Large scale data movement in R&E networks: the CERN experiments

Tim Chown, tim.chown@jisc.ac.uk

hp-wan BoF, IETF 121, Dublin, 4 Nov 2024

The Worldwide Large Hadron Collider Computing Grid (WLCG)

The CERN experiments produce multiple PB of data per day

The LHC data is distributed to over 170 sites in over 40 countries for storage and analysis by multiple experiments - ATLAS, Alice LHCb, CMS, …

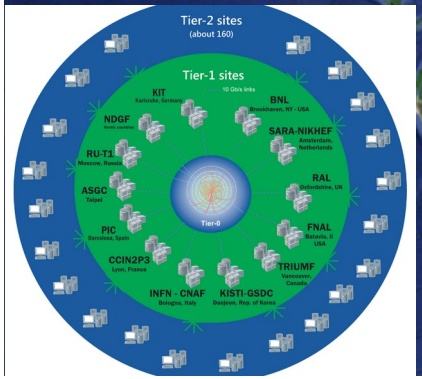
The WLCG has a general but not strict tiered model:

- CERN -> Tier 1s, well-provisioned **dedicated** private optical networks (LHCOPN) - where the largest flows happen, generally very effectively utilised and not a bottleneck
- Tier 1s -> Tier 2s, largely use a L3VPN/VRF overlay (LHCONE) - largely provisioned over the R&E networks (**shared**), with minimal use of QoS
- Other data transfers happen over general R&E networks (**shared**) managed by the National Research and Education Networks (NRENs), used for all types of R&E traffic

It’s also the big **IPv6** use case for R&E - over 90% of transfers are now IPv6

Worldwide LHC computing Grid (WLCG)

2024:
- 170 sites
- 42 countries



The WLCG uses the worldwide R&E infrastructure, not commercial / public networks

WLCG infrastructure and administration

Network:

- LHCOPN and LHCONE - coordinated by CERN, assisted by the NRENs
- Other IP (R&E networks) - managed by the NRENs - generally well-provisioned

Campus infrastructure

- Connecting local WLCG resources to the campus' NREN backbone
- Operated by local campus IT teams

Storage and compute

- Run by local WLCG teams with HPC expertise, usually independent of campus IT

Important to note the l**arge number of different administrative teams** supporting the WLCG, rather than one single company or organisation doing so

Local Tier-2 architectures - ‘Science DMZ’

Campus Tier-2 facilities have evolved over time to be performant for data movement

Their architectures generally match the “Science DMZ” **principles** written up by ESnet in 2012: https://fasterdata.es.net/science-dmz/

- A local network architecture optimised and tuned for high-performance applications, distinct from the general purpose network, typically an “on ramp” at the campus edge
- Use of appropriate software tools for data transfer
- Well-tuned (i/o and network) dedicated data transfer nodes (DTNs)
- Persistent network monitoring - typically using perfSONAR (https://www.perfsonar.net)
- Appropriate security implementation supporting the performance mission - thus generally ACL-based rather than (expensive at scale) stateful DPI firewalls

For optimal performance, both ‘ends’ and the WAN need good configuration

Transport?

Wide use of CERN’s own software - Rucio (orchestrator) and FTS (transfer engine)

FTS can use many transfer tools (e.g., XRootD) and will use many parallel TCP streams

Systems generally use the default production Linux distribution CCA (so not BBRv3)

Systems might run with tuned network parameters, but not a requirement, and optimal settings may be different for different receivers/peers with different capacities, RTT, …

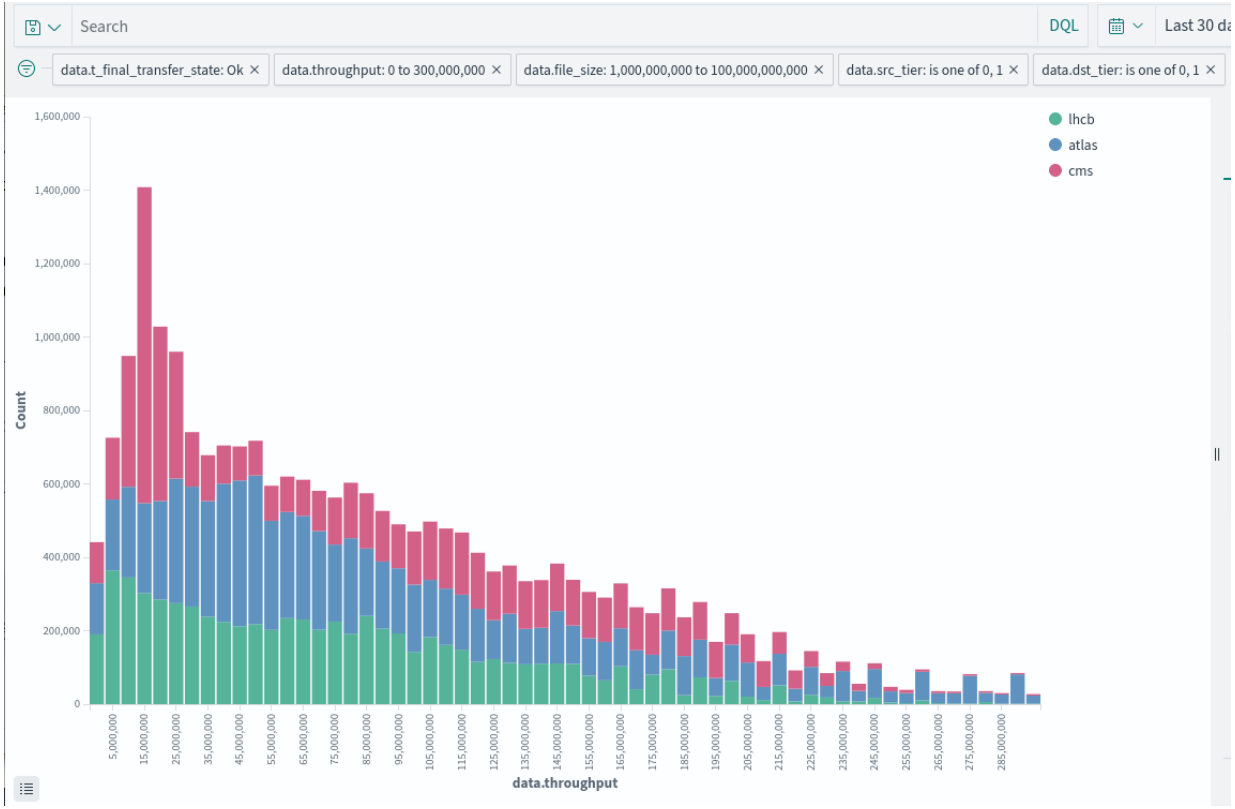
Heavy use of IPv6 means working PMTUD matters for wider adoption of 9000 MTU

perfSONAR shows at least 20-30 Gbit/s single TCP is possible intercontinental, but that’s not the nature of transfers with FTS

perfSONAR also allows network tuning testing - MTU (MSS), CCA, windows, pacing, ...

Application-oriented (disk-to-disk) performance is key - how good is the i/o?

CERN T0 to T1 traffic flow profile (*Bytes* per second)



The most common flow is ~100 Mbit/s, and the bulk of flows are under 1 Gbit/s

The long tail in this chart is around 2 Gbit/s

WLCG future

The LHC will enter its ‘High Luminosity’ phase in 4-5 years

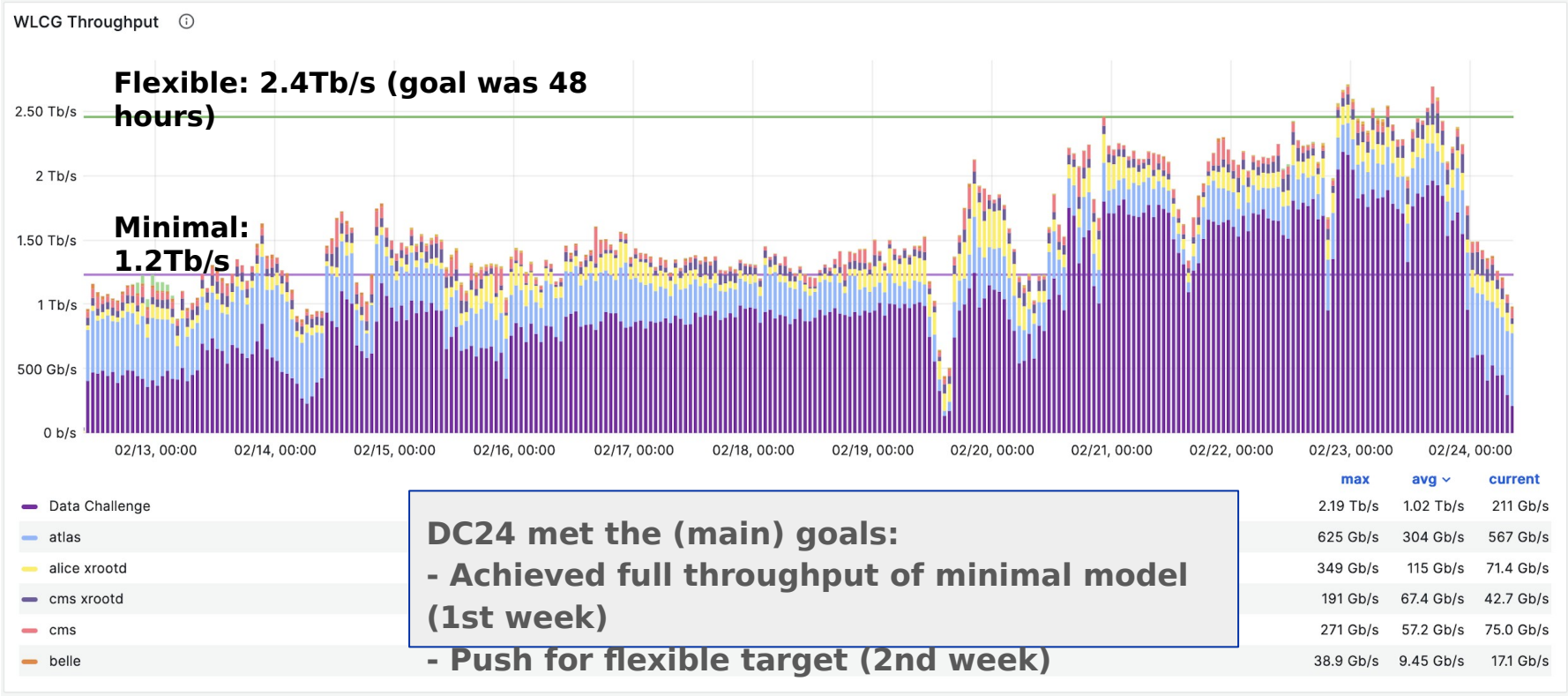
Significant ramp up in data volumes

Testing being done in advance - Data Challenge 2024 was at 25% of the volume

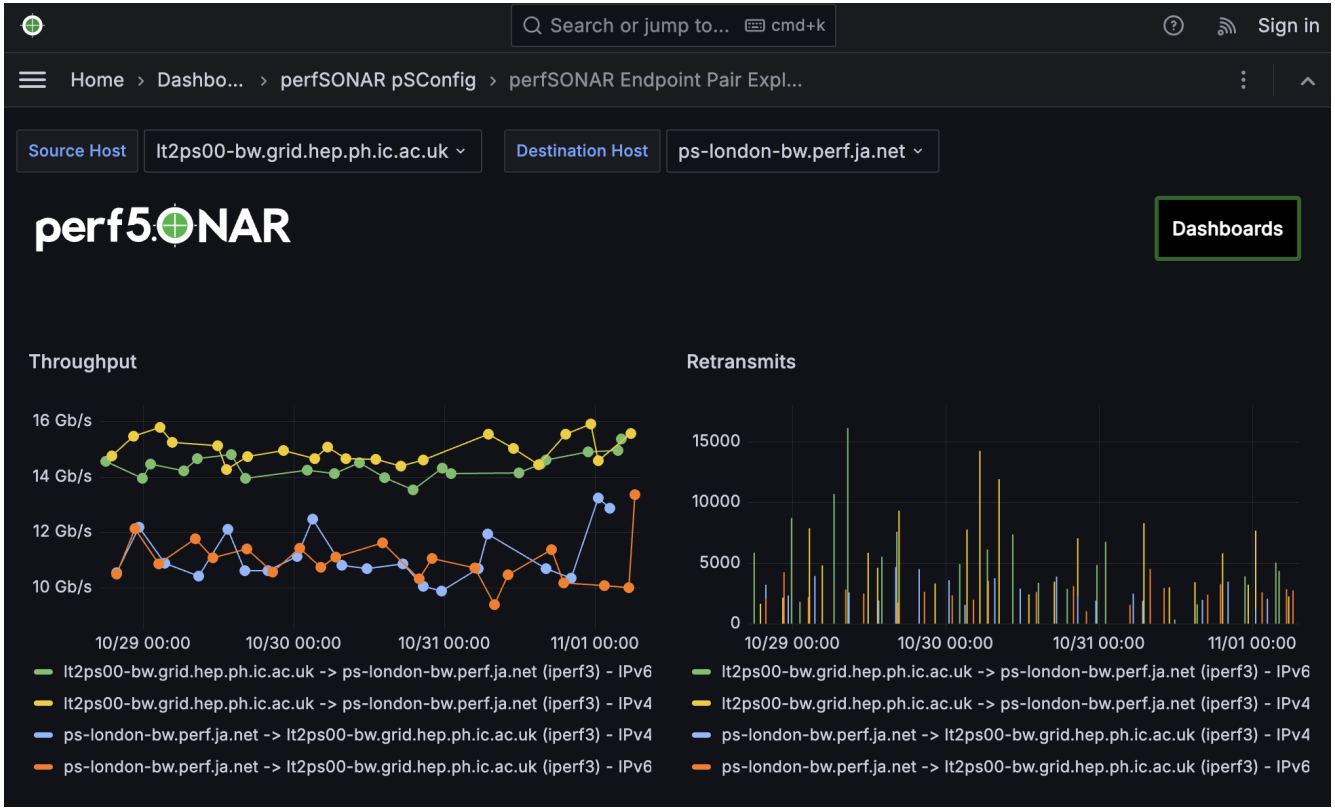
WLCG net-wg tests various things: both new hardware platforms but also CCAs, MTUs, buffers, pacing, … some of this is within the IETF remit

[The wg also does some work on ‘routing’: Rucio multi-hop, NOTED, VRF steering, (IPv6) packet marking, …]

Data Challenge 24 - overall throughput by experiment



perfSONAR persistent monitoring



pscheduler testing : London -> CERN : CUBIC / BBRv3

$ **pscheduler task throughput --dest pse01-gva.cern.ch --source ps-london-bw.perf.ja.net --congestion cubic**
Submitting task...
Task URL:
https://ps-london-bw.perf.ja.net/pscheduler/tasks/c5e9e329-77d2-43a8-bc26-3d52c70dac02
Running with tool 'iperf3'
Fetching first run...
Next scheduled run:
https://ps-london-bw.perf.ja.net/pscheduler/tasks/c5e9e329-77d2-43a8-bc26-3d52c70dac02/runs/eb9cb082-2ac8-4f6a-9aaa-5103e7d58403
Starts 2024-11-03T15:28:21+00:00 (~5 seconds)
Ends 2024-11-03T15:28:40+00:00 (~18 seconds)
Waiting for result...
* Stream ID 5

Interval	Throughput	Retransmits	Current Window
0.0 - 1.0	114.19 Mbps	0	625.46 KBytes
1.0 - 2.0	1.90 Gbps	1642	3.07 MBytes
2.0 - 3.0	1.67 Gbps	0	3.24 MBytes
3.0 - 4.0	1.58 Gbps	0	3.38 MBytes
4.0 - 5.0	1.74 Gbps	0	3.49 MBytes
5.0 - 6.0	1.79 Gbps	0	3.57 MBytes
6.0 - 7.0	1.77 Gbps	0	3.63 MBytes
7.0 - 8.0	1.76 Gbps	407	2.59 MBytes
8.0 - 9.0	1.34 Gbps	0	2.74 MBytes
9.0 - 10.0	1.39 Gbps	0	2.85 MBytes

Summary

Interval	**Throughput**	**Retransmits**	**Receiver Throughput**
0.0 - 10.0	**1.51 Gbps**	**2049**	**1.47 Gbps**

$ **pscheduler task throughput --dest pse01-gva.cern.ch --source ps-london-bw.perf.ja.net --congestion bbr**
Submitting task...
Task URL:
https://ps-london-bw.perf.ja.net/pscheduler/tasks/bd140b4c-29bb-4016-85bc-61996d70acb9
Running with tool 'iperf3'
Fetching first run...
Next scheduled run:
https://ps-london-bw.perf.ja.net/pscheduler/tasks/bd140b4c-29bb-4016-85bc-61996d70acb9/runs/5e50c358-1e69-4037-8c87-08d4e010dddf
Starts 2024-11-03T15:29:02+00:00 (~5 seconds)
Ends 2024-11-03T15:29:21+00:00 (~18 seconds)
Waiting for result...
* Stream ID 5

Interval	Throughput	Retransmits	Current Window
0.0 - 1.0	11.58 Gbps	3526	112.46 MBytes
1.0 - 2.0	13.83 Gbps	0	57.65 MBytes
2.0 - 3.0	14.11 Gbps	0	61.21 MBytes
3.0 - 4.0	14.09 Gbps	0	60.38 MBytes
4.0 - 5.0	14.24 Gbps	0	60.23 MBytes
5.0 - 6.0	13.95 Gbps	0	60.04 MBytes
6.0 - 7.0	14.14 Gbps	0	60.36 MBytes
7.0 - 8.0	14.18 Gbps	0	60.68 MBytes
8.0 - 9.0	14.22 Gbps	0	62.56 MBytes
9.0 - 10.0	14.14 Gbps	0	61.52 MBytes

Summary

Interval	**Throughput**	**Retransmits**	**Receiver Throughput**
0.0 - 10.0	**13.85 Gbps**	**3526**	**13.82 Gbps**

Note: latency (and loss) is a separate perfSONAR test type (using OWAMP)

Summary

The WLCG is an infrastructure distributed over the worldwide R&E networks - both private optical (dedicated) and overlay/general R&E (shared)

Large volumes of data are moved successfully (“well enough”) on a daily basis

There are a lot of lower throughput flows rather than a small number of higher throughput flows - but flows of 20-30 Gbit/s+ are possible over a wide area

Local application of ‘Science DMZ’ principles are important, at both ‘ends’

Beneficial techniques are not (yet) widely adopted - e.g., 9000 MTU, TCP BBRv3, …

Good monitoring is in place to view network characteristics and test tuning

A big ramp up in data volumes coming with LHC-HL phase in 4-5 years - the community is interested in anything that can help make transfers more efficient