

OIN Operating Innovative Networks

ESnet

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 – and its important you know this going into all conversations).
3. A 1U/2U data center/racklan network device ***should never be in the path*** of your data intensive network use case. You have lost immediately, if one lives there.
4. Non-passive (e.g. stateful) security devices are the same for buffering, and are actually worse due to the processing overhead (e.g. delayed TCP is worse than dropped TCP
5. Anytime you jump around the OSI stack – add friction (e.g. routing when you don't need to, application layer inspection, etc.)

OIN Operating Innovative Networks

ESnet

All About That Buffer (No Cut Through)

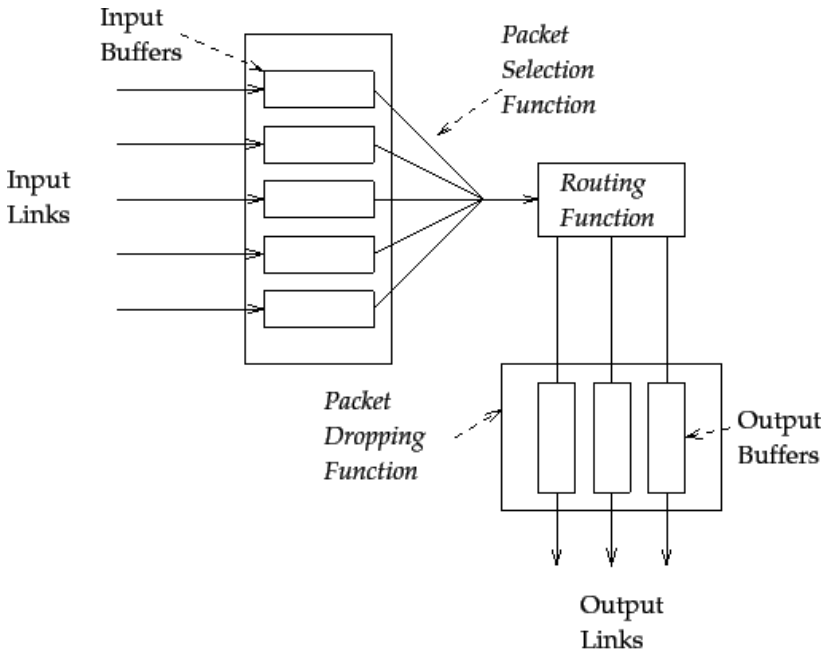


Figure 1: Basic Router Architecture

All About That Buffer (No Cut Through)

- Data arrives from multiple sources

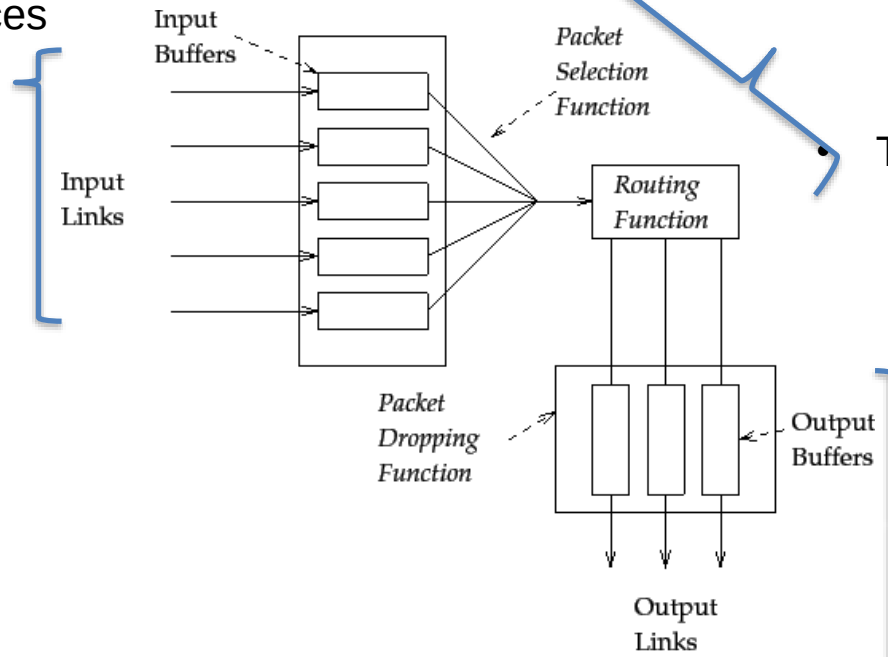
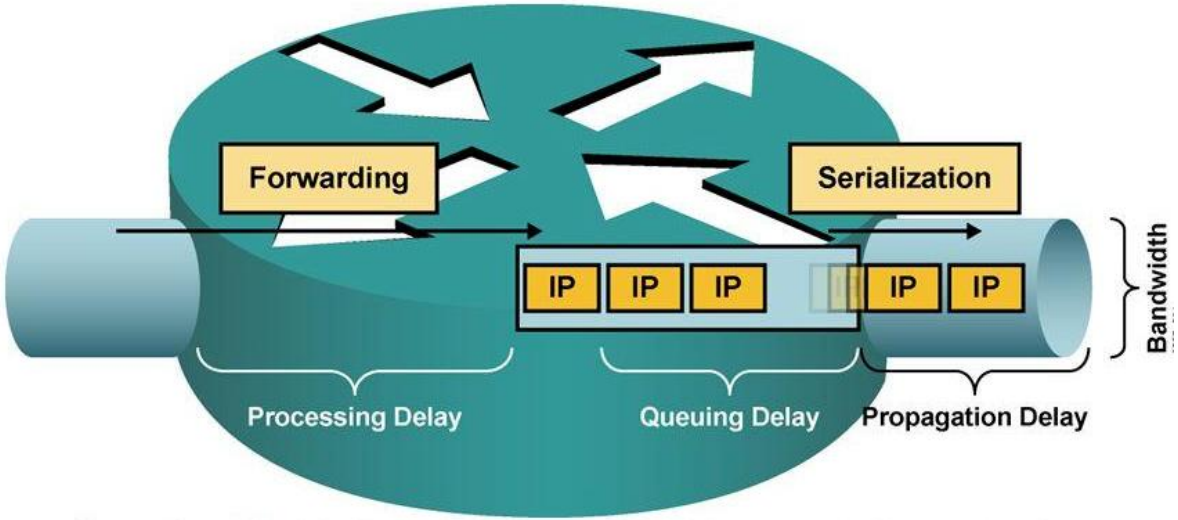


Figure 1: Basic Router Architecture

- 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)

All About That Buffer (No Cut Through)



- **Processing delay:** The time it takes for a router to take the packet from an input interface, examine it, and put it into the output queue of the output interface.
- **Queuing delay:** The time a packet resides in the output queue of a router.
- **Serialization delay:** The time it takes to place the "bits on the wire."
- **Propagation delay:** The time it takes for the packet to cross the link from one end to the other.

OIN Operating Innovative Networks

ESnet

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.

OIN Operating Innovative Networks

ESnet

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

OIN Operating Innovative Networks

ESnet

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 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.

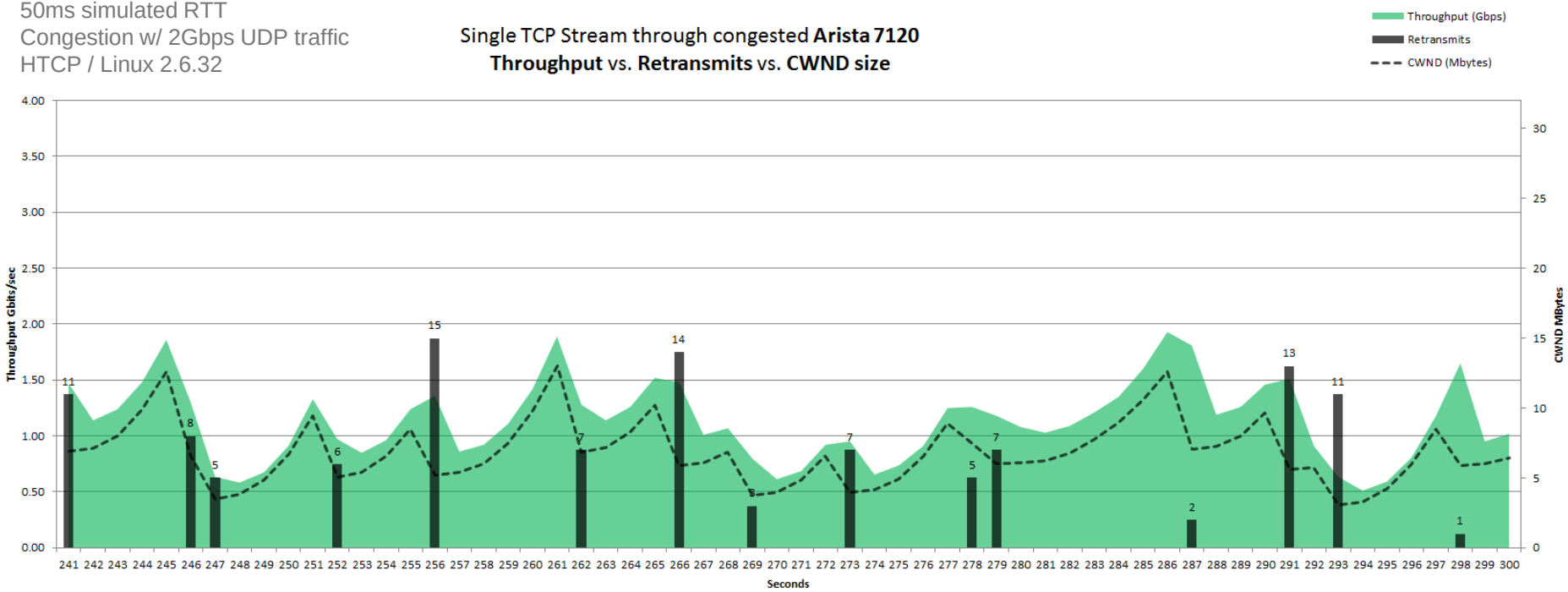
```
[ ID] Interval           Transfer     Bandwidth       Retr  Cwnd
[ 14]   0.00-1.00   sec   524 KBytes  4.29 Mbits/sec    0    157 KBytes
[ 14]   1.00-2.00   sec  3.31 MBytes  27.8 Mbits/sec    0    979 KBytes
[ 14]   2.00-3.00   sec  17.7 MBytes   148 Mbits/sec    0   5.36 MBytes
[ 14]   3.00-4.00   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
```

OIN Operating Innovative Networks

ESnet

TCP's Congestion Control

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



OIN Operating Innovative Networks

Slide from Michael Smitasin, LBLnet

ESnet

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*

OIN Operating Innovative Networks

ESnet

Decoding Specifications

- So lets say the spec sheet says this:

VOQ buffer	72 MB per module

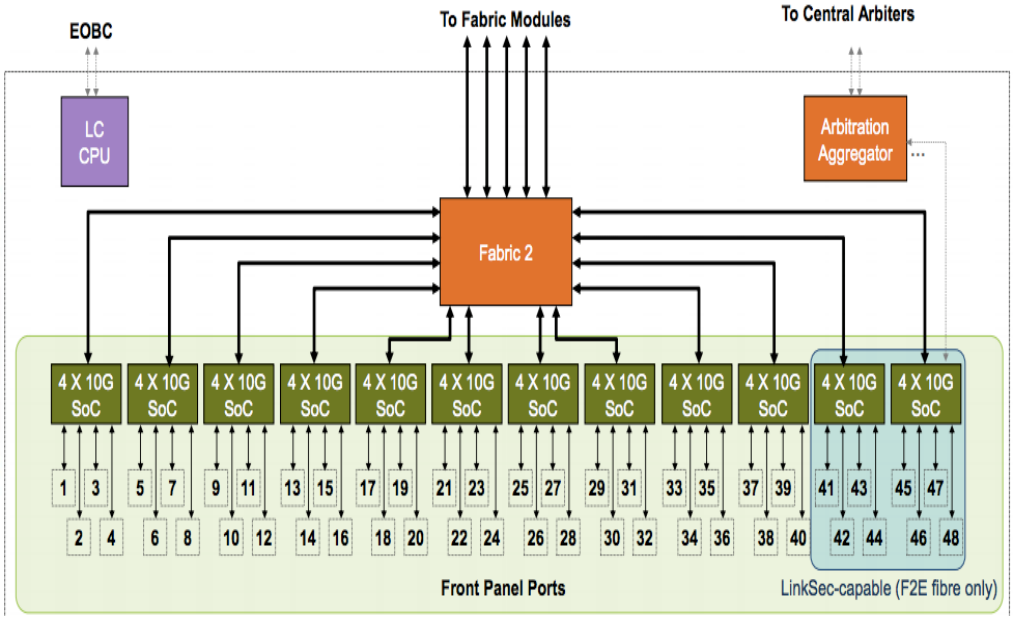
- What does ‘module’ mean?
  - Typically this means the amount of memory for the entire switch (if it’s a single unit) or a blade (if the chassis supports more than one.
  - BUT … this memory can be allocated in a number of different ways:
    - ***Shared between all ports***
    - ***Dedicated (smaller)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

OIN Operating Innovative Networks

ESnet

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 is in use, it can use 6M
  - Additional memory is often available in a ‘burst buffer’ in the fabric



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

OIN Operating Innovative Networks

ESnet

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

TCP Performance

OIN Operating Innovative Networks

ESnet

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

OIN Operating Innovative Networks

ESnet

More Takeaways

- There is no single “correct” way build a Science DMZ
  - These are design patterns, not rules
- It depends on things like:
  - site requirements
  - existing resources
  - availability of dark fiber
  - budget

- The main point is to reduce the opportunities for packet loss, and be able to find loss if it’s present

Know your (Platform) Limitations

- We have seen significant limitations in 100G equipment from all vendors with a major presence in R&E
  - 100G single flow not supported
    - Channelized forwarding plane
    - Unexplained limitations
    - Sometimes the senior sales engineers don't know!
  - Non-determinism in the forwarding plane
    - Performance depends on features used (i.e. config-dependent)
    - Packet loss that doesn't show up in counters anywhere
- If you can't find it, nobody will tell you about it
  - Vendors don't know or won't say
  - Watch how you write your procurements
- Second-generation equipment seems to be much better
- Vendors have been responsive in rolling new code to fix problems

They Don’t Test For This Stuff

- Most sales engineers and support engineers don’t have access to 100G test equipment
  - It’s expensive
  - Setup of scenarios is time-consuming
- R&E traffic profile is different than their standard model
  - IMIX (Internet Mix) traffic is normal test profile
    - Aggregate web browsers, email, YouTube, Netflix, etc.
    - Large flow count, low per-flow bandwidth
    - This is to be expected – that’s where the market is
  - R&E shops are the ones that get the testing done for R&E profile
    - SCinet at Supercomuting conference provides huge value
    - But, in the end, it’s up to us – the R&E community

New Technology, New Bugs

- Bugs happen.
  - Data integrity (traffic forwarded, but with altered data payload)
  - Packet loss
  - Interface wedge
  - Optics flaps
- Monitoring systems are indispensable
- Finding and fixing issues is sometimes hard
  - Rough guess – difficulty exponent is degrees of freedom
    - Vendors/platforms, administrative domains, time zones
- Takeaway – don’t skimp on test gear (at least maintain your perfSONAR boxes)

OIN Operating Innovative Networks

ESnet