DC24 Site Perspectives: North America

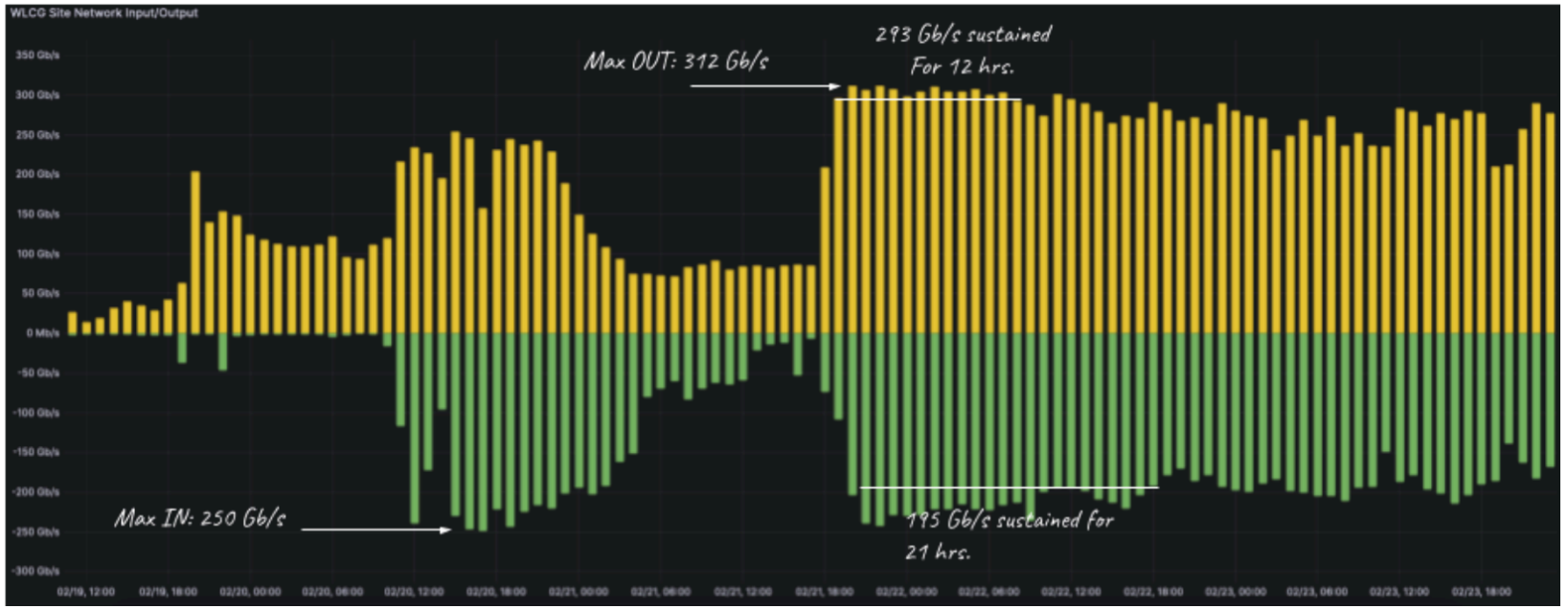
Ofer Rind

WLCG/HSF Workshop

DESY, Hamburg, May 14th, 2024

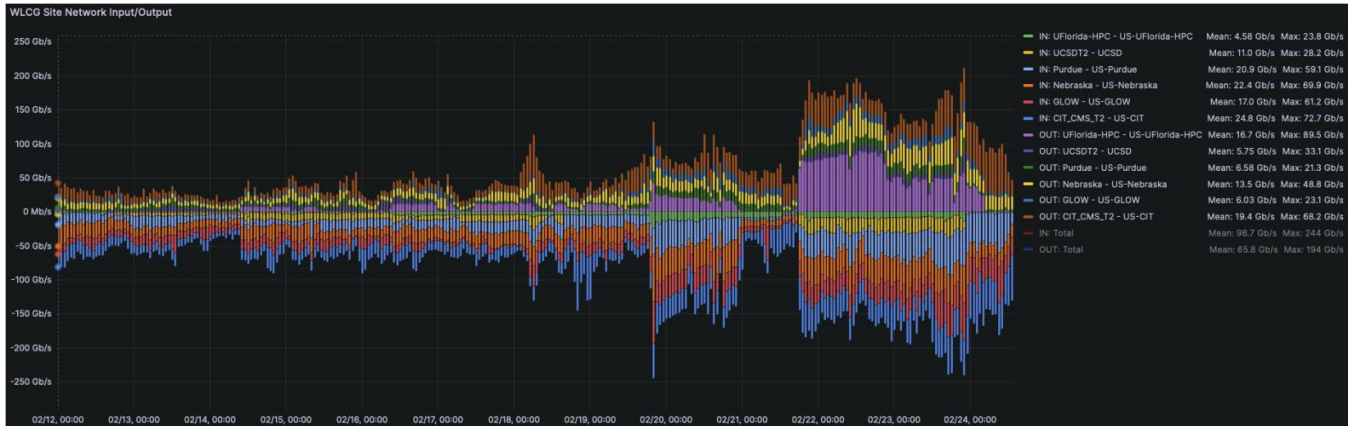
US CMS Tier-1

- Successful exercise: max rates achieved at FNAL on the last day were 248 Gb/s write (target: 209), 312 Gb/s read (target: 299)
- Sustained rates
  - 195 Gb/s write for 21 hrs
  - 293 Gb/s read for 12 hrs



US CMS Tier-2

- Eight Tier-2 sites in the US, all were involved in six different scenarios after day 3 of the challenge
- All sites were able to achieve throughput in excess of original injection rates (x1.3-4.8 write, x2.1-4.9 read)
- On Day 12 of the challenge, US CMS requested a large injection increase
  - No sites were able to reach the 100 Gb/s bi-directional target (Max write 89.5 Gb/s Florida, Max read 68.2 Gb/s CalTech)
- 4 sites enabled support for scitokens (Caltech, Florida, UCSD and Wisconsin)



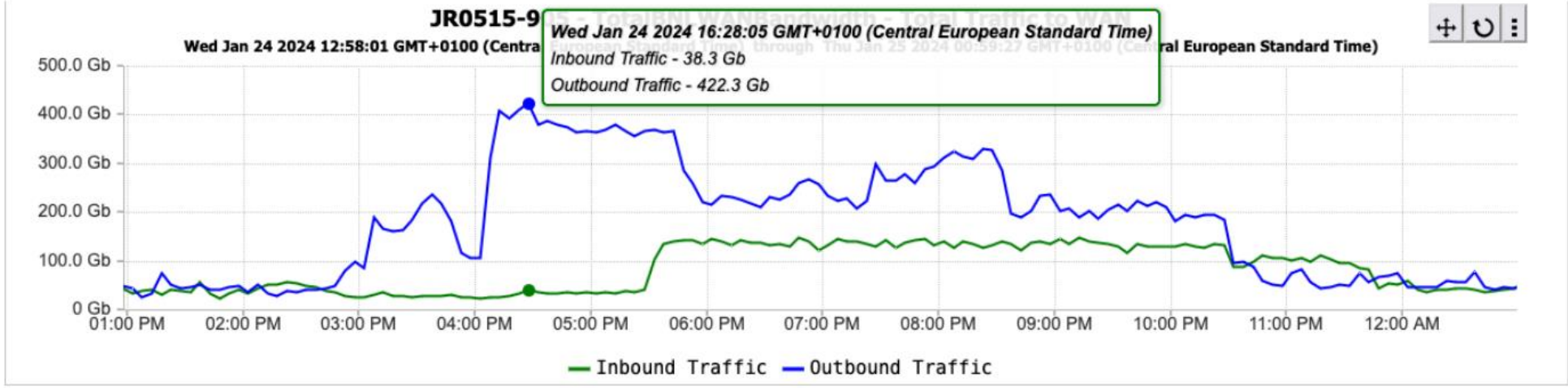
US T2s except MIT and Vanderbilt

US ATLAS Tier-1

- Prior to DC24, BNL network capacity was updated to 2 x 800 Gb/s, with 800 Gb/s available on both LHCOPN and LHCONE
  - 800 Gb/s WAN capacity for ATLAS dCache DTNs (plus 75 Gb/s Belle-II, 100 Gb/s DUNE)
  - BNL target rates were very low in comparison and there were no site network or storage bottlenecks observed
- BNL and the US Cloud were served by the CERN, rather than BNL, FTS instance during DC24 to make use of the recently updated storage token capability
- The US Cloud ran multiple pre- and post-DC24 tests, including a Joint US ATLAS-CMS stress test, using a test suite developed by Hiro Ito.

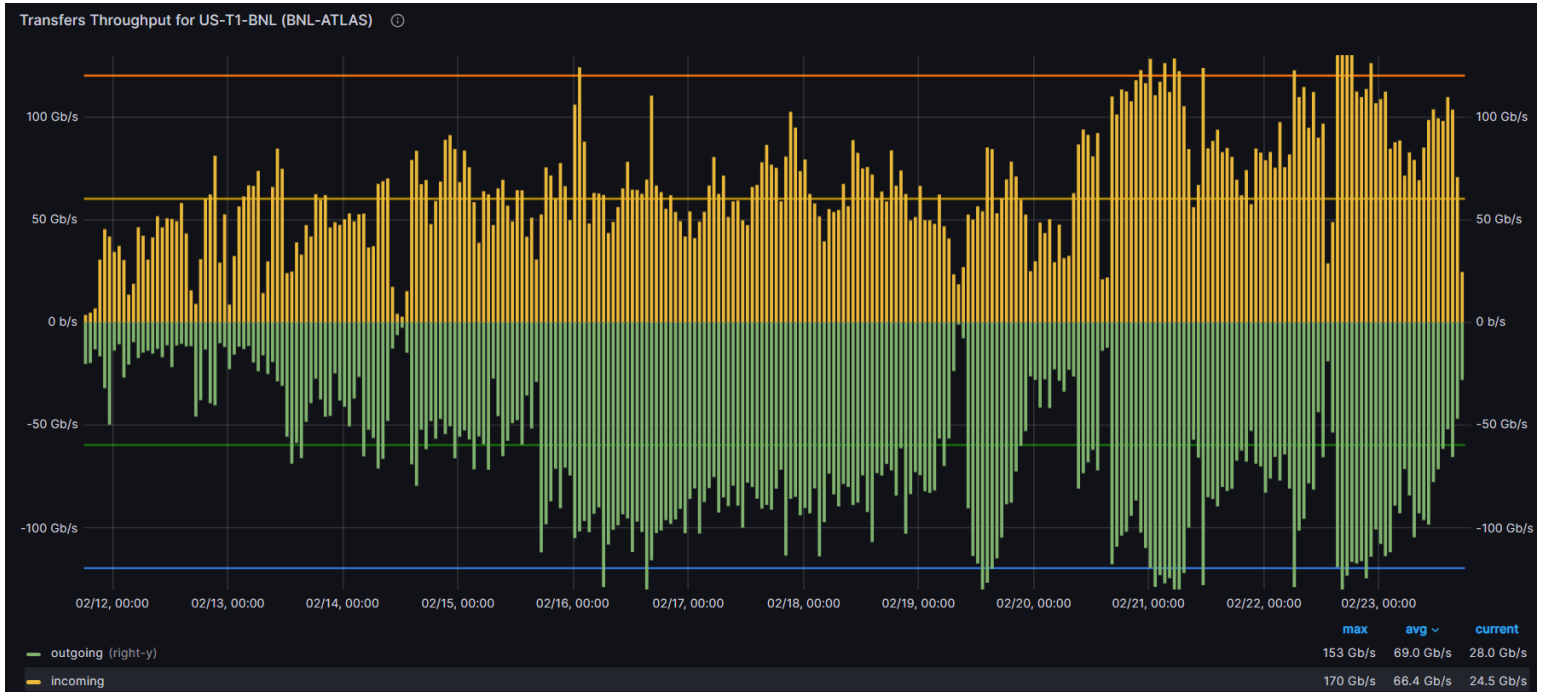
US ATLAS Pre-DC24 Tests

- Test full end-to-end transfer capacity limits at US sites – establish a baseline
- Identify bottlenecks, misconfigurations; tune storage parameters
- Joint test with US CMS – identify points of network contention
  - Involved BNL, Michigan, Chicago for ATLAS

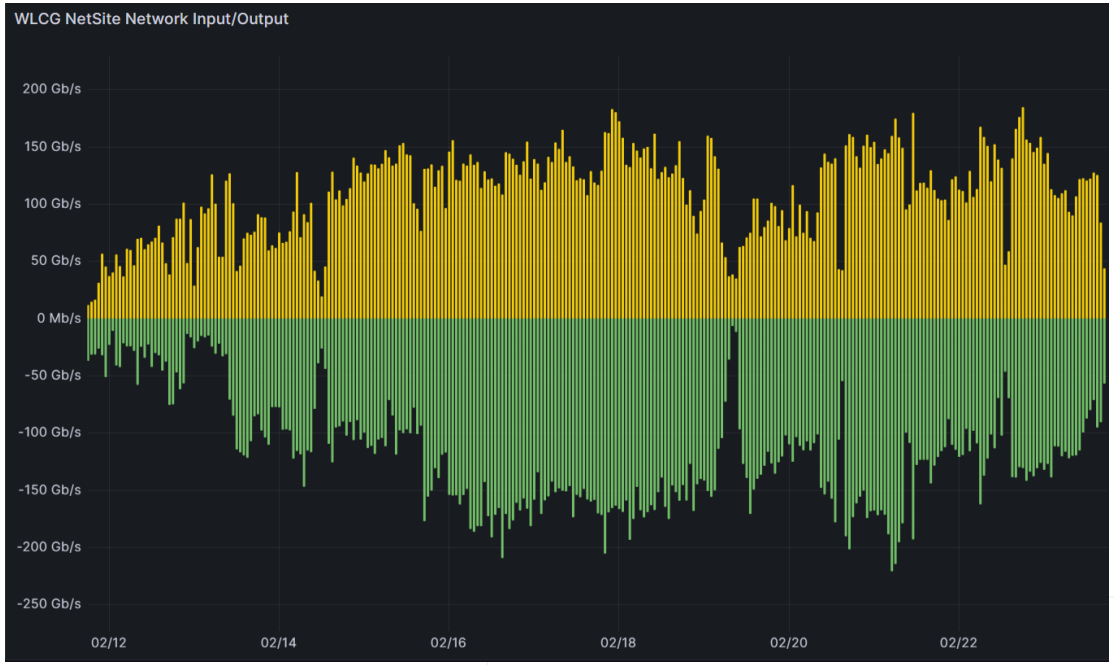


US ATLAS Tier-1

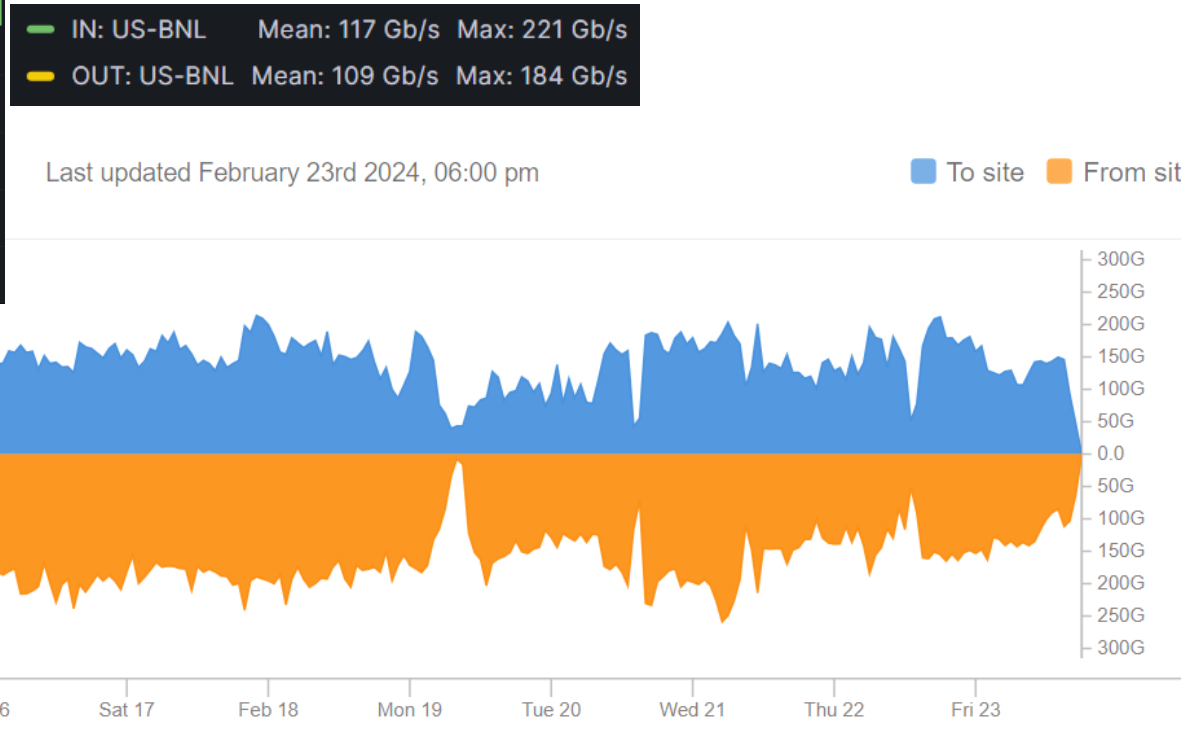
- WLCG FTS monitoring indicates that BNL was able to exceed the 48 hr average minimal, but not flexible, scenario targets during DC24
- This was not a site limitation, as evident from pre-DC24 exercise



US ATLAS Tier-1

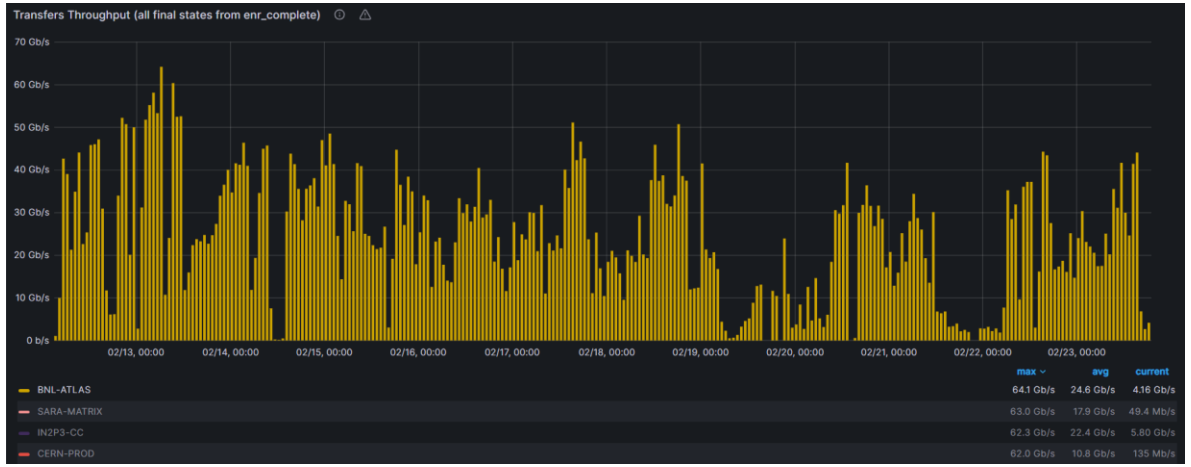


ESNET and WLCG Site Network Monitoring agree reasonably well with each other, and indicate higher rates than FTS monitor



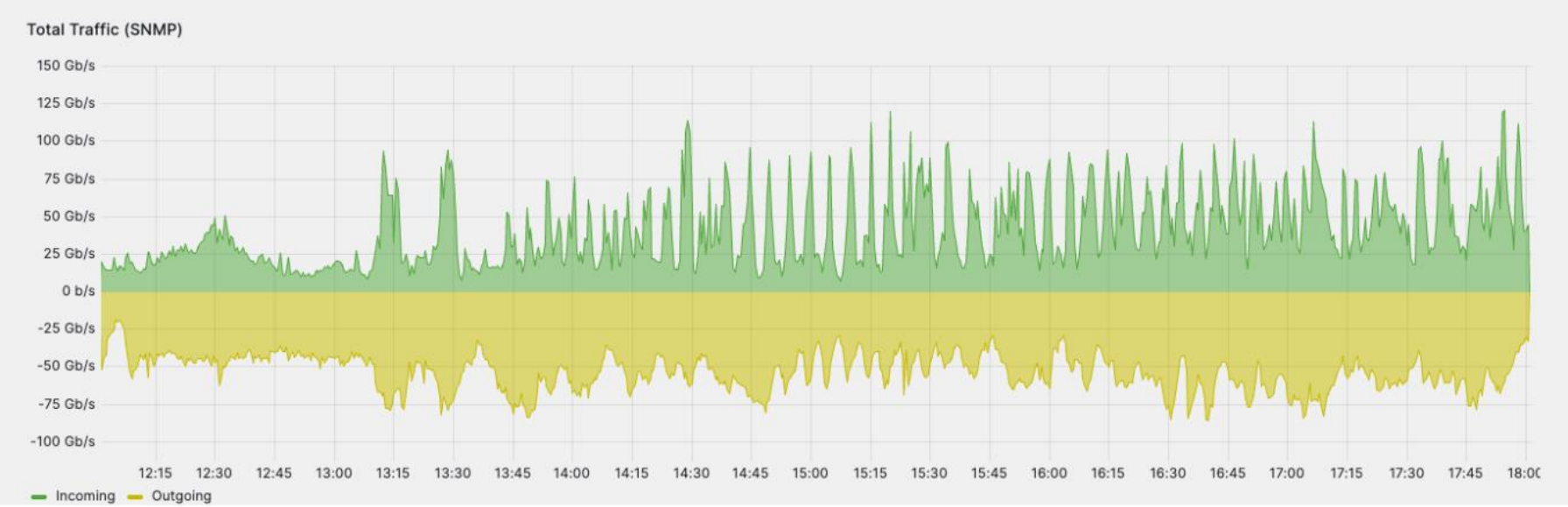
T0 Export Tests to US ATLAS Tier-1

- T0 export to BNL reached a max of 64 Gb/s, highest among all T1s
  - Briefly exceeded 60 Gb/s target but not sustained



- Retested post-DC24 as part of dedicated T0 export test to each T1
  - Max export rate of 81.3 Gb/s
  - 62 Gb/s average (target was 68.4)
  - Sawtooth pattern evident in both DDM and site network monitoring, not understood

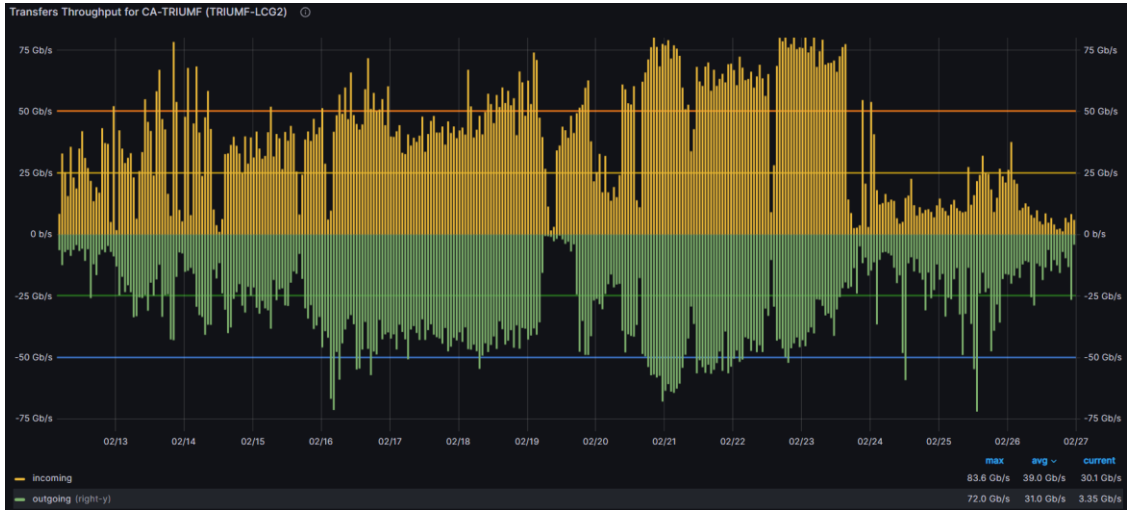
Transfer Structures



- Transfer rate structures observed during DC24
  - Visible in higher resolution ESNET monitoring
  - Data injection and transfer finished quickly before next cycle – artifact of the injection tool

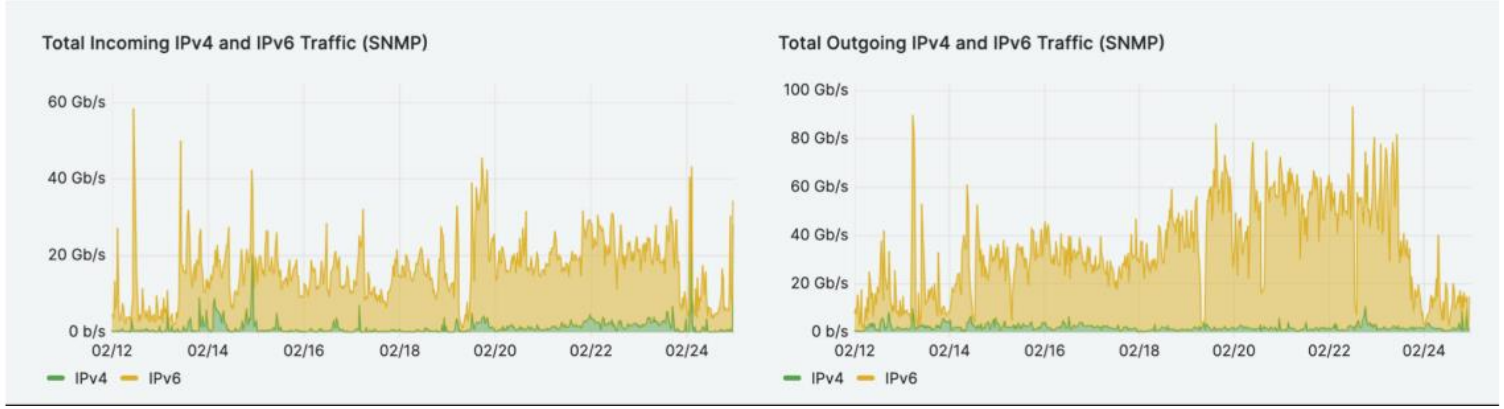
Canadian ATLAS Tier-1

- TRIUMF-LCG2
  - Successful, met throughput goals
  - No network bottleneck – new 400G link
  - Token auth requirement reduced available mover concurrency (only external webdav doors)
  - High load on namespace due to deletion rate
  - Tape system experienced knock-on effects (high load, failed transfers, leftover FTS pins, filled buffer) – follow ups planned



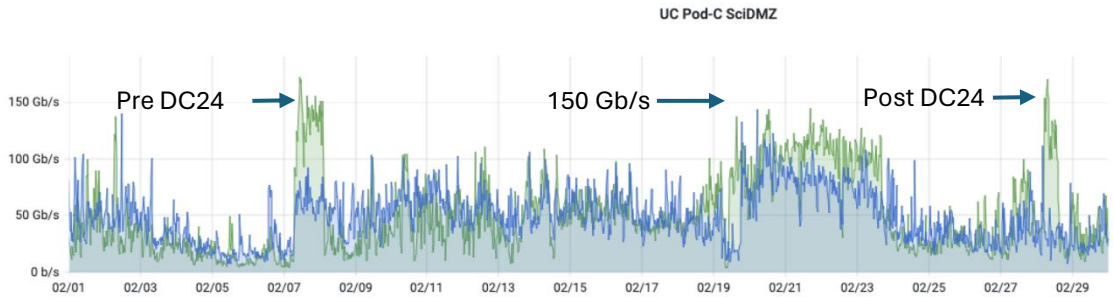
US ATLAS Tier-2 – AGLT2

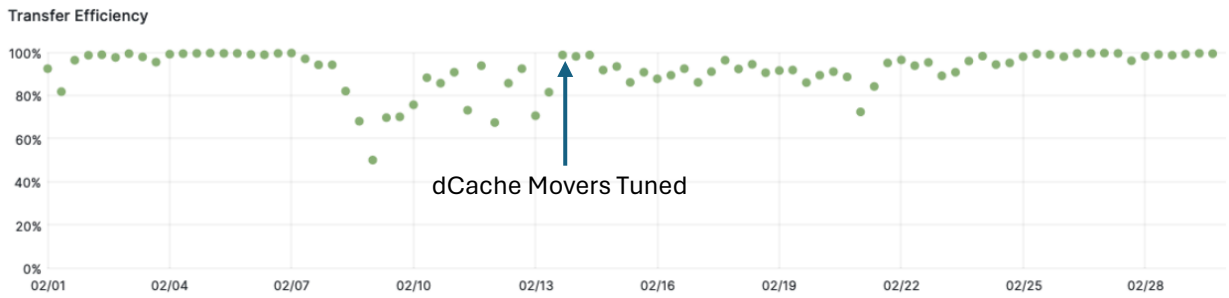
- Network and site in general were not stressed during DC24
- Pre-DC24 stress testing had a larger impact
  - Bandwidth limited due to known 2x40G bottleneck between AGLT2-UM and UM Campus – to be removed in network upgrade, possibly by end of 2024
  - Some older storage servers with dual 60x8T drive shelves got overloaded – demonstrated validity of already existing plan to downsize to single shelf
- Plan to test flow labeling (when dCache is ready), packet marking (in EL9), and network utilization optimization



US ATLAS Tier-2 – MWT2

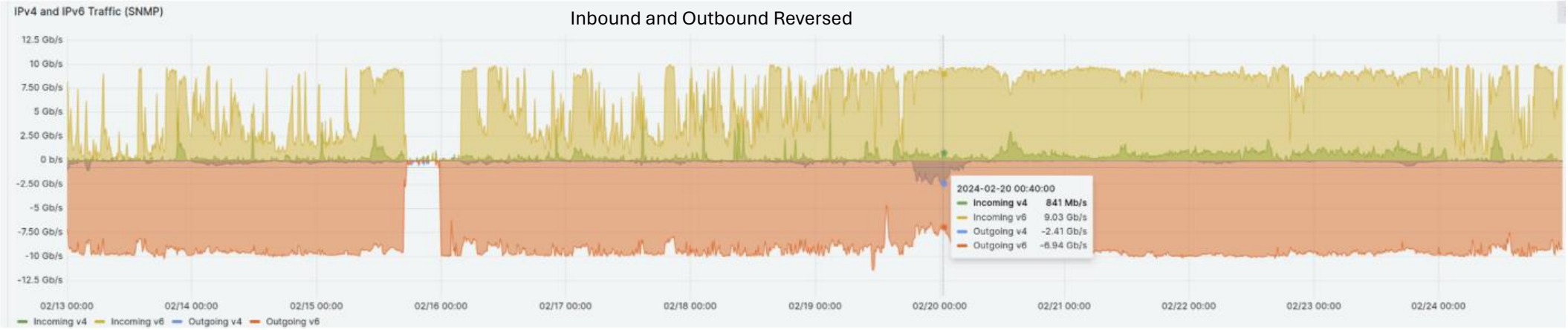
- Site was not stressed overall during DC24 save for some older high density storage servers
  - Network throughput peaked ~150 Gb/s (UC site had 200G connectivity)
  - I/O issues on older high density MD3460 pools led to high rate of inbound transfer failures early in DC24
    - Fixed by reverting number of dCache movers per pool to default 100 (from 2000)
    - Older storage scheduled for replacement soon
- There were concerns when site fell below 1% free disk space, but deletion was fast enough to keep the site from filling completely





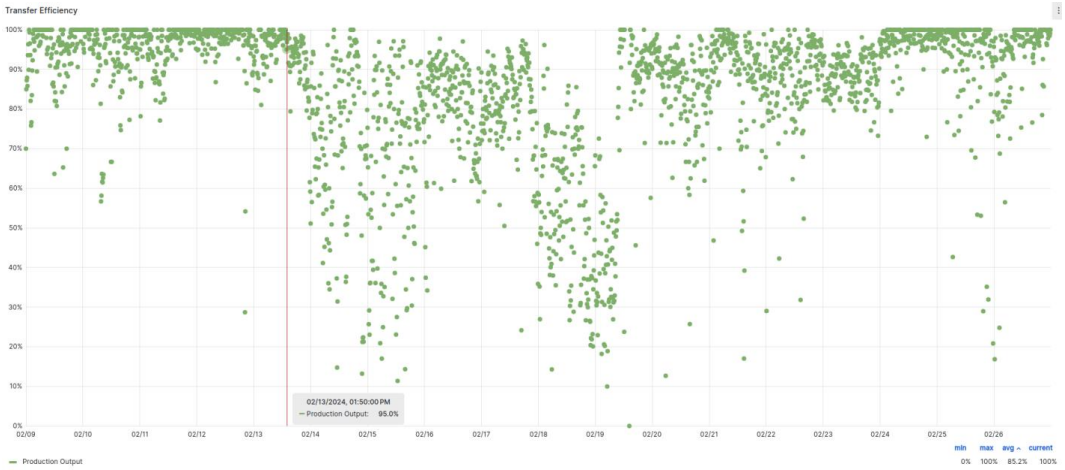
US ATLAS Tier-2 – NET2

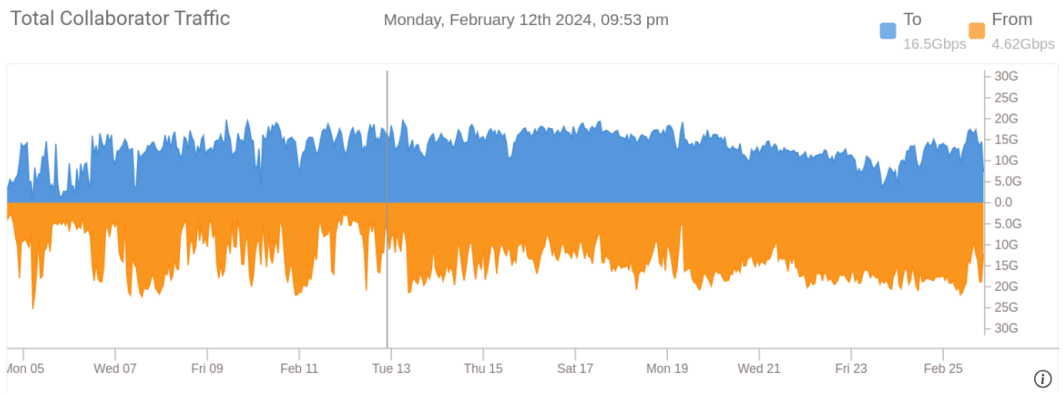
- Relatively new Tier-2 with only 10 Gb/s connection to LHCONE
  - Incoming transfers saturated the connection for the entire DC24 period
  - Low transfer efficiency at times, especially when site transfer concurrency limits were increased – these limits should take bandwidth into account
  - Network upgrade to multiple 200G connections planned for this Summer



US ATLAS Tier-2 – SWT2

- UTA experienced transfer efficiency problems and was unable to reach its network throughput limit
  - Attributed to FTS limitations and insufficient deletion rate (90-95% watermark mismatch for higher throughput)
- Transfer rates at OU were capped at 20 Gb/s due to dual-25G NIC link aggregation issues
  - Old Xrootd storage servers also experienced issues, causing transfer failures – these will be replaced 100G DTN, ceph-based storage this Summer





Summary

- DC24 was successful in identifying some bottlenecks along with software/hardware issues that need to be addressed and monitoring that needs to be understood
- Most target rates were achieved and results will be instrumental for resource planning
- A successful program of US Cloud stress testing proved useful and will likely continue in the interim between challenges

Thank you to the many people who contributed the information and analysis for this presentation!