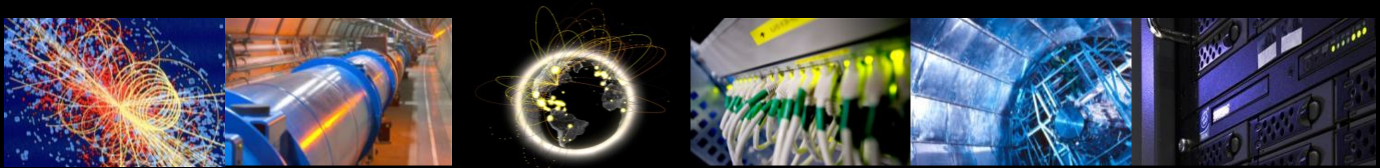
DC24 Network Activities & Results

Shawn McKee / University of Michigan
Presented by Edoardo Martelli / CERN
on behalf of WLCG Network Throughput WG and DOMA

May 14, 2024

WLCG/HSF Workshop 2024

https://indico.cern.ch/event/1369601



Worldwide LHC Computing Grid

Open Science Grid

Acknowledgements

Thanks to all those who contributed content:

Edoardo Martelli, Tim Chown, Mattias Wadenstein, Dave Kelsey, Bruno Hoeft, Harvey Newman

And from authors of various source presentations:

Katy Ellis, Alessandra Forti, Christoph Wissing, Mario Lassnig, Carmen Misa Moreira

Outline

- DC24 Network Overview
- Application / Technology Testing during DC24
- Planning for HL-LHC and DC26

WLCG Network Planning for DC24

DC24 Preparations started shortly after DC21 and ramped up significantly in late 2022

Some of the important **network** related areas targeted for DC24:

- **Network planning**: make sure our sites and their local and regional networks are aware of WLCG requirements and timeline, and plan accordingly.
- Update and utilize **perfSONAR** to clean up links & fix problems before DC24.
- Instrument and document **site networks**, for at least our largest sites.
- **IPv6** should be enabled everywhere not just because of packet marking, but because it will allow us to get back to a single stack sooner!
- **Optimize** our ability to utilize the network (jumbo frames, protocols, pacing)
- Improve net traffic **visibility** (SciTags)
- Explore and demonstrate the value of **Network Orchestration/SDN**.

Pre-DC24 Testing

Very important for DC24 was the significant amount of **pre-testing** and **mini-challenges** that occurred during the year prior.

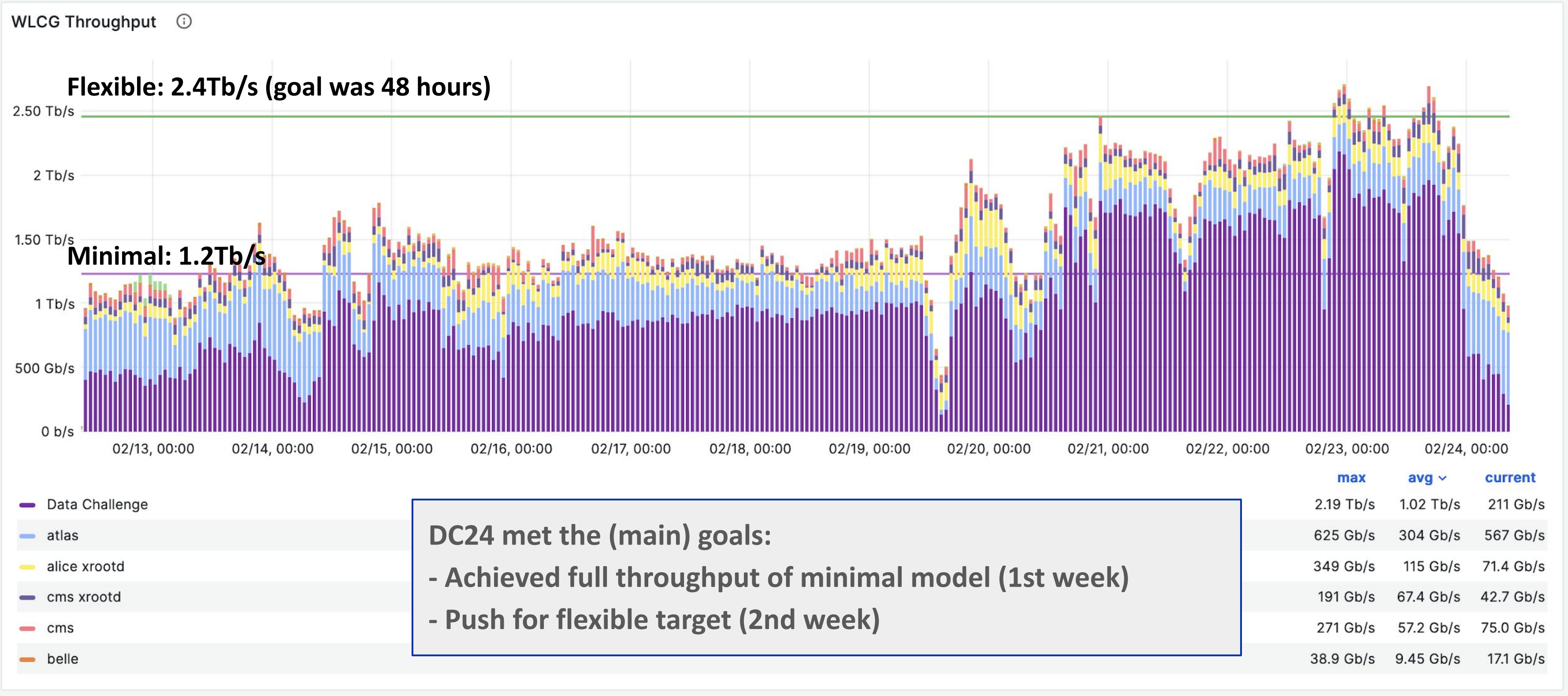
These mini challenges typically involved small sets of sites, specific regions and/or specific WLCG experiments and sometimes non-production equipment.

This testing was critical to identifying bottlenecks well in advance of DC24.

- For many sites pre-testing actually stressed their infrastructure far beyond what was needed for the flexible model requirements of DC24.

perfSONAR testing and follow-ups also helped ensure the wide-area R&E networks were operating cleanly prior to DC24.

DC24 Throughput Summary Plot



Flexible: 2.4Tb/s (goal was 48 hours)

Minimal: 1.2Tb/s

	max	avg ⌄	current
Data Challenge	2.19 Tb/s	1.02 Tb/s	211 Gb/s
atlas	625 Gb/s	304 Gb/s	567 Gb/s
alice xrootd	349 Gb/s	115 Gb/s	71.4 Gb/s
cms xrootd	191 Gb/s	67.4 Gb/s	42.7 Gb/s
cms	271 Gb/s	57.2 Gb/s	75.0 Gb/s
belle	38.9 Gb/s	9.45 Gb/s	17.1 Gb/s

DC24 met the (main) goals:

- Achieved full throughput of minimal model (1st week)
- Push for flexible target (2nd week)

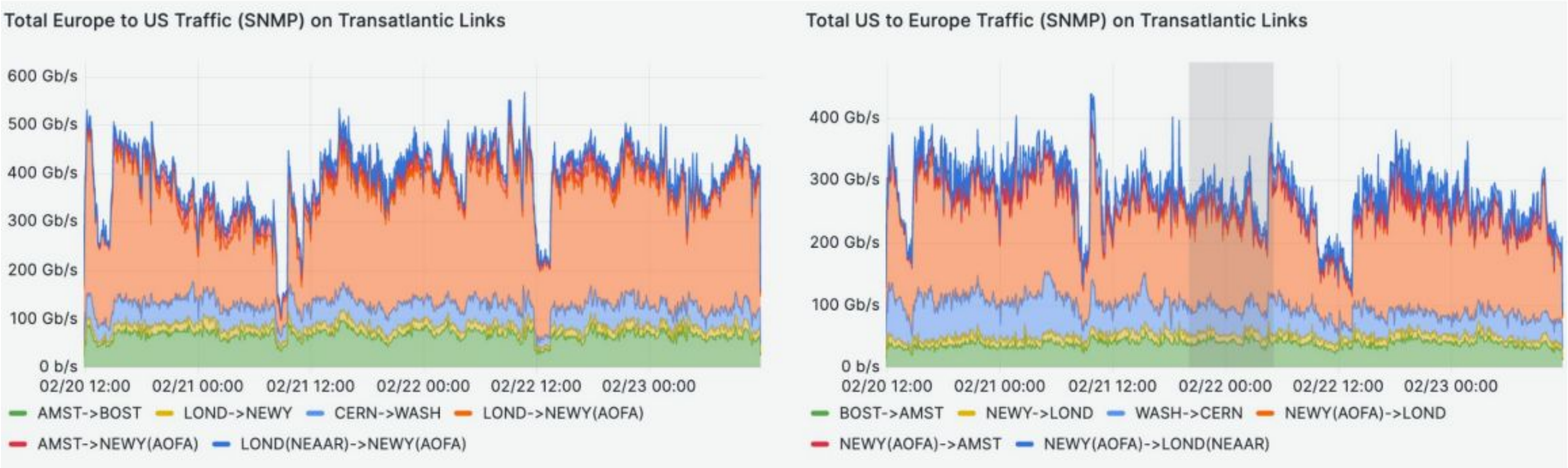
Research & Education (R&E) Networks

The global set of R&E networks underlie the whole of WLCG’s distributed infrastructure and their connectivity, reliability and capability directly impact its performance.

Transatlantic bandwidth was a concern just prior to DC24 because of delayed deployment of 2x400G links and the loss of the only 400G link but it was restored in time.

CERN-RAL OPN failed during DC24 but was transparently re-routed via GEANT-Janet

In general, the R&E networks performed very well for DC24 and were not a bottleneck for any of the experiments. (See LHCONE ESnet presentation)



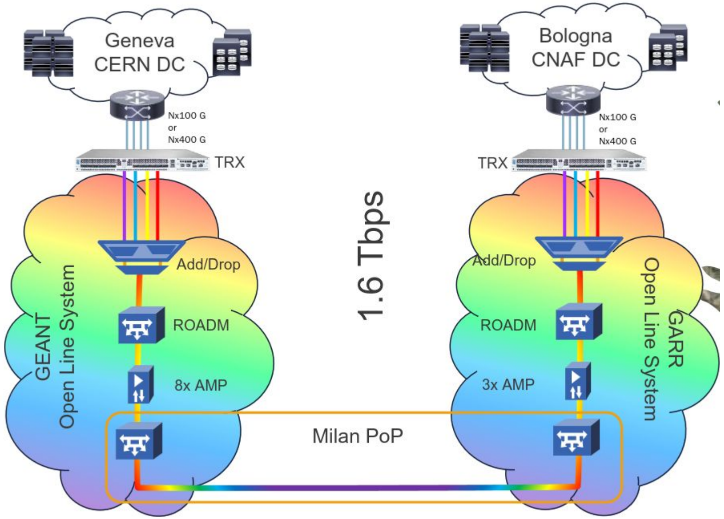
Transmission Testing

CNAF-CERN DCI

- spectrum sharing over GEANT and GARR dark fibres

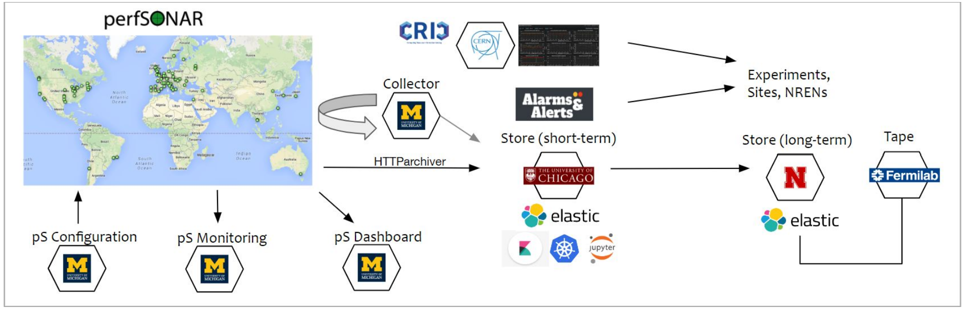
- 4x100Gbps links between CERN and CNAF used for DC24 and now in production

Cost effective technique to get >1Tbps LHCOPN links already today



DC24 perfSONAR

WLCG/OSG-LHC have deployed and maintained a perfSONAR infrastructure to monitor our site network connectivity to each other.



For DC24 the plan was to utilize the testing to identify any problems in the networks and fix them prior to DC24

No major misconfigurations or network degradations were found in the lead up to DC24.

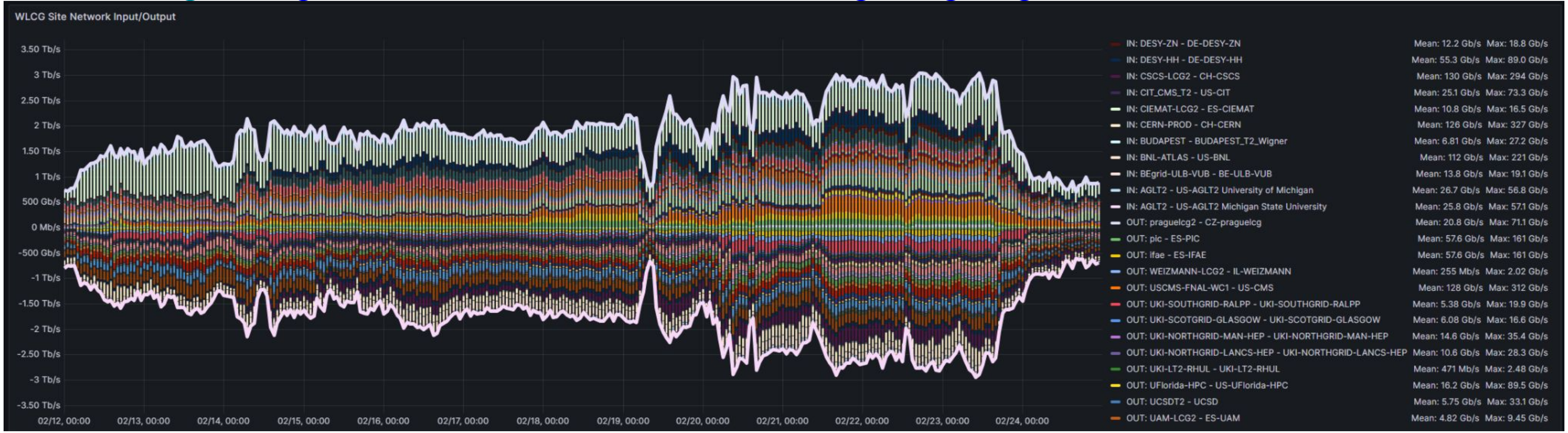
- Some site issues were identified and fixed
- R&E networks were given a clean bill of health

The process of identifying and localizing network problems remains a challenge and work continues to better automate this via AI/ML

DC24 Site Network Monitoring

During DC21, we didn’t have sufficient data about how much traffic each contributing site was experiencing and it made the analysis to identify bottlenecks difficult. For DC24 the WLCG Monitoring Task Force was charged with filling in this missing information.

The larger sites participating in DC24 were given GGUS tickets to enable site network monitoring, using SNMP to monitor total incoming/outgoing traffic.



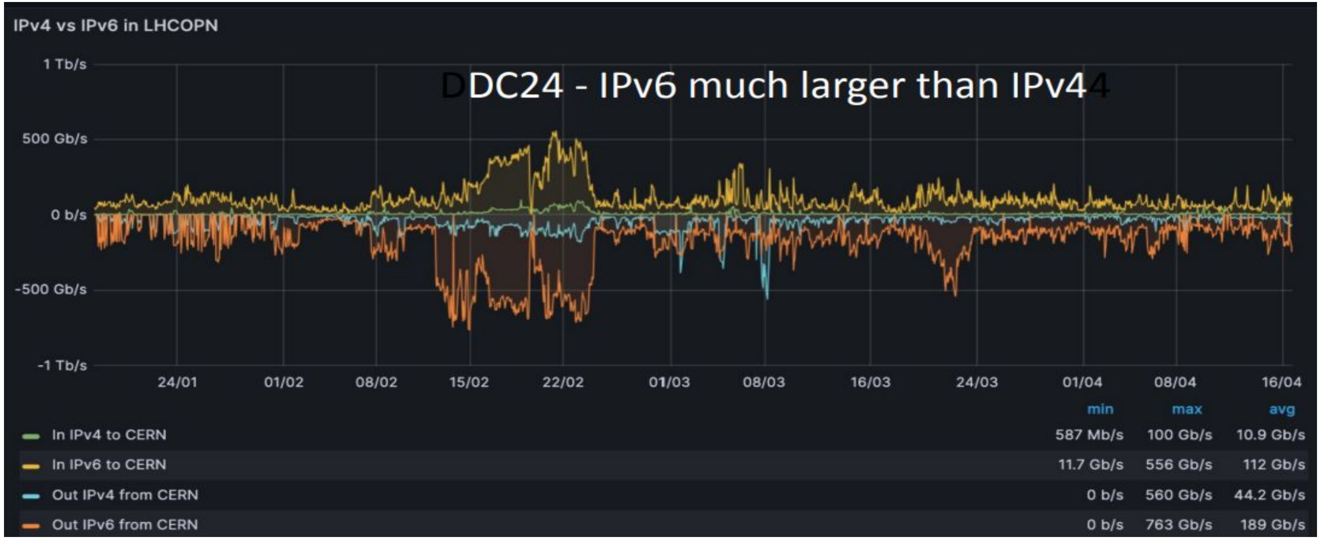
IN: DESY-ZN - DE-DESY-ZN	Mean: 12.2 Gb/s	Max: 18.8 Gb/s
IN: DESY-HH - DE-DESY-HH	Mean: 55.3 Gb/s	Max: 89.0 Gb/s
IN: CSCS-LCG2 - CH-CSCS	Mean: 130 Gb/s	Max: 294 Gb/s
IN: CIT_CMS_T2 - US-CIT	Mean: 25.1 Gb/s	Max: 73.3 Gb/s
IN: CIEMAT-LCG2 - ES-CIEMAT	Mean: 10.8 Gb/s	Max: 16.5 Gb/s
IN: CERN-PROD - CH-CERN	Mean: 126 Gb/s	Max: 327 Gb/s
IN: BUDAPEST - BUDAPEST_T2_Wigner	Mean: 6.81 Gb/s	Max: 27.2 Gb/s
IN: BNL-ATLAS - US-BNL	Mean: 112 Gb/s	Max: 221 Gb/s
IN: BEgrid-ULB-VUB - BE-ULB-VUB	Mean: 13.8 Gb/s	Max: 19.1 Gb/s
IN: AGLT2 - US-AGLT2 University of Michigan	Mean: 26.7 Gb/s	Max: 56.8 Gb/s
IN: AGLT2 - US-AGLT2 Michigan State University	Mean: 25.8 Gb/s	Max: 57.1 Gb/s
OUT: praguelcg2 - CZ-praguelcg	Mean: 20.8 Gb/s	Max: 71.1 Gb/s
OUT: pic - ES-PIC	Mean: 57.6 Gb/s	Max: 161 Gb/s
OUT: ifae - ES-IFAE	Mean: 57.6 Gb/s	Max: 161 Gb/s
OUT: WEIZMANN-LCG2 - IL-WEIZMANN	Mean: 255 Mb/s	Max: 2.02 Gb/s
OUT: USCMS-FNAL-WC1 - US-CMS	Mean: 128 Gb/s	Max: 312 Gb/s
OUT: UKI-SOUTHGRID-RALPP - UKI-SOUTHGRID-RALPP	Mean: 5.38 Gb/s	Max: 19.9 Gb/s
OUT: UKI-SCOTGRID-GLASGOW - UKI-SCOTGRID-GLASGOW	Mean: 6.08 Gb/s	Max: 16.6 Gb/s
OUT: UKI-NORTHGRID-MAN-HEP - UKI-NORTHGRID-MAN-HEP	Mean: 14.6 Gb/s	Max: 35.4 Gb/s
OUT: UKI-NORTHGRID-LANCS-HEP - UKI-NORTHGRID-LANCS-HEP	Mean: 10.6 Gb/s	Max: 28.3 Gb/s
OUT: UKI-LT2-RHUL - UKI-LT2-RHUL	Mean: 471 Mb/s	Max: 2.48 Gb/s
OUT: UFlorida-HPC - US-UFlorida-HPC	Mean: 16.2 Gb/s	Max: 89.5 Gb/s
OUT: UCSDT2 - UCSD	Mean: 5.75 Gb/s	Max: 33.1 Gb/s
OUT: UAM-LCG2 - ES-UAM	Mean: 4.82 Gb/s	Max: 9.45 Gb/s

This new monitoring has become part of our regular infrastructure which is also part of the purpose of the data challenges: to evolve our infrastructure to what is needed for HL-LHC!

DC24 IPv6 Activities

One of the longstanding challenges for WLCG has been the transition to IPv6, which has been going on for more than a decade; **DC24 gave us a chance to see where we are.** Sites have been requested to provide perfSONAR and storage **dual-stacked** since November 2017.

Recently the request to dual stack services and worker nodes has been made, targeting June 2024



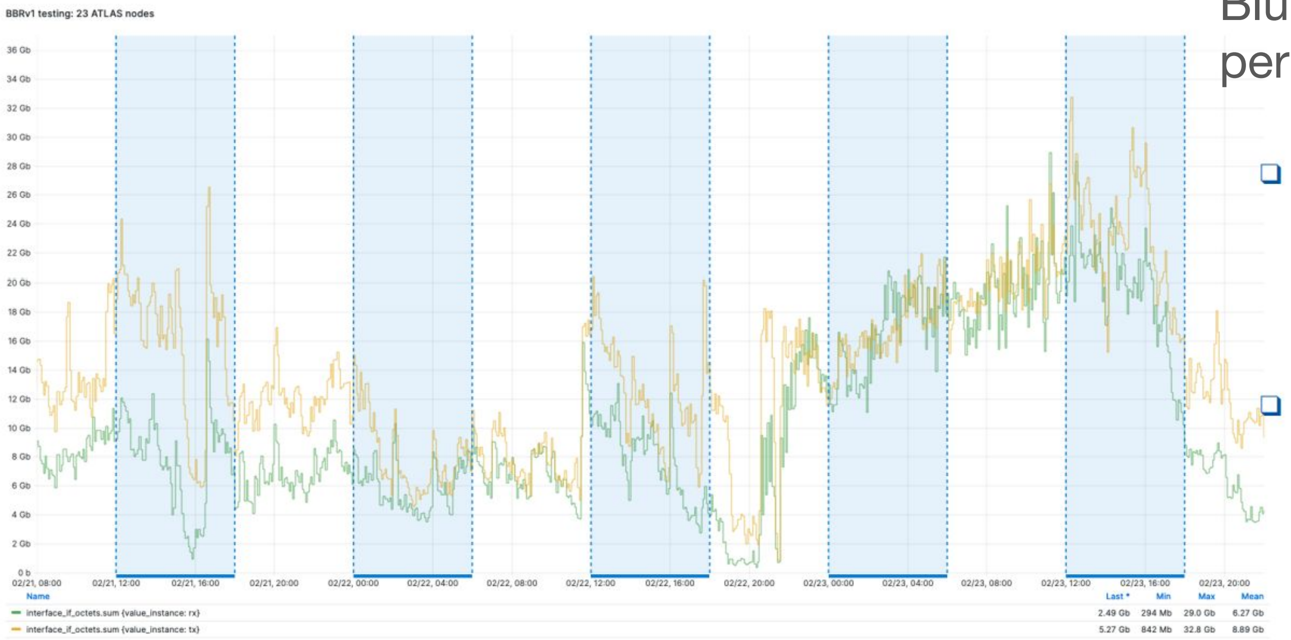
	min	max	avg
🟩 In IPv4 to CERN	587 Mb/s	100 Gb/s	10.9 Gb/s
🟨 In IPv6 to CERN	11.7 Gb/s	556 Gb/s	112 Gb/s
🟦 Out IPv4 from CERN	0 b/s	560 Gb/s	44.2 Gb/s
🟧 Out IPv6 from CERN	0 b/s	763 Gb/s	189 Gb/s

While there is still significant amounts of IPv4 traffic, it is heading in the right direction.
All T1 and 97% of T2 storage is IPv6 capable; some sites are transitioning to IPv6-only

DC24 TCP Protocol Explorations

For WLCG data transfers, optimizing the TCP protocols may lead to higher throughput for individual transfers and better sharing of available bandwidth.

BBR demo at DC24



Name	Last *	Min	Max	Mean
🟩 interface_if_octets.sum {value_instance: rx}	2.49 Gb	294 Mb	29.0 Gb	6.27 Gb
🟧 interface_if_octets.sum {value_instance: tx}	5.27 Gb	842 Mb	32.8 Gb	8.89 Gb

Blue highlighted time periods are BBRv1

- ❑ During WLCG DC24 23 ATLAS nodes swapped every 6 hours between CUBIC and BBRv1 congestion protocol
- ❑ No evidence of gain nor loss using BBRv1

DC24 Jumbo Frames Activities

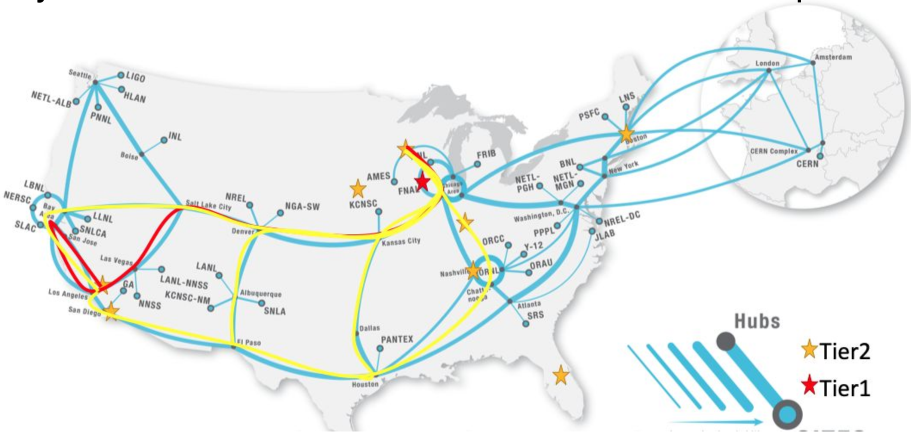
Benefits of jumbo frames are evident on long distance transfers, less on the short distance
Operational issues are also evident, but they can be mitigated by sharing deployment experiences

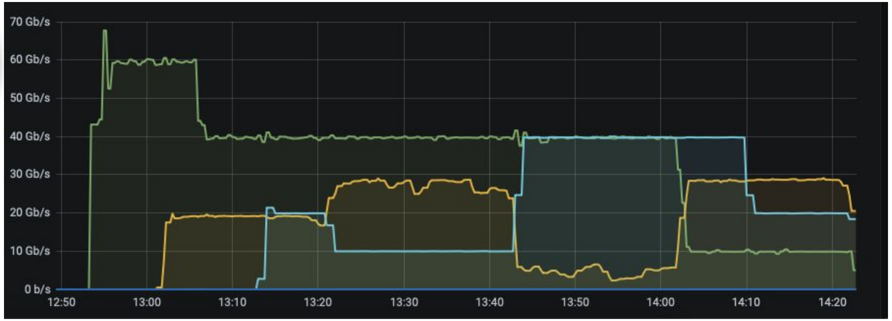
No special test was done on Jumbo frames. To be noted that several sites have been happily using Jumbo for years and had no problem with it during DC24

The preparation of DC26 will focus on a wider use of jumbo frames

DC24 SENSE/Rucio (Network Orchestration)

The objective was to provide Rucio with capabilities to request network services via SENSE in order to: ***a) improve accountability, b) increase predictability, and c) isolate and prioritize transfer requests.*** This project used a dedicated Rucio as well as XRootD instances so it would not interfere with Production systems. Data was transferred across a mix of production and next generation network paths.





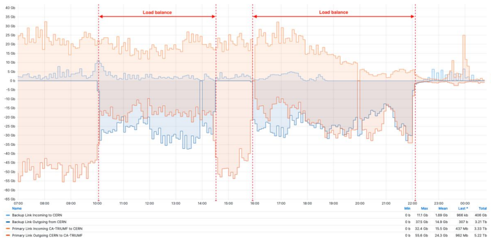
Between Fermilab, Caltech, UCSD Rucio-DMM/SENSE-FTS-XRootD multiple Rucio-triggered data flows were managed between multiple pairs of sites; The modify feature of DMM was used to change bandwidth allocation on the fly in response to Rucio requests. The following Quality of Service policies were demonstrated: Hard QoS / Soft QoS on Server; Hard QoS at the network level. DMM Real time API-driven FTS tuning was used to adjust active/max transfers settings. Additional US-CMS Tier2 sites will be evaluated for deployment.

DC24 NOTED (Software Defined Networking)

NOTED (Network Optimized Transfer of Experimental Data) is an intelligent network controller to improve the throughput of large data transfers in FTS (File Transfer Services) by handling dynamic circuits.

During DC24 traffic was monitored and balanced for DE-KIT (between its LHCONE and LHCOPN links), CA-TRIUMF (between LHCOPN and its backup link) and ES-PIC (between LHCONE and LHCOPN links).

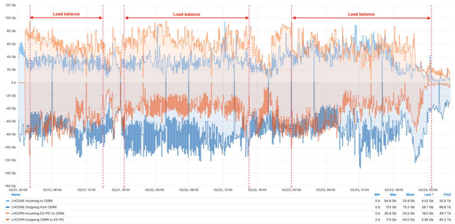
NOTED demo at DC24 (LHCOPN, LHCONE versions)



Name	Min	Max	Mean	Last *	Total
Backup Link Incoming to CERN	0 b	11.1 Gb	1.89 Gb	966 kb	406 Gb
Backup Link Outgoing from CERN	0 b	37.5 Gb	14.9 Gb	307 b	3.21 Tb
Primary Link Incoming CA-TRIUMF to CERN	0 b	32.4 Gb	15.5 Gb	437 Mb	3.33 Tb
Primary Link Outgoing CERN to CA-TRIUMF	0 b	55.6 Gb	24.3 Gb	962 Mb	5.22 Tb

- CA-TRIUMF load balancing between LHCOPN and its backup link (from 21st to 23rd of February 2024)

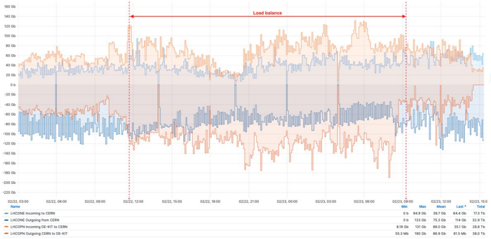
NOTED demo at DC24 (LHCOPN, LHCONE versions)



Name	Min	Max	Mean	Last *	Total
LHCONE Incoming to CERN	0 b	94.9 Gb	33.9 Gb	4.02 Gb	30.9 Tb
LHCONE Outgoing from CERN	0 b	133 Gb	73.2 Gb	28.7 Gb	66.8 Tb
LHCOPN Incoming ES-PIC to CERN	0 b	95.9 Gb	54.5 Gb	16.0 Gb	49.7 Tb
LHCOPN Outgoing CERN to ES-PIC	0 b	113 Gb	44.0 Gb	5.36 Gb	40.2 Tb

- ES-PIC load balancing between LHCOPN and LHCONE (from 21st to 23rd of February 2024)

NOTED demo at DC24 (LHCOPN, LHCONE versions)



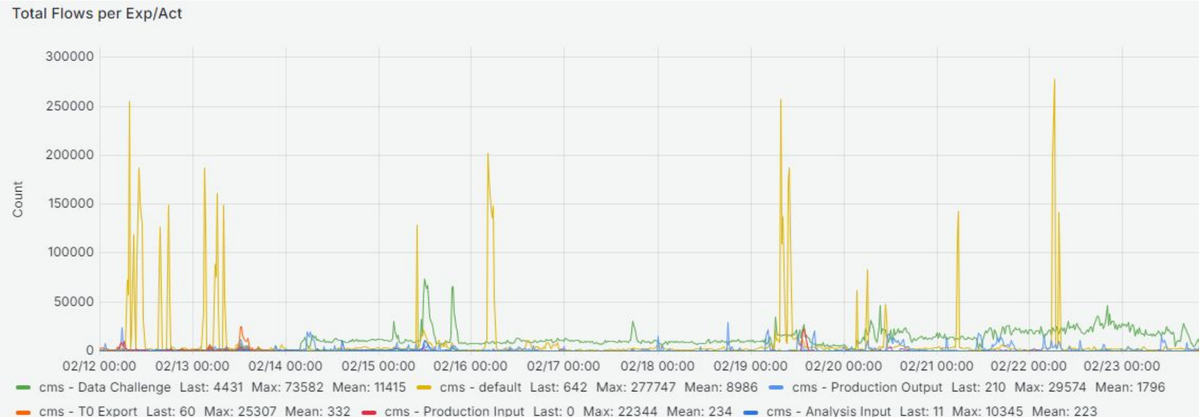
Name	Min	Max	Mean	Last *	Total
LHCONE Incoming to CERN	0 b	94.9 Gb	39.7 Gb	64.4 Gb	17.3 Tb
LHCONE Outgoing from CERN	0 b	133 Gb	75.3 Gb	114 Gb	32.9 Tb
LHCOPN Incoming DE-KIT to CERN	8.19 Gb	131 Gb	66.0 Gb	35.1 Gb	28.8 Tb
LHCOPN Outgoing CERN to DE-KIT	55.3 Mb	190 Gb	86.9 Gb	81.5 Mb	38.0 Tb

- DE-KIT load balancing between LHCOPN and LHCONE (from 22nd to 23rd of February 2024)

Further work is underway to enable NOTED as a possible production service and it may become part of our toolkit by DC26

DC24 Scitags (Packet & Flow Marking)

In the January 2020 LHCONE/LHCOPN meeting at CERN, the experiments converged on three areas of interest for networking: visibility, optimization and orchestration. The SciTags Initiative has been working on making R&E network traffic "visible" anywhere in the network.



During **DC24** we had over **80% of the CERN EOS CMS** instance as well as the **USCMS Nebraska Tier-2** sending fireflies for their associated data flows.

ESnet dash: https://dashboard.stardust.es.net/goto/RrVzQwLIg?orgId=2

Currently have EOS and Xrootd SciTags-capable and are working on Storm and dCache

Target: have all production traffic labeled by DC26

Next Steps and Planning for DC26

The two data challenges we have already completed have been both very successful and very useful for guiding our path towards HL-LHC. Beneficial improvements are folded into production

DC21(10%) provided first visibility into infrastructure capabilities and shortcomings, giving us an opportunity to improve our existing infrastructure and plan.

DC24(25%) allowed us to understand how DC21 recommended changes worked and where critical infrastructure bottlenecks exist.

- Perhaps most valuable was the series of pre-DC24 tests and mini-challenges which pushed our infrastructure locally beyond DC24

DC26(50%) will require us to operate at 50% of expected HL-LHC scale; **mini-challenges** for load, technologies and visibility will be **critical** for success.

Summary

- Our global set of Research & Development (R&D) networks demonstrated more than sufficient capacity and reliability during DC24 and were NOT a bottleneck for any of the experiments
  - Some sites did identify local network bottlenecks or non-optimal architectures to support 25% of HL-LHC scale
- Various network technologies were very successfully tested during DC24 and show promising results, motivating the effort to put them into production.
  - Efforts to better monitor our networks has already become part of our operational toolkits
- DC24 has been an opportunity to push needed networking capabilities forward, guiding development based upon operational experiences.
- While the R&E networks were not a bottleneck for DC24, as other application, storage infrastructures and middleware are improved, this could change quickly.
  - We need to continue regular mini-challenges to track progress and prepare for DC26

Questions / Discussion?

General Acknowledgements

We would like to thank the **WLCG**, **HEPiX**, **perfSONAR** and **OSG** organizations for their work on the topics presented.

In addition we want to explicitly acknowledge the support of the **National Science Foundation** which supported this work via:

- OSG: NSF MPS-1148698
- IRIS-HEP: NSF OAC-1836650

Useful URLs

- OSG/WLCG Networking Documentation
  - https://opensciencegrid.github.io/networking/
- The SciTags Initiative: https://www.scitags.org/
- Rucio/SENSE in DC24
- The Research Networking Technical WG
- The Global Network Advancement Group: https://www.gna-g.net/
- WLCG DOMA DC24 plans
  - https://indico.cern.ch/event/1225415/contributions/5155042/attachments/2593516/4476291/Data%20Challenge%202024.pdf
- Grafana dashboards
  - http://monit-grafana-open.cern.ch/
- ATLAS Alerting and Alarming Service: https://psa.osg-htc.org/
- The perfSONAR Dashboard application: https://ps-dash.uc.ssl-hep.org/
- ESnet WLCG Stardust Dashboard: https://public.stardust.es.net/d/XkxDL5H7z/esnet-public-dashboards?orgId=1

Sources for this Talk

ESnet presentation at LHCONE/LHCOPN meeting

DC24 presentation LHCOPN/LHCONE workshop

See WLCG DC24 Network Content

Backup Slides Follow

WLCG “Network” Data Challenges

A planned series of **end-to-end** tests to demonstrate the progress of **infrastructure**, **applications** and **middleware** of WLCG sites and experiments to reach the capacity and capability required for HL-LHC scale operations.

These data challenges provide many benefits, allowing sites, networks and experiments to evaluate their progress, motivate and validate their developments in hardware and software and show readiness of technologies at suitable scale.

The first data challenge (Data Challenge 2021; DC21) targeted **10%** of HL-LHC

DC24 was targeting **25%** of **HL-LHC** operations scale for both a minimal and flexible model and took place from February 12-23, 2024.

DC24 Packet Pacing Testing

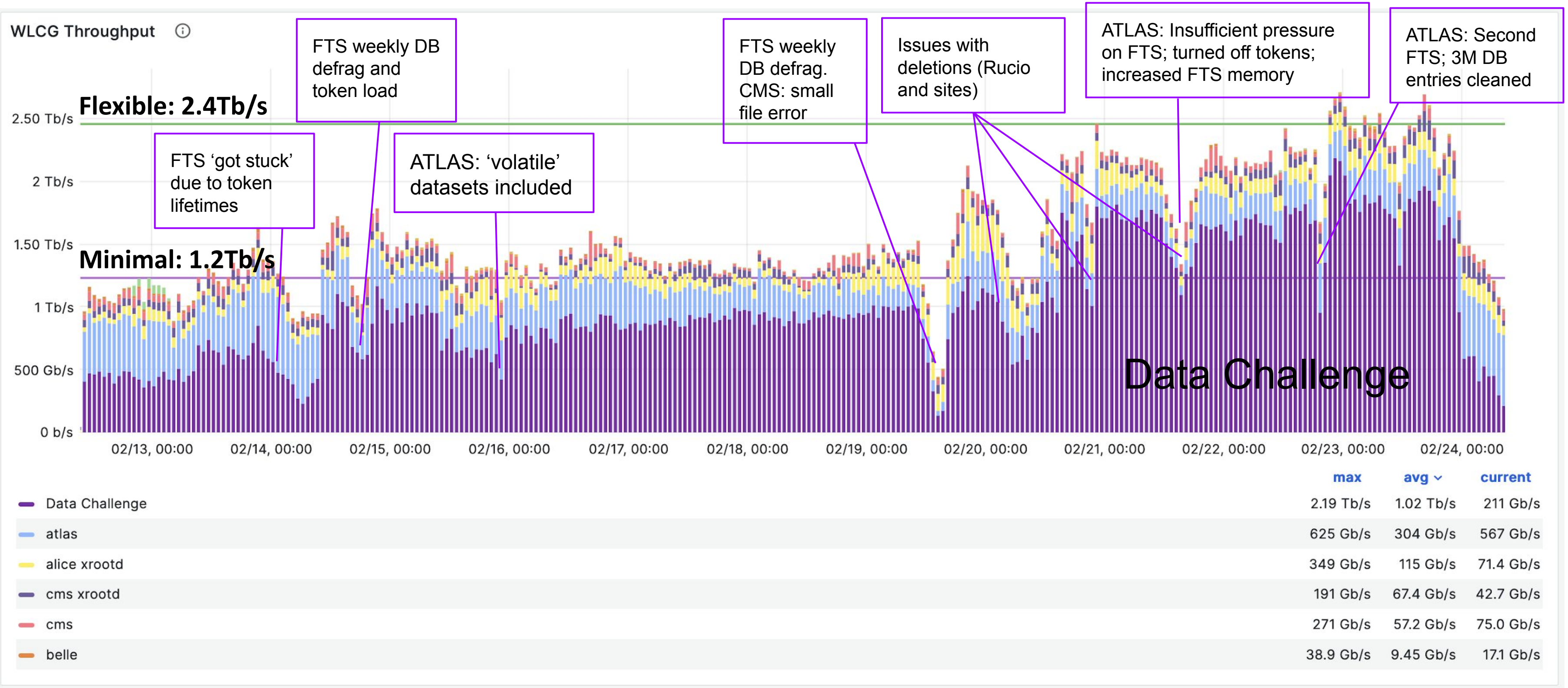
Packet pacing at the network interface card (NIC) level can improve performance of data transfers. More resiliency to packet loss

- Better bandwidth utilization in buffer-constrained networks
- Reducing load on sending hosts
- More effectively sharing available bandwidth

The linux ‘tc’ command can be used to have the NIC emit packets at a specified rate to match the minimum of source-read, destination-write and available network bandwidth. Note BBR uses pacing internally.

While packet pacing was not extensively tested prior to DC24 it is on the list of areas being explored for DC26.

Identifying Issues during DC24



Flexible: 2.4Tb/s

Minimal: 1.2Tb/s

FTS 'got stuck' due to token lifetimes

FTS weekly DB defrag and token load

ATLAS: 'volatile' datasets included

FTS weekly DB defrag. CMS: small file error

Issues with deletions (Rucio and sites)

ATLAS: Insufficient pressure on FTS; turned off tokens; increased FTS memory

ATLAS: Second FTS; 3M DB entries cleaned

Data Challenge

	max	avg	current
Data Challenge	2.19 Tb/s	1.02 Tb/s	211 Gb/s
atlas	625 Gb/s	304 Gb/s	567 Gb/s
alice xrootd	349 Gb/s	115 Gb/s	71.4 Gb/s
cms xrootd	191 Gb/s	67.4 Gb/s	42.7 Gb/s
cms	271 Gb/s	57.2 Gb/s	75.0 Gb/s
belle	38.9 Gb/s	9.45 Gb/s	17.1 Gb/s

WLCG Network Throughput Support Unit

Support channel where sites and experiments can report potential network performance incidents:

- Relevant sites, (N)RENs are notified and perfSONAR infrastructure is used to narrow down the problem to particular link(s) and segment. Also tracking past incidents.
- Feedback to WLCG operations and LHCOPN/LHCONE community

Most common issues: MTU, MTU+Load Balancing, routing (mainly remote sites), site equipment/design, firewall, workloads causing high network usage

As there is no consensus on the MTU to be recommended on the segments connecting servers and clients, LHCOPN/LHCONE working group was established to investigate and produce a recommendation.