Scientific Computing Operations Updates

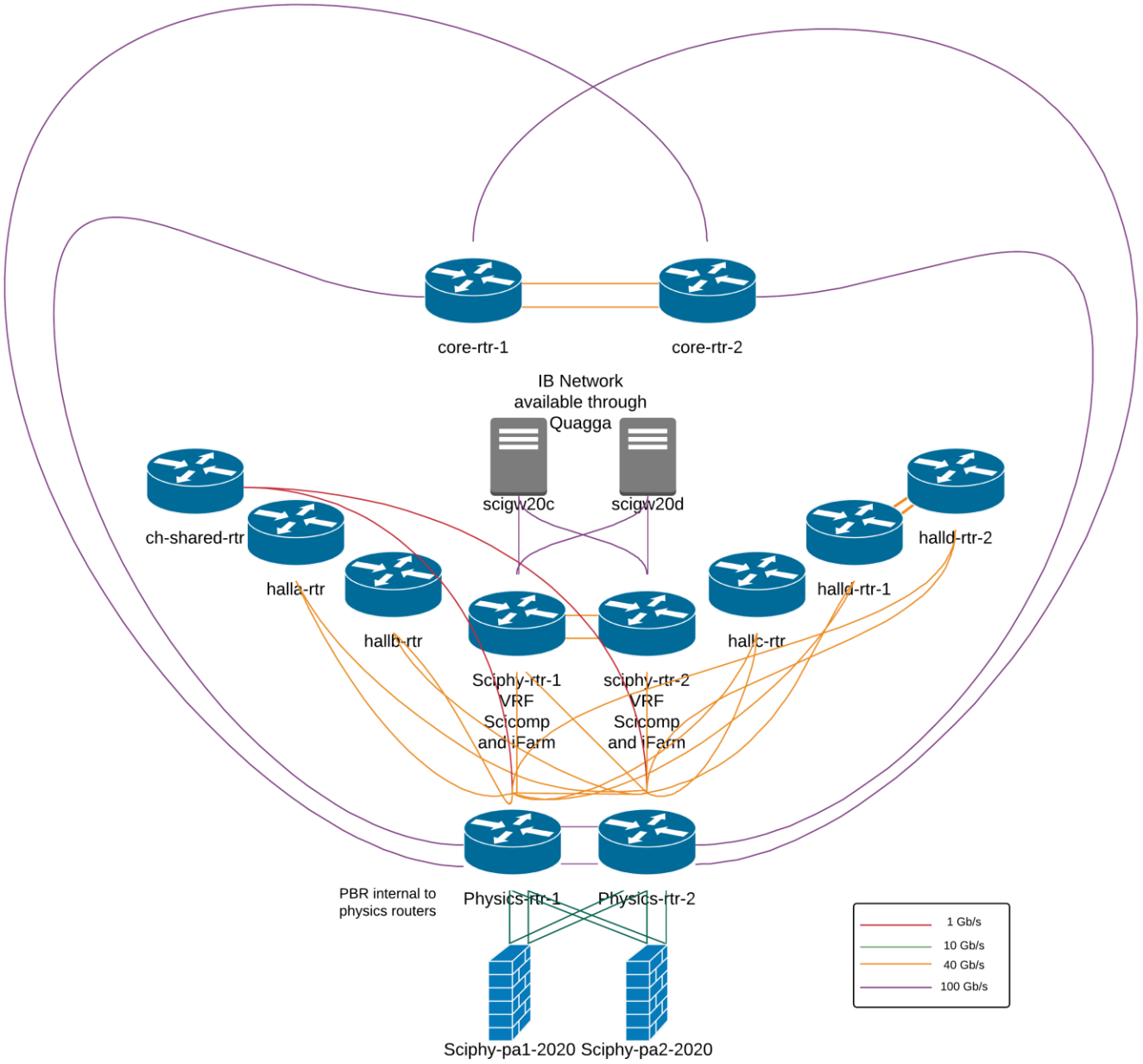
March 2021

Bryan Hess

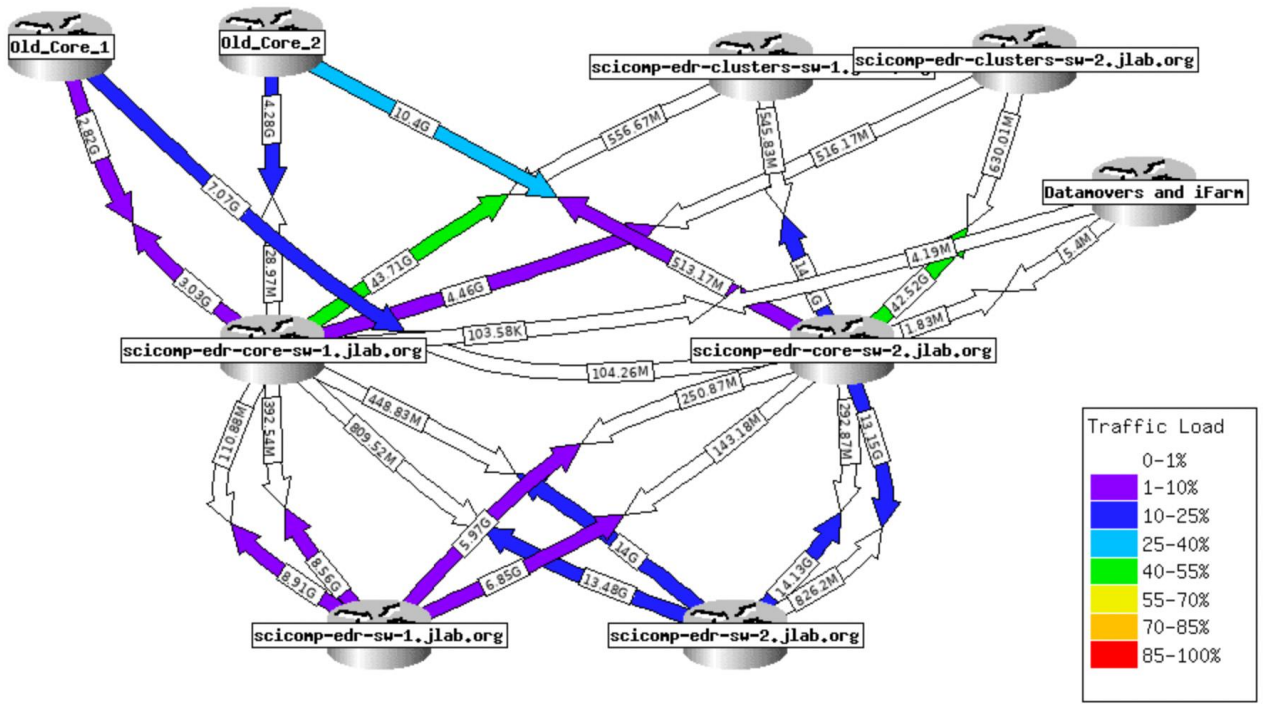
Networking Updates

Network Re-Design

- Changes (in progress) for Scientific Computing
- Separation between Online and Offline at the Network layer with physics-rtr and sciphy-rtr
- Facilitates better traffic management and Cybersecurity policy/segmentation
- Adding two routers between Ethernet and InfiniBand
  - LNET – limited non-IB on access to Lustre
  - BGP – limited non-IB access to IPoIP hosts (simplifies some cases)
- Slide credit: Carl Bolitho, Networking

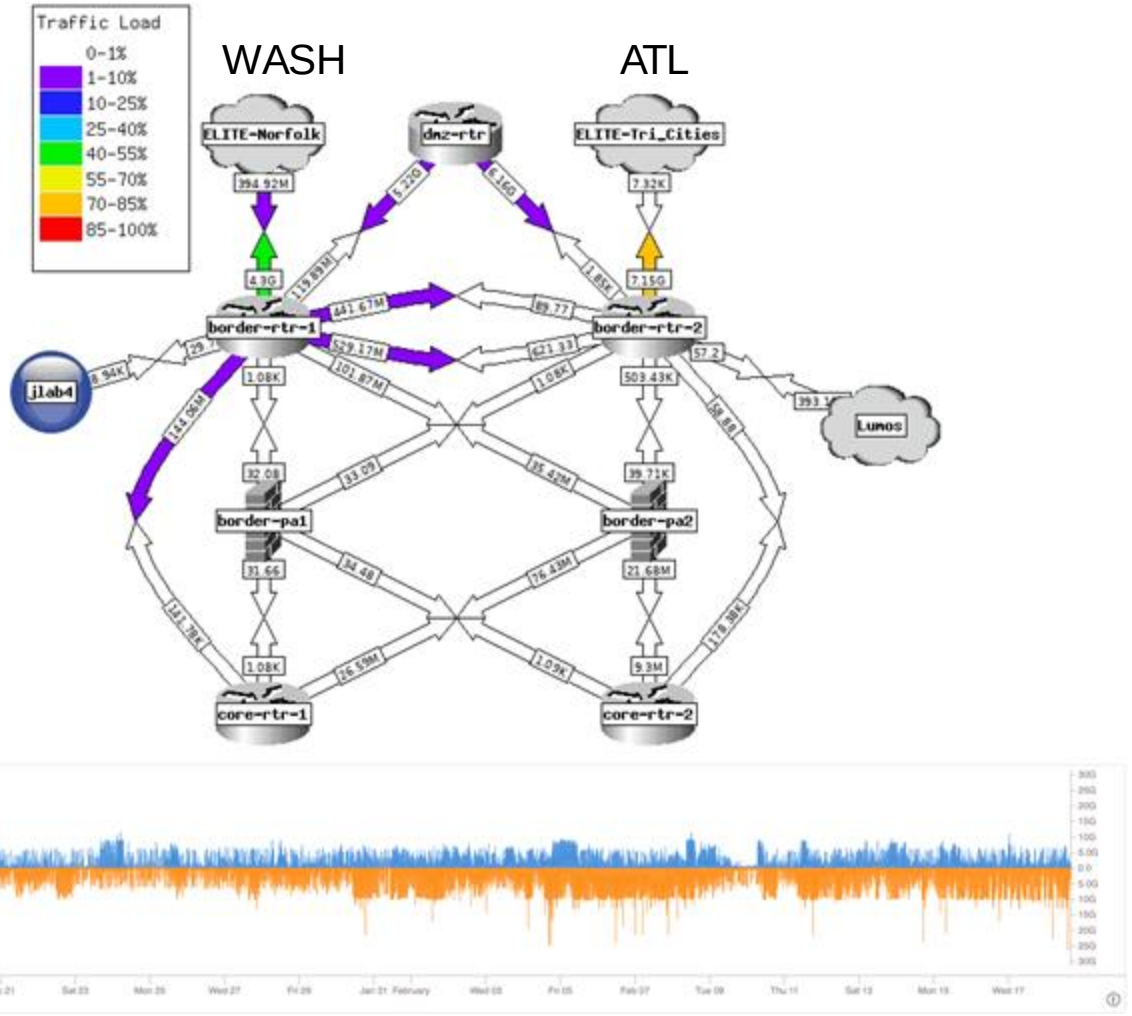


InfiniBand: Managed EDR Core & Aggregation



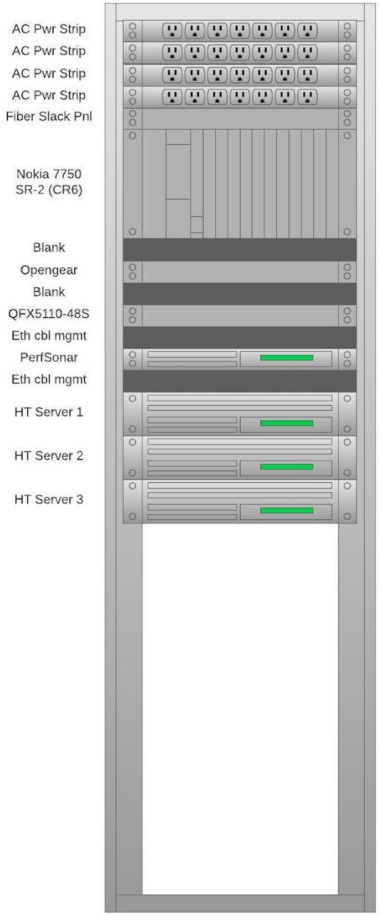
- IB Now Integrated into Network Management
- We now know things like---
  - When important links go down
  - When links are saturated
  - When links have errors
- "Old Core" was unmanaged QDR

Science DMZ Routing Policy



- JLab has two optically diverse 10Gbit links to ESNet at Atlanta and Washington.
- A Science DMZ network outside the firewall facilitates OSG traffic flows
- Science traffic began to regularly saturate the Atlanta link in early 2021.
- BGP traffic steering policies were used in coordination with ESNet to dynamically redirect certain well-known flows (NERSC, UConn) to achieve >10Gbps aggregate
- Looking ahead, ESNet 6 will provide 100Gbit/sec capability, removing the need for flow balancing

ESNet 6: 100Gbit for JLab



- We frequently exceed 10Gbps in aggregate
- With the ESNet 6 deployment, we will have 100Gbit/sec, with optical switching capability
- We are in the installation planning phase now
- Recent network upgrades laid groundwork for this
- Full 100Gbit/sec path from scicomp ethernet core to Science DMZ to internet

Storage Updates

Lustre: /cache and /volatile

- Lustre is self-supported, open source software, and therefore requires in-house expertise, which we are working to deepen
- Recent user-visible issues have been in three categories
  - Hardware failures (items replaced under warranty) - with failover pairs, this is less problematic that it once was
  - Software bugs - working to develop a testbed system this spring/summer for testing software changes. Lustre+Kernel+ZFS are tightly coupled.
  - Configuration and monitoring issues – we are formalizing monitoring (e.g. automatic monitor to make sure failover software is in a good state; scrub disks; check for avoidable hangs)
- Recent hires bring Lustre expertise

New ENP Work Fileserver

- New work file servers are in procurement, should be awarded soon.
- These will be a pair of ZFS file servers in an active/active failover pair
  - 2 shelves, 14TB * 60 drives each = 1.68PB raw
  - After overhead for ZFS and shapshots, ~1PB user visible
  - Network is redundant EDR/100Gigabit
- New features to implement (if they prove stable)
  - SSD read cache layer
  - SSD write cache layer
  - User-accessible snapshots

IBM LTO8 Tape Issue

Tape Issue End Game

- Bracketing the problem:
- Ongoing Recovery Measures:
  - 28 tapes shipped to IBM for recovery
  - 10 returned, reinserted, still to be reconciled
  - List of Purged Files to be Determined
- Data Integrity Measures:
  - Wrote 629TB of "Just In Case" duplicates of all new on 58 new LTO tapes, now starting the verify pass.
  - Read-back of LTO8 tapes written from July – December ongoing
- We owe you lists of affected files; Short staffing (read Jury Duty) has slowed this.

Hardware

Farm Changes

- Sixteen more "farm19" era nodes are in receiving. Purchased with Theory funds. FairShare will be adjusted to reflect this
- Hybrid Farm21 node purchase coming – specification being developed now
  - Likely to be AMD Rome again
  - Similar CPU:Memory:Disk space balance
  - Add one GPU for some farm nodes. Some for inference, a few for training
  - Moving to Data Center class NVIDIA GPUs.
  - Expect to finalize this month
- Aside: We support GPU jobs in slurm, and Just added JupyterHub GPU support in certain notebooks.
- Farm{13,14} are now in "Do Not Resusciate" mode. We will remove them as they fail, but meanwhile they continue to provide cycles

Software

OSG "flow back" jobs

- We now allow OSG jobs for the JLAB VO to run on the farm, filling in space as available
- We have several controls for this
  - Fairshare
  - Slurm QOS for OSG jobs
  - OSG submitter limits
- Although this is a modest number of cpu hours, it keeps all the machinery in running order

The ifarm "slowness" problem

- Slow interactive response seems to comprise several layers
- Done – peeled layers
  - Replace the overloaded firewall
  - Solve NIS dropped packet issue to ypserver
- In progress – to be peeled
  - Credential caching: need nscd or sssd or cache getent() lookups. Complication is that password hashes are now in ldap, but netgroups remain in NIS which can cause cache flush/reload at inopportune times. Symptom: "ls –l" latency varies widely.
  - There appears to be a packet lost/congestion issue with ESC Server NICs that is subtle, but affects UDP flows in particular. Working on that with Systems and networking. We rely on UDP-encapsulation for certain traffic over VXLAN.

Miscellanous Notes

- Scicomp-2020.jlab.org continues to add features. Feedback to ychen@jlab.org
- Swif2 needs testers, feedback to larrieu@jlab.org
- GitHub plan remains in place. Information from Microsoft about changing plans is muddy at best.
- No clarity yet for the follow on to CentOS 7, but it is an active topic of discussion at many labs.