http://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/

CC* Data Movement Workshop and Exhibition – Data Transfer Hardware

Jason Zurawski, Eli Dart
zurawski@es.net, dart@es.net
ESnet / Lawrence Berkeley National Laboratory

Dr. Jennifer M. Schopf
jmschopf@indiana.edu
Indiana University International Networks

CC/CICI PI Meeting Pre-Workshop*
September 22nd 2019

National Science Foundation Award #1826994

https://epoc.global

Outline

- ***Problem Statement & Expected Outcomes***
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Outline

- ***Problem Statement & Expected Outcomes***
  - ***Where We Started***
  - Where We Need Go
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Problem Statement – Start

- DTN History & Purpose:
  - Original concept came from initial Science DMZ Design (~2012)
  - Basic idea:
    - Host(s) dedicated to the task of data movement (and only data movement)
    - Limited application set (data movement tools), and users (rarely shell access)
    - Specific security policy enforced on the switch/router ACLs
      - Ports for data movement tools, most in a ‘closed wait’ state
      - Nothing to impact the data channel
    - Typically 2 footed:
      - Limited reach into local network (e.g. ‘control channel’: shared filesystem, instruments)
      - WAN piece that the data tools use (e.g. ‘data channel’)
  - Position this, and the pS node, in the DMZ enclave near the border

Outline

- ***Problem Statement & Expected Outcomes***
  - Where We Started
  - ***Where We Need Go***
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Problem Statement – Destination

- Building a DTN is largely the same
  - ***Understanding user needs***
  - Design/Specification/purchase/construction of hardware
  - Tuning and measuring performance implications
  - Integrating with users tools, and suggesting new approaches
- Some technology adaptations in the past ~8 years are significant
  - SSD technology
  - CPU ability to handle advanced functionality
  - Intelligence of OS
  - Integration with other workflow components (storage, compute)

EPOC
Engagement and Performance Operations Center

Problem Statement – Destination

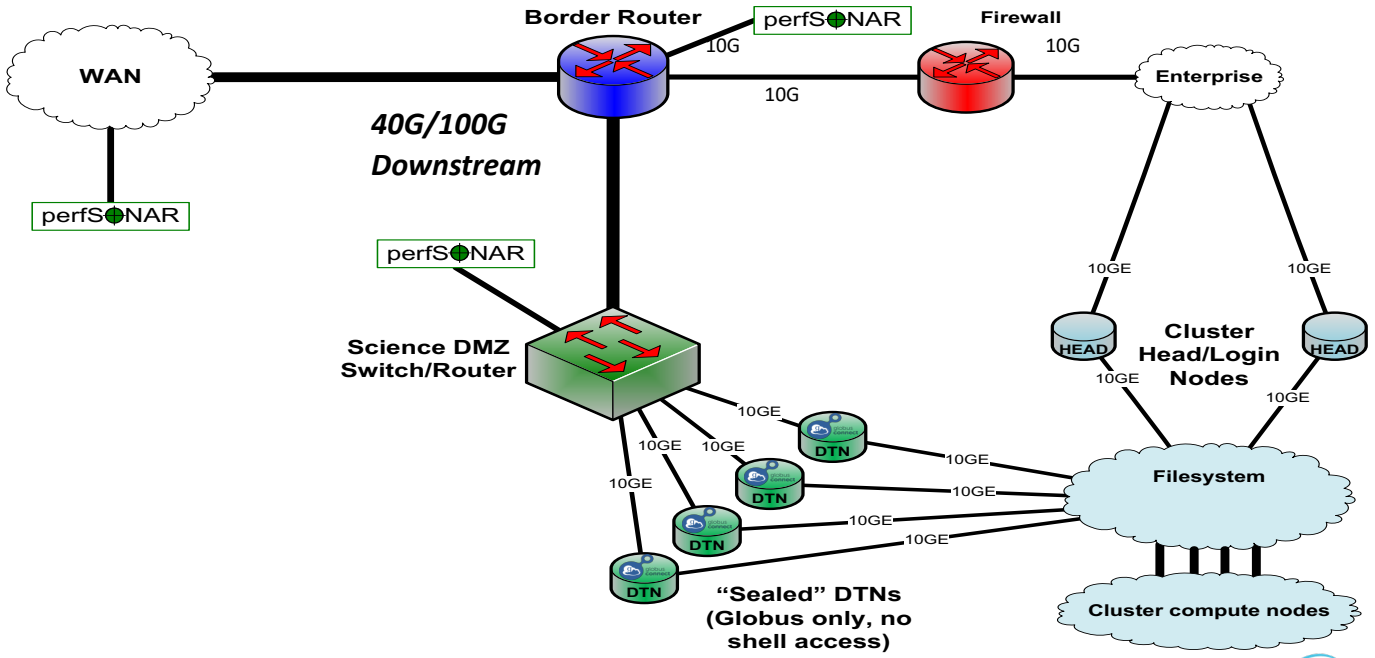
- Some technology adaptations in the past ~8 years add a layer of distraction:
  - Lots of different approaches to do the same thing, abundance of choice is sometimes worse than no choice (e.g. config management)
  - Distraction from root problem of data mobility
  - "Build it big" regularly conflicts with "build it right"
    - E.g. 100G DTN is an exciting capability – there are a small single digit integer of collaborations that are capable of using this collaboration regularly.
    - 10G/40G capability remains the most cost effective, scalable, supportable, usable, and network friendly tools to deploy
  - Integration with other workflow components (e.g. storage, compute)
    - "Once you have a hammer, everything looks like a nail". The technology becomes the enemy when we don't understand the implications of use.

EPOC
Engagement and Performance Operations Center

Problem Statement – Destination

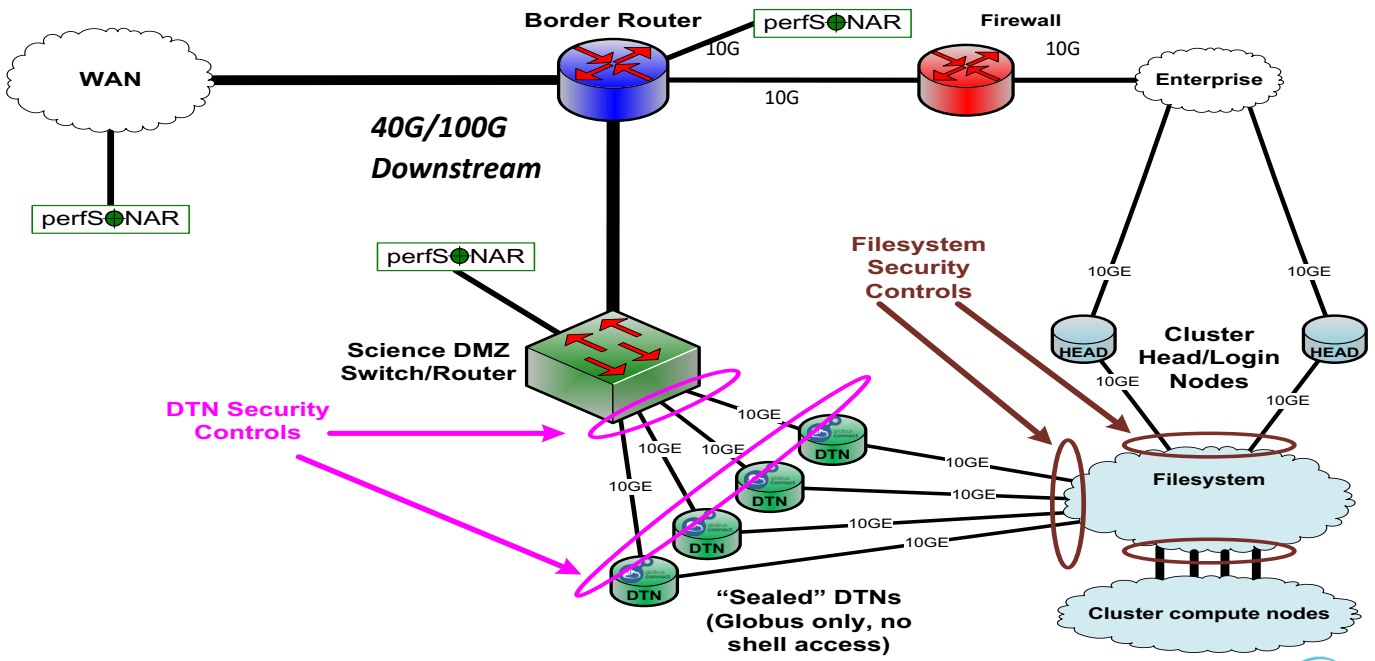
- Understanding use cases is critical (and often skipped)
  - How the user interacts with the tool is critical
  - If a tool is constructed that doesn't fit their usage profile, they won't use it
  - Spend the extra time on the front end, to avoid the mistakes of a poorly designed back-end tool
  - This means:
    - Understanding how it will be used today
    - How it will be used in the future
    - Knowing the types of data, and requirements on how the data is used/migrated
    - Suggesting the right sized technology solution to go with what you learned

Problem Statement – End Result



EPOC
Engagement and Performance Operations Center

Problem Statement – End Result



Outline

- Problem Statement & Expected Outcomes
- ***Operating Environment***
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- ***Operating Environment***
  - ***Rack & Form Factor***
  - Power
  - Cooling
  - Networking & Proximity to Resources
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Ops Env – Rack & Form Factor

- Server Form Factor:
  - Calculated using the same U measurements.
  - Typical sizes are between 1U and 4U. Larger ‘possible’, but these are dying out as the use of spinning HDDs are fading toward smaller SSDs
  - Depth varies – make sure you consider this vs. rack size
  - Considerations:
    - Components (e.g. NVMe’s) are getting smaller, thus moving to smaller U
    - Heat dissipation is still challenging, favoring more room in the case
    - Expansion (adding more stuff) is always an option
    - Don’t forget aboud PSU needs

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- ***Operating Environment***
  - Rack & Form Factor
  - ***Power***
  - Cooling
  - Networking & Proximity to Resources
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Ops Env – Power

- Server Power Supply Unit (PSU) Requirements:
  - Typically ‘sized’ for a specific U chassis (e.g. 1U, 2U, etc)
  - Come with multiple fans
  - Options for redundancy within a single unit, or by installing multiples (if your motherboard supports it)
  - Rated to a specific power delivery (e.g. 1000W, etc)

EPOC
Engagement and Performance Operations Center

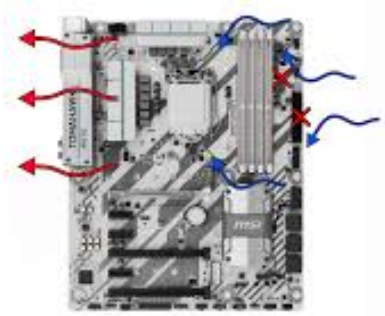
Outline

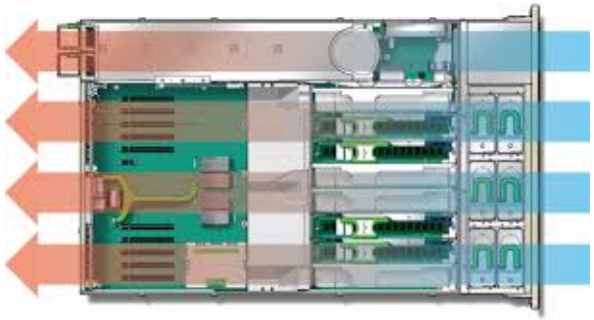
- Problem Statement & Expected Outcomes
- ***Operating Environment***
  - Rack & Form Factor
  - Power
  - ***Cooling***
  - Networking & Proximity to Resources
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

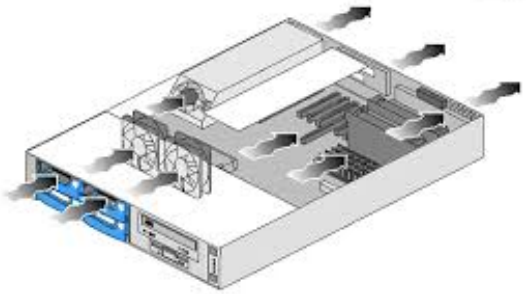
EPOC
Engagement and Performance Operations Center

Ops Env – Cooling

- Server Heat Management:
  - Servers typically have one direction of airflow: front to back, or back to front
  - Internal design is such that airflow is maximized:
    - Vents/Intake Fans
    - Heat Producing Components
      - Motherboard + CPU/RAM/Daughter cards
      - Disk(s) or NVMe
      - PSU(s)
    - Vents/Exhaust Fans







EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- ***Operating Environment***
  - Rack & Form Factor
  - Power
  - Cooling
  - ***Networking & Proximity to Resources***
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

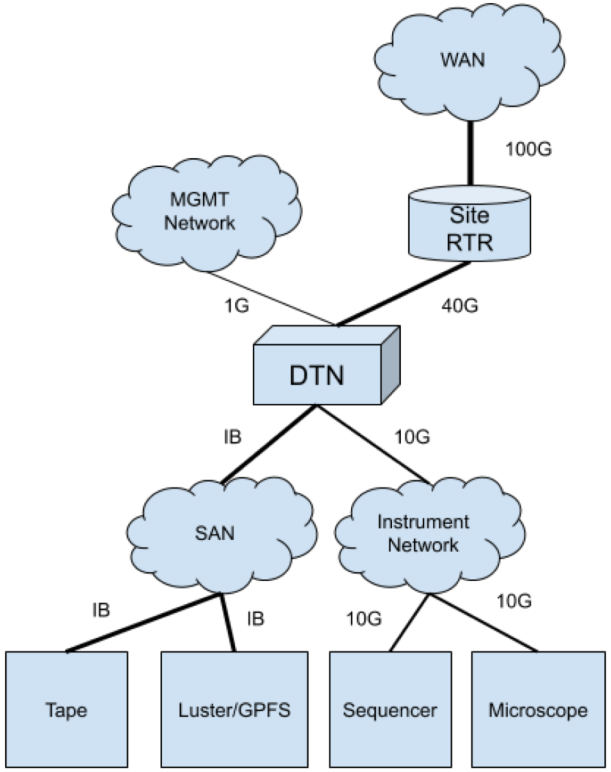
Ops Env – Network & Resource Proximity

- Connectivity Considerations:
  - Need to have enough ‘outbound’, and try not to create too many bottlenecks
  - E.g. 40 x 1Gs into a single 10G is bad when all are talking at once.
  - Compute/Storage network requirements vary, caveat emptor
- Oversubscription can be a very bad thing
  - E.g. a DTN is ‘bursty’, bulk data movement may require line rate (10G/40G etc.) performance for hours.
    - This can push out other things
    - Other things can impact this
- Suggestion: DTN data plane should be ‘home run’ to the building switch
  - Prevents local ToR congestion
  - Improved performance to carrier-class networking gear
  - Avoids in-line security devices that may impact performance

EPOC
Engagement and Performance Operations Center

Ops Env – Network & Resource Proximity

- DTN can end up existing in different networks depending on use case
- Support requires:
  - Adequate space in chassis
  - Power/Cooling to support additional technology
  - Security profile for each of these connections



EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- ***Hardware***
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Hardware Interlude

- Reference Implementation:
  - https://fasterdata.es.net/science-dmz/DTN/reference-implementation/
  - 3 Builds exist today:
    - Experimental 10G/40G/100G Platform (Purchased Sept 2019, not in Production)
    - Existing 10G/40G Platform (Purchased Oct 2016, 4 x Deployed on ESnet)
    - 100G Platform (Experiment on ESnet testbed in ~2017. Deployed, but not public)
  - This material will focus on the 2019 iteration above. Note that we don't have hard performance numbers available as of Sept 2019, but will publish to fasterdata when we do (expected early 2020).

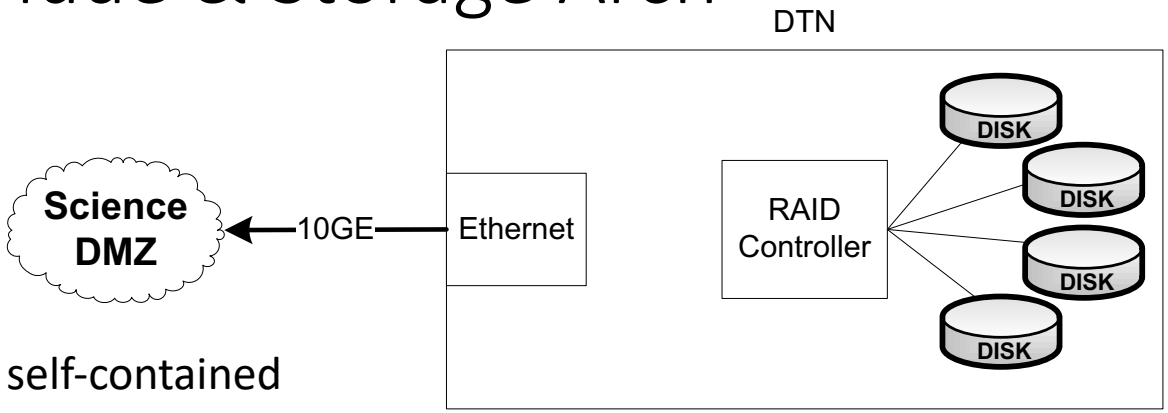
EPOC
Engagement and Performance Operations Center

Hardware Interlude

- 2019 Reference Implementation:
  - Base System: Gigabyte R281-NO0 dual socket P 2U server
    - Onboard: VGA, 2 x GbE RJ45 Intel i350, IPMI dedicated LAN
    - 24 x front access U.2 hotswap bays
    - 2 x rear access 2.5” SATA hotswap bays
    - Dual redundant hotswap 1600W PSU
  - 2 x Intel Cascade lake Xeon Gold 6246
    - 12 cores each
    - 3.3GHz 165W TDP processor
  - 12 x 16G DDR4 2933 ECC RDIMM (192G total)
  - 10 x Intel P4610 1.6TB U.2/2.5” PCIe NVMe 3.0 x4 Drives (connect directly to CPU for VROC)
  - 2 x Enterprise 960G 2.5“ SATA SSD (OS, onboard Intel SATA Raid 1)
  - VIntel® Virtual RAID On CPU (VROC) RAID 0, 1, 10, 5
  - Mellanox ConnectX-5 EN MCX516A-CCAT 40/50/100GbE dual-port QSFP28 NIC

EPOC
Engagement and Performance Operations Center

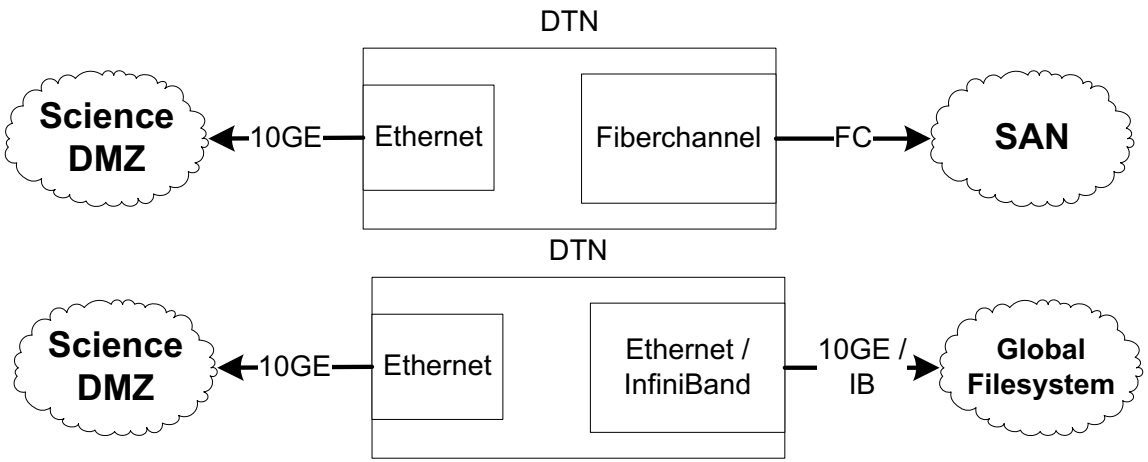
Hardware Interlude & Storage Arch



- 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

Hardware Interlude & Storage Arch



- 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

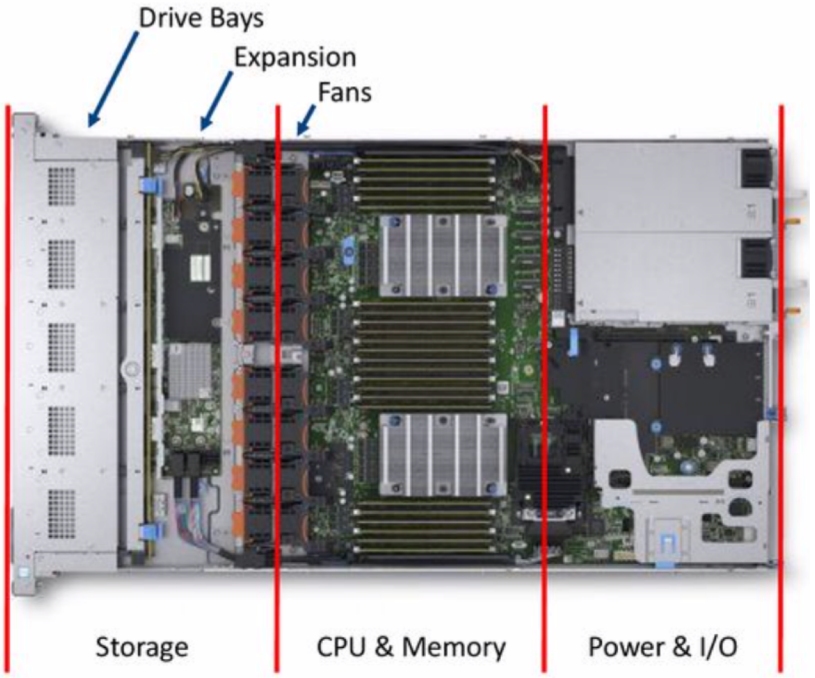
Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- ***Hardware***
  - ***Motherboard & CPU***
  - Main Memory
  - Disk Subsystem
  - Networking (Internal & External)
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Hardware – Chassis, PSU, Fans

- Layout:
  - Typical rack server could be divided into thirds
  - Front third typically for storage and supporting infrastructure
  - Need to consider:
    - Cooling
    - Power
    - Storage performance
    - Memory performance
    - Networking performance
    - I/O capability



EPOC
Engagement and Performance Operations Center

Hardware – Motherboard & CPU

- Without getting into too much of a battle …
  - As of early 2019, Cascade Lake is the latest iteration of the Intel microarchitecture (x86-64)
    - Replaces ‘Sky Lake’, ’Copper Lake’ is on Deck for 2019/2020
  - As of Late 2018, EPYC 2 (Rome) is the latest iteration of the AMD microarchitecture (Zen 2)
    - Replaces ‘Opteron’, ‘Zen 3’ product on Deck for 2019/2020
- Going with one or the other is a balancing act optimizing on:
  - Clock Speed & Core Count
  - PCI Support
  - Cache
  - Multi-processor support

EPOC
Engagement and Performance Operations Center

Hardware – Motherboard & CPU

- There are pros and cons with either approach. Our observations:
  - Intel:
    - Higher base clock rate, fewer cores (3.8 Ghz, between 4-56 cores / 128 threads)
    - 14nm Thinness
    - Fewer PCI Lanes (48)
    - Gen 3 PCI Support
  - AMD:
    - Lower base clock rate, more cores (3.2 GHz, up to 128 cores / 256 threads)
    - Going Thinner (14nm, going to 7nm)
    - More PCI Lanes (128)
    - Gen 3 PCI support (gen 4 'soon')
- AMD is more loved in the cloud space (e.g. lots of little things), Intel lends itself more toward a general purpose single/small # of use cases machine

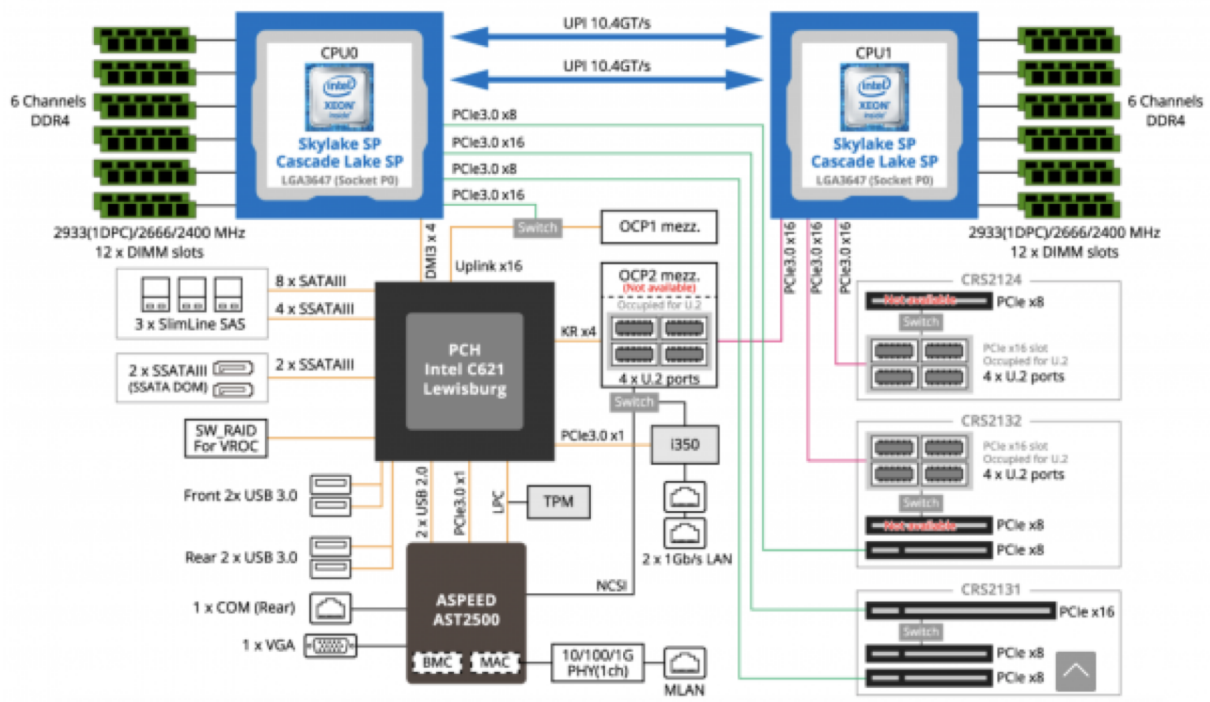
EPOC
Engagement and Performance Operations Center

Hardware – Motherboard & CPU

- Our workload still favors the faster clock – thus we build Intel
  - TCP streams are heavily dependent on the clock.
    - Faster the clock, more capable we are at dealing with TCP behavior (e.g. data sending/retransmitting)
    - TCP Parallelization can be done as the cores increase – but you are playing a resource allocation game typically. E.g. more streams means more possibility of locally induced loss
  - Extra cores will be needed to handle the I/O to and from the NVMe's
- Approach:
  - Fast clock (>= 3.2 Ghz)
  - Sensible number of cores (<= 12)
  - Number of processors (1 – 2)
  - Balance peripherals across the available PCIe resources on each core (e.g. networking cards, NVMes
  - Tune interrupts, etc. as needed

EPOC
Engagement and Performance Operations Center

Hardware – Motherboard & CPU



EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- ***Hardware***
  - Motherboard & CPU
  - ***Main Memory***
  - Disk Subsystem
  - Networking (Internal & External)
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Hardware – Main Memory

- Double Data Rate 4 Synchronous Dynamic Random-Access Memory (DDR4 SDRAM) is still the standard (since 2014).
  - DDR5 expected to start appearing in the wild in late 2019/2020.
  - "double data rate" interface, operates at clock rates between 800–2133 MHz and consumes between 1.2 – 1.4V of power
- When purchasing, you will see a number, e.g. "DDR4-2933", this is the 'data rate', which is measured in megatransfers per second (MT/s)
  - Higher is better
  - Higher is more expensive
  - Find a balance between size and speed that fits the budget

EPOC
Engagement and Performance Operations Center

Hardware – Main Memory

- Typical Strategy:
  - Max out the available DIMM (Dual In-Line Memory Module) slots with ‘something’ on the motherboard.
    - Spreads memory consumption equally around the processor
    - Ideally these all look the same (size, speed)
  - Try to get the largest / fastest you can afford
- Ref Implementation:
  - 2 x Intel Cascade lake Xeon Gold 6246
    - 12 x 16G DDR4 2933 ECC RDIMM (192G total)
    - 6 per Processor

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- ***Hardware***
  - Motherboard & CPU
  - Main Memory
  - ***Disk Subsystem***
  - Networking (Internal & External)
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

- Old Days:
  - HDD was mechanical (spinning disks) w/ set form factor/size (e.g. 2.5” or 3.5” were common, and based on history)
  - Connectivity:
    - Serial ATA (SATA) – works with SSDs and HDDs. Transfer rates between 1.5 Gb/s (Gen 1) and 6 Gb/s (Gen 3)
    - Serial Attached SCSI (SAS) – higher performance (up to 12 Gb/s) and limited backward compatibility/interoperability with earlier SCSI/SATA
  - To get a larger storage chunk – RAID (Redundant Array of Independent/Inexpensive Disks) was employed
    - Typically a hardware-based controller.
    - Used to manage HDDs/SSDs to work as a logical unit

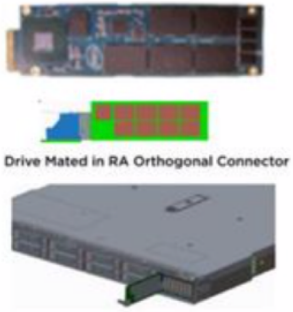
EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

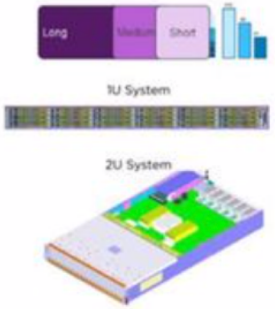
- New Days:
  - SSD form factors ***don't have to*** look like HDDs, they can be compact
    - Many still support the old style of connection (e.g. SATA) to facilitate a drop in replacement, but M.2 and U.2 are becoming more common due to their smaller form factor/cost/performance
  - Connectivity:
    - PCI Express® (PCIe®) - originally designed for peripherals
      - Offers minimal latency, high bandwidth
      - Allows NAND flash and Non-Volatile Memory Express (NVMe)
    - RAID is done via software, with hardware assist on chip
    - E.g. Intel Virtual RAID on CPU (VROC)
      - Allows for NVMe SSDs to be connected directly to the CPU via PCIe
      - Simpler/Faster RAID, no additional hardware needed
      - Requires a 'license'

EPOC
Engagement and Performance Operations Center

Potential NVMe SSD Form Factors



SFF-TA-1006



SFF-TA-1008

SFF-TA-1007

NGSFF

AIC

2.5in U.2

m.2

EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

- NVMe: host controller interface/storage protocol
  - Allows for use of SSDs over PCIe versus the older technologies (SATA/SAS)
  - Faster/lower latency, more compact, more energy efficient. Downside (?) is that older machines designed for big HDD RAIDs waste a lot of that space.
  - Underlying SSDs technology can vary (NAND Flash, 3D XPoint, etc.)
  - Form factor matters less too, although most are moving away from HDD on new machine purchase
- NVMe functions by:
  - Maps I/O commands and responses to shared memory
  - Supports parallel I/O with multicore processors

EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

- M.2
  - Large/Flexible in terms of form factors (drive replacement or SFF card)
  - Can do PCIe/NVMe, SATA, USB (e.g. compatible with prior tech)
    - In the typical use case, uses PCIe 3.0 (x4 interface) vs. SATA
  - Potential bandwidth of 32Gb/s
  - Same cable for power and data
  - Does not allow 'hot swap'
  - Supports 3.3V power

EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

- U.2
  - Uses PCIe 3.0 (x4 interface) and is is mechanically the same as SATA Express.
  - Can connect to M.2 with an adapter
  - Allows 'hot swap'
  - Support for higher capacity SSDs (e.g. 2.5” form factor can be densely populated)
  - Supports 3.3V and 12V power

EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

- Design Architecture/Layout Considerations:
  - Balance Drive Size, Quantity, Space, Power/Cooling, PCIe Availability
  - Current generation typically requires PCIe 3.0 x4 lanes.
  - Processors have a limited set of PCIe lanes. E.g. recent Xeons have ~48.
  - Single CPU Worksheet (Assumes 1-armed NIC):
    - 40GB/100GB NIC = x16 PCI Slot
    - 8 NVMe Drives = x32 PCI Slots (@ x4 each)
  - Dual CPU Worksheet (Assumes 1-armed NIC):
    - 1:
      - 40GB/100GB NIC = x16 PCI Slot
      - 8 NVMe Drives = x32 PCI Slots (@ x4 each)
    - 2:
      - 12 NVMe Drives = x48 PCI Slots (@ x4 each)

EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

- Intel Virtual RAID on CPU (VROC)
  - hybrid RAID solution
  - Designed for NVMe SSDs connected directly to the CPU
  - CPU feature on certain Intel Xeon Chipsets: Volume Management Device (VMD)
    - Enhance the 48 PCIe lanes for NVMe connections
    - e.g. simpler RAID that requires no additional hardware
  - Use requires a “License” provided by a “hardware key” (dongle). Typically you buy this up front.
    - We highly recommend you pay the tax and do this at purchase time (its harder to ‘upgrade’ if you forget)

EPOC
Engagement and Performance Operations Center

Hardware – Disk Subsystem

- Reference System:
  - 2U server
    - 24 x front access U.2 hotswap bays
    - 2 x rear access 2.5” SATA hotswap bays
  - 10 x Intel P4610 1.6TB U.2/2.5” PCIe NVMe 3.0 x4 Drives (connect directly to CPU for VROC)
  - 2 x Enterprise 960G 2.5“ SATA SSD (OS, onboard Intel SATA Raid 1)
  - Virtual RAID On CPU (VROC) Premium for RAID 0, 1, 10, 5
- Decoding:
  - 16TB Base Storage (lose some due to RAID setup) provided by 10 x 2.5” U.2 Drives
    - Consumes 40 PCIe slots
    - Also can cram into a 1U enclosure if you try hard, we went the 2U route since cooling is a factor
  - 2 x 2.5” SATA SSDs for OS, segregated from data filesystem
  - VROC Premium – buy this up front (its not much $$) so you can configure RAID how you want it. Comes w/ hardware key

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- ***Hardware***
  - Motherboard & CPU
  - Main Memory
  - Disk Subsystem
  - ***Networking (Internal & External)***
- Software
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Hardware – Networking Topics

- The Network Interface Card (NIC) is critical to DTN operation:
  - The serer may have 1Gbps (RJ45 connector/1000Base-T) connections on the motherboard already. These are useful for management connectivity
  - Higher-speed is delivered via daughter cards, connected to the PCIe bus/slots
    - 10GBase
    - 40GBase
    - 100Base
    - 25G-Base/50GBase (less common, but available)

EPOC
Engagement and Performance Operations Center

Hardware – Networking Topics

- Reference Implementation: Mellanox ConnectX5
  - Single/Dual-Port Adapter
    - Note, 2 port operation is typically ‘hot spare’, and not 2x performance
  - Supports 100Gb/s (and lower) Ethernet Standards (read the glossy to verify)
  - NVMe Offload & RoCE Support
  - PCIe Gen 4, with prior gen compatibility
    - x16 PCI Lanes, some models (lower speed) can function w/ x8 PCI lanes. Auto-negotiates depending on use case
  - SFP28 / QSFP28 Ports (depends on model)
    - Facilitates connectivity via optics / fiber or live-cable
    - N.B. what you connect your NIC to upstream may be challenging, not all routers/switches communicate well over live cables that aren’t brand specific

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- ***Software***
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- ***Software***
  - ***Operating System***
  - ”Helper” Infrastructure
  - Measurement / Monitoring / Maintenance
  - Data Transfer
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Software – Operating System

- We use varieties of Linux, specifically Red Hat derivatives (e.g. CentOS). There are other Linux varieties out there (Debian derivatives), as there are other different OSs (Microsoft)
- General statement – "use what you are comfortable supporting"
  - N.B. driver support may force your hand.
- CentOS 6.x/3.x Kernel is deployed on our production machines
- CentOS 7.x is what we are experimenting with, and we have enabled the 4.x Kernel (this is not default).
  - TCP enhancements, as well as some additional driver and filesystem support in new kernels
  - For those looking to play with 'BBR TCP', you will need to go the 4.x route.

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- ***Software***
  - Operating System
  - ***"Helper" Infrastructure***
  - Measurement / Monitoring / Maintenance
  - Data Transfer
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Software – Helper Infrastructure

- Configuration Management (There are way too many options here …)
  - Puppet - https://puppet.com
  - Ansible - https://www.ansible.com
  - Chef - https://www.chef.io
  - CFEngine - https://cfengine.com
- Why use config management?
  - Automation simplifies management of devices (e.g. security patching, common packages, accounts, etc.)
  - Allows for easier 'special case' management too. For example, a DTN may require software that the DNS server doesn't need
- General statement – "use what you are comfortable supporting"

EPOC
Engagement and Performance Operations Center

Software – Helper Infrastructure

- Security Protections:
  - Standard host-based firewall (e.g. iptables/ip6tables) is a good idea.
  - Host or Network-based IDS if you want to go that route:
    - OSSEC - https://www.ossec.net
    - Suricata - https://suricata-ids.org
    - Zeek (Bro) - https://www.zeek.org
  - Logging is always a good plan
    - Send your syslog to a central log server
    - Use analysis (ELK stack, or pay for something like Splunk)

EPOC
Engagement and Performance Operations Center

Software – Helper Infrastructure

- Sample Security Profile:
  - IPMS (out of band) for management access segregated from main network
  - MGMT Interface:
    - Private/internal IP if you can
    - SSH listening specifically on this interface (or only allowing logins via designated networks/jump hosts). Keep shell account access small
  - Data Interface:
    - ***NO GENERAL LOGIN***
    - Interface to data tools
      - Contact (public) address – in/out TCP traffic on designated ports
      - Data channels (TCP or UDP) in/out on designated port range. Could be 1000 or so, but note these sit in a closed-wait
      - Control channel communications to known endpoints (e.g. Globus cloud connection point, etc.)
    - Measurement tools (perfSONAR ports, etc) – if desired
    - Portal access (e.g. HTTP/HTTPS) - if enabled

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- ***Software***
  - Operating System
  - ”Helper” Infrastructure
  - ***Measurement / Monitoring / Maintenance***
  - Data Transfer
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Software – Meas/Mmgt/Maintenance

- It's a good idea to have some host monitoring in place:
  - Ensures that the host is passing basic tests
  - Sending back some telemetry on operation (CPU, memory, network, etc.)
  - You are probably doing this anyway …
- Some options:
  - Ganglia - http://ganglia.sourceforge.net
  - Nagios - https://www.nagios.org
  - Solar Winds - https://www.solarwinds.com
- Integrate this onto your 'management' network/VLAN
  - Should/could/would also check the data interface, if separate
- General statement – "use what you are comfortable supporting"

EPOC
Engagement and Performance Operations Center

Software – Meas/Mmgt/Maintenance

- Performance Testing:
  - perfSONAR tools are a good addition
  - You don't need to be running a 'full' toolkit
  - Options:
    - https://docs.perfsonar.net/install_options.html
    - Consider the 'tools' at a minimum, or the 'testpoint' if you do want to enable regular testing
- "Will testing impact data transfer"? … or … "Will data transfer impact testing?"
  - In the general case, the standard perfSONAR bandwidth test is done using TCP, so it will back off if there is a production data transfer occuring
  - Data transfer can be TCP or UDP – the former will back off, the latter won't
  - Suggested approach:
    - Install tools to allow for testing as needed
    - Do not schedule regular tests on the DTN
    - Have a nearby perfSONAR node to test the network path

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- ***Software***
  - Operating System
  - ”Helper” Infrastructure
  - Measurement / Monitoring / Maintenance
  - ***Data Transfer***
- Configuration & Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

Software – Data Transfer

- Oodles of Options out there
- Older/non sophisticated:
  - SCP/RSYNC (please apply patches to SSH: https://www.psc.edu/hpn-ssh)
  - FTP/SFTP
- Newer/sophisticated:
  - Globus - https://www.globus.org
  - Aspera - https://asperasoft.com
- Scientific Use Case/VO Specific:
  - XRootD - https://xrootd.slac.stanford.edu
  - PhEDEx - http://cmsdoc.cern.ch/cms/aprom/phedex/
  - FDT - http://monalisa.cern.ch/FDT/
  - Rucio - https://rucio.cern.ch

EPOC
Engagement and Performance Operations Center

Software – Data Transfer

- Functionality varies
  - Some are command line, some are graphical, some are tied to advanced workflow software
  - All use different protocols (TCP, UDP)
  - All have different port in/out requirements
  - Some require shell access to the machine, some are invoked via other known protocols (HTTP/HTTPS), others can be run 3rd party
- Common themes to a 'good' tool:
  - Parallelism
  - Checksumming
  - Aggressive (application layer) tuning
  - API that allows for integration into higher-level software

EPOC
Engagement and Performance Operations Center

Software – Data Transfer (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

Software – Data Transfer (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

Software – Data Transfer

- Some considerations as you test performance of hardware and software:
  - Application and filesystem interaction is crucial
  - Some are more tuned for high performance than others
  - Trade offs:
    - Block Size on the storage (small vs. large)
    - Handling of metadata (directories, nesting)
    - Single vs. distributed filesystem (e.g. single host vs. multiple host [cloud/grid])
    - Support for EXT(4), LUSTRE, GPFS, XFS, ZFS
- The ‘mission’ software (e.g. Rucio, FDT) is very closely tied to the workflow of those use cases (big block size, distributed FS) and can work in a highly parallel environment
- More general purpose software (Globus, Aspera) make no assumptions about what lives beneath and will offer performance gains that span configurations

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- ***Configuration & Tuning***
- Management & Maintenance
- Integration With Users

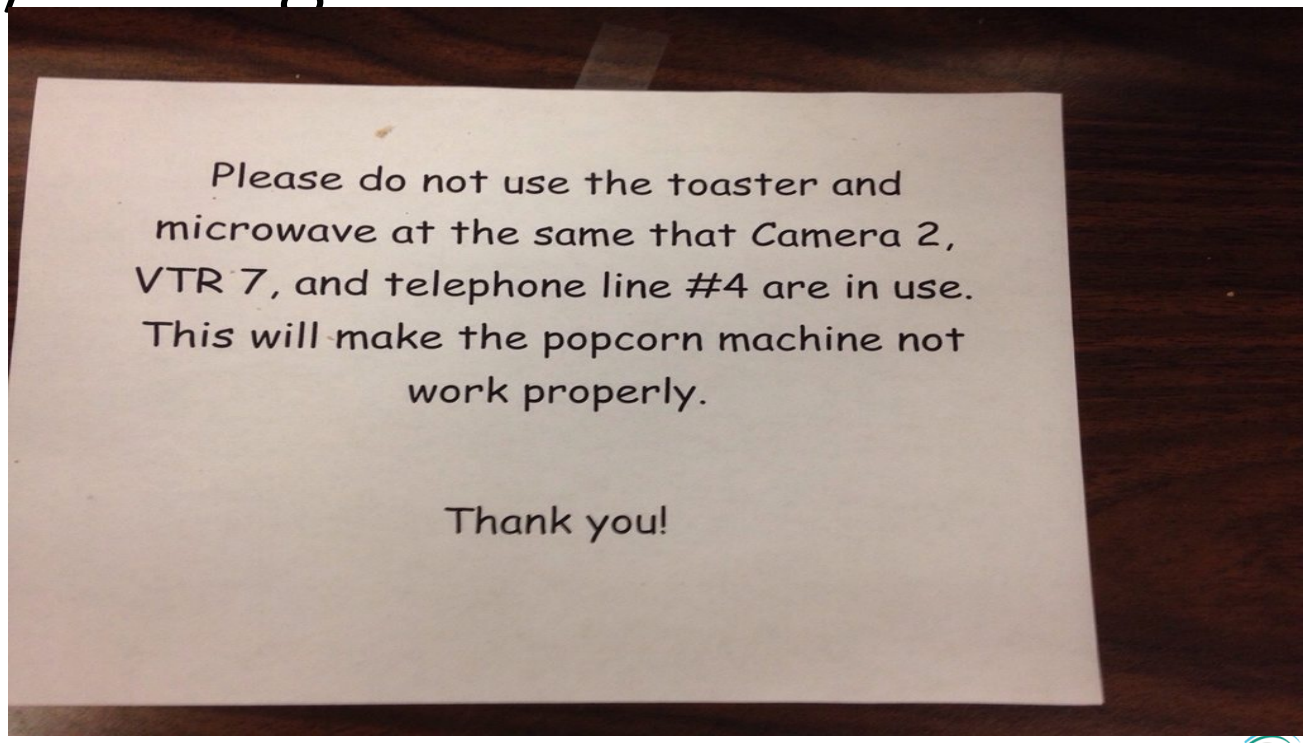
EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- ***Configuration & Tuning***
  - ***Operating System Tuning***
  - Disk / Memory Tuning
  - Network Hardware Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

CFG/Tuning = Art & Science



EPOC
Engagement and Performance Operations Center

CFG/Tuning – Operating System

- BIOS
  - There are still a number of knobs here (differs by system)
  - “Hyperthreading” is good to enable , so is “Performance mode” (e.g. turbo boost) too - “Power Saving” isn’t
    - Always try these steps first, since the BIOS influences the OS
  - “Software pre-fetch” can assist with data access
  - Dell has a good guide with lots of options: https://www.dell.com/support/article/us/en/04/sln311501/high-performance-computing?lang=en

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Operating System

- IO Scheduler
  - The default scheduler on some versions of Linux 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

CFG/Tuning – Operating System

- CPU Governor
  - By default Linux uses 'powersave'
  - Throughput can rise ~25% by changing to the 'performance' governor
  - The command to set this is:
    - For RHEL/CentOS systems:

      cpupower frequency-set -g performance
    - For Debian/Ubuntu systems:

      cpufreq-set -r -g performance
  - To watch the CPU governor in action, you can do this:

    watch -n 1 grep MHz /proc/cpuinfo
  - Note that the BIOS also has some control here (see prior slide), if you make the changes there, these may not have an impact

EPOC
Engagement and Performance Operations Center

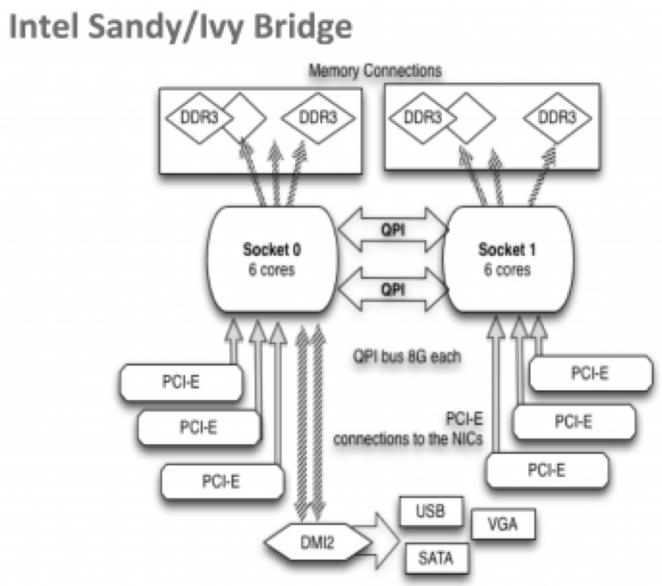
CFG/Tuning – Operating System

- Interrupts
  - 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

CFG/Tuning – Operating System

- EX:
  - The PCI slot for the NIC is directly attached to only Socket 1
  - There is a large performance penalty if either the interrupts or the application running on Socket 0
    - E.g. everything must cross the QPI bus
  - TCP and UDP performance can be up to 2x slower if you are using cores on the wrong CPU socket
  - ***<u>It is important that both the NIC IRQs and the application are using the correct CPU socket.</u>***



EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- ***Configuration & Tuning***
  - Operating System Tuning
  - ***Disk / Memory Tuning***
  - Network Hardware Tuning
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Disk & Memory

- Filesystem
  - We recommend using the ***ext4*** file system
  - Increasing the amount of "readahead" usually helps where workflow is mostly sequential reads.
  - Setting readahead should be done at system boot time. For example, add something like this to ***/etc/rc.local***:
    - ***/sbin/blockdev --setra 262144 /dev/sdb***

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Disk & Memory

- Virtual memory Subsystem
  - Setting dirty_background_bytes and dirty_bytes improves write performance. For our system, the settings that gave best performance was:
    - ***echo 1000000000 > /proc/sys/vm/dirty_bytes***
    - ***echo 1000000000 > /proc/sys/vm/dirty_background_bytes***
  - On heavily used DTN's we've seen cases where the host will run out of memory and give an error such as:
    - ***SLUB: Unable to allocate memory on node***
  - Reserving about 5% of the RAM for the VM subsystem using vm.min_free_kbytes seems to fix the problem.
  - For example, for a host with 96MB of RAM, add the following to /etc/sysctl.conf to set min_free to 4MB:
    - ***vm.min_free_kbytes = 4096000***
- Swap
  - To prolong SSD lifespan, do not swap on an SSD. In Linux you can control this using the sysctl variable vm.swappiness. For example, add this to ***/etc/sysctl.conf***:
    - ***vm.swappiness=1***

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- ***Configuration & Tuning***
  - Operating System Tuning
  - Disk / Memory Tuning
  - ***Network Hardware Tuning***
- Management & Maintenance
- Integration With Users

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Network Hardware

- General Node – use this for tuning @ 1Gbps or higher. Note that
  - Most of the tuning settings described here will ***decrease*** performance of hosts connected at rates less than 1Gbps (e.g. home users)
- The Bandwidth Delay Product is an important calculation to remember, e.g. a 10Gbps flow across a 100ms network requires 120MB of buffering (see https://www.switch.ch/network/tools/tcp_throughput/)
  - The settings discussed are not meant to get full line rate with a single flow
  - We are assuming that the data transfer tools can support parallel streams, thus there will be sharing and higher memory requirements
  - E.g. if you request that sockets consume 256M of memory, ensure you have that much to spend for all the possible sockets that may be open
- General Approach
  - To check what setting your system is using, use
    - 'sysctl name' (e.g.: 'sysctl net.ipv4.tcp_rmem')
  - To change a setting use 'sysctl -w'.
  - To make the setting permanent add the setting to the file 'sysctl.conf'.

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Network Hardware

- For a host with a 10G NIC, optimized for network paths up to 100ms RTT, and for friendlyness to single and parallel stream tools, add this to /etc/sysctl.conf

```
# allow testing with buffers up to 64MB
net.core.rmem_max = 67108864
net.core.wmem_max = 67108864
# increase Linux autotuning TCP buffer limit to 32MB
net.ipv4.tcp_rmem = 4096 87380 33554432
net.ipv4.tcp_wmem = 4096 65536 33554432
# recommended default congestion control is htcp
net.ipv4.tcp_congestion_control=htcp
# recommended for hosts with jumbo frames enabled
net.ipv4.tcp_mtu_probing=1
# recommended for CentOS7+/Debian8+ hosts
net.core.default_qdisc = fq
```

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Network Hardware

- For a host with a 10G NIC optimized for network paths up to 200ms RTT, and for friendliness to single and parallel stream tools, or a 40G NIC up on paths up to 50ms RTT:

```
# allow testing with buffers up to 128MB
net.core.rmem_max = 134217728
net.core.wmem_max = 134217728
# increase Linux autotuning TCP buffer limit to 64MB
net.ipv4.tcp_rmem = 4096 87380 67108864
net.ipv4.tcp_wmem = 4096 65536 67108864
# recommended default congestion control is htcp
net.ipv4.tcp_congestion_control=htcp
# recommended for hosts with jumbo frames enabled
net.ipv4.tcp_mtu_probing=1
# recommended for CentOS7+/Debian8+ hosts
net.core.default_qdisc = fq
```

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Network Hardware

- MTU (Maximum Transmission Unit)
  - The world is built around 1500 Bytes. As a result most things 'expect this' (and are not really built to be changed away from it)
  - 9000 Bytes gives you performance gains (especially at 40G/100G)
    - Send less data packets
    - Less for CPU/NIC to process
    - ~6x faster recovery from loss events
  - ***Moving from 1500 is not for the faint of heart. In fact it should be carefully considered since this change has the potential to really impact other traffic***
  - If you do … observations:
    - Can increase performance by a factor of 2-4 on 10G paths on older hardware
    - Can see 9.9Gbps vs 5Gbps for UDP, and 9.9Gbps vs 9.3Gbps for TCP.
    - For 40G (single stream) TCP can see gains of 25Gbps vs. 10Gbps. UDP testing can show improvement of 15Gbps vs. 8Gbps.

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Network Hardware

- MTU (Maximum Transmission Unit)
  - ***All hosts in a single broadcast domain must to be configured with the same MTU***, and this can be difficult and error-prone.
  - ***Ethernet has no way of detecting an MTU mismatch*** - this is a Layer 3 function that requires ICMP signaling in order to work correctly.
    - Unfortunately some sites block ICMP, which breaks path MTU discovery
    - If tracepath fails, that is likely what is happening.
  - A good approach is to create a new jumbo frame enabled subnet for your high-speed data transfer hosts.

- ping can be used to verify the MTU size, learn this trick since you will need to use it at least once:

  ping -s 8972 -M do -c 4 $IPADDRESS

EPOC
Engagement and Performance Operations Center

CFG/Tuning – Network Hardware

- MTU – Rules:
  - ***DO NOT BLINDLY ENABLE 9000 BYTE MTU ON HOSTS OR NETWORK DEVICES***
  - ***IF you do use 9000 MTUs …***
    - ***Think carefully about the choice, and what needs to change (e.g. entire broadcast domain!!!)***
    - ***NETWORKING DEVICES (switches, routers) SHOULD BE SET TO MAX PER INTERFACE (e.g. 9126, etc)***
      - This accounts for spill from VLANs/QinQ, etc
    - ***NETWORK HOSTS SHOULD BE SET TO 9000, NO MORE***
- ***Be conservative in what you send, liberal in what you accept***

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- ***Management & Maintenance***
- Integration With Users

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- ***Management & Maintenance***
  - ***Replacement Strategy***
  - Security / Data Flow Profile
- Integration With Users

EPOC
Engagement and Performance Operations Center

Mgmt/Maint – Replacement Strategy

- Hardware Scales to Multiple Years on Average
- Some things will require more frequent attention:
  - SSD Hardware
    - Expect that there will be performance degradation as the r/w cycle count increases
    - ~6-12mo of constant use will impact hardware
    - Buy extras of similar size
  - Networking (as up-stream hardware changes)
    - Some cards will facilitate a swap-in of other optics/live cables for different ethernet protocols
    - Check documentation to see what is supported
  - Main Memory augmentation
    - Make sure all the DIMM sticks are fully populated, and of the same 'size' (helps with performance)
    - As cost drops, and density increases, a full replacement of all sticks could be a cheaper alternative to upgrade

EPOC
Engagement and Performance Operations Center

Mgmt/Maint – Replacement Strategy

- If you are running out of capacity, there are two ways to scale:
  - Horizontal
    - Add more DTNs, of similar make/model speed
    - Facilitates more users, spread to a larger pool
    - Doesn't overwhelm the existing WAN connections
    - Doesn't increase the size of a single flow, but does scale data reads/writes and a larger number of users
  - Vertical
    - Increase the network capacity/size on a new DTN (e.g. moving from 10G to 40G, or 40G to 100G)
    - Can facilitate larger single flow speed
    - Can also facilitate more concurrent smaller flows
    - May result in CPU/memory bottlenecks as user numbers grow

EPOC
Engagement and Performance Operations Center

Outline

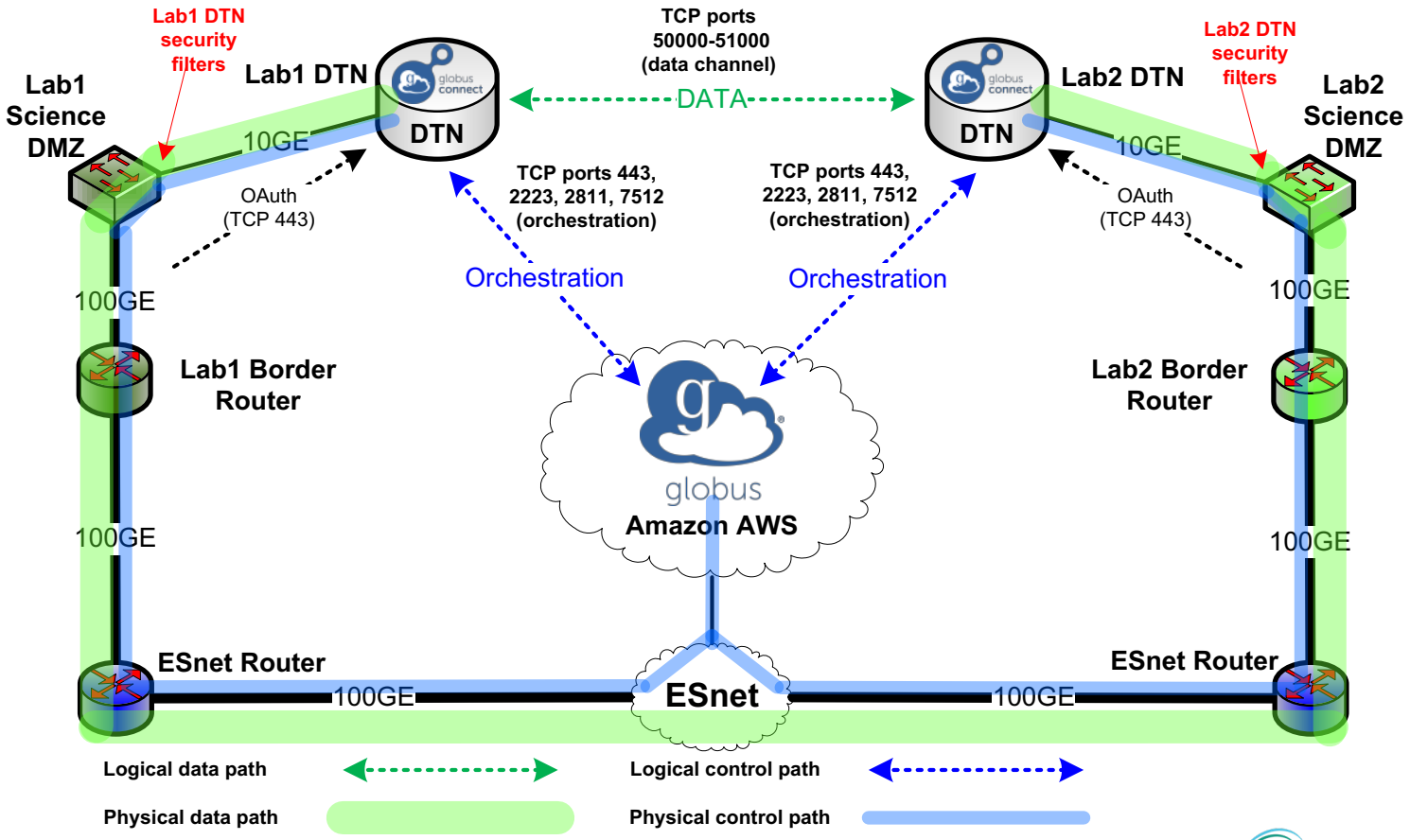
- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- ***Management & Maintenance***
  - Replacement Strategy
  - ***Security / Data Flow Profile***
- Integration With Users

EPOC
Engagement and Performance Operations Center

Mgmt/Maint – Security / Data Flow

- An overlapping concern is understanding how data will flow in/out, and how you can manage security without impacting performance.
  - Each tool will have different requirements for ports
    - Understand the protocol (TCP/UDP) the port numbers
    - Understand the amount (e.g. be prepared for ‘many’, and don’t let this scare you). Understand if they are open, or are in closed/wait state
  - Each user may have different requirements for endpoints
    - R&E or non-R&E?
    - Other DTNs, or random machines?

EPOC
Engagement and Performance Operations Center



EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- ***Integration With Users***

EPOC
Engagement and Performance Operations Center

Outline

- Problem Statement & Expected Outcomes
- Operating Environment
- Hardware
- Software
- Configuration & Tuning
- Management & Maintenance
- ***Integration With Users***
  - ***"What Do You Need" vs. "What/How/Why Are you Doing"***

EPOC
Engagement and Performance Operations Center

Integration – User Engagement

- Typical IT approach to data movement:
  - "What do you Need?"
  - This assumes that:
  - A) the user is knowledgeable about several things:
    - Available tech options (hardware, software)
    - That the way they want to solve the problem is the correct/accepted way
    - That what they are doing today will scale to how things change in the future with their science/workflow
  - B) the user can make a 'snap' decision to give you a concrete action
  - C) there is enough trust that they would answer your question to begin with
- We propose a different set of questions: "What/How/Why Are you Doing?"

Integration – User Engagement

- ”What Are you Doing?”
  - How is data movement handled today?
  - Is it working?
- ”How Are you Doing?”
  - Can the workflow be fully articulated?
  - What pieces are the bottlenecks?
  - What things just don’t work?
- ” Why Are you Doing?”
  - Is this an intermediate step to do something else?
  - Does it scale for the use case that is defined?

E.g. do the social part of this first, then figure out the tech part

EPOC
Engagement and Performance Operations Center

Conclusions

- Hardware Changes since ~2012
  - New features
  - Reduced cost, improved reliability
- Use case is still the same: safe/secure/easily integrated into scientific workflows
  - Once you understand the workflows – do this first
- Tempting to ‘build it big’, ask you to consider ‘scope it right’

EPOC
Engagement and Performance Operations Center

http://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/

CC* Data Movement Workshop and Exhibition – Data Transfer Hardware

Jason Zurawski, Eli Dart
zurawski@es.net, dart@es.net
ESnet / Lawrence Berkeley National Laboratory

Dr. Jennifer M. Schopf
jmschopf@indiana.edu
Indiana University International Networks

CC/CICI PI Meeting Pre-Workshop*
September 22nd 2019

National Science Foundation Award #1826994

https://epoc.global