http://bit.ly/EPOC-USC-CI

End devices in Science DMZs: DTNs

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

Training Workshop for Network Engineers and Educators on Tools and Protocols for High-Speed Networks
University of South Carolina
July 22-23, 2019

National Science Foundation Award #1826994

Data Transfer Node

- A DTN server is made of several subsystems. Each needs to perform optimally for the DTN workflow:
  - **Storage:** capacity, performance, reliability, physical footprint
  - **Networking:** protocol support, optimization, reliability
  - **Motherboard:** I/O paths, PCIe subsystem, IPMI
  - **Chassis:** adequate power supply, extra cooling

- **Note: the workflow we are optimizing for here is sequential reads/write of large files, and a moderate number of high bandwidth flows.**

- We assume this host is dedicated to data transfer, and not doing data analysis/manipulation

EPOC
Engagement and Performance Operations Center

Motherboard and Chassis selection

- Chassis
  - Extra cooling (for future expansion, unless you buy the system full)
  - Make sure the power supply is adequate
- Motherboard/CPU
  - Ivy Bridge/SkyLake/Cascade Lake or newer CPU architecture (e.g. for 10/40G, you can get away with a reasonably modern machine)
    - High clock rate better than high core count for DTNs – max this out
    - Faster QPIC for communication between processors
  - PCI Gen 3 or newer (40G and 100G requires PCI Gen 3)
  - Memory speed – faster is better, more is better
    - We recommend 128GB of RAM for a DTN node
  - IPMI for remote management (optional)

EPOC
Engagement and Performance Operations Center

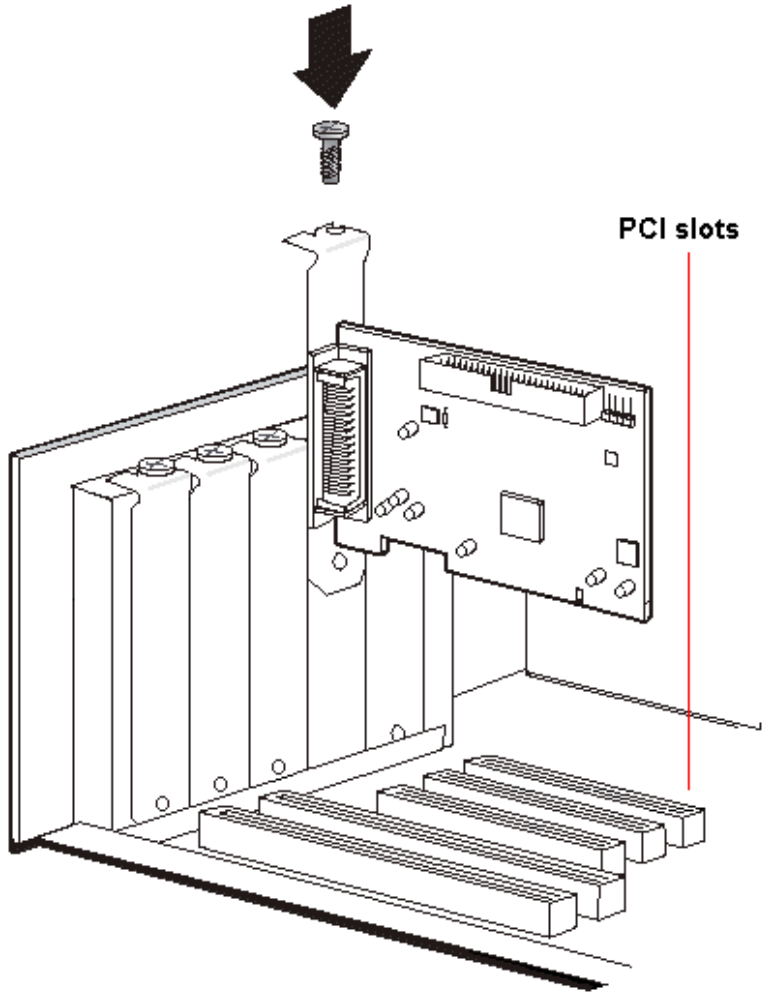
PCI Slot Considerations

- PCI slots are defined by:
  - Slot width:
    - Physical card and form factor
    - Max number of lanes
  - Lane count:
    - Maximum bandwidth per lane
    - Most cards will run slower in a slower slot
    - Not all cards will use all lanes

EPOC
Engagement and Performance Operations Center

PCI Slot Considerations

- Card being inserted:
  - x8 Slot Width
- Slot being used:
  - x8 Slot Width
- Wider Slot (bottom):
  - x16 Slot Width



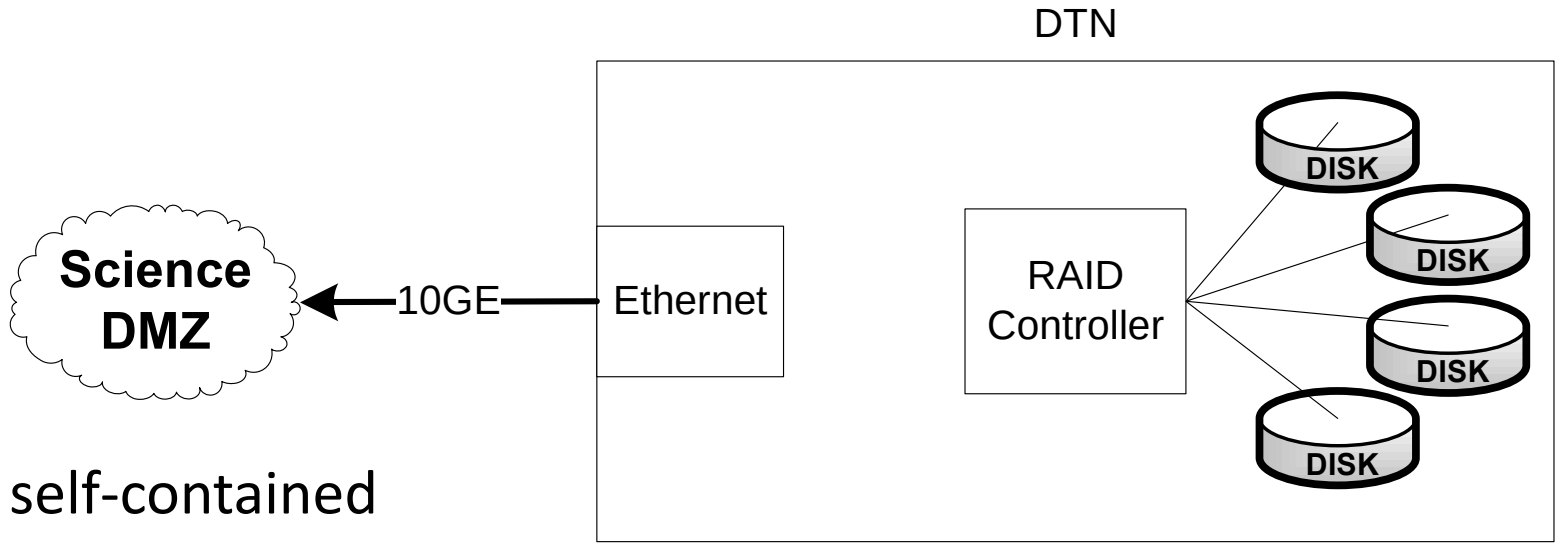
EPOC
Engagement and Performance Operations Center

PCI Bus Considerations

- Example:
  - 10GE NICs require an 8 lane PCIe-2 slot
  - 40G/QDR NICs require an 8 lane PCIe-3 slot
  - 100G NICs (ex. Mellanox) requires a 16 lane PCIe-3 slot
  - Most RAID controllers require an 8 lane PCIe-2 slot
  - Fusion-IO cards may require a 16 lane PCIe-3 slot

EPOC
Engagement and Performance Operations Center

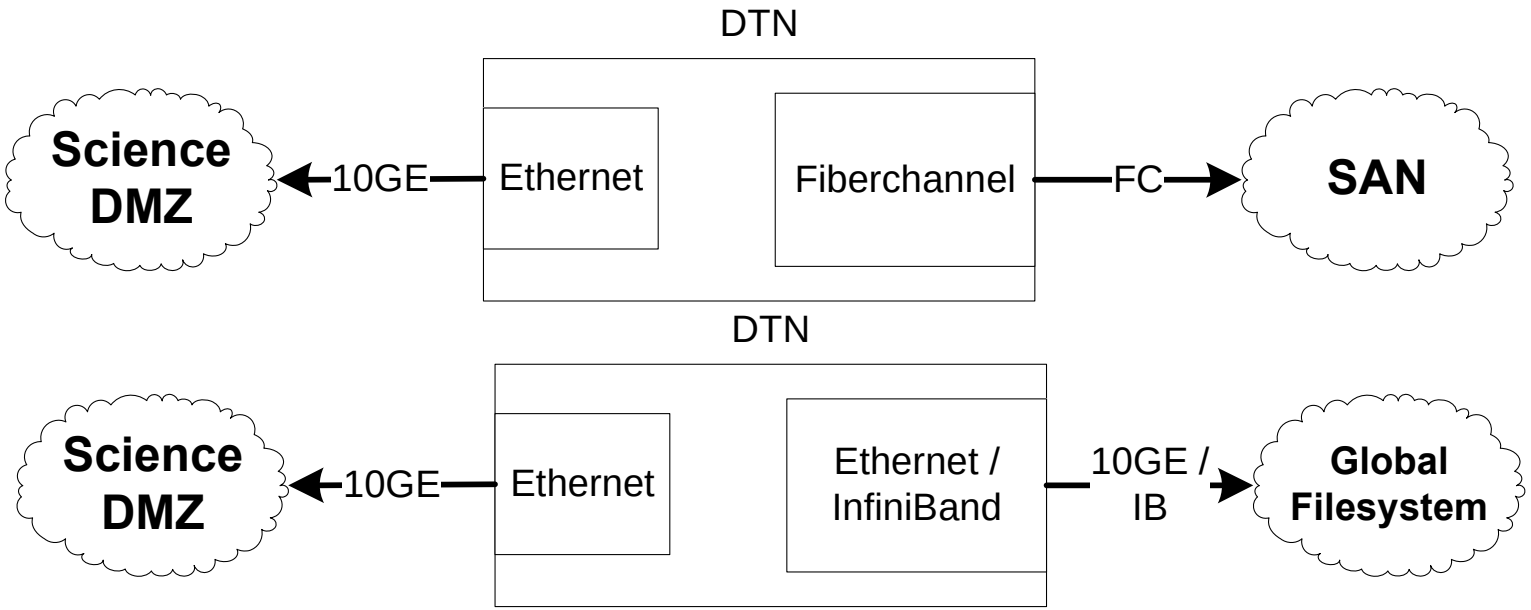
Storage Architectures - Internal



- DTN with internal RAID is self-contained
  - Same CPU, RAM, etc. as DTN with external storage
  - No external dependencies for storage
  - Deployable anywhere
  - Limited scalability
  - Storage managed locally (you get whatever tools the RAID controller gives you)

EPOC
Engagement and Performance Operations Center

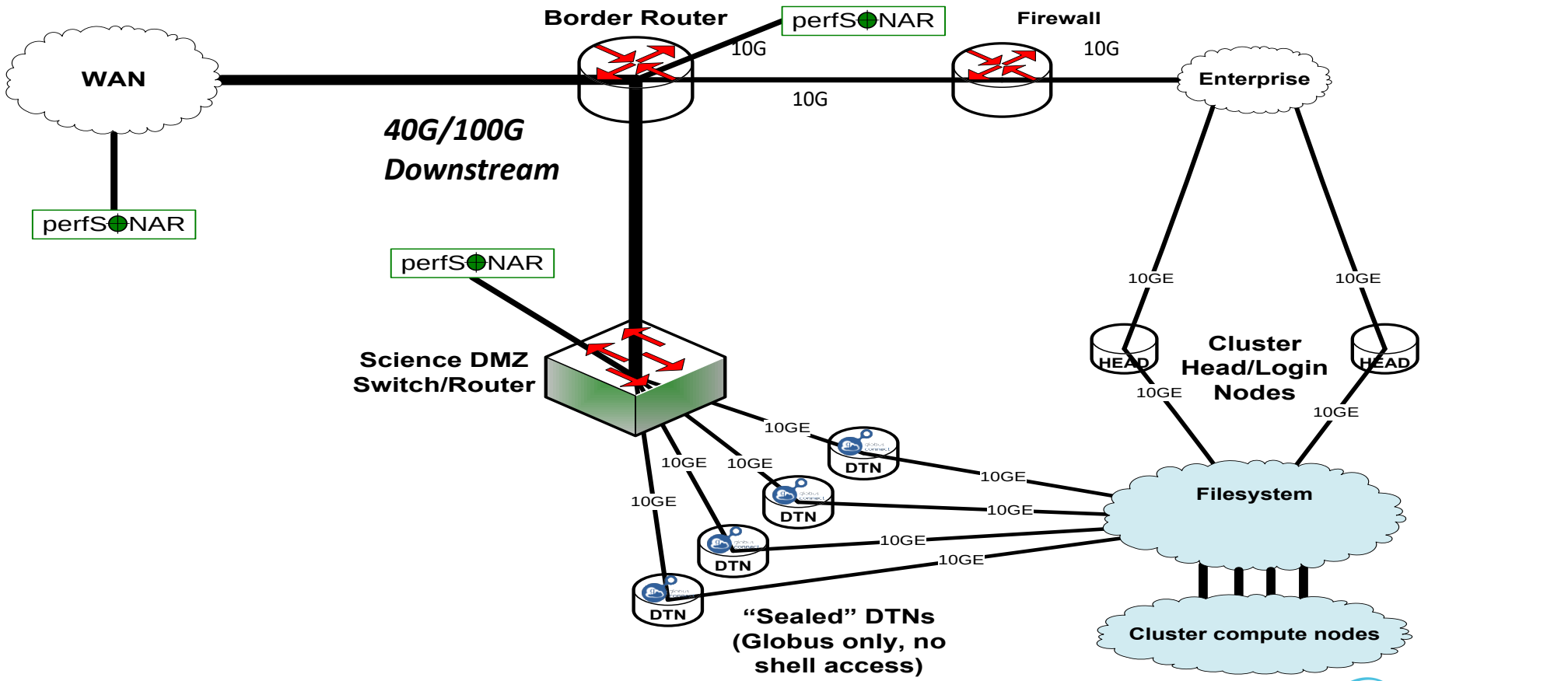
Storage Architectures - External



- These are essentially the same from a DTN host design perspective
  - IB, Ethernet, or Fibrechannel card connects to external storage
  - Other system components (CPU, RAM, etc.) the same
  - Central storage management, greater flexibility
  - Integration with other large-scale resources (e.g. HPC)

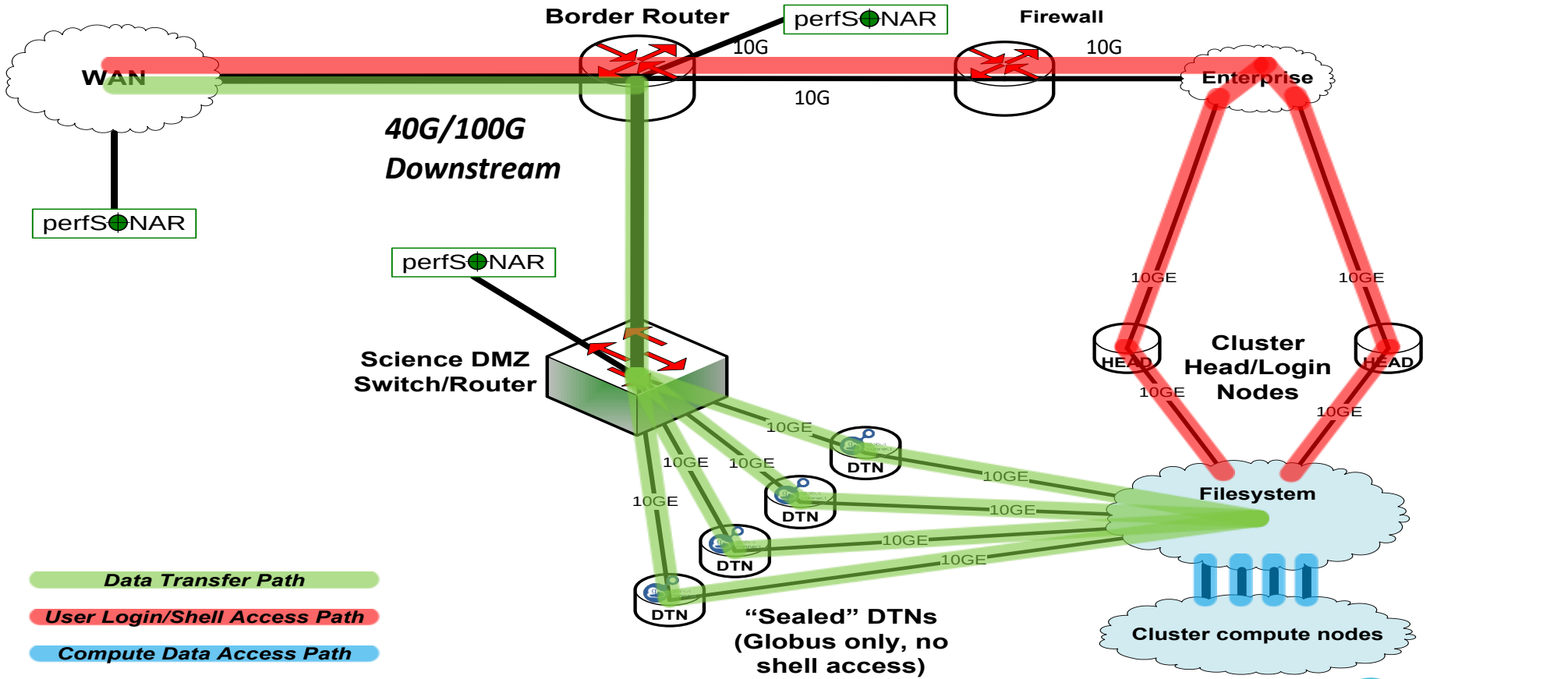
EPOC
Engagement and Performance Operations Center

DTNs in a Facility



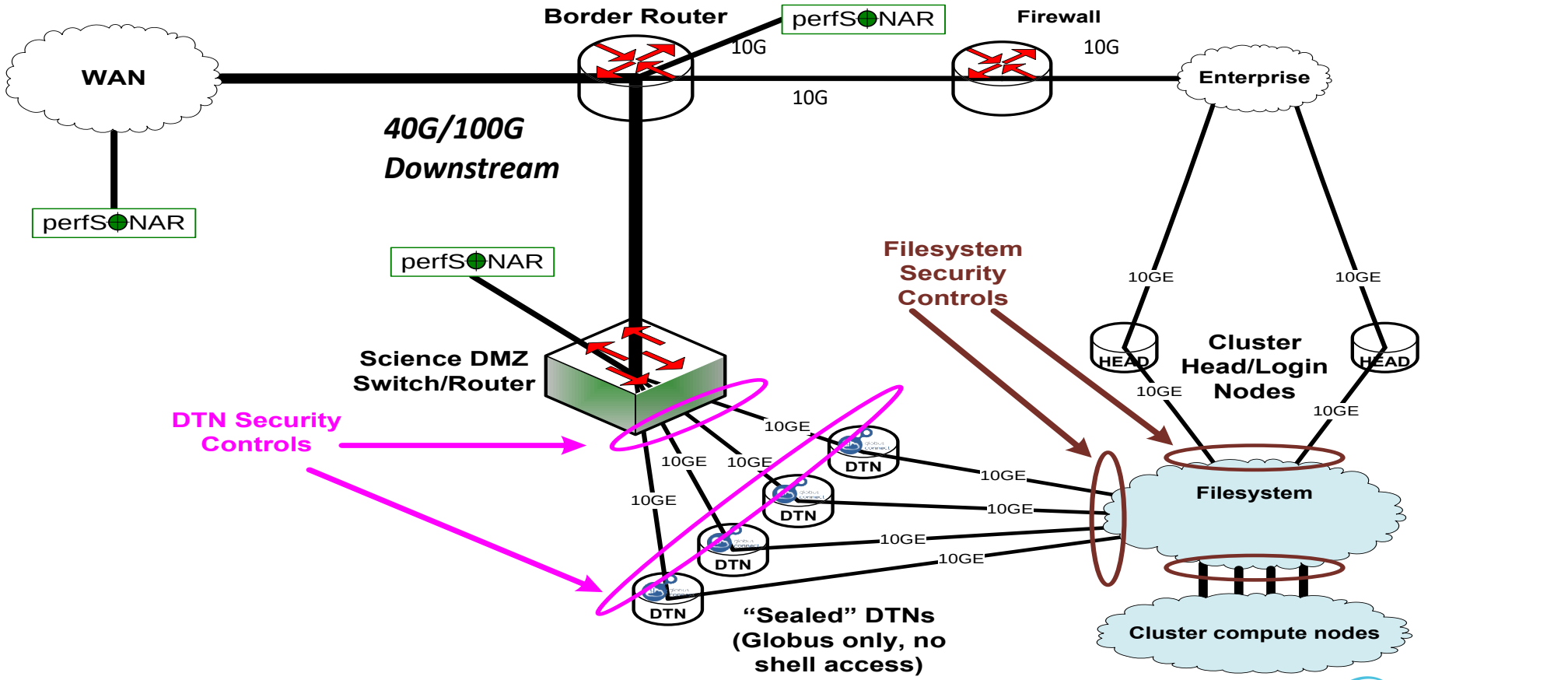
EPOC
Engagement and Performance Operations Center

DTNs in a Facility – Network Paths



EPOC
Engagement and Performance Operations Center

DTNs in a Facility – Security



EPOC
Engagement and Performance Operations Center

Storage Subsystem Selection

- Deciding what storage to use in your DTN is based on what you are optimizing for:
  - performance, reliability, capacity, and/or cost
- SATA disks historically have been cheaper and higher capacity, while SAS disks typically have been the fastest.
  - Technologies have been converging (and its hard to keep up)
- Do what you can support well (ditto for filesystems – ZFS, ext4, etc.)

EPOC
Engagement and Performance Operations Center

SSDs and HDs

- SSD storage costs much more than traditional hard drives (HDs), but are much faster. They come in different styles:
  - PCIe card: some vendors (Fusion-IO) build PCI cards with SSDs.
    - These are the fastest type of SSD: up to several GBytes/sec per card.
      - Note that this type of SSD is typically not hot-swapable.
  - HD replacement: several vendors now sell SSD-based drives that have the same form factor as traditional drives such as SAS and SATA.
    - The downside to this approach is that performance is limited by the RAID controller, and not all controllers work well with SSD.
      - ***Be sure that your RAID controller is "SSD capable"***.

- Note that the price of SSD is coming down quickly, so an SSD-based solution may be worth considering for your DTNs.

EPOC
Engagement and Performance Operations Center

SSD Form Factors

EPOC
Engagement and Performance Operations Center

RAID Controllers

- Often optimized for a given workload, rarely for performance.
- RAID0 is the fastest of all RAID levels but is also the least reliable.
- The performance of the RAID controller is a factor of the number of drives and its own processing engine.

EPOC
Engagement and Performance Operations Center

RAID Controller

- Be sure your RAID controller has the following:
  - >= 1GB of on-board cache
  - PCIe Gen3 support (and your board can support that)
  - dual-core RAID-on-Chip (ROC) processor ***if you will have more than 8 drives***

EPOC
Engagement and Performance Operations Center

Network Subsystem Selection

- There is a huge performance difference between cheap and expensive 10G/40G NICs.
  - E.g. please don't cheap out on the NIC – it's important for an optimized DTN host.
- NIC features to look for include:
  - support for interrupt coalescing
  - support for MSI-X
  - TCP Offload Engine (TOE)
  - ***support for zero-copy protocols such as RDMA (RoCE or iWARP)***
- Note that many 10G/40G/100G NICs come in dual ports, but that ***does not mean if you use both ports at the same time you get double the performance***. Often the second port is meant to be used as a backup port.
  - Myricom 10G-PCIE2-8C2-2S
  - Mellanox MCX312A-XCBT
  - Mellanox ConnectX-3 , ConnectX-4

EPOC
Engagement and Performance Operations Center

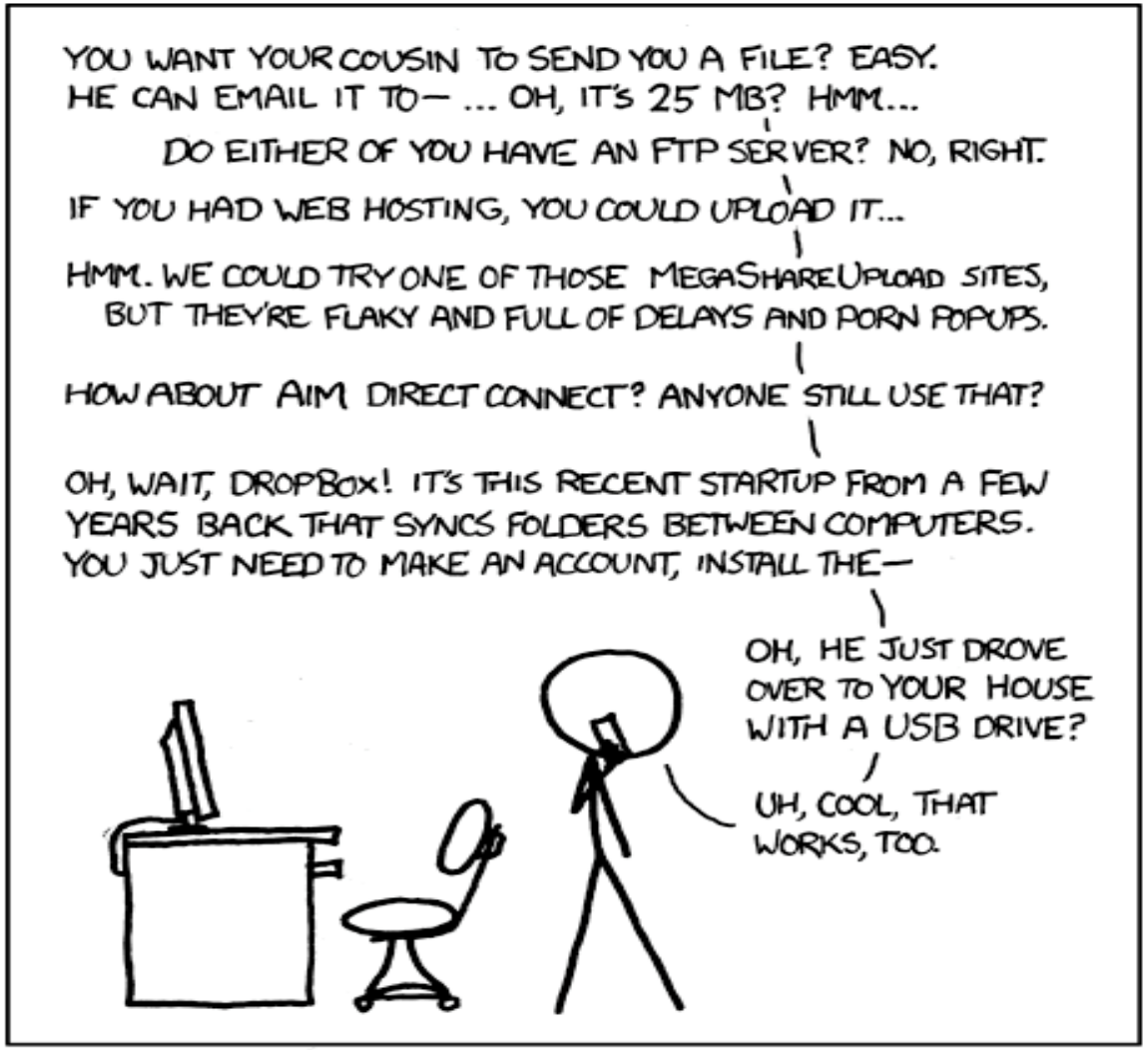
Reference Implementation (Oct 2014)

- Hardware description
  - Motherboard: SuperMicro X9DRi-F
  - CPU: 2 x Intel(R) Xeon Ivy Bridge E5-2643V2 3.5GHz 6 Cores (Total 12 Cores)
  - Memory: 96G ((12) 8GB DDR3-1866MHz RAM ECC/REG)
  - RAID: Adaptec ASR-81605ZQ (16 ports)
  - Network Controller: Myricom 10G-PCIE2-8C2-2S or Mellanox MCX312A-XCBT
  - Other recommended 10G NICs include Chelsio T5, and the Intel X520.
- System Configuration
  - We use the most recent CentOS-6 distribution of Linux, and have configured the data drives as RAID6.
- Performance Results for this configuration (Back-to-Back Testing using GridFTP)
  - memory to memory, 1 10GE NIC: 9.9 Gbps
  - memory to memory, 4 10GE NICs: 39.5 Gbps
  - disk to disk: 9.2 Gbps (1.2 GBytes/sec) using a single large file

EPOC
Engagement and Performance Operations Center

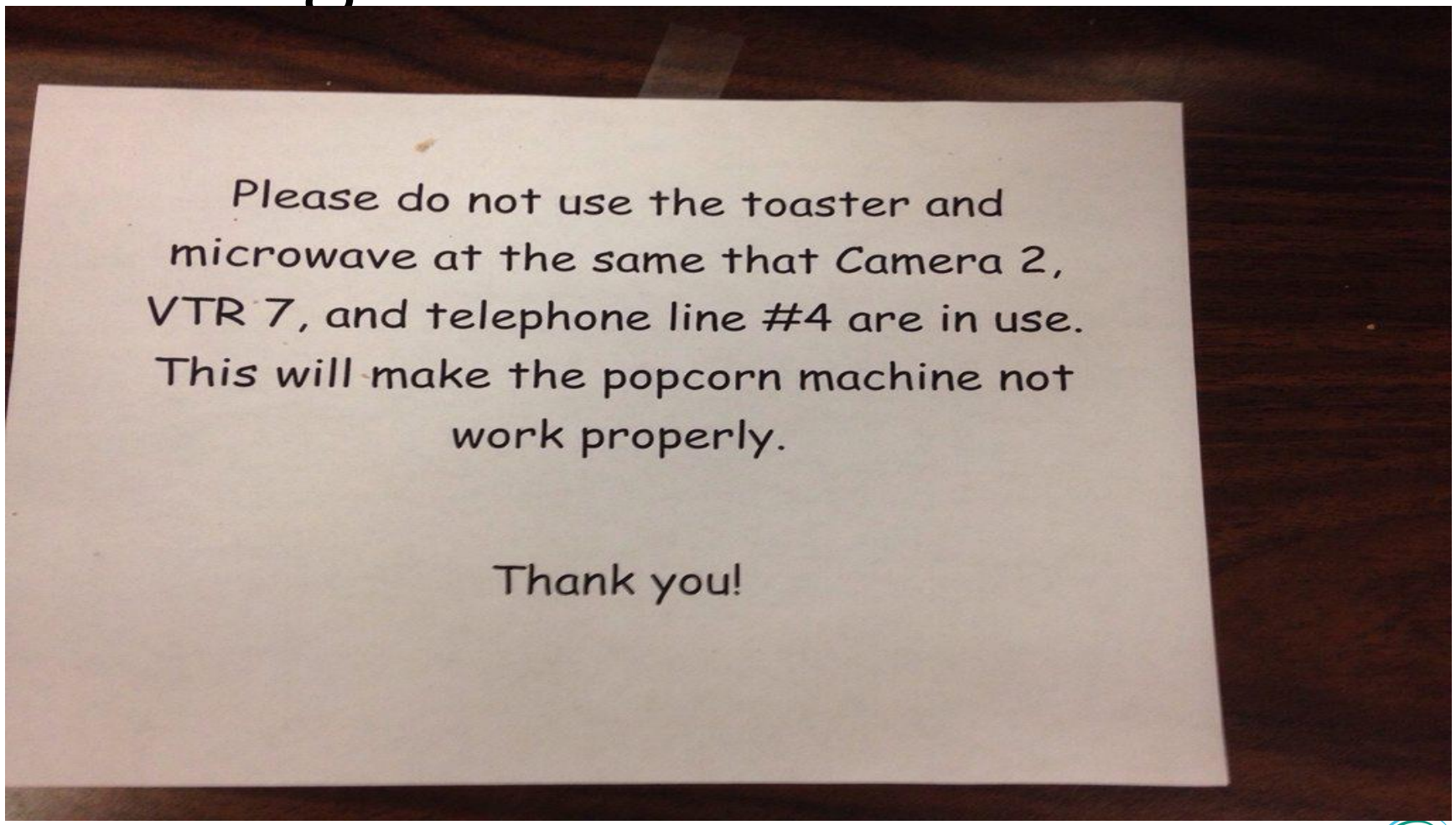
Take Home Points

- The 2 key “take homes” for hardware are:
  - Needs to be ***expandable*** and
  - Needs to be ***supportable***
- ***Needs to be able to seamlessly support data mobility***



I LIKE HOW WE'VE HAD THE INTERNET FOR DECADES, YET "SENDING FILES" IS SOMETHING EARLY ADOPTERS ARE STILL FIGURING OUT HOW TO DO.

DTN Tuning is Art & Science



EPOC
Engagement and Performance Operations Center

DTN Tuning

http://fasterdata.es.net/science-dmz/DTN/tuning

- Defaults are not appropriate for performance.
  - What needs to be tuned:
  - BIOS
  - Firmware
  - Device Drivers
  - Networking
  - File System
  - Application

EPOC
Engagement and Performance Operations Center

DTN Tuning

- Tuning your DTN host is extremely important. We have seen overall IO throughput of a DTN more than double with proper tuning.
- Tuning can be as much art as a science. Due to differences in hardware, its hard to give concrete advice.
- Here are some tuning settings that we have found do make a difference.
- This tutorial assumes you are running a RedHat-based Linux system, but other Unix flavors should have similar tuning knobs.
- Note that you should always use the most recent version of the OS, as performance optimizations for new hardware are added to every release.

EPOC
Engagement and Performance Operations Center

Network Tuning

```
# add to /etc/sysctl.conf
net.core.rmem_max = 67108864
net.core.wmem_max = 67108864
net.ipv4.tcp_rmem = 4096 87380 33554432
net.ipv4.tcp_wmem = 4096 65536 33554432
net.core.netdev_max_backlog = 250000

Add to /etc/rc.local to increase send queue depth
# increase txqueuelen
/sbin/ifconfig eth2 txqueuelen 10000
/sbin/ifconfig eth3 txqueuelen 10000
```

This info (and more) is available on fasterdata, formatted for cut and paste

```
# make sure cubic and htcp are loaded
/sbin/modprobe tcp_htcp
/sbin/modprobe tcp_cubic
# set default to CC alg to htcp
net.ipv4.tcp_congestion_control=htcp

# with the Myricom 10G NIC
# using interrupt coalencing helps a lot:
/usr/sbin/ethtool -C ethN rx-usecs 75
```

Please consider jumbo frames, but don't just blindly turn them on

EPOC
Engagement and Performance Operations Center

I/O Scheduler

- The default Linux scheduler is the "fair" scheduler. For a DTN node, we recommend using the "deadline" scheduler instead.
- To enable deadline scheduling, add "elevator=deadline" to the end of the "kernel' line in your /boot/grub/grub.conf file, similar to this:
- `kernel /vmlinuz-2.6.35.7 ro root=/dev/VolGroup00/LogVol00 rhgb quiet elevator=deadline`

EPOC
Engagement and Performance Operations Center

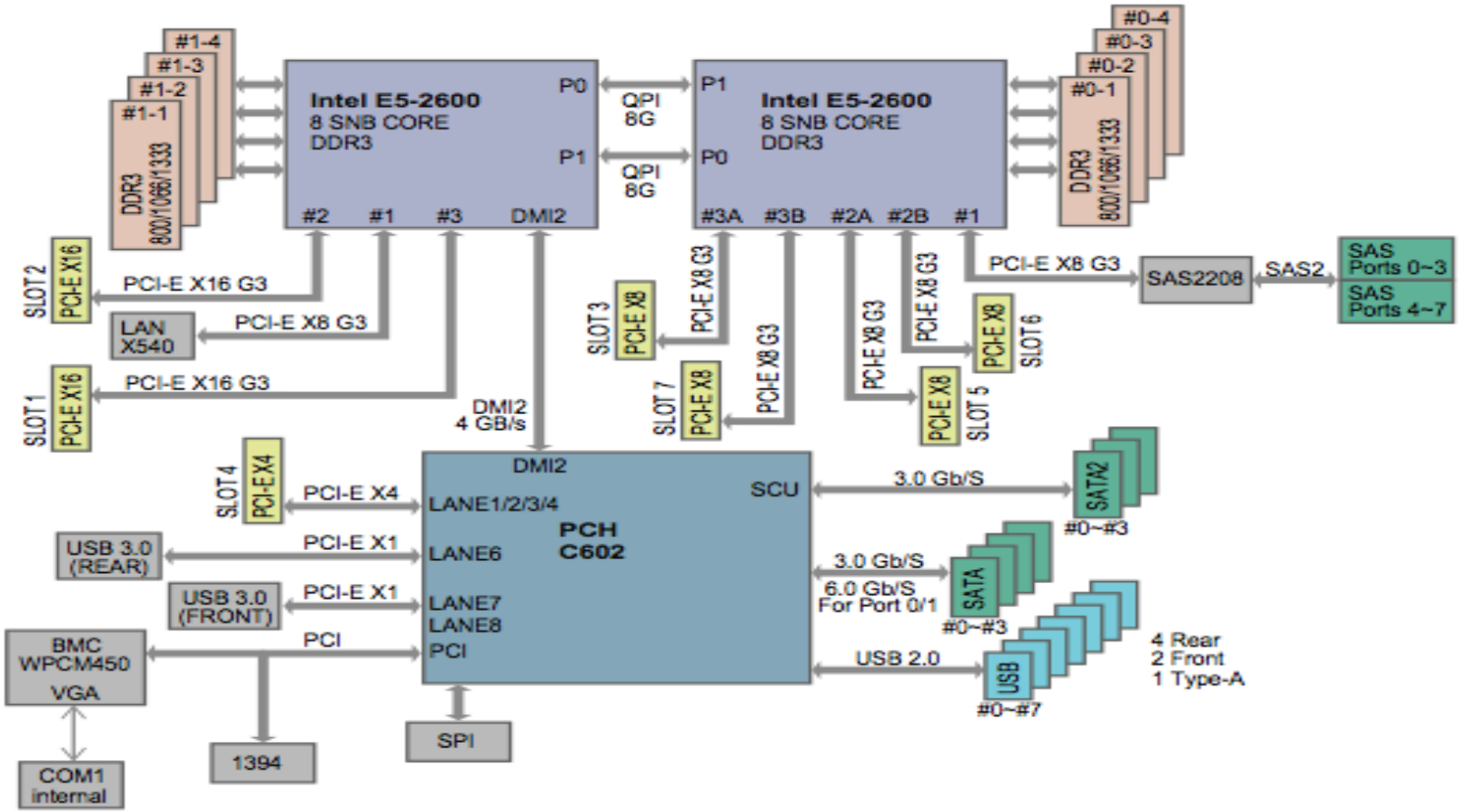
Interrupt Affinity

- Interrupts are triggered by I/O cards (storage, network). High performance means lot of interrupts per second
- Interrupt handlers are executed on a core
  - Interrupt handler is just code – it gets run for every interrupt
  - Cache effects matter (with lots of I/O we're going to run that code a lot)
- Depending on the scheduler, core 0 gets all the interrupts, or interrupts are dispatched in a round-robin fashion among the cores: both are bad for performance:
  - Core 0 get all interrupts: with very fast I/O, the core is overwhelmed and becomes a bottleneck
  - Round-robin dispatch: very likely the core that executes the interrupt handler will not have the code in its L1 cache.
  - Two different I/O channels may end up on the same core.

EPOC
Engagement and Performance Operations Center

Know Your Layout

- Where are your cards?
- Where are your cores?
- Understand what your bindings mean physically
- Performance of a given config can be tightly coupled to physical layout
- Don’t be afraid to experiment



EPOC
Engagement and Performance Operations Center

SSD Issues – Write Sparingly

- Tuning your SSD is more about reliability and longevity than performance
  - Each flash memory cell has a finite lifespan that is determined by the number of "program and erase (P/E)" cycles
  - Without proper tuning, SSD can die within months.
  - ***never do "write" benchmarks on SSD: this will damage your SSD quickly.***
- TRIM
  - Trim informs the SSD when the filesystem no longer needs space
  - Important to prolong the life of SSDs
  - Modern SSD drives and modern OSes should all include TRIM support
  - Only the newest RAID controllers included TRIM support as of late 2012
- Swap
  - To prolong SSD lifespan, do not swap on an SSD
  - In Linux you can control this using the sysctl variable vm.swappiness. E.g.: add this to /etc/sysctl.conf:
    - vm.swappiness=1
    - This tells the kernel to avoid unmapping mapped pages whenever possible.
- Avoid frequent re-writing of files (for example during compiling code from source), use a ramdisk file system (tmpfs) for /tmp /usr/tmp, etc.

EPOC
Engagement and Performance Operations Center

Benchmarking

- Single-threaded sequential file write:
  - `$ dd if=/dev/zero of=/storage/data1/file1 bs=4k count=33554432`
- Single-threaded sequential file read:
  - `$ dd if=/storage/data1/file1 of=/dev/null bs=4k`
- Use more to simulate parallel workload
  - Use `oflag=direct` to disable caching

EPOC
Engagement and Performance Operations Center

Sample Data Transfer Results (2005)

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

Tool	Throughput
scp:	140 Mbps
HPN patched scp:	1.2 Gbps
ftp	1.4 Gbps
GridFTP, 4 streams	5.4 Gbps
GridFTP, 8 streams	6.6 Gbps

  - Note that to get more than 1 Gbps (125 MB/s) disk to disk requires RAID.

EPOC
Engagement and Performance Operations Center

Say NO to SCP (2016)

- Using the right data transfer tool is 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 Tools

- Parallelism is key
  - It is much easier to achieve a given performance level with four parallel connections than one connection
  - Several tools offer parallel transfers
- Latency interaction is critical
  - Wide area data transfers have much higher latency than LAN transfers
  - Many tools and protocols assume a LAN
  - Examples: SCP/SFTP, HPSS mover protocol

EPOC
Engagement and Performance Operations Center

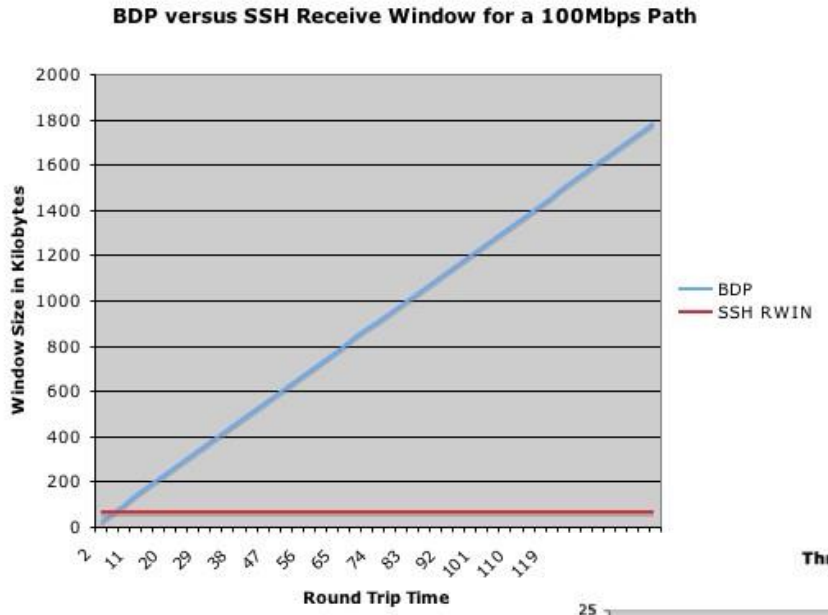
Why Not Use SCP or SFTP?

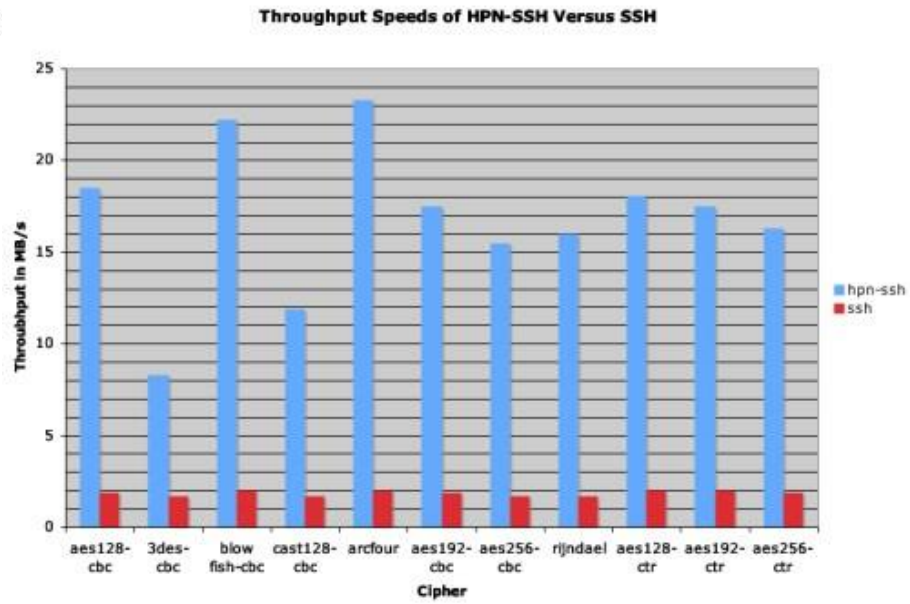
- Pros:
  - Most scientific systems are accessed via OpenSSH
  - SCP/SFTP are therefore installed by default
  - Modern CPUs encrypt and decrypt well enough for small to medium scale transfers
  - Credentials for system access and credentials for data transfer are the same
- Cons:
  - The protocol used by SCP/SFTP has a fundamental flaw that limits WAN performance
  - CPU speed doesn't matter – latency matters
  - Fixed-size buffers reduce performance as latency increases
  - It doesn't matter how easy it is to use SCP and SFTP – they simply do not perform
- Verdict: Do Not Use Without Performance Patches

EPOC
Engagement and Performance Operations Center

A Fix For scp/sftp

- PSC has a patch set that fixes problems with SSH
  - http://www.psc.edu/networking/projects/hpn-ssh/
- Significant performance increase (allows the TCP window to open up if the host is tuned)
- Advantage – this helps rsync too





sftp

- Uses same code as scp, so don't use sftp WAN transfers unless you have installed the HPN patch from PSC
- But even with the patch, SFTP has yet another flow control mechanism
  - By default, sftp limits the total number of outstanding messages to 16 32KB messages.
  - Since each datagram is a distinct message you end up with a 512KB outstanding data limit.
  - You can increase both the number of outstanding messages ('-R') and the size of the message ('-B') from the command line though.
- Sample command for a 128MB window:
  - sftp -R 512 -B 262144 user@host:/path/to/file outfile

EPOC
Engagement and Performance Operations Center

Commercial Data Transfer Tools

- There are several commercial UDP-based tools
  - Aspera: http://www.asperasoft.com/
  - Data Expedition: http://www.dataexpedition.com/
  - TIXstream: http://www.tixeltec.com/tixstream_en.html

- These should all do better than TCP on a lossy path
  - advantage of these tools less clear on an clean path

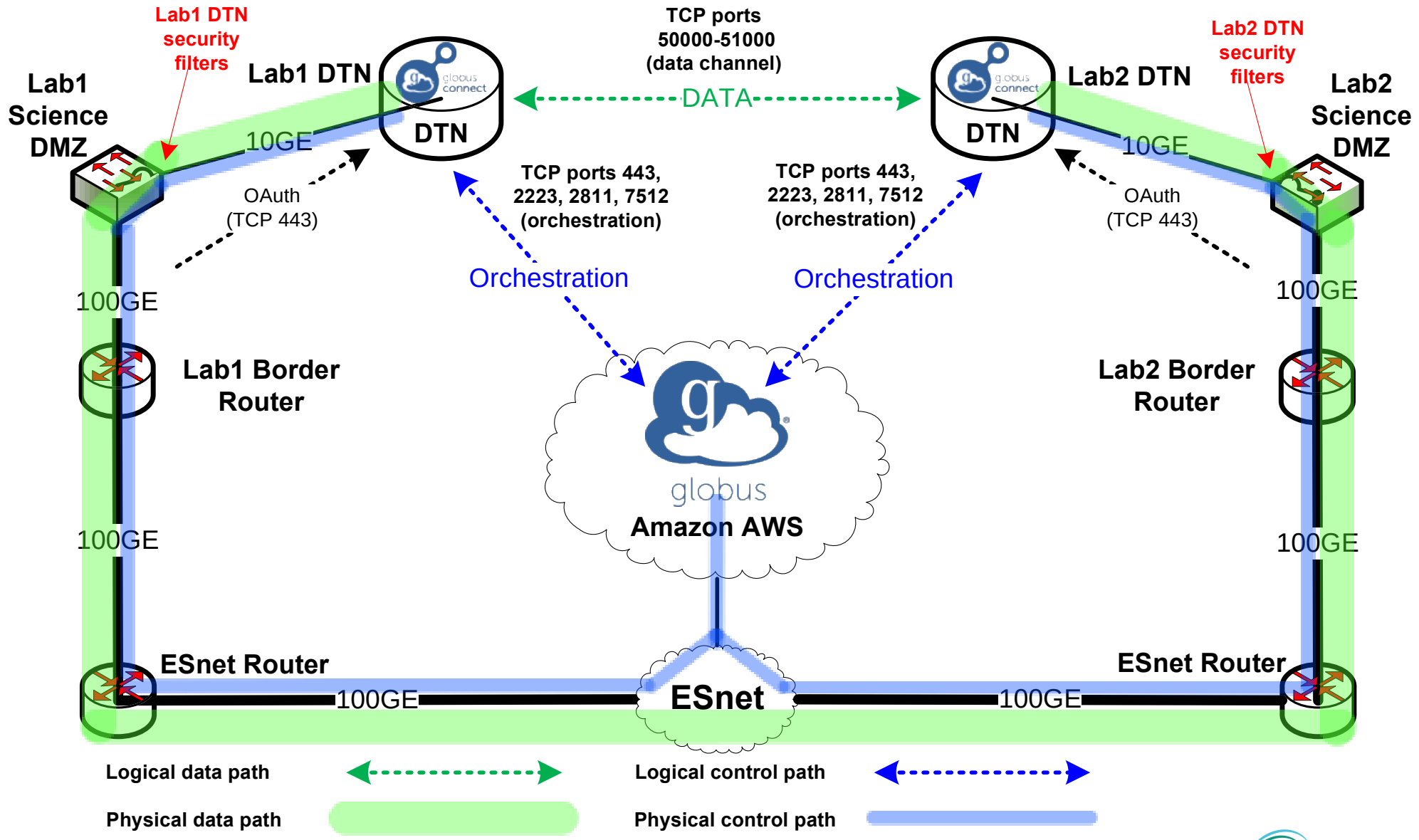
- They all have different, fairly complicated pricing models

EPOC
Engagement and Performance Operations Center

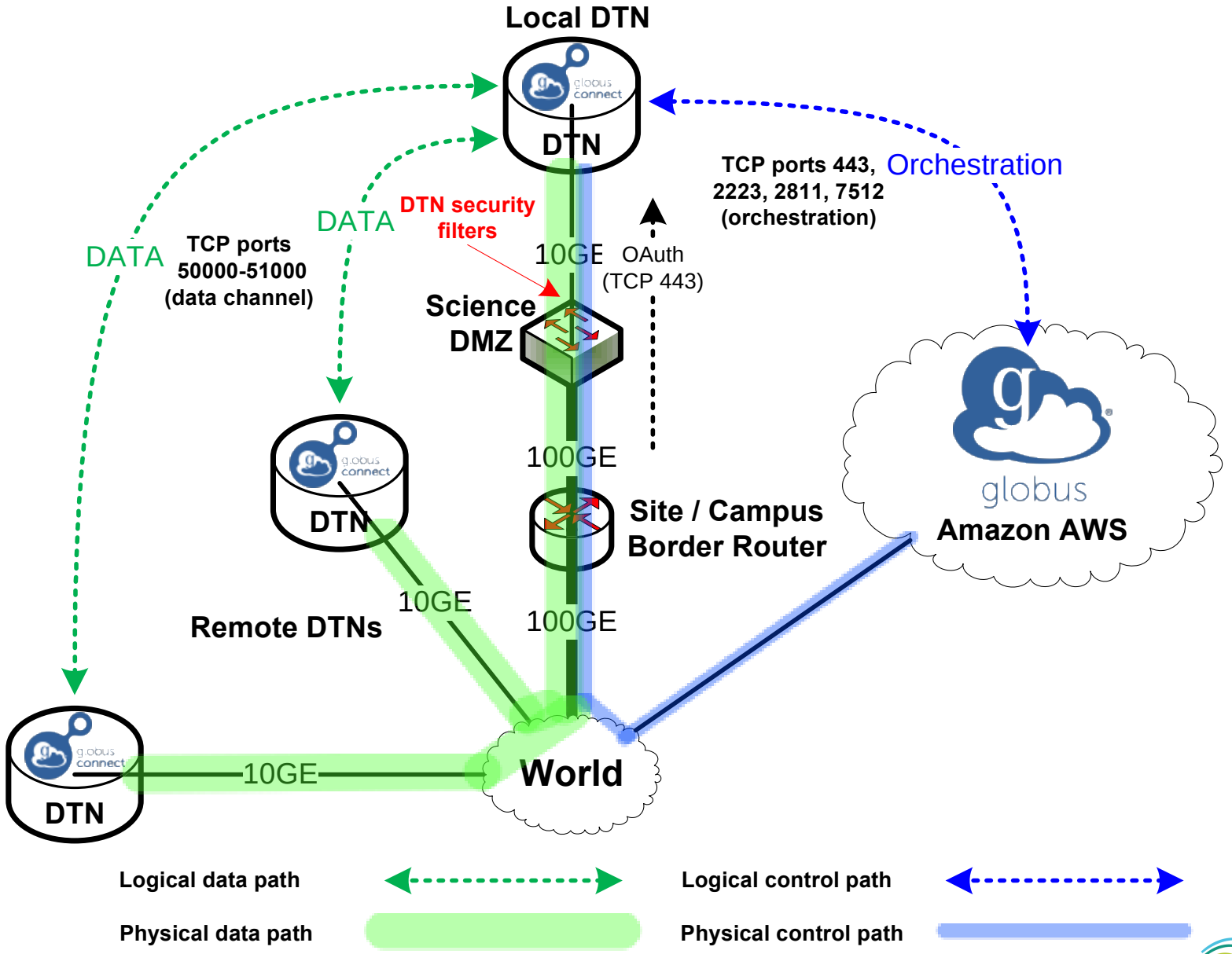
GridFTP

- GridFTP from ANL has features needed to fill the network pipe
  - Buffer Tuning
  - Parallel Streams
- Supports multiple authentication options
  - Anonymous
  - ssh
  - X509
- Ability to define a range of data ports
  - helpful to get through firewalls
- Partnership with ESnet and Globus Online to support Globus Online for use in Science DMZs

EPOC
Engagement and Performance Operations Center



EPOC
Engagement and Performance Operations Center



EPOC
Engagement and Performance Operations Center

http://bit.ly/EPOC-USC-CI

End devices in Science DMZs: DTNs

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

Training Workshop for Network Engineers and Educators on Tools and Protocols for High-Speed Networks
University of South Carolina
July 22-23, 2019

National Science Foundation Award #1826994