EPOC
Engagement and Performance Operations Center

DTN Hardware Elements and Tuning

Ken Miller ken@es.net
EPOC - Performance Chaser
ESnet - Science Engagement Team

ESnet
ENERGY SCIENCES NETWORK

INDIANA UNIVERSITY

National Science Foundation Award #1826994

Cyberinfrastructure for Research Data Management Workshop

Get to know you:

- Quick poll
  - Creates or is responsible for data?
  - Collaborates with data?
  - Uses SneakerNet™ in any part of data movement?
  - Uses dedicated hardware for data movement?

EPOC
Engagement and Performance Operations Center

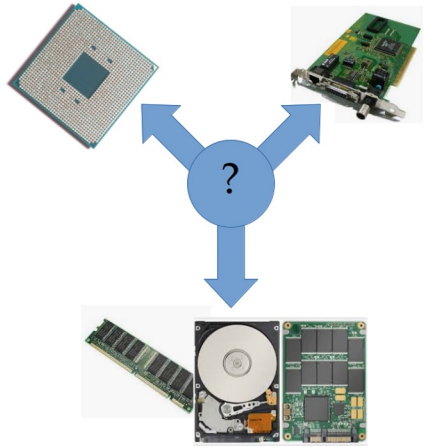
ESnet DTN Scorecard

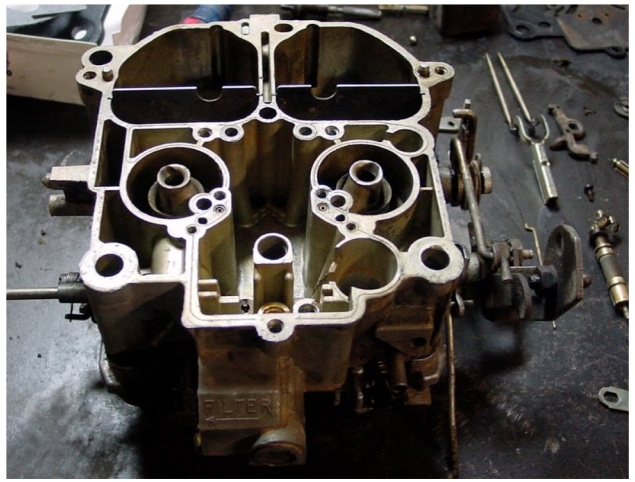
	10G DTN				x10G, 25G, 40G, 100G DTNs			x400G
DTN host Transfer Rates	1/6 PetaScale	1/3 PetaScale	1/2 PetaScale		PetaScale: 1 PB/wk	PetaScale: 1 PB/day		Future ExaScale: 1 XB/month
Data Transfer Volume (Researcher)	1 TB/hr	2 TB/hr	3 TB/hr		5.95 TB/hr	41.67 TB/hr		33.33 PB/day
Network Transfer Rate (Network Admin)	2.22 Gb/s	4.44 Gb/s	6.67 Gb/s		13.23 Gb/s	92.59 Gb/s		3.09 Tb/s
Storage Transfer Rate (Sys/Storage Admin)	277.78 MB/s	555.54 MB/s	833.33 MB/s		1.65 GB/s	11.57 GB/s		385.80 GB/s

EPOC
Engagement and Performance Operations Center

Elements of Data Transfer Nodes (DTNs)

- What and Why
- Elements that build a DTN
- DTN Network Placement
- Tuning to make science go
- Balance to solve problems and hit goals





EPOC
Engagement and Performance Operations Center

What is a DTN

- Simply put, a **Data Transfer Node**
  - Serves *science* first
  - Designed, and tuned, for data movement
  - Solve a specific problem first, then have fun
  - Extremely fun to tune (see perfSonar talks!)
  - Don’t get lost in the speeds and feeds.

EPOC
Engagement and Performance Operations Center

Why have a DTN

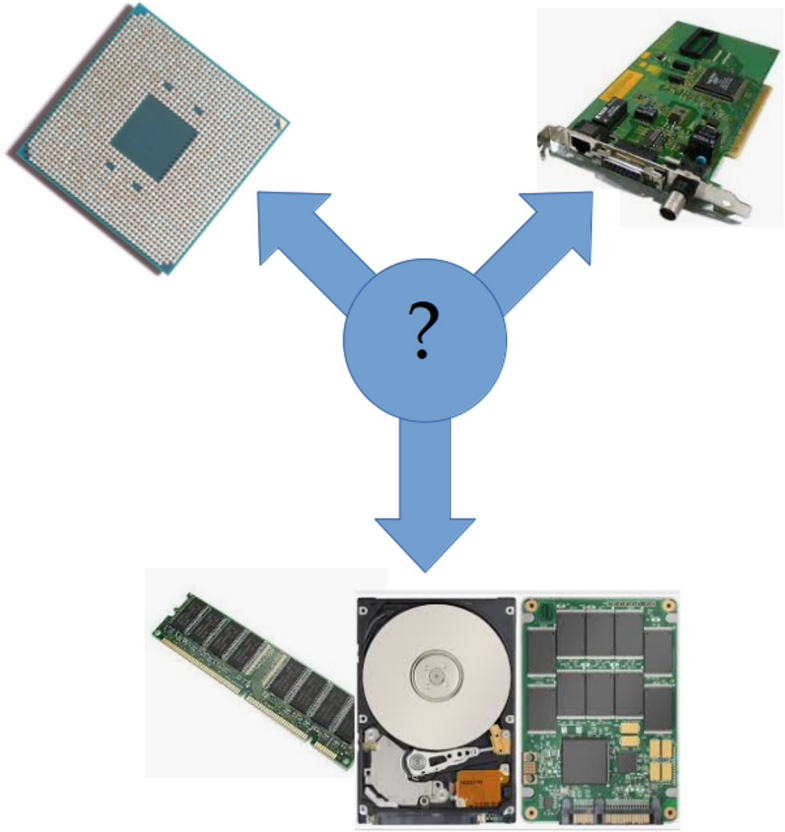
- Purpose built for a closed-set problem definition
  - Non-compete with other services or resources
  - Excellent security posture
  - Time on the wire is time not doing science!
  - Easily translate (and ship) instrument, or cluster attached storage.
  - Can be scheduled

EPOC
Engagement and Performance Operations Center

Elements that build a DTN

- Problem definitions help focus design
- Find the Balance.

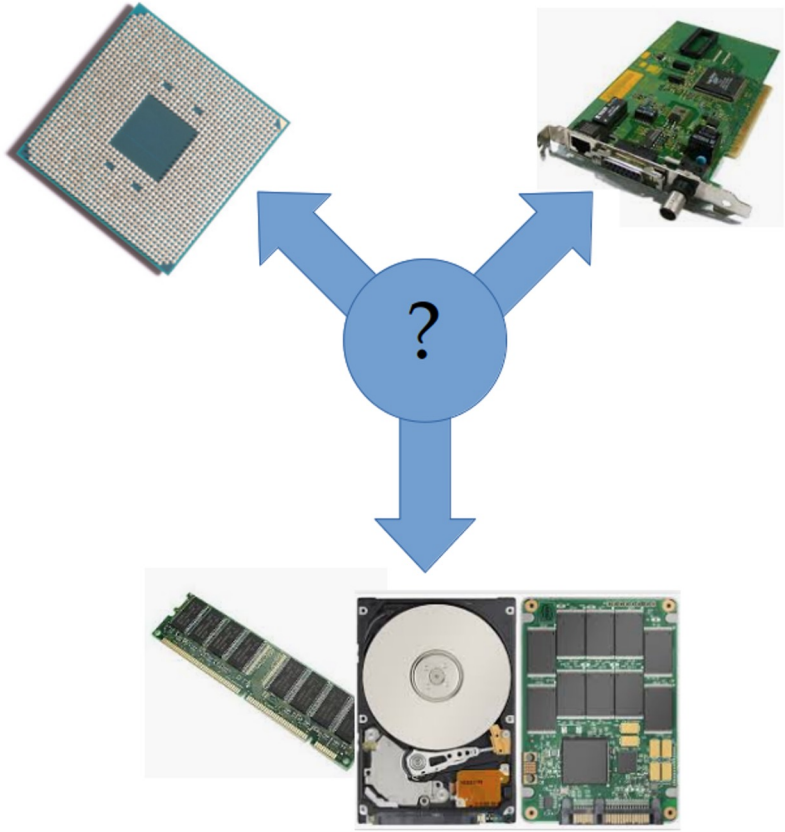
- Three major adjustments:
  - CPU
  - Storage
  - Network



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Problem definition.

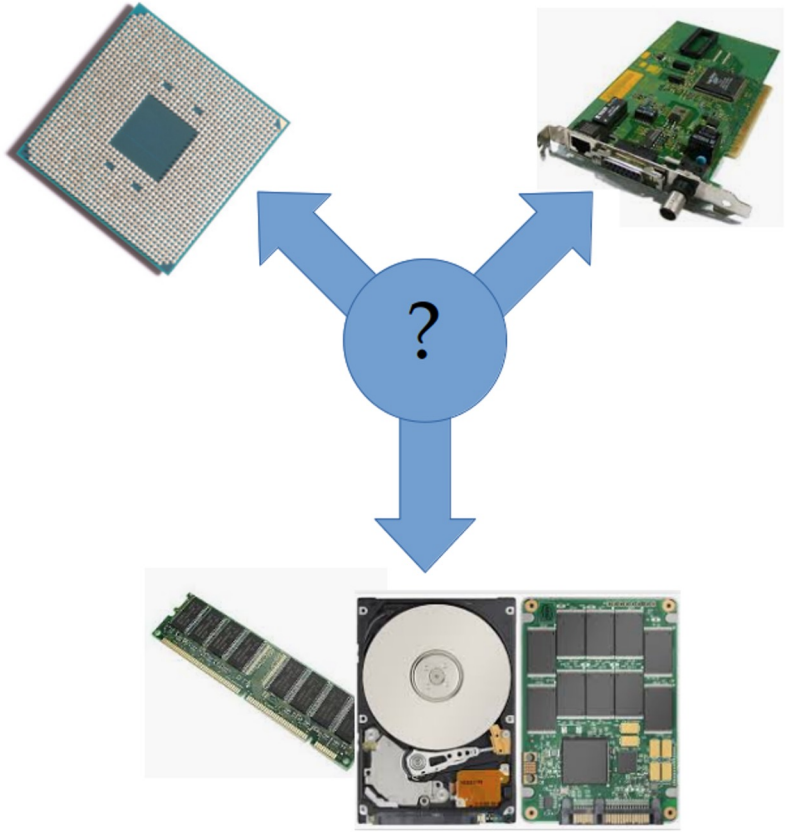
- Problem definitions help focus design
  - Collaboration between *ACME portable hole company* and *RoadRunner High Energy Physics* is slow because USB attached drives are used for data movement.
  - Risky, USB transfer speeds, single copy.
  - What else?



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Problem definition.

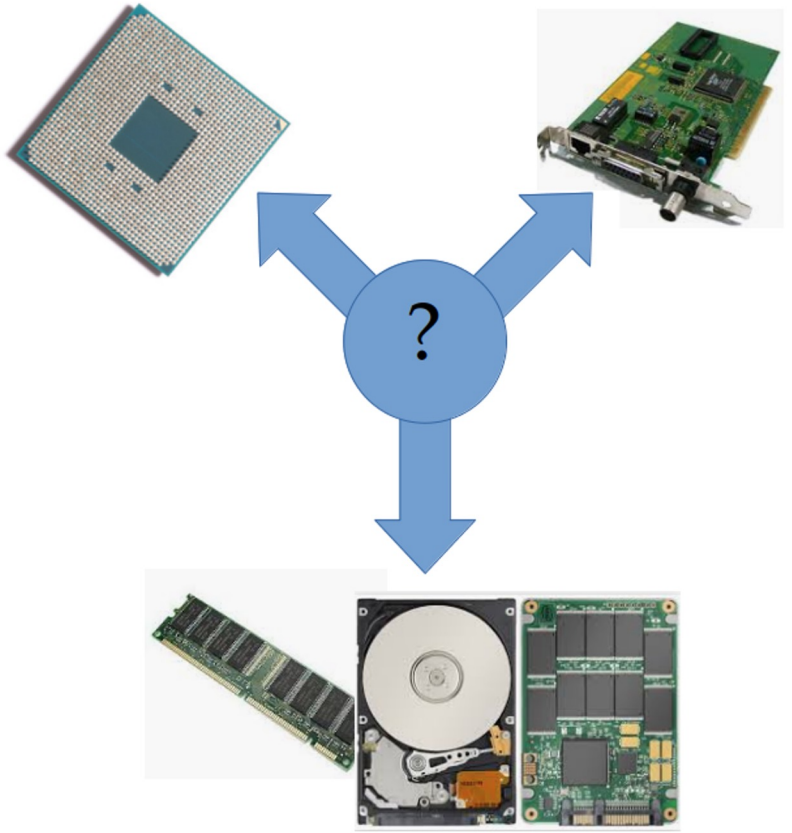
- Problem definitions help focus design
  - Collaboration with our team and *Spacely-Sprockets* is slow with our current 100Mbps network connection.
  - Simple bottle neck or can the rest of the path keep up.
  - Another statement once resolved?



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Problem definition.

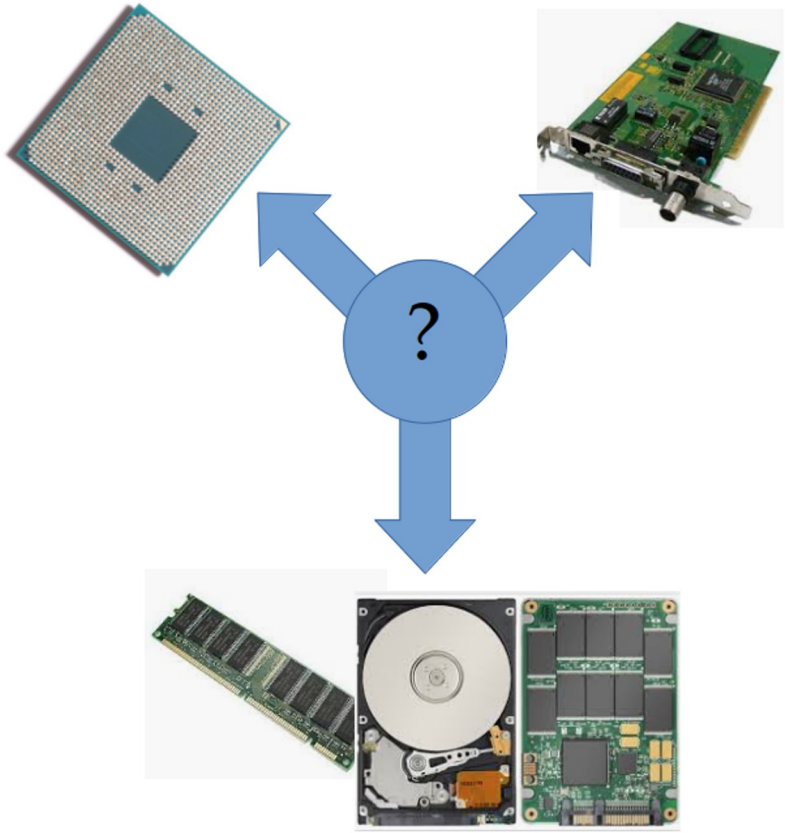
- Problem definitions help focus design
  - Our researchers lost the primary copy of our data because we had a drive fail on our RAID0 and we had to re-transfer 2TiB of unsorted files. Again…
  - (Be honest with the workflow)



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Problem definition.

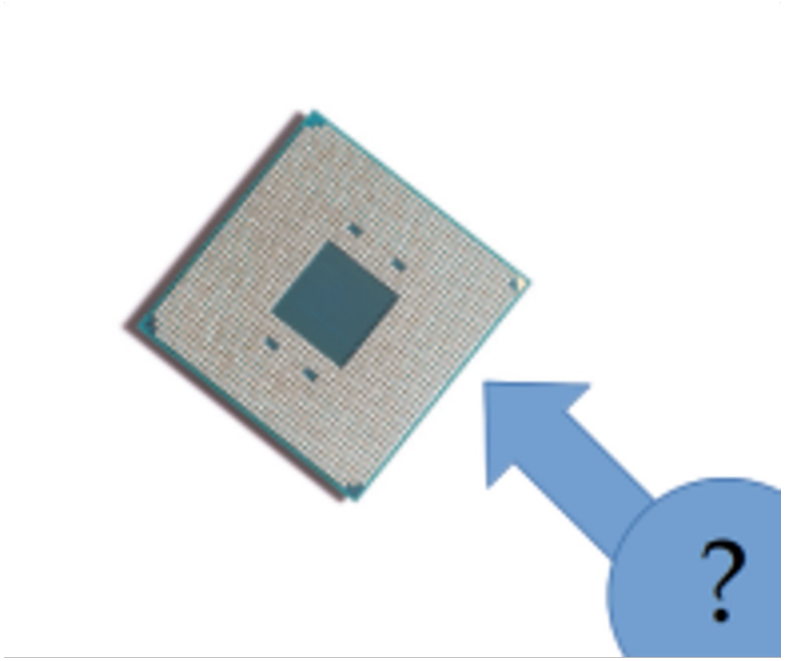
- Problem definitions help focus design
  - Iterate
  - Collaborate
  - Improve
  - Hit the goals by solving the problems



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - CPU

- DTN’s three major adjustments:
  - **CPU**
  - Storage
  - Network

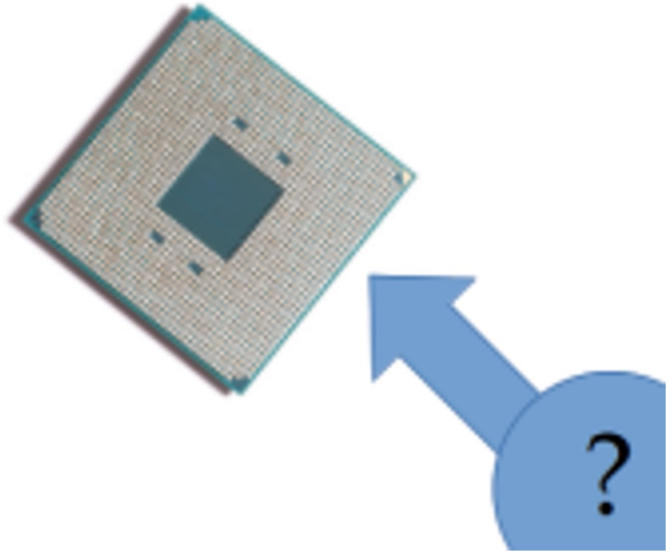


EPOC
Engagement and Performance Operations Center

Elements that build a DTN - CPU

- CPU - Central Processing Unit

- The CPU Balancing act contains:
  - Sockets
  - Cores
  - Clock
  - Threads.

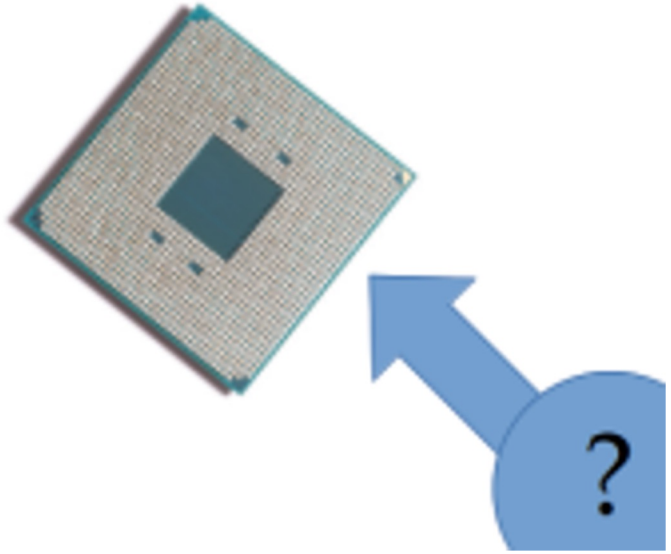


EPOC
Engagement and Performance Operations Center

Elements that build a DTN - CPU

- CPU - Central Processing Unit

- Sockets:
  - Actual CPU processor that one can socket into a server.
  - More may not always be better.
  - Many to select from!

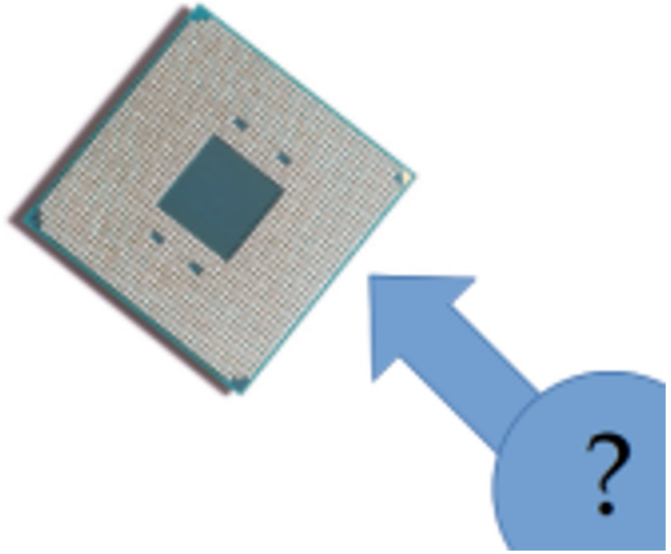


EPOC
Engagement and Performance Operations Center

Elements that build a DTN - CPU

- CPU - Central Processing Unit

- Cores:
  - Package within the CPU socket that can be dedicated by the Operating System as a logical processing unit.
    - NOTE:
      - Simultaneous multithreading can increase hardware threads presenting additional cores (logical doubling of physical cores)

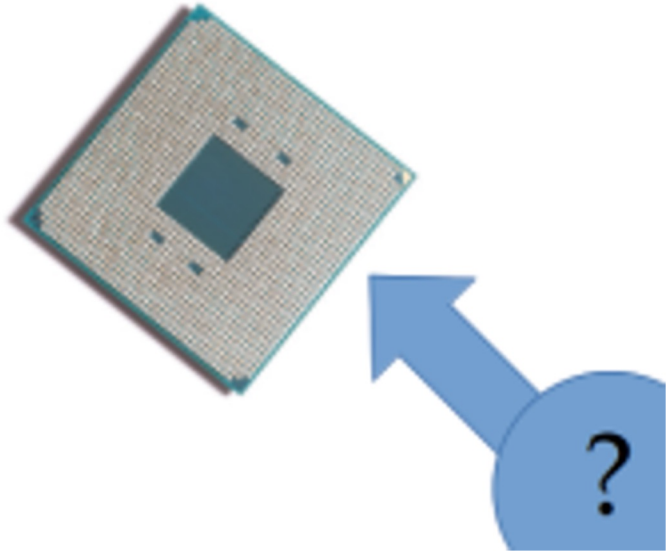


EPOC
Engagement and Performance Operations Center

Elements that build a DTN - CPU

- CPU - Central Processing Unit

- Clock(s):
  - Digital cadence of when state change or calculations happen.
  - Bursts of higher clock rates
  - Tunable
  - Trade off of power vs clock
  - Select higher clock rate over CPU cores

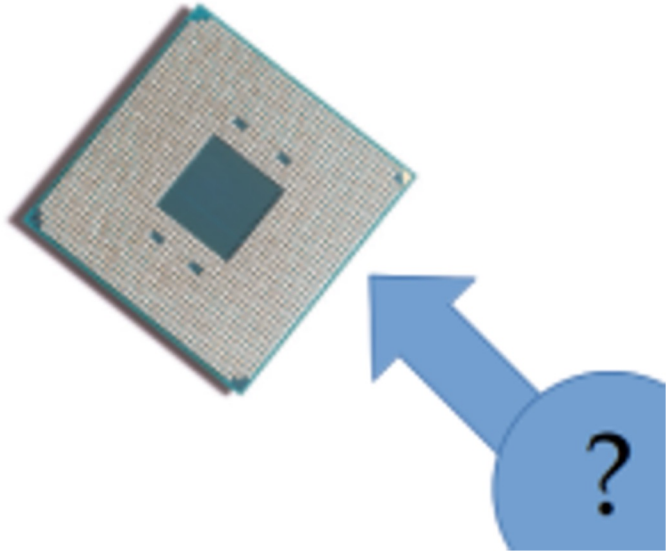


EPOC
Engagement and Performance Operations Center

Elements that build a DTN - CPU

- CPU - Central Processing Unit

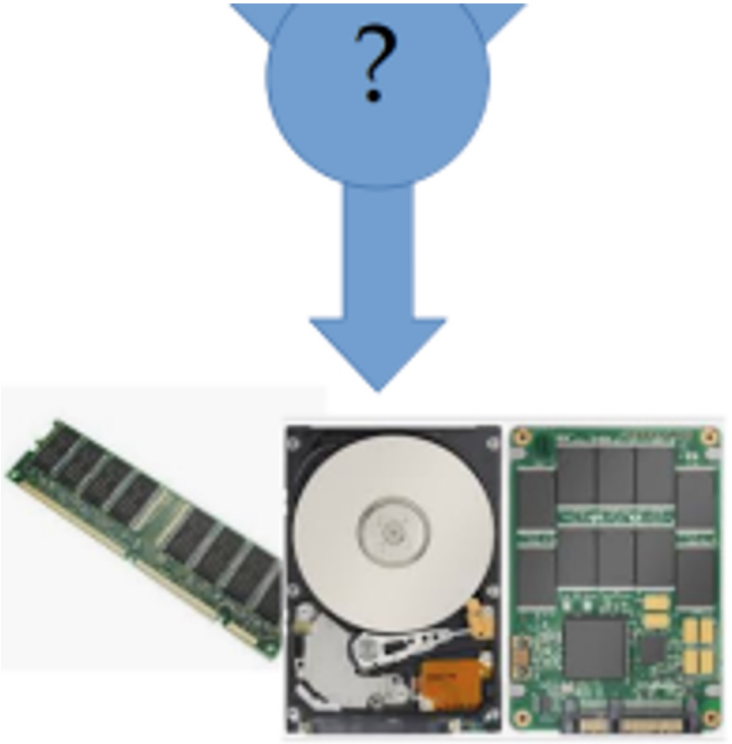
- Threads:
  - Dedicated pipelines for ‘work’.
  - Direct interaction with rest of the hardware.



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - STORAGE

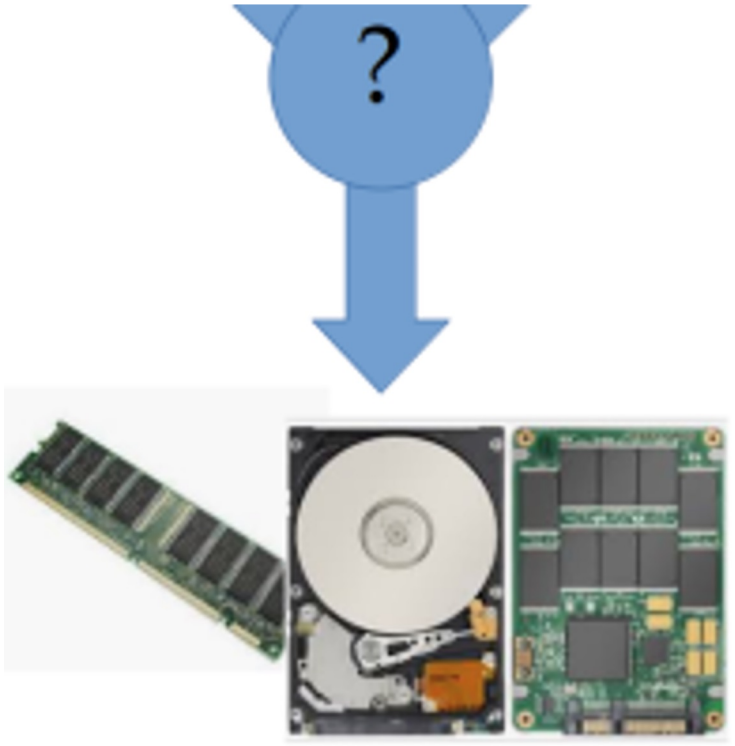
- DTN’s three major adjustments:
  - CPU
  - **Storage**
  - Network



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Storage

- Balance follows the familiar “only pick two”
  - Speed
  - Capacity
  - Cost



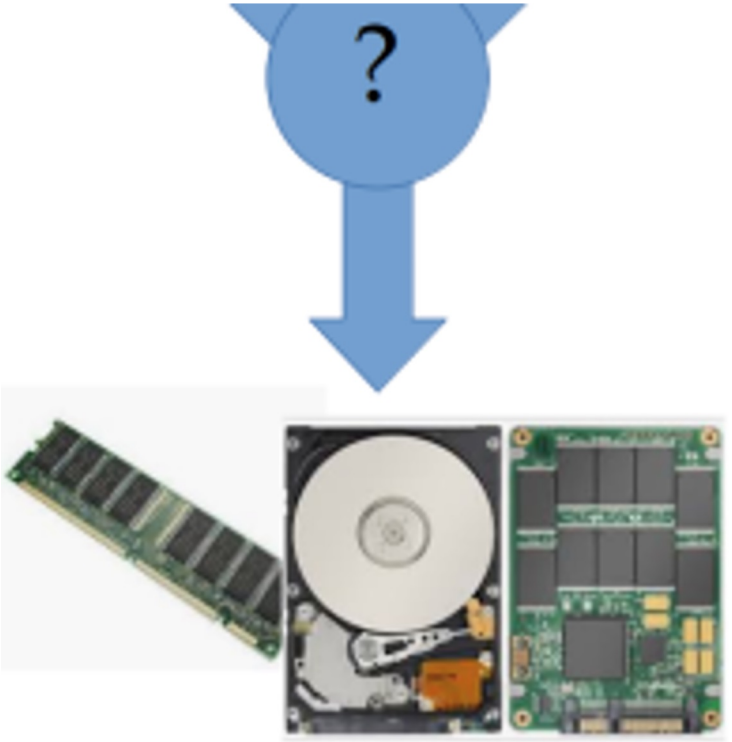
- RAM ***is*** part of the storage picture
  - are your channels fully populated? :-)
  - Keep the threads busy!
- Spinning rust still has place

EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Storage (RAID)

- Problem definition
  - **Speed | Capacity |Cost**

- Hardware RAID
  - Speed - Redundant with speed
    - RAID 10 (read and write speedups)
  - Capacity - Often Plug and play
  - Cost - Can be spicy (nvme)



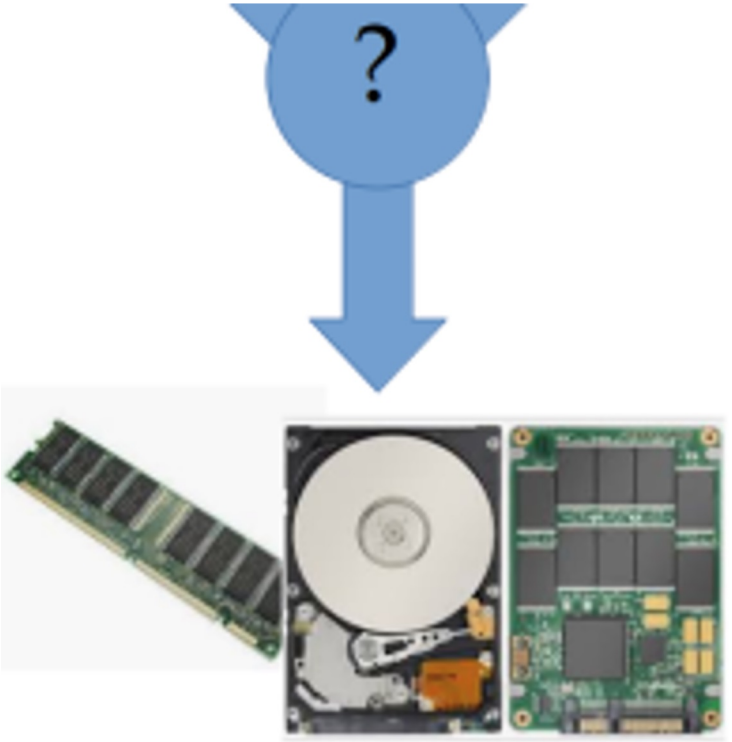
EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Storage (Filesystems)

- Problem definition
  - **Speed | Capacity |Cost**

Software RAID:
ZFS / mdadm / LVM:

- Redundant
- Plug and play**
- Performance, and capacity
- RAM truly is storage
- CPU / RAM consumption cost

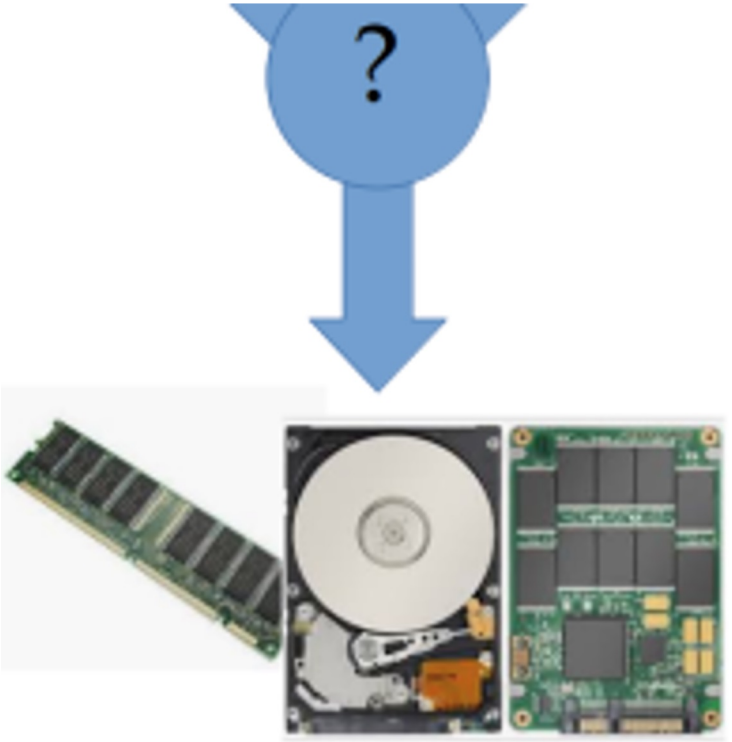


EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Storage (appliances)

- Problem definition
  - **Speed |Capacity| Cost**

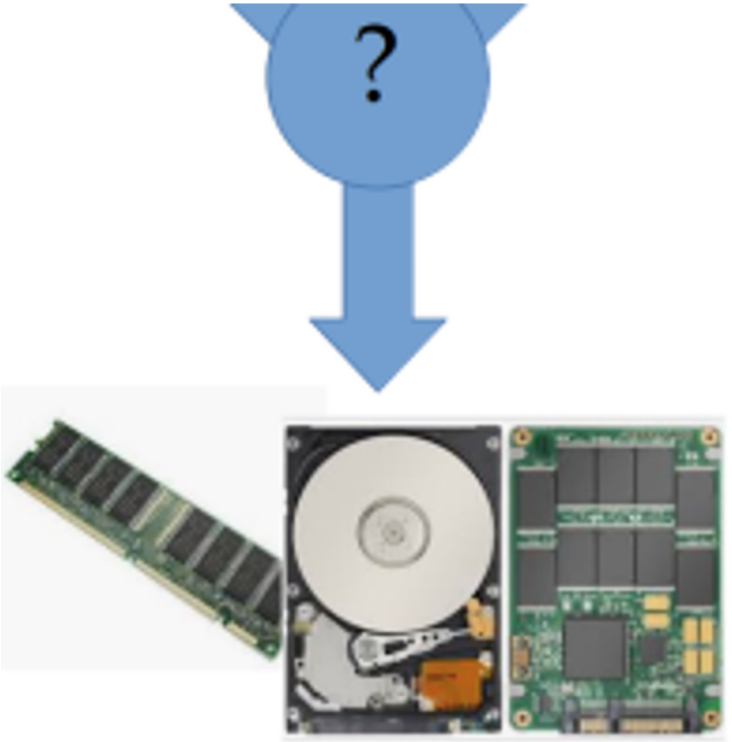
- Purpose built appliances
  - Redundant
  - Plug and play
  - Performance, and capacity
  - Painted into a corner



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Network

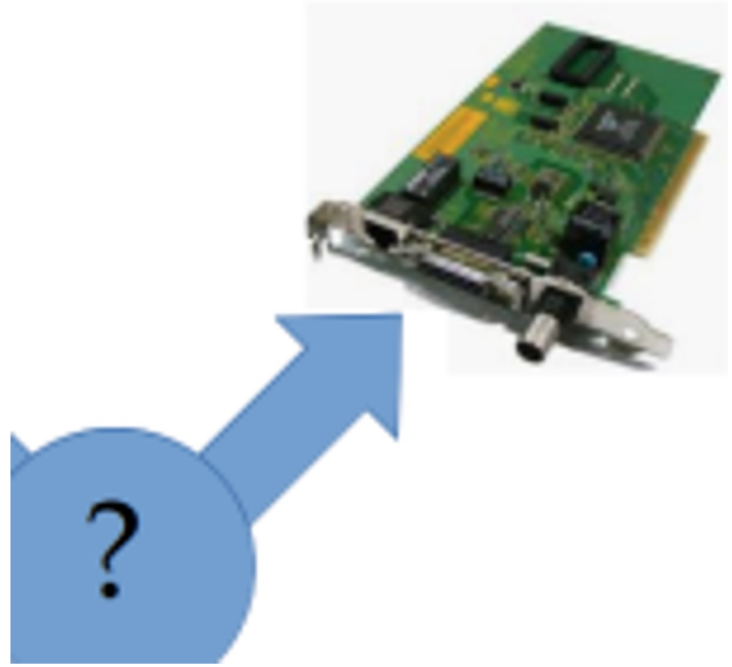
- DTN’s three major adjustments:
  - CPU
  - Storage
  - **Network**



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Network

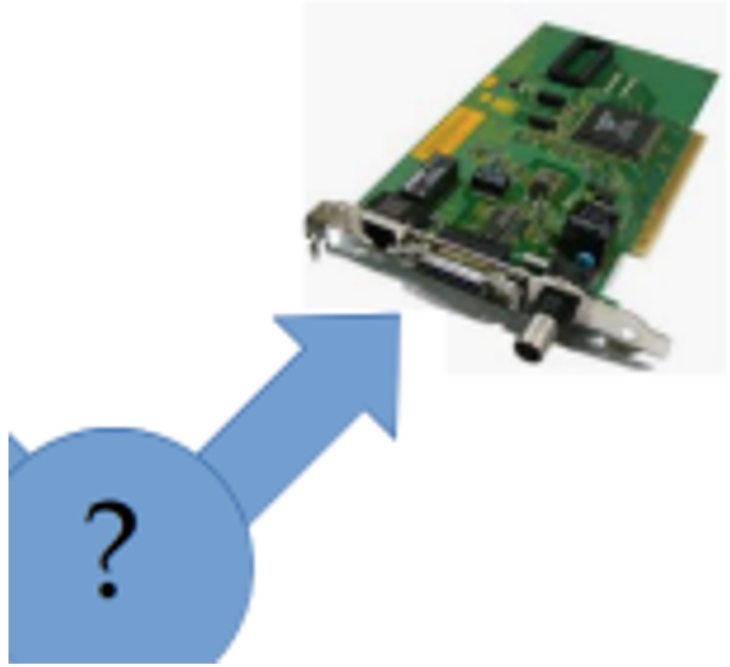
- Does one truly need 100G DTN?!
  - Vendor Options
  - Speeds and feeds lots of options
    - e.g. Look at the types of fiber
  - LAN / WAN scale
  - Instrument to/from cluster(s)
  - Media
    - Fiber / DAC / Copper



EPOC
Engagement and Performance Operations Center

Elements that build a DTN - Network

- On a Science DMZ
  - Scalable well known and referenced pattern
  - Get science to and from peers
  - (see perfSonar / DME talks!)

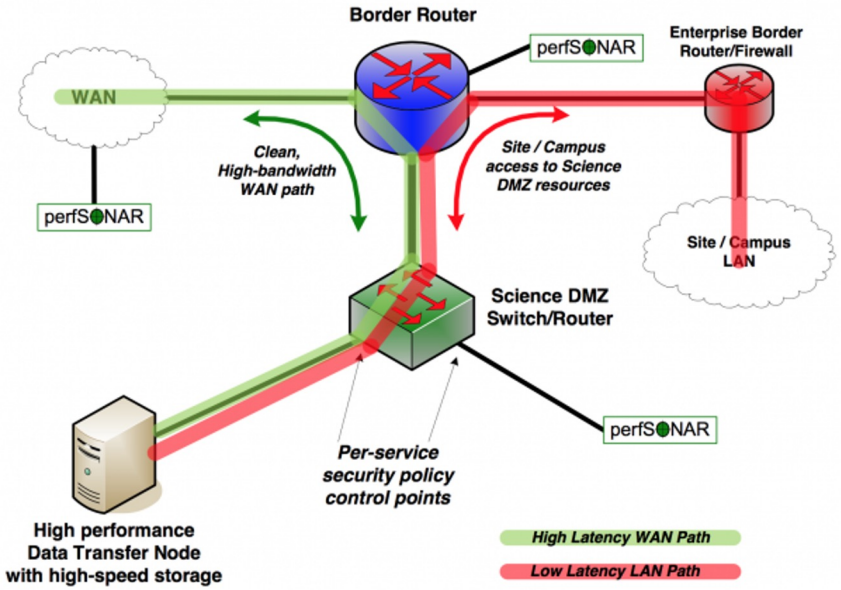


EPOC
Engagement and Performance Operations Center

So where do we put our DTN?

- Problem statement and what is being solved
  - Sense of scale
  - RE: DME talks:
    1TiB a day / an hour / more

- Example Possibilities
  - Elegant case:
    - data to the fiber (rack scale).
  - Complex cases
    - Cluster based file systems scheduled I/O to clusters of DTNs to Science DMZ
    - Long haul global scale (science knows no bounds)

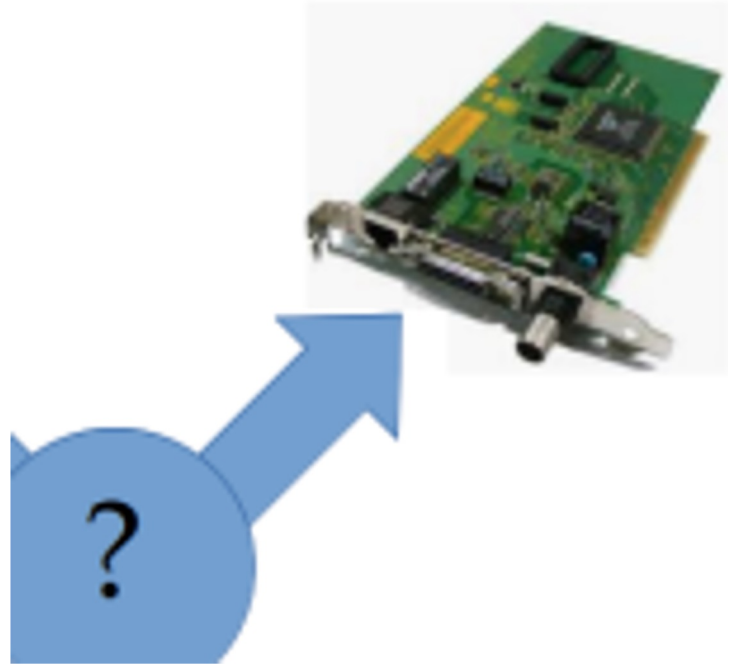


EPOC
Engagement and Performance Operations Center

So where do we put our DTN? (Elegant case)

- Elegant case:
  - data direct to fiber.

  - Not always so elegant:
    - Cadence of transfers
    - Time for data validation
      - (many elegant software solutions here.)
    - Simple flows can take strange paths!

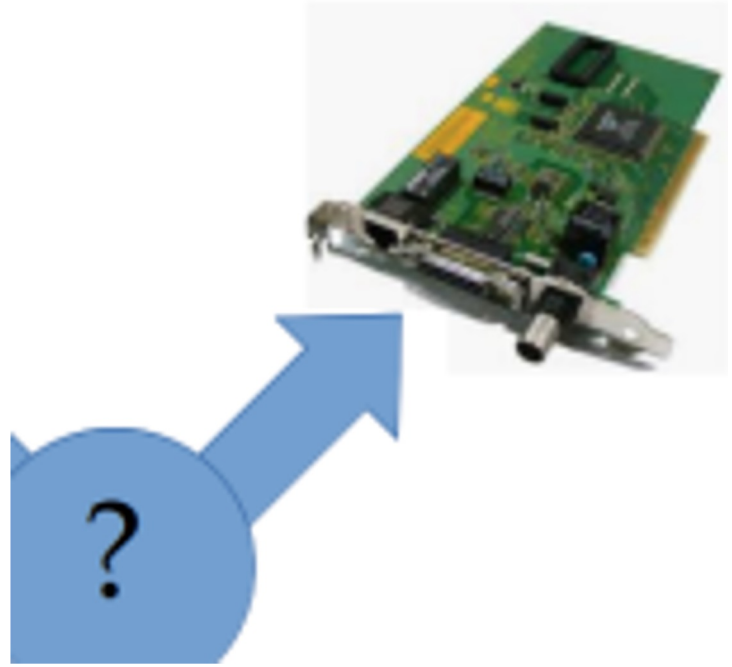


EPOC
Engagement and Performance Operations Center

So where do we put our DTN? (Elegant case)

- Elegant case:
  - data direct to fiber.

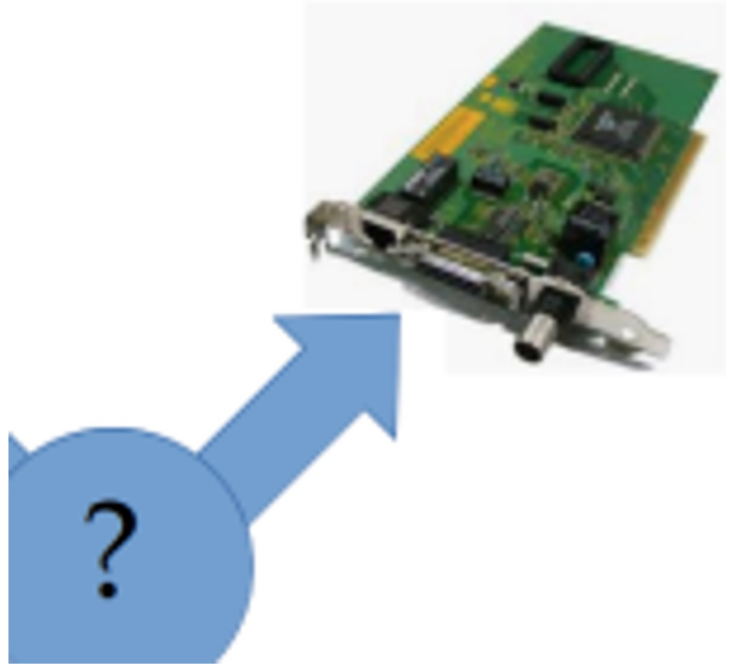
  - Not always so elegant:
    - Capacity woes
      - Network
      - Storage
      - CPU



EPOC
Engagement and Performance Operations Center

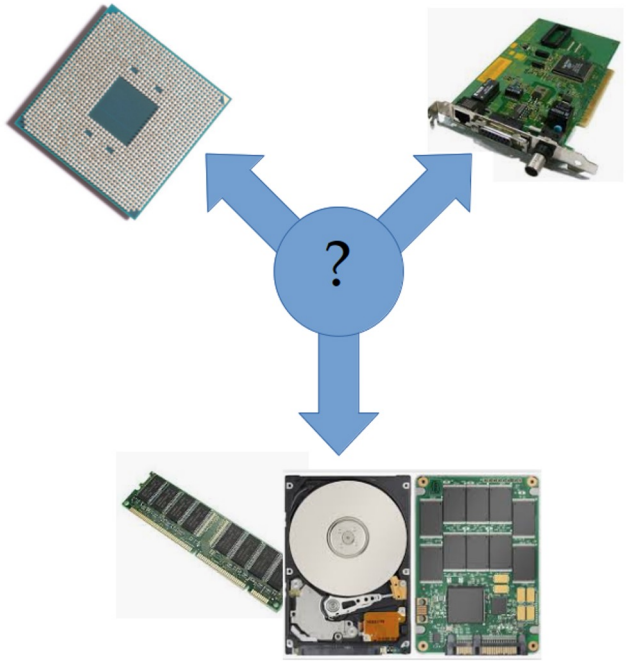
So where do we put our DTN? (Complex case)

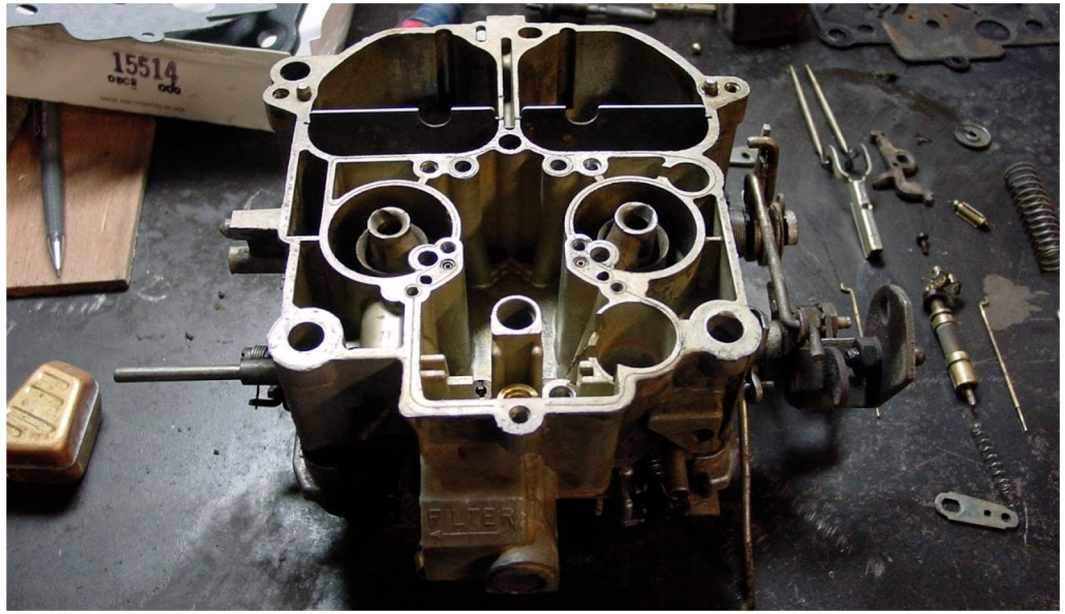
- Complex case:
  - Our instrument to WAN and beyond!
  - Sometimes elegant:
    - Cluster or instrument schedulers
    - Just another filesystem
  - Capacity scales up too:
    - Let’s be excellent to each other
    - Communicate with our peers
    - Calls for assistance are great



EPOC
Engagement and Performance Operations Center

Tuning a DTN





EPOC
Engagement and Performance Operations Center

High Level Scope

- Linux network performance tuning scope:
  - Existing hardware
  - Minor changes
  - Measure the results

EPOC
Engagement and Performance Operations Center

Bespoke Tuning

- Know (reasonable) desired results to guide the tuning:
  - Measure.
  - Coarse -> Fine tuning steps.
  - One Change at a Time.
  - Divide the problem for what next.
  - Reset and do it again.
  - Measure.
  - Performance improvement didn’t happen if you didn’t document or share.

EPOC
Engagement and Performance Operations Center

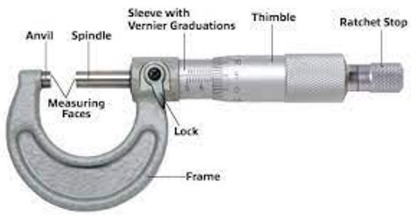
Measure - Capture trends.

- Raw data is ok
  - Look back at it, refine a copy later.
- Command line logging:
  - “ >> logfile.txt ”
  - “ | `tee -a logfile.txt` “
    - Works great for raw collection.
    - Easily scriptable.
    - Insanely low effort.
- What do I think this adjustment will do?
  - What did it really do?
- Trends
  - Bad results can be *good*.
  - Just make sure why is answered.

EPOC
Engagement and Performance Operations Center

Coarse to fine adjustments.

“ >> “



EPOC
Engagement and Performance Operations Center

One Change At A Time.

- Make it easy:
  - Focus on coarse and easy first (80 / 20 rule).
    - Four unknowns need four equations.
    - Stick with one equation.
  - Measurement results reasonable?

EPOC
Engagement and Performance Operations Center

One Change At A Time (Suggested scope: DME).

- Adjustments (Coarse to Fine adjustments):
  - Patches and Updates
  - Path
    - MTU / Buffers
  - Kernel Parameters
    - congestion control
      - BBR anyone?
    - memory
      - sysctl's net.*
    - qlen
    - ring buffers

EPOC
Engagement and Performance Operations Center

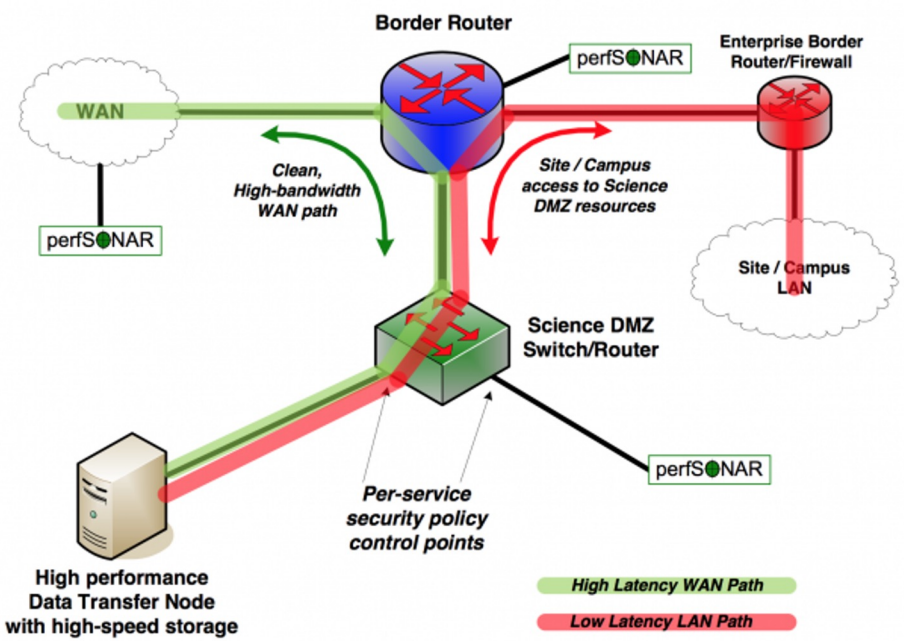
Patches and Updates

- Well patched and maintained systems purr (don’t forget firmware).
  - Flaws fixed
    - Updated modules
    - Newer kernel support features
      - (BBR vs RENO :-))
  - Collaboration
    - Trivial to reset problem
    - Easier to reproduce results

EPOC
Engagement and Performance Operations Center

Path (rocky road or smooth highway)

- What is out there?
- Where are you going?
  - (rack / lan / wan / farther)
- What is in the way?
- Can EPOC help?
  - (shameless plug)



EPOC
Engagement and Performance Operations Center

Ring Buffers

- What are ring buffers?
  - ethtool {-g | -k }
- Why would we mess with them?
  - Let's make the NIC work for us.
  - Turn off ‘extras’
  - Turn on performance

EPOC
Engagement and Performance Operations Center

Ring Buffers (-g | -G)

- Have your interrupt ready to work with a queue.
- ethtool -g <interface here>
  - read current settings
- ethtool -G < interface here>
  - Drop the clutch!
  - Defaults are for compatibility


```
root@tars:~# ethtool -g enp9s0
Ring parameters for enp9s0:
Pre-set maximums:
RX:                 8184
RX Mini:  n/a
RX Jumbo: n/a
TX:                 8184
Current hardware settings:
RX:                 2048
RX Mini:  n/a
RX Jumbo: n/a
TX:                 4096
```

EPOC
Engagement and Performance Operations Center

Ring Buffers (-k | -K)

- Flip the switches on the NIC feature set.
  - Defaults again are for compatibility or even other OS performance optimizations.
- ethtool -k < interface here>
  - read current configs
- ethtool -K < interface here>
  - to offload or not to offload?
  - It depends...

```
root@tars:~# ethtool -k enp9s0
Features for enp9s0:
    …. SNIP ….
generic-segmentation-offload: on
generic-receive-offload: on
large-receive-offload: on
rx-vlan-offload: on
tx-vlan-offload: on
ntuple-filters: on
receive-hashing: on
    …. SNIP ….
```

EPOC
Engagement and Performance Operations Center

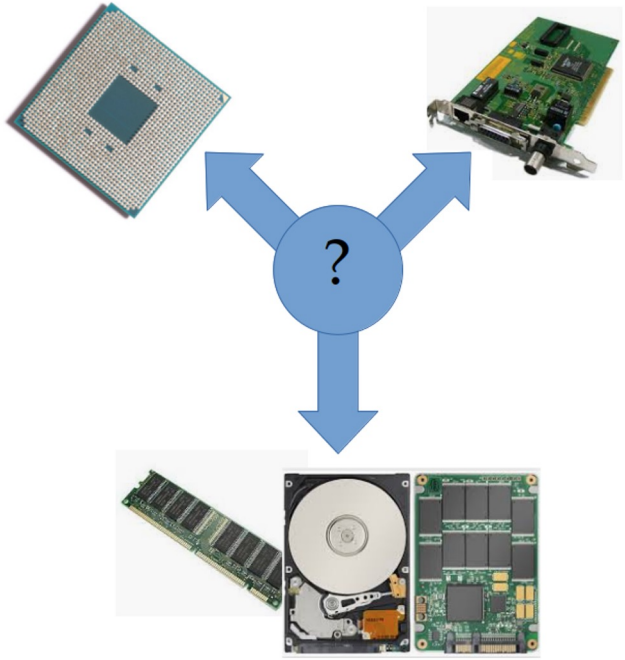
Rough (non-persistent) sysctl script as a toy:

```
#!/bin/bash

sysctl -w net.ipv4.tcp_mtu_probing=1
sysctl -w net.ipv4.tcp_sack=1
sysctl -w net.ipv4.tcp_tw_reuse=1

sysctl -w net.core.default_qdisc=fq
sysctl -w net.ipv4.tcp_congestion_control=bbr

sysctl -w net.core.rmem_max=134217728
sysctl -w net.ipv4.tcp_rmem="20480 12582912 67108864"
sysctl -w net.core.wmem_max=134217728
sysctl -w net.ipv4.tcp_wmem="20480 12582912 67108864"

sysctl -w net.ipv4.udp_mem="20480 12582912 67108864"
sysctl -w net.ipv4.udp_rmem_min=16384
sysctl -w net.ipv4.udp_wmem_min=16384

sysctl -w net.core.netdev_max_backlog=65536
sysctl -w net.ipv4.tcp_fin_timeout=15
sysctl -w net.ipv4.tcp_max_syn_backlog=10240

sysctl -w net.ipv4.tcp_low_latency=1
ip link set qlen 8333 eth0  ##<<<## CHANGE TO YOUR ENVIRONMENT
```

EPOC
Engagement and Performance Operations Center

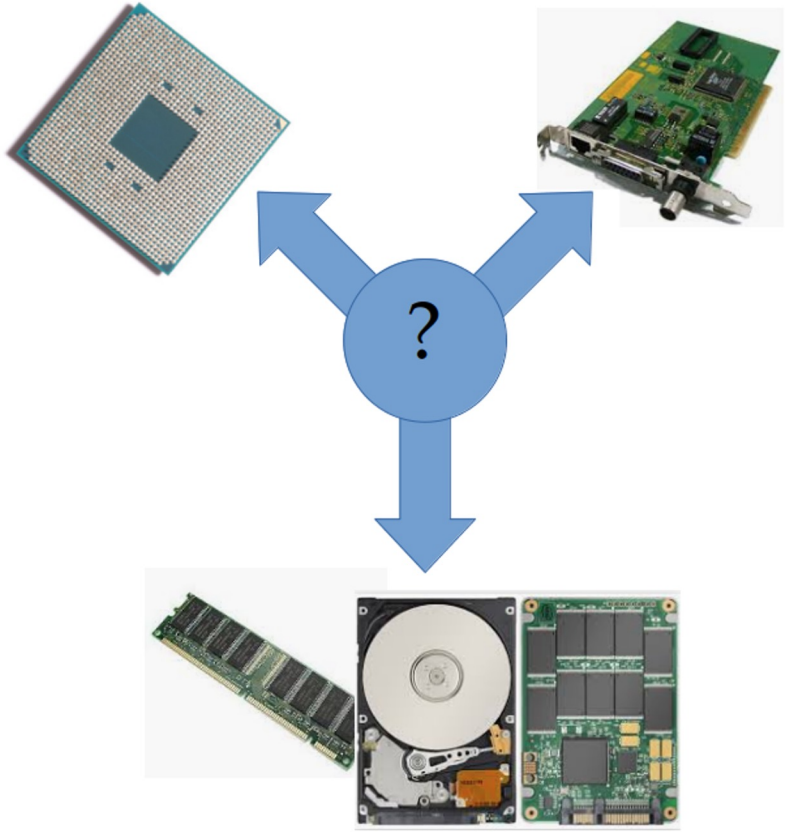
Tune our DTN

- Find the balance
- Problem statements help focus design



Tuning one element at a time. (Network)

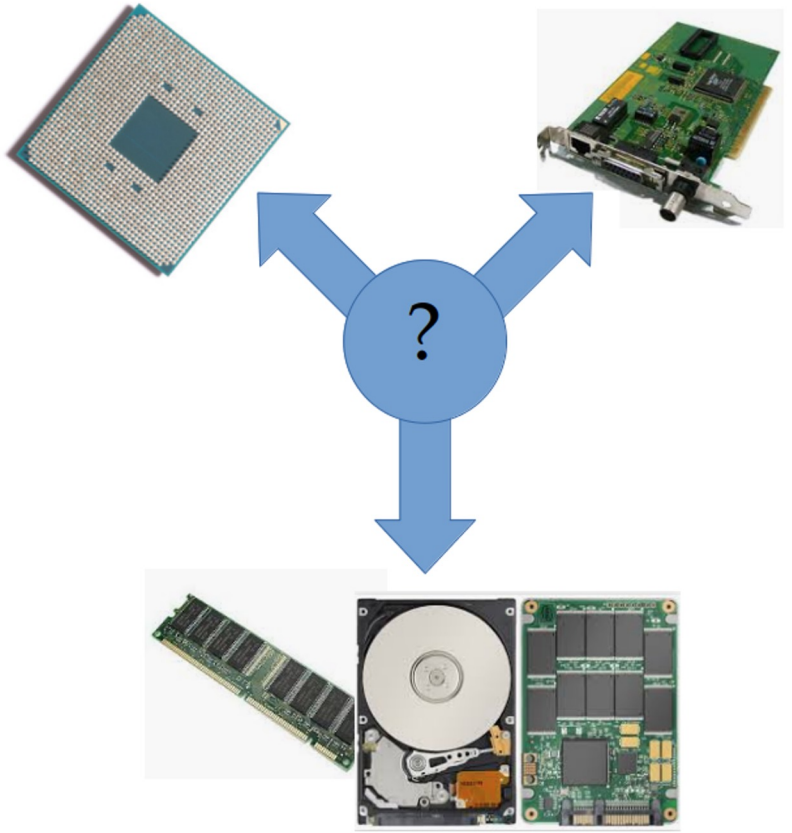
- Network Tuning
  - Firmware / Drivers
  - Congestion control
    - RE: talks on BBR
  - Ring buffers
  - CPU pin to network card
  - Path
    - MTU
    - Work with the network not against it.



EPOC
Engagement and Performance Operations Center

Tuning one element at a time. (CPU)

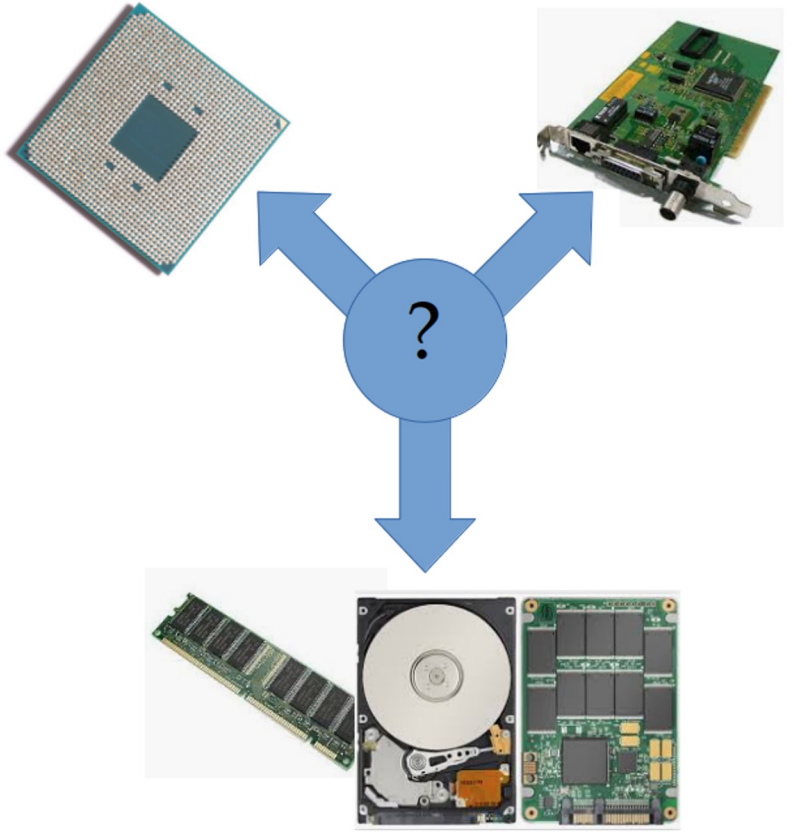
- CPU as the heart of the beast.
  - Take a look at your BIOS settings
    - Vendors often have guides
    - Power or thermal concerns?
    - Monitoring and baselines
      - (vast topic)
    - Interactions with storage
      - Are your RAM channels optimized?



EPOC
Engagement and Performance Operations Center

Tuning one element at a time. (Storage)

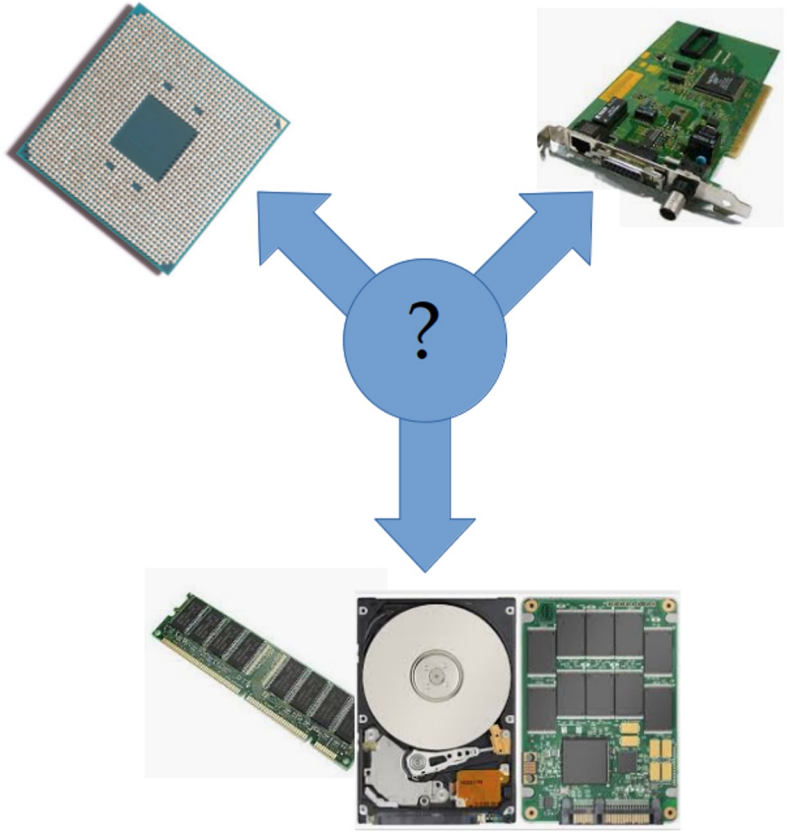
- All elements are key
  - Latest revisions
    - RAID firmware (Dedicated vs. HBA)
    - Filesystem/HBA balance
    - OS (appliance)
  - Maintenance concerns
    - Resilver / Rebuild
    - Failure does happen!



EPOC
Engagement and Performance Operations Center

Tuning one element at a time. (Community)

- All together now
  - Share failures and successes!
    - Community participation
    - Maintenance concerns
    - Get to know your network folks
      - Local, regional, and beyond.
    - Ask for help, we all have the same question(s)!
    - DATA Movement plans and policies can help.
    - Collaborate and enjoy.



EPOC
Engagement and Performance Operations Center

ESnet DTN Scorecard

	10G DTN				x10G, 25G, 40G, 100G DTNs			x400G
DTN host Transfer Rates	1/6 PetaScale	1/3 PetaScale	1/2 PetaScale		PetaScale: 1 PB/wk	PetaScale: 1 PB/day		Future ExaScale: 1 XB/month
Data Transfer Volume (Researcher)	1 TB/hr	2 TB/hr	3 TB/hr		5.95 TB/hr	41.67 TB/hr		33.33 PB/day
Network Transfer Rate (Network Admin)	2.22 Gb/s	4.44 Gb/s	6.67 Gb/s		13.23 Gb/s	92.59 Gb/s		3.09 Tb/s
Storage Transfer Rate (Sys/Storage Admin)	277.78 MB/s	555.54 MB/s	833.33 MB/s		1.65 GB/s	11.57 GB/s		385.80 GB/s

EPOC
Engagement and Performance Operations Center

Data Transfer performance 1TB/min?

$ numactl -m 1 ~/repos/xfer_test/xfer_test -c 10.10.1.10 -t 120 -i 2 -o 22 -a 1 -r

Using a SLaBS buffer of size 4194304 with 1 partitions of size 4194304

1099364: | port=18515 | ib_port=1 | tx_depth=16 | sl=0 | duplex=0 | cma=1 |

Created SLAB buffer with SIZE: 4194304 PARTITIONS: 1

raddr: 0x7f41ce66f000, laddr: 0x7f892c4aa000, size: 4194304

Metadata exchange complete

[0.0-2.0 sec]	29.92 GB	119.67 Gb/s
[2.0-4.0 sec]	32.12 GB	128.49 Gb/s
[4.0-6.0 sec]	35.36 GB	141.42 Gb/s
[6.0-8.0 sec]	43.22 GB	172.86 Gb/s
[8.0-10.0 sec]	44.83 GB	179.32 Gb/s
[10.0-12.0 sec]	43.38 GB	173.50 Gb/s
[12.0-14.0 sec]	44.83 GB	179.32 Gb/s
[14.0-16.0 sec]	45.34 GB	181.34 Gb/s
[16.0-18.0 sec]	45.70 GB	182.78 Gb/s
[18.0-20.0 sec]	46.62 GB	186.49 Gb/s
[20.0-22.0 sec]	47.00 GB	188.00 Gb/s
[22.0-24.0 sec]	47.35 GB	189.41 Gb/s
[24.0-26.0 sec]	48.07 GB	192.26 Gb/s
[26.0-28.0 sec]	48.05 GB	192.17 Gb/s
[28.0-30.0 sec]	48.08 GB	192.29 Gb/s

…

[0.0-60.6 sec]	1385.80 GB	182.89 Gb/s	bytes: 1385802235904

EPOC
Engagement and Performance Operations Center

ESnet Reference DTNs

2023 ESnet6 25/50/100 Gb/s Capable DTN Design

The total cost of this server was around $25K in late 2022. These systems be deployed to ESnet in late 2022 and into 2023 for ESnet6. Please note that specifics on configuration will be available after full evaluation.

Base System: Supermicro 2124US-TNRP 2U dual AMD socket SP3 server

- onboard VGA, dual10G RJ45, dual10G SFP+, onboard dedicated IPMI RJ45
- 1 PCI-E 4.0 x16 slot,
- 24 front access NVME hotswap bays
- dual redundant hotswap 1200W PSU
- 2x AMD EPYC Milan 73F3
- 16 cores each
- 3.5Ghz 240W TDP processor
- 256 GB RAM - 16x 16G DDR4 3200 ECC RDIMM
- 800G System Disk: 2x Micron 7300 MAX 800G U.2/2.5" NVME
- 25TB Data Disk: 10x Micron 9300 MAX 3.2TB U.2/2.5" NVME
- NVIDIA MCX613106A-VDAT ConnectX-6 EN Adapter Card 200GbE
- Mellanox MMA1L10-CR Optical Transceiver 100GbE QSFP28 LC-LC 1310nm LR4 up to 10km
- OOB license for IPMI management
- 2x 1300W -48V DC PSU OR 2x 1200W AC PSU

EPOC
Engagement and Performance Operations Center

ESnet and Globus Reference Datasets

- DTNs 10G, 25G+
  - ds08 1TB - 50 x 10GB; 350 x 1GB; 1,000 x 100MB; 5,500 x 10MB; 23,176 x 1MB files in single directory
  - ds10 1TB - 100 x 10GB files in single directory
  - ds16 1TB - 4 x 250GB files in single directory
- DTNs 1-10G+
  - Climate-Huge - 2 files, 1 folder, 245GB
  - Climate-Large - 11 files, 1 folder, 245GB
  - Climate-Medium - 1117 files, 1 folder, 245GB
  - Climate-Small - 1496 files, 305 folders, 245 GB
- DTNs <1G
  - 100M.dat single file
  - ds01 100MB - 10,000 x 10KB files in single directory

EPOC
Engagement and Performance Operations Center

THANK YOU!

EPOC
Engagement and Performance Operations Center

LINKS

ESnet - https://es.net/
EPOC - https://epoc.global/
FasterData - https://fasterdata.es.net/
EPOC YouTube - https://www.youtube.com/c/EPOC-IU-ESnet
Data Mobility Expo (DME) - https://fasterdata.es.net/science-dmz/learn-more/2019-2020-data-mobility-workshop-and-exhibition/

EPOC
Engagement and Performance Operations Center