Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

Throughput Load Testing for Data Challenges in USATLAS

Hironori Ito

Brookhaven National Laboratory

@BrookhavenLab

Preparing Sites for Data Challenge

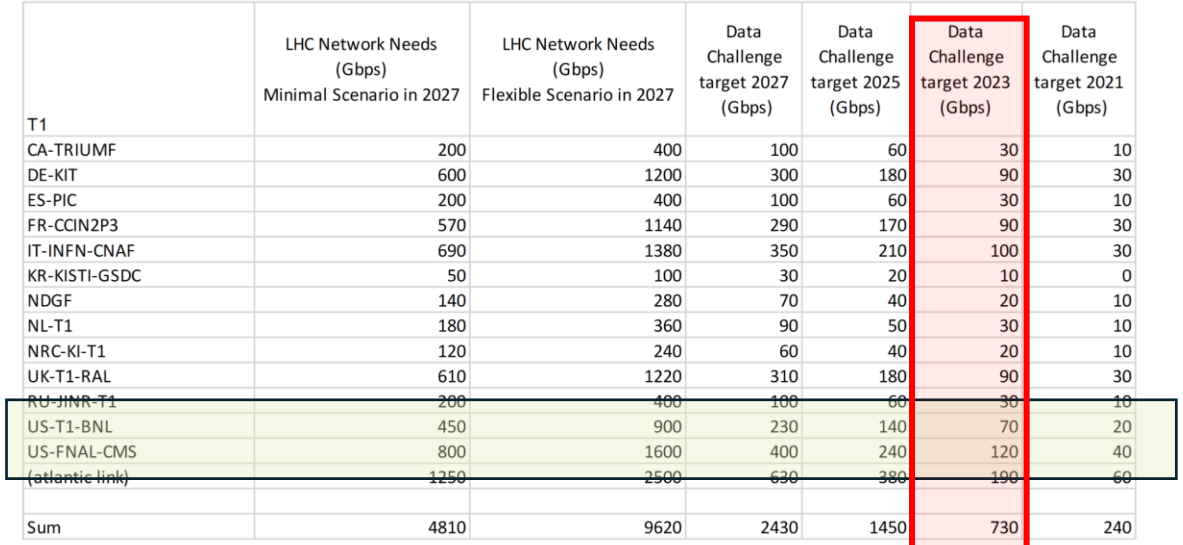
- Data Traffic between sites is expected to increase significantly for HL-LHC
- WLCG has plans to test the preparedness of the participating sites for higher data demands in every few years before the start of HL-LHC
- To help prepare sites for these tests, more frequent evaluations are seen as useful.

- More frequent tests will provide
  - The existing, production capabilities of data services at sites.
  - Highlight the bottleneck if any.
  - The information for site admins and mangers to identify the issues and improve if necessary.

These tests are not meant for criticizing sites. But rather, the sites should use the information to improve the data throughput capabilities of the sites.

Original plan before the revised LHC schedule

WLCG data challenges for HL-LHC - 2021 planning



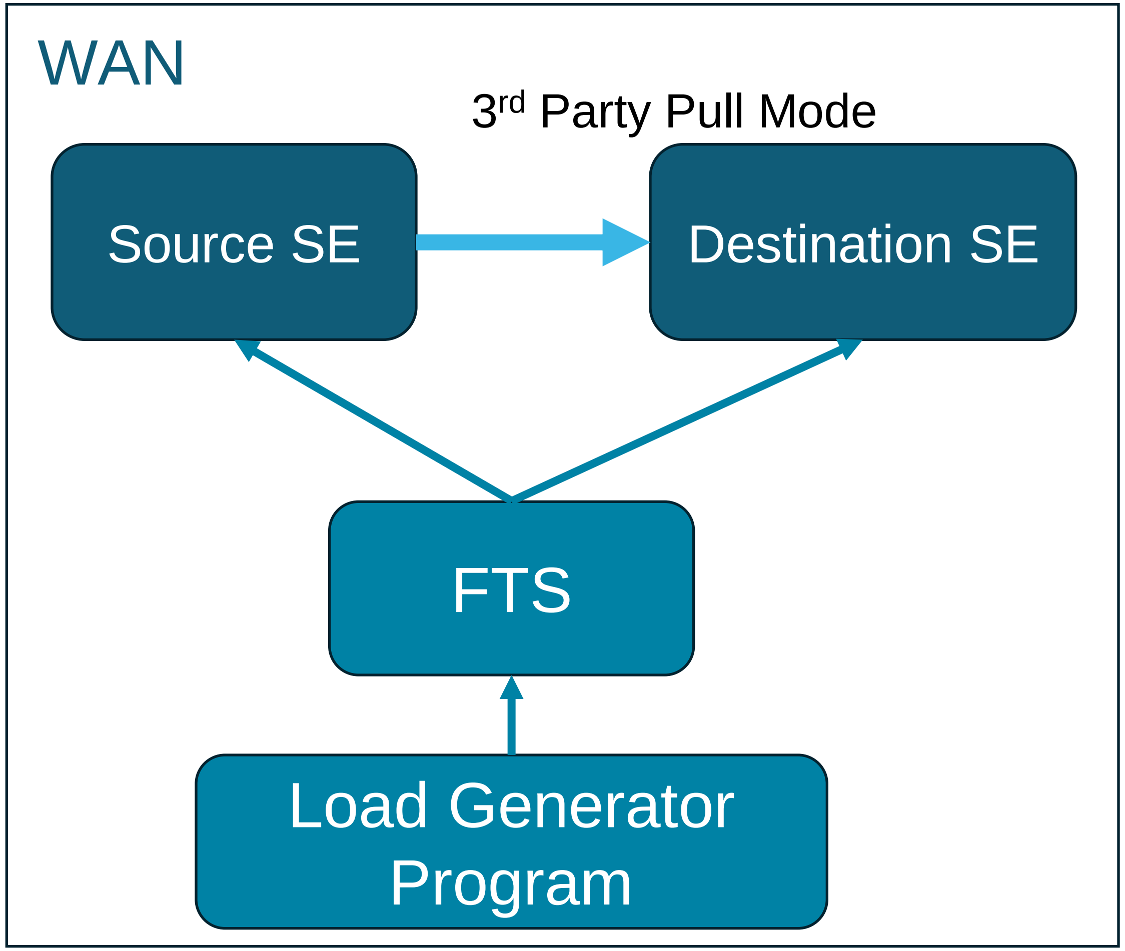
T1	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027	Data Challenge target 2027 (Gbps)	Data Challenge target 2025 (Gbps)	Data Challenge target 2023 (Gbps)	Data Challenge target 2021 (Gbps)
CA-TRIUMF	200	400	100	60	30	10
DE-KIT	600	1200	300	180	90	30
ES-PIC	200	400	100	60	30	10
FR-CCIN2P3	570	1140	290	170	90	30
IT-INFN-CNAF	690	1380	350	210	100	30
KR-KISTI-GSDC	50	100	30	20	10	0
NDGF	140	280	70	40	20	10
NL-T1	180	360	90	50	30	10
NRC-KI-T1	120	240	60	40	20	10
UK-T1-RAL	610	1220	310	180	90	30
~~RU-JINR-T1~~	~~200~~	~~400~~	~~100~~	~~60~~	~~30~~	~~10~~
US-T1-BNL	450	900	230	140	70	20
US-FNAL-CMS	800	1600	400	240	120	40
~~(atlantic link)~~	~~1250~~	~~2500~~	~~630~~	~~380~~	~~190~~	~~60~~
Sum	4810	9620	2430	1450	730	240

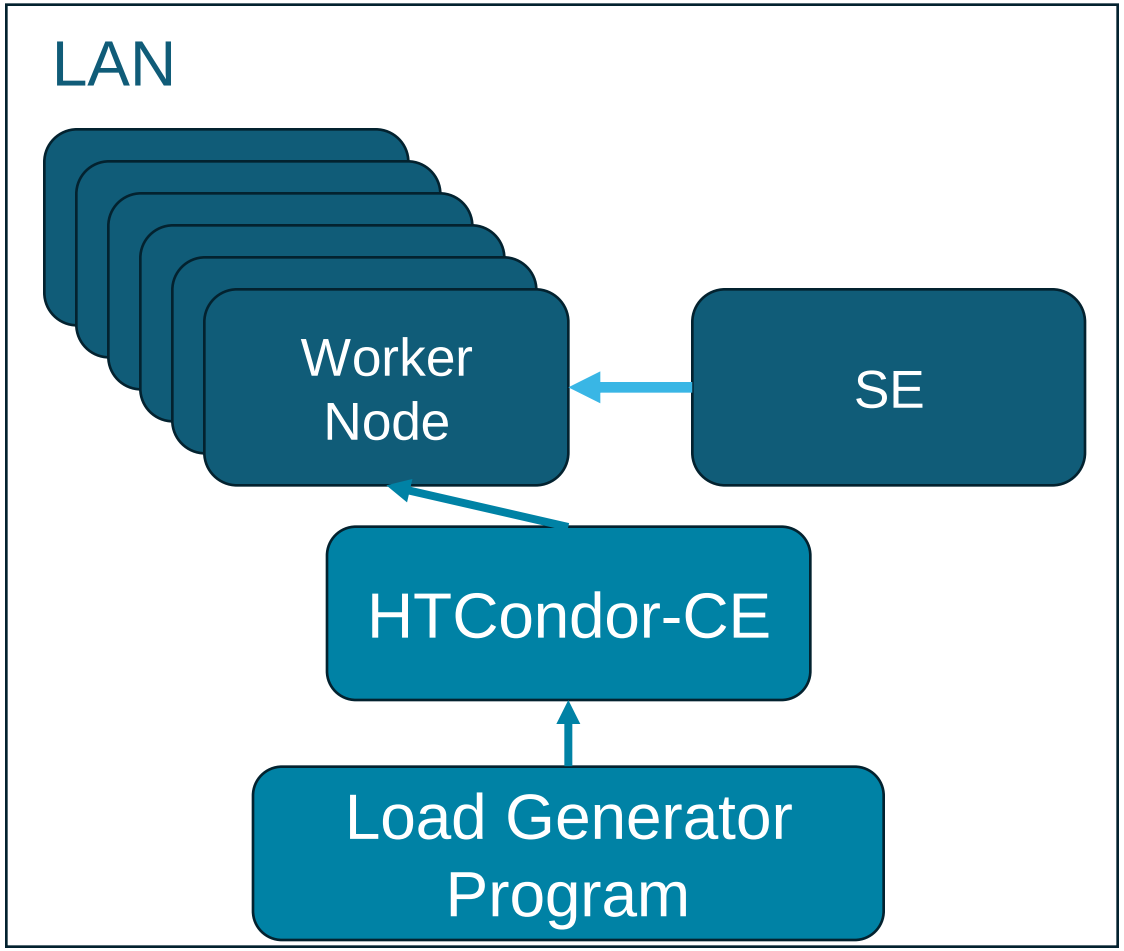
Revised according to new LHC schedule.
2nd Test on Feb 2024 with 25% of the target

Types of Load Tests

- WAN Throughput Load Tests
  - One site to one site data transfers
    - Relatively large size file ~3GB are sent using **FTS** between sites to achieve the high throughput
      - Do we need to test high transaction rate?
      - High transaction will likely have more negative impact to the performance of the existing, production storage than high bandwidth.
    - Multiple protocols; **Davs/https**, **XRootD**, etc…
    - Checksum off and on.
      - Some storage service might see impact on checksum
    - Identify the existing WAN Writes rates
      - Identify any site-specific limiting factors.
  - Simultaneous multi sites data transfers
    - Conduct transfers to/from multiple sites
      - Test both read and write. → Generate X2 load on the storage
      - Identify any issues between the sites.
- LAN Data Tests
  - Typically, the total LAN rate at a site is much higher than that of WAN. (can be factor of 2 or 3 higher)
  - They will increase linearly with the total available CPUs and WAN rate in general
  - To mimic higher IO condition in HL-LHC, we can run High IO jobs from the worker nodes.

Tests


WAN
3rd Party Pull Mode
Source SE
Destination SE
FTS
Load Generator Program


LAN
Worker Node
SE
HTCondor-CE
Load Generator Program

Plans for USATLAS load tests until the next WLCG Data challenge

Quarterly tests

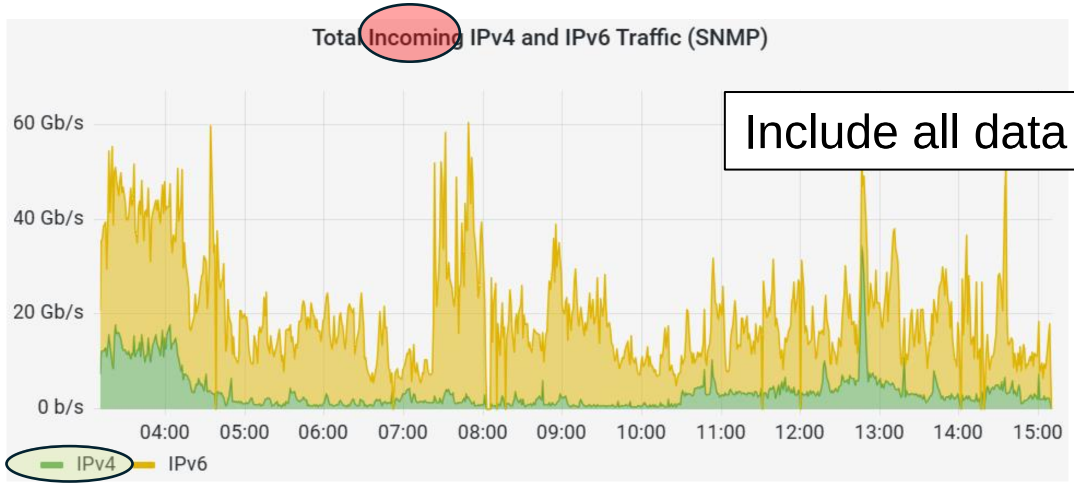
1. Spring 2023 (Done)
    1. WAN WebDAV test without checksum
    2. Setup and check monitors
2. Summer 2023 (in progress)
    1. WAN XRootD test
    2. WAN WebDAV tests with checksum
3. Fall 2023
    1. LAN Test
    2. Multi-sites WAN test
4. Winter 2023-24
    1. Multi-sites WAN and LAN test
5. Spring 2024
    1. **WLCG Data challenge**

Monitors

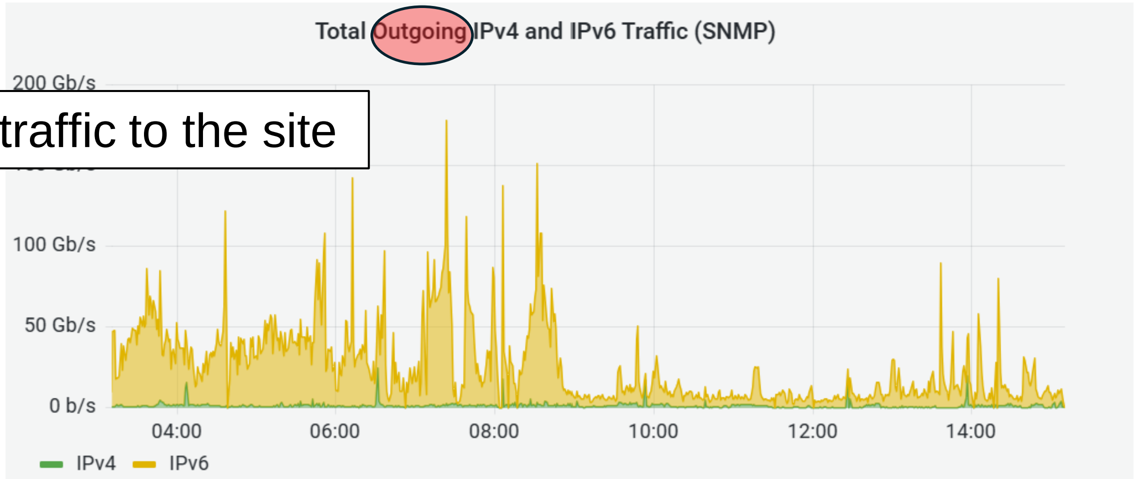
- ESNet monitor https://public.stardust.es.net/d/u5qX95N7k/lhc-data-challenge-sites
  - The monitor shows "in" and "out" in the reverse with respect to the site.
  - The monitor shows all traffic regardless of clients. It includes more than those from the data transfers.
  - US centric. Needs something for non-US sites.
  - It shows IPv4 and IPv6 separately.
- BNL FTS Monitor https://monitoring.sdcc.bnl.gov/pub/grafana/d/A4JjYk24k/usatlas-lhc-wan-write-throughput
  - It shows all FTS transfers to the target site from all SEs including own if the site has multiple SEs.
  - One can look at the specific source and destination pair.
  - Due to the time record used (unix time at UTC), the time is off (ahead) by 4 hours. One must input the right time (+5h as the current time)

Site Throughput Monitors

Outgoing

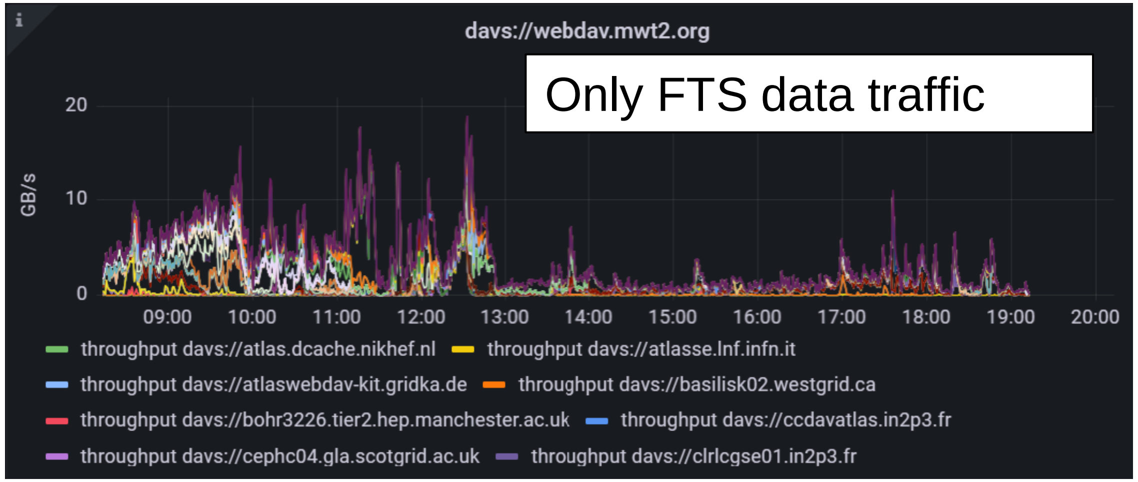


Incoming



Include all data traffic to the site

IPv 4 or 6: determined by the destination SE
(if both are dual stack)



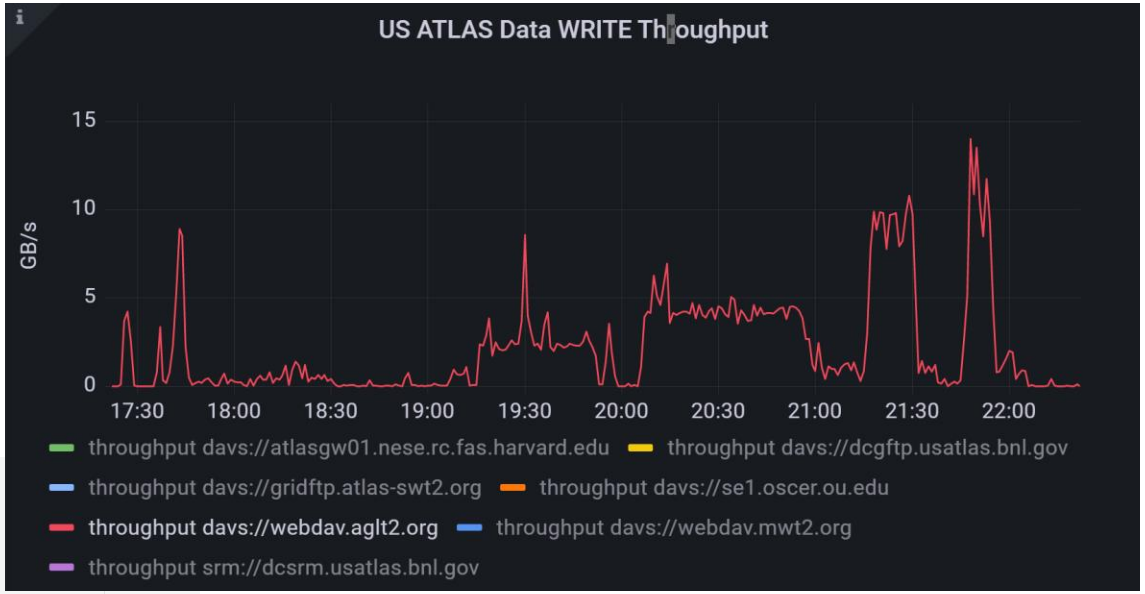
Only FTS data traffic

Checking Validity of Monitors and Load Generators

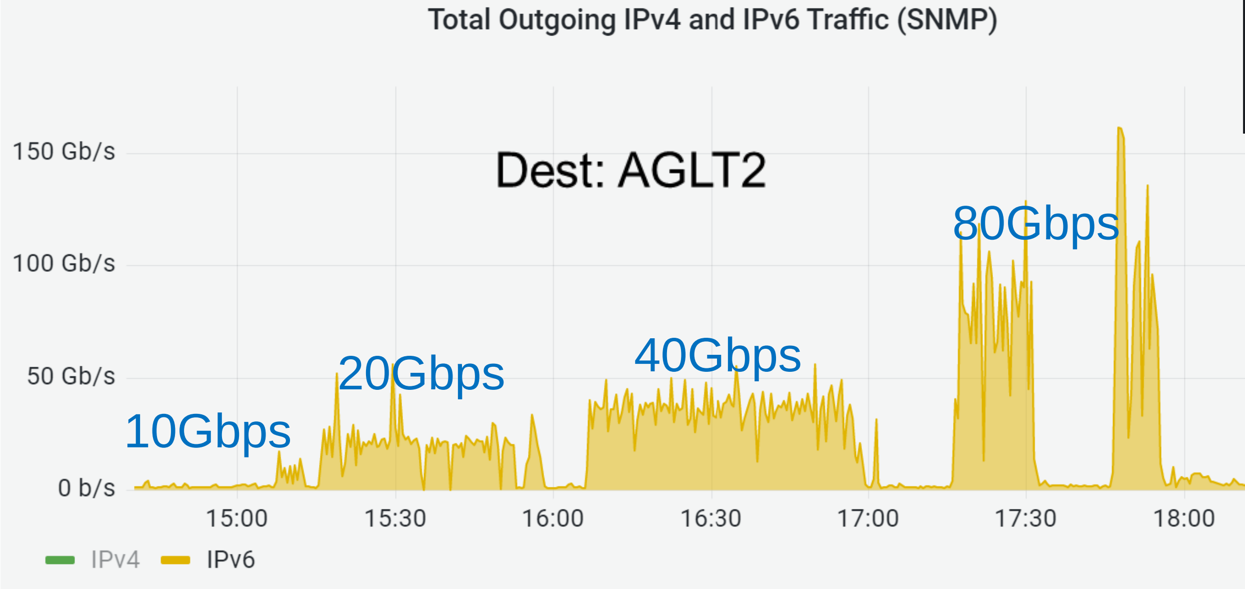
Both monitors shows the same level of the throughput.

Load generator can target the specific level of throughput.

FTS Monitor



ESNet Monitor

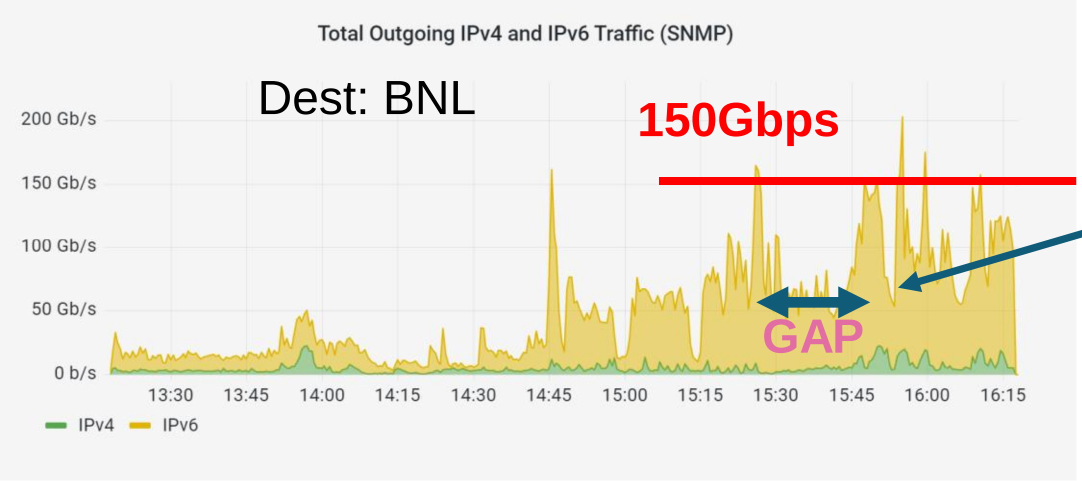


- FTS Monitor
- Can identify the load generator throughput without inclusion of the other data traffic
- No identification of IPv6 or IPv4 yet

- ESNet Monitor
- Doesn't distinguish the throughput by load generator from all the transfers to the site.
- Identify IPv4 or IPv6

Round 1. WebDAV WAN test

- Slow production transfers are taking up queues from load generation, resulting in drop of overall throughput.
- Could have adjusted the size of concurrent transfers to get maximum throughput.
  - Not a limitation of storage or FTS.
  - How do we address the slow transfers/sites?

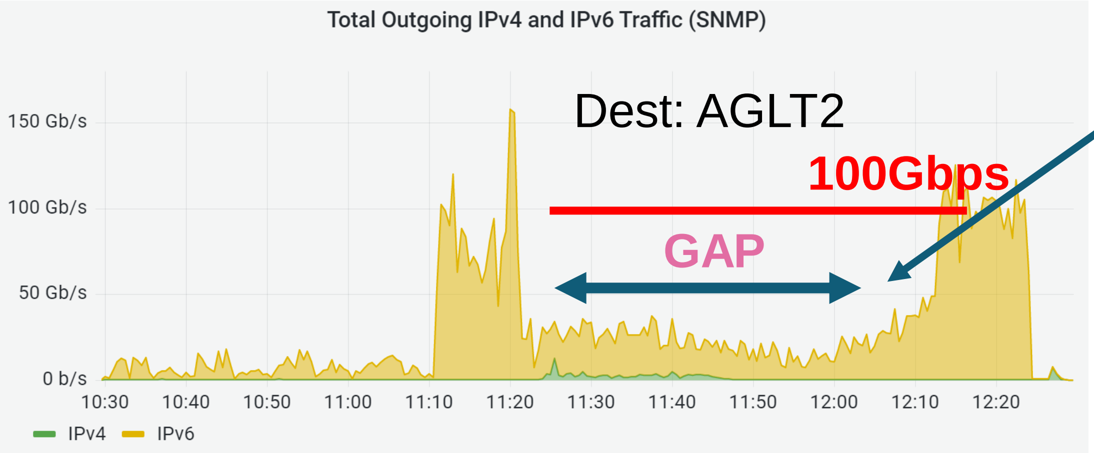


Total Outgoing IPv4 and IPv6 Traffic (SNMP)

Dest: BNL

150Gbps

GAP

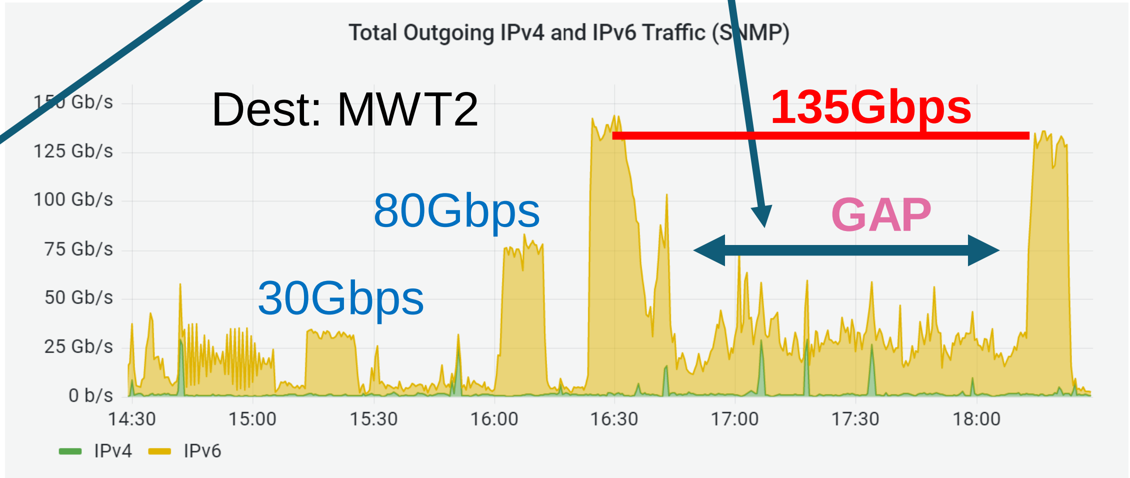


Total Outgoing IPv4 and IPv6 Traffic (SNMP)

Dest: AGLT2

100Gbps

GAP



Total Outgoing IPv4 and IPv6 Traffic (SNMP)

Dest: MWT2

135Gbps

80Gbps

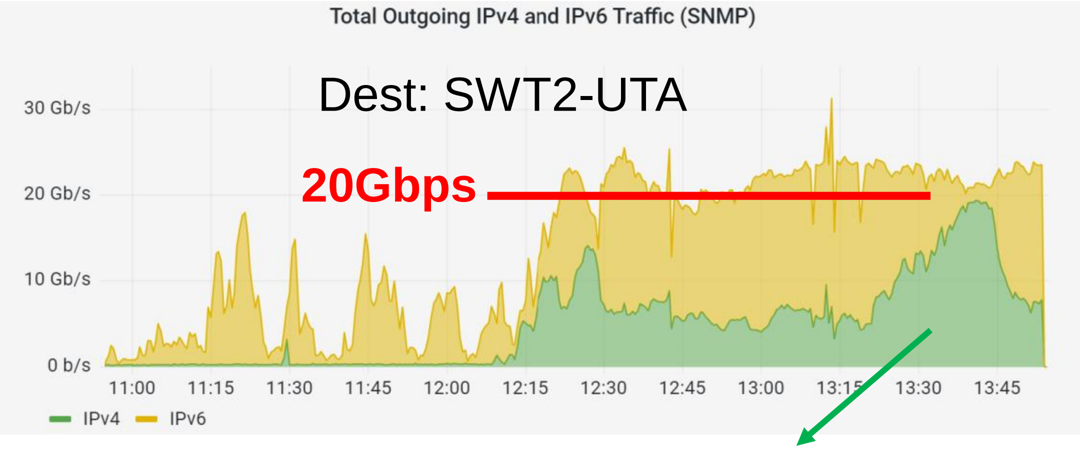
30Gbps

GAP

Round 1. WebDAV WAN test continues…

- **SWT2 UTA**: Middle of network reconfiguration
- **SWT2 OU**: Middle of the storage deployment.
  - Testing Ceph
  - Noticed that the data is not shown in ESNet monitor

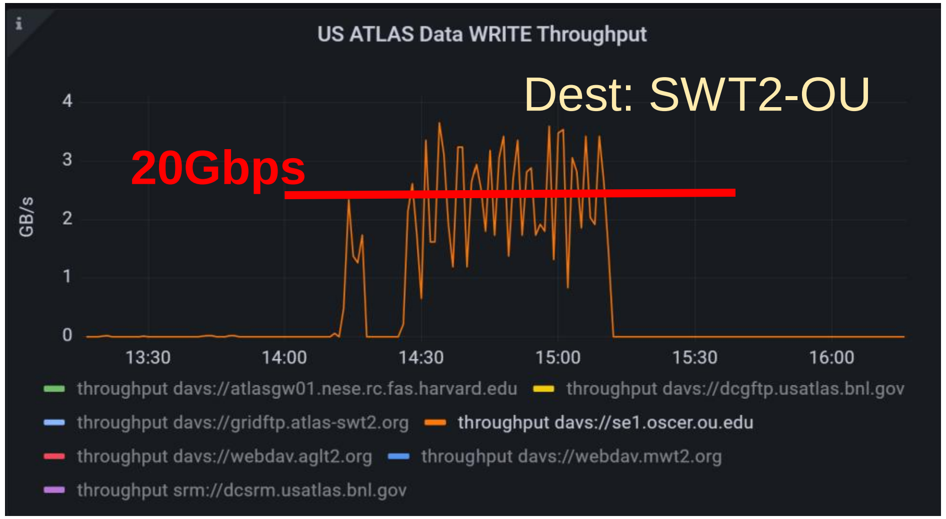
NET2: Working on the deployment of new storage. Run functional tests successfully.



Dest: SWT2-UTA

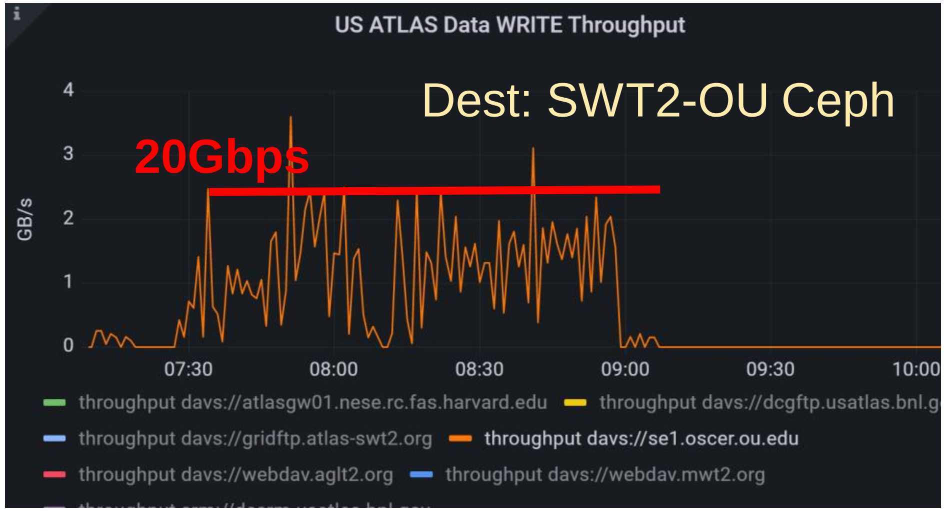
20Gbps

Production source site (lcg-se1.sfu.computecanada.ca) has IPv4 Only.



Dest: SWT2-OU

20Gbps



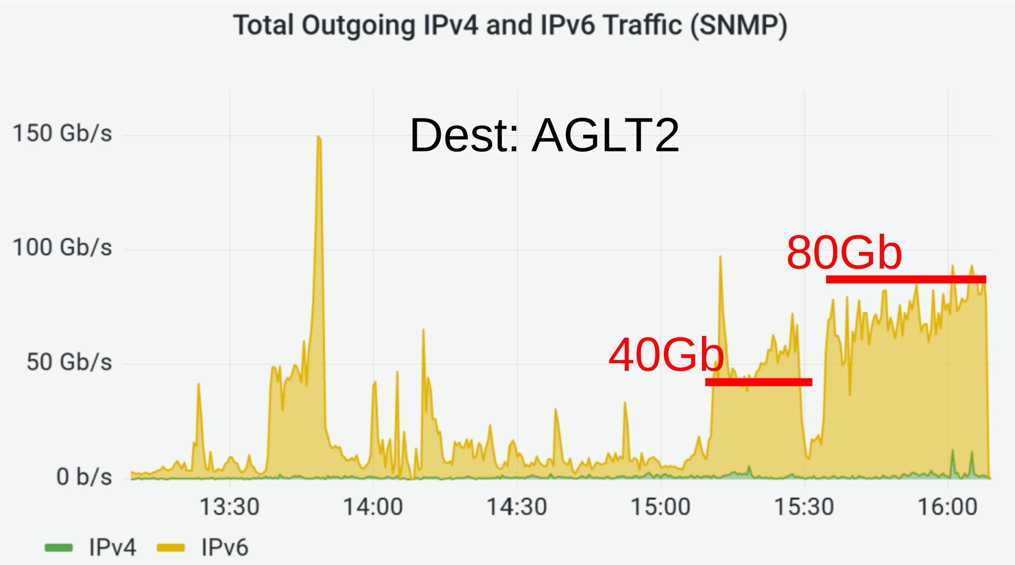
Dest: SWT2-OU Ceph

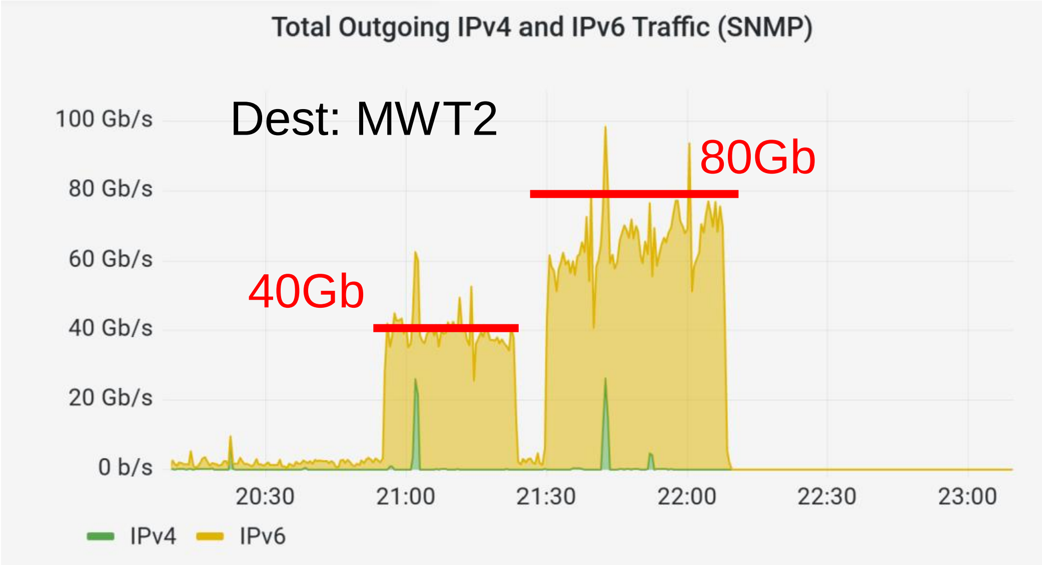
20Gbps

Round 2. XRootD WAN

- The transfer by XRootD might be behaving little differently
  - Not reaching 80Gbps easily like under WebDAV for AGLT2 and MWT2
  - However, no detail analysis was conducted.
- Although XRootD is not widely used for SE to SE transfers in ATLAS, XRootD is used in many (if not the most) jobs as well as XCache

- BNL: 3rd Party XRootD doesn't function
  - Under investigation
- SWT2-UTA:
  - Waiting for the completion of network reconfiguration
- SWT2-OU:
  - Waiting for the deployment of the storage
- NET2: Waiting for new storage





Brookhaven National Laboratory

Conclusion

- We have multiple monitors to measure the relevant WAN throughputs for the target site(s)
- We have a load generator that can generate the desired throughputs for source-destination pair.
- Tests can be conducted by requests besides quarterly.
- Tests have already identified the issues and help to resolve them before WLCG data challenge
  - Monitoring issues
  - Throughput limitations
  - Functional issues
- Test can be conducted for any sites (not limited to US ATLAS sites)
  - Helpful if a site has monitor like ESNet.
    - The monitoring URL for a site should be in CRIC?
  - The script is being cleaned up for use by the others
  - If non-BNL FTS is being used, one can add similar Grafana (or like) monitor.