Towards IPv6-only in the Worldwide Large Hadron Collider Computing Grid (WLCG)

Tim Chown (Jisc), tim.chown@jisc.ac.uk
IPv6 side meeting, IETF120, Vancouver, 25 July 2024

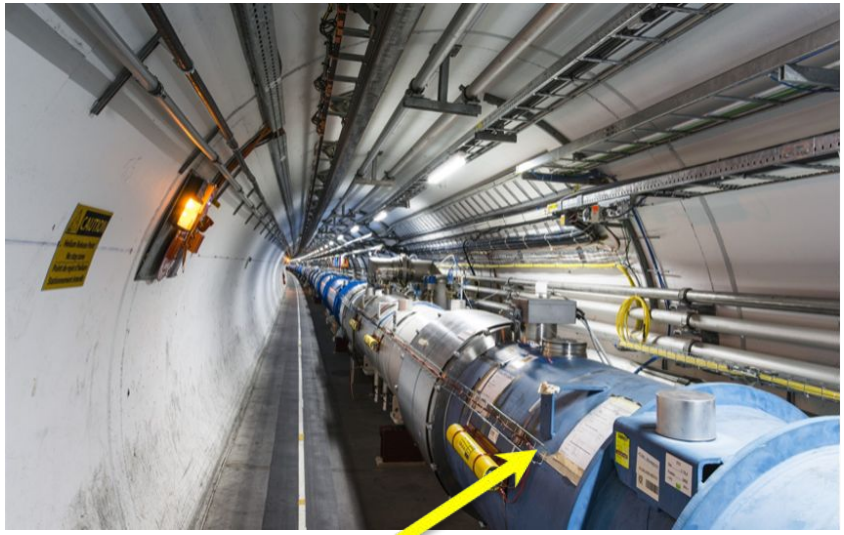
Agenda

Topics:

- About the WLCG - the LHCOPN and LHCONE networks
- The IPv6 requirement
- HEPiX IPv6 WG activity, and phases, 2011 to today
- WLCG Data Challenge 2024
- SciTags - packet marking
- Related activity, topics and pointers

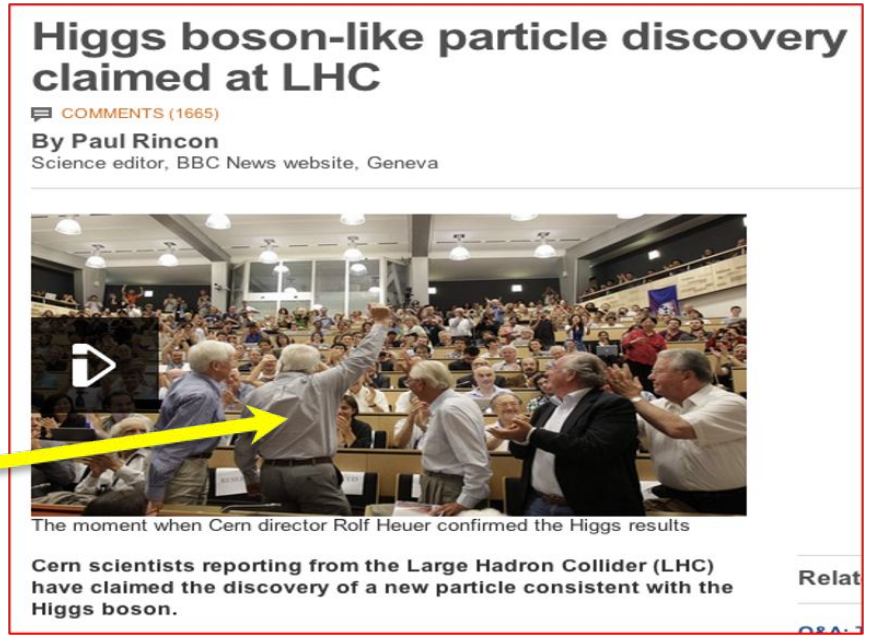
Acknowledgements:

With thanks to Dave Kelsey (STFC), Andrea Sciabà (CERN), Edoardo Martelli (CERN), Bruno Hoeft (KIT), Mihai Patrascoiu (CERN), Shawn McKee (U. Michigan) and Chris Walker (Jisc) for material included



4 July 2012

Nobel Prize in Physics 2013: F. Englert & P. Higgs



How do you get from this to this?

Worldwide Large Hadron Collider Computing Grid (WLCG)

The WLCG is a global collaboration

More than 170 computing centres in 42 countries

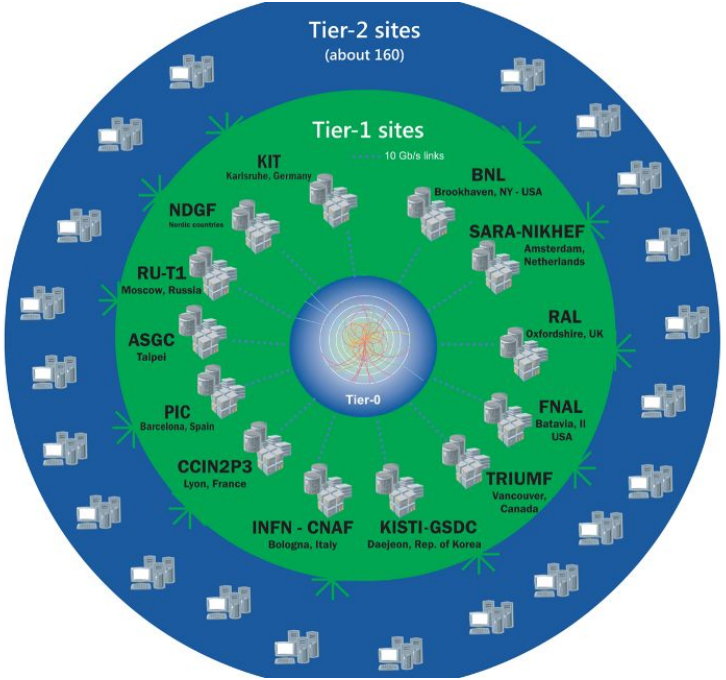
Many experiments: ATLAS, Alice, LHCb, CMS, …

Mission to **store, distribute** and **analyse** the data from the LHC experiments

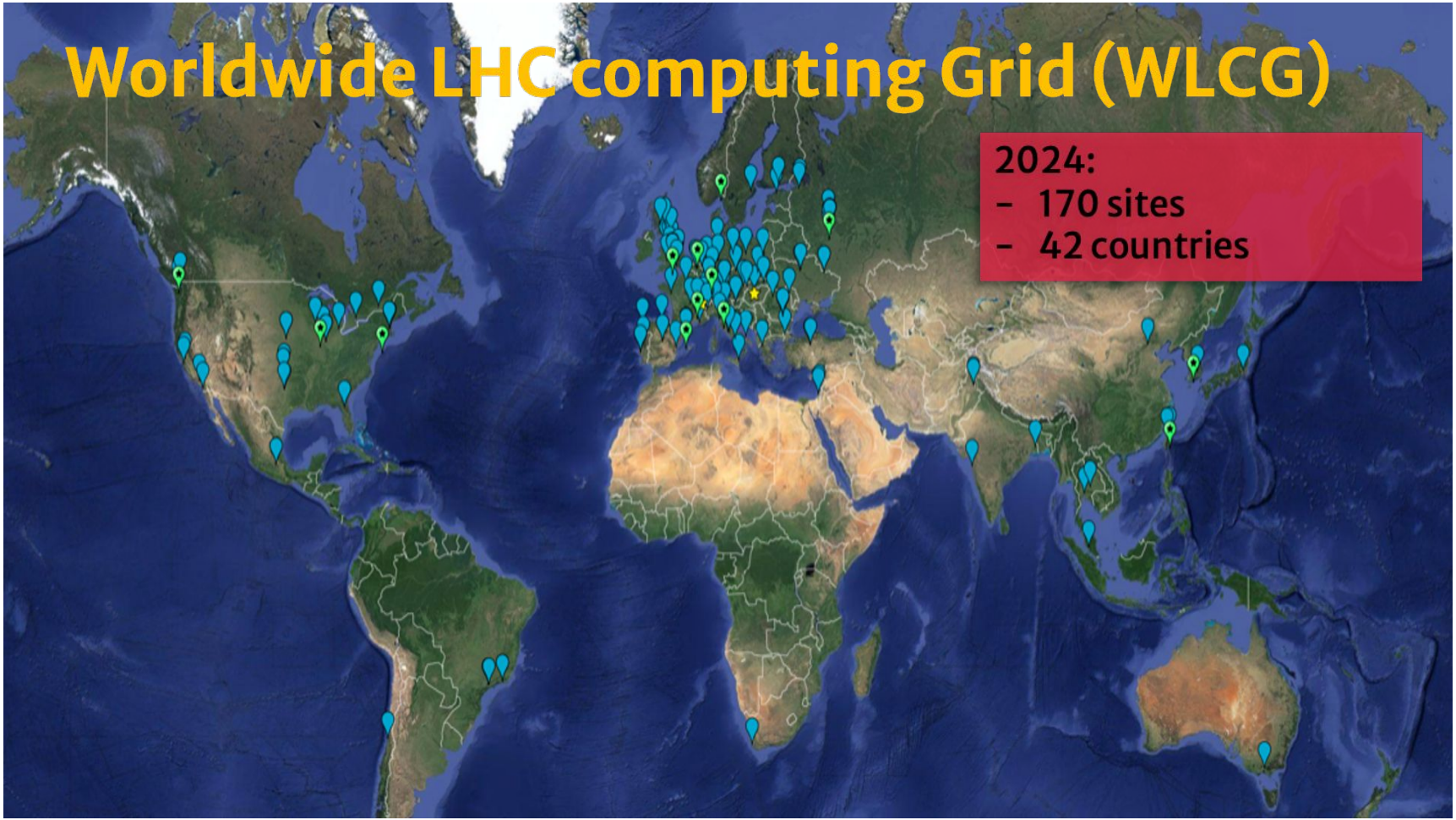
Sites in three tiers:

- Tier-0: CERN, home of the LHC
- Tier-1s: 14 significant national laboratories
- Tier-2s: 160 university physics departments

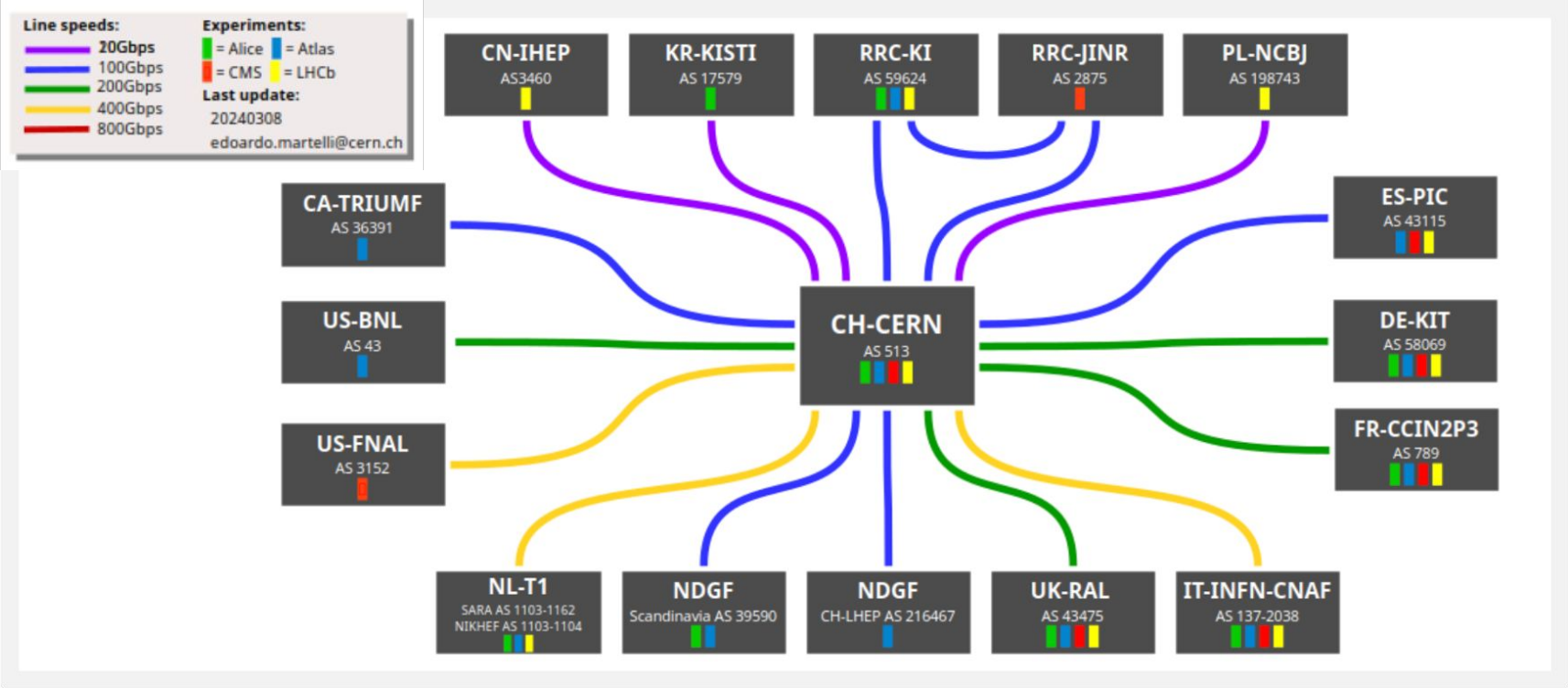
Two main networks used: LHCOPN and LHCONE



Worldwide LHC computing Grid (WLCG)



LHCOPN - Optical Private Network - Tier-0 to Tier-1s



See https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

LHCONE - the LHC Open Network Environment

A global L3VPN providing paths between sites via VRFs over the general research and education (R&E) IP network, used by Tier-1s and Tier-2s

CERN maintains prefix lists for traffic allowed on LHCONE (for IPv4 and IPv6)

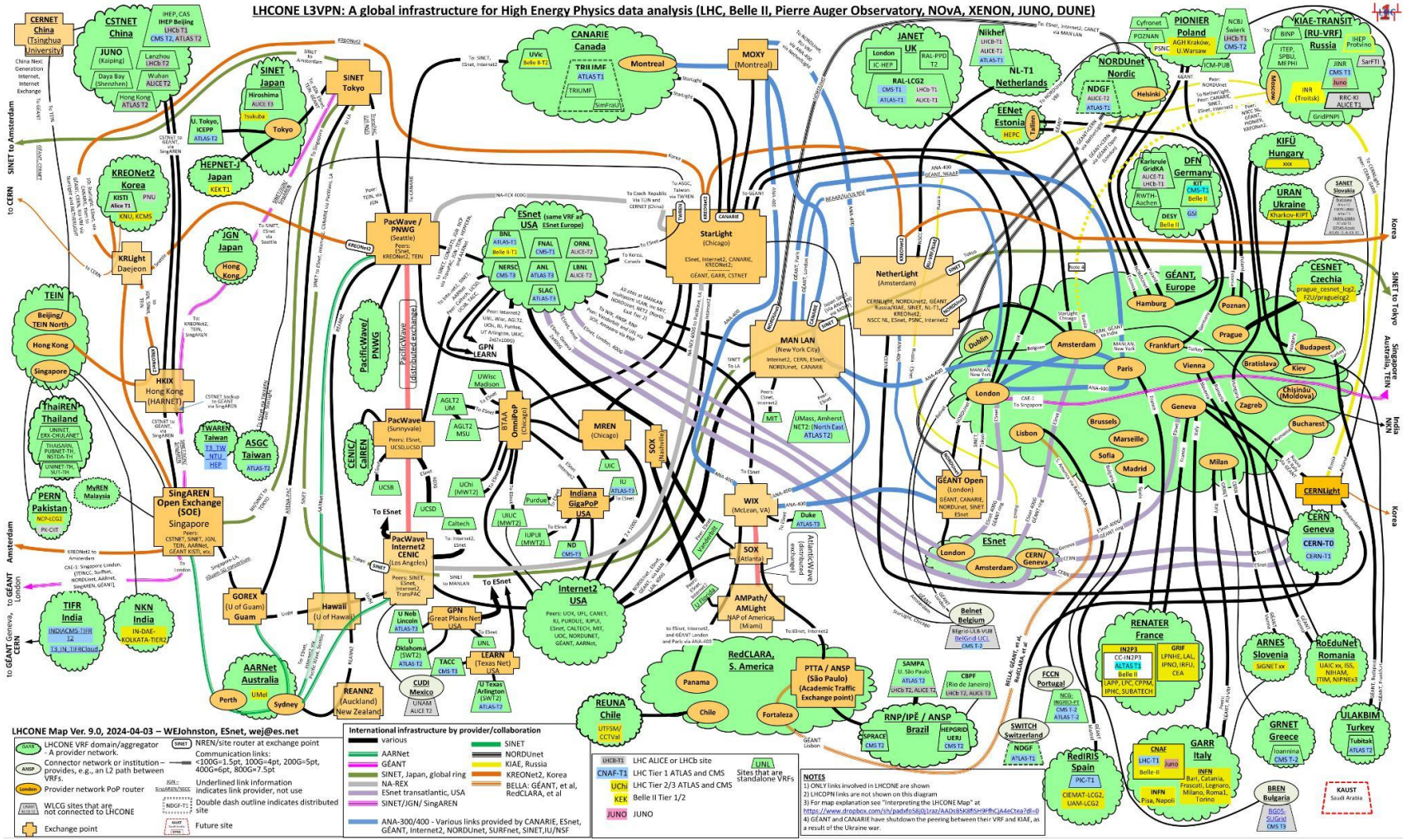
Acts as a trust network as well as enabling traffic engineering where needed

Sites not required to join LHCONE, but it's generally advantageous to do so

Tier-2 sites not on LHCONE use the general R&E IP network

See https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneVRF

(Other research communities have asked to use LHCONE, discussion is ongoing)


LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO, DUNE)
LHCONE Map Ver. 9.0, 2024-04-03 – WEJohnston, ESnet, wej@es.net
International infrastructure by provider/collaboration
NOTES
1) ONLY links involved in LHCONE are shown
2) LHCOPN links are not shown on this diagram
3) For map explanation see "Interpreting the LHCONE Map" at https://www.dropbox.com/sh/padxfo58jlj1rag/AADs85K8fiSH9FfhCjA4eCtea?dl=0
4) GÉANT and CANARIE have shutdown the peering between their VRF and KIAE, as a result of the Ukraine war.

UK context: GridPP

A collaboration of UK institutes providing data-intensive distributed computing resources for the UK High Energy Physics community

RAL is the UK WLCG Tier-1

Connected via Janet, the UK National Research and Education Network (NREN), which is operated by Jisc (who I work for)

Janet connects to the global R&E network via GÉANT, see https://map.geant.org/.

Janet backbone is up to 800G, its peering to GÉANT (for R&E IP including LHCONE) is 400G.

High-level view of WLCG data flows

The WLCG consists of network, storage and compute elements

CERN is at the heart, as the Tier-0 and point of raw data capture for experiments

Tier-1s are significant facilities, Tier-2s are generally at university campuses

The network is largely LHCOPN (private optical) and LHCONE (L3VPN/VRF)

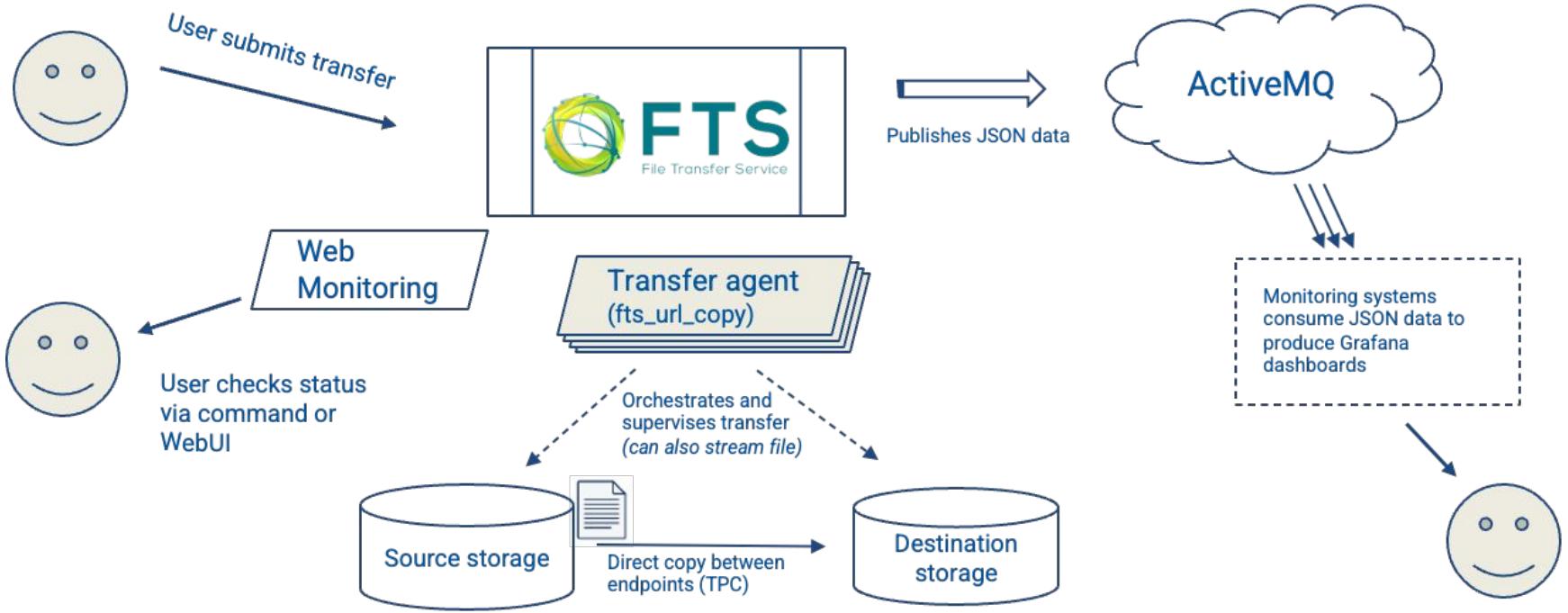
Currently very limited use of public cloud or commercial networks

Data generally flows Tier-0 -> Tier-1 -> Tier-2, but the original ‘strict’ hierarchy was relaxed

Compute resources may pull data directly from another site’s storage

Data movement orchestrated and managed by Rucio and CERN’s FTS software

Overview of CERN File Transfer Service (FTS)



WLCG infrastructure administration

Network:

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

Campus infrastructure

- Connecting local WLCG campus 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 large number of different administrative teams supporting the WLCG

Local Tier-2 architectures

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, dedicated data transfer nodes (DTNs) - TCP buffers, CCAs, MTUs,...
- Appropriate security implementation supporting the performance mission - thus generally ACL-based rather than (expensive at scale) stateful DPI firewalls

Note that WLCG sites do **not** require IPv6 be enabled on the whole campus; IPv6 can be, and often is, just enabled to/from and within the Tier-2 system elements

General history of IPv6 deployment in R&E networks

The NREN backbones have had dual-stack IPv6 since the early 2000's

But campuses are well behind the commercial ISPs, just like most corporate enterprises, nowhere near the 40-45% worldwide level

To date, arguments for deploying on campuses have not led to significant deployment, be that to support teaching and research, to secure the IPv6 that is present in an "IPv4 only" network, or to facilitate innovation and smart campus technology that may use IPv6

However, participation in WLCG **is** a higher priority reason for sites to deploy IPv6, for at least the part of the network where the WLCG resources are hosted

While WLCG can use the existing IPv6 in the NREN backbones, it needs to coordinate with both the campus IT teams and local WLCG teams for successful deployment

The origins of IPv6 interest in the WLCG

WLCG ran a survey in 2011 on IPv6 readiness for its community

Triggered by the IANA statement on IPv4 exhaustion (13 years ago!)

NRENs were IPv6-ready in 2011, university / research sites generally not

Some sites were running out of IPv4 (though most had a long-standing Class B)

WLCG noted that opportunistic offers of IPv6-only CPU resources could arise at any time, and that the middleware, software, technology and tools were generally not IPv6-capable

To address this, the HEPiX IPv6 WG was formed to move IPv6 adoption forward

It was expected back then it would take a long time to resolve all the issues

See https://twiki.cern.ch/twiki/bin/view/LCG/WlcgIpv6

Additional IPv6 rationale for WLCG

US government directive M-21-07. This applies to the WLCG experiment facilities at Fermilab/FNAL (CMS) and Brookhaven/BNL (ATLAS)

- See https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-07.pdf
- Everyone benefits from vendors implementing IPv6 support in their products in response to the directive

Avoidance of NATs and proxies

SciTag per-packet marking - only supported by IPv6 (using the Flow Label)

Ability to scale: expand sites and/or introduce new sites

HEPiX IPv6 WG Phase 1: 2011-2016

Carried out a full analysis of work to be done

Applications, middleware, system and network tools, security

Created and operated a distributed test-bed

Initial plan to be able to support IPv6-only clients drawn up in 2014

Test the important data transfer protocols, technology and data storage / file systems for IPv6 readiness

Fixing the storage and data transfers took more than 5 years

Required working with both campus and WLCG teams at organisations

Aside: Ticketing

WLCG uses the GGUS ticketing system (private to WLCG participants, sorry!)

Allows tickets to be raised for any issue at any site, including IPv6-specific ones

GGUS helps drive campaigns that all sites are encouraged to respond to, e.g., to IPv6-enable their storage elements

Campaigns are also tracked on the WLCG wikis, where GGUS tickets for each site can be linked for easy reference

Important tool for HEPiX IPv6 WG members to target help where it’s needed

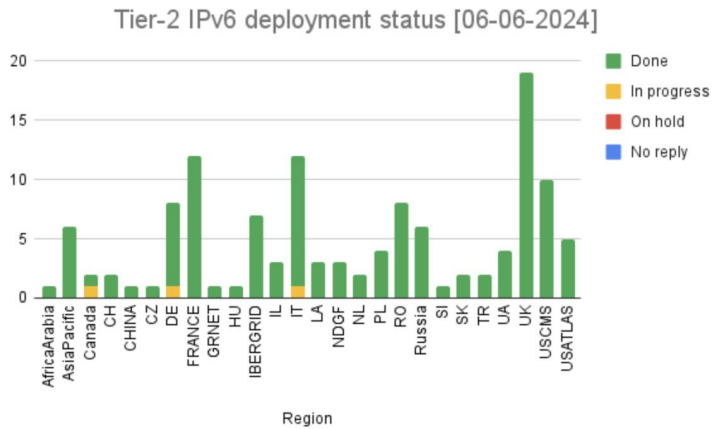
HEPiX IPv6 WG Phase 2: 2017 onwards

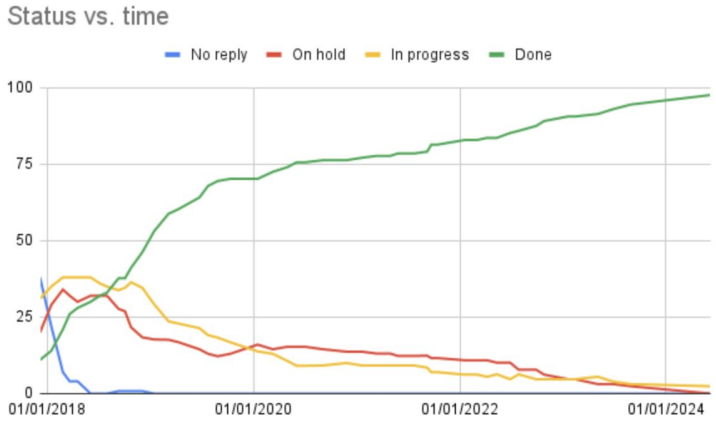
Campaign: enabling IPv6 for Tier-2 storage from Nov 2017

Current status shows > 98% of storage is IPv6-enabled (dual-stack)

VO	T2 storage on IPv6 (%)
ALICE	94
ATLAS	97
CMS	100
LHCb	100
WLCG	98

(checked on 06-06-2024)

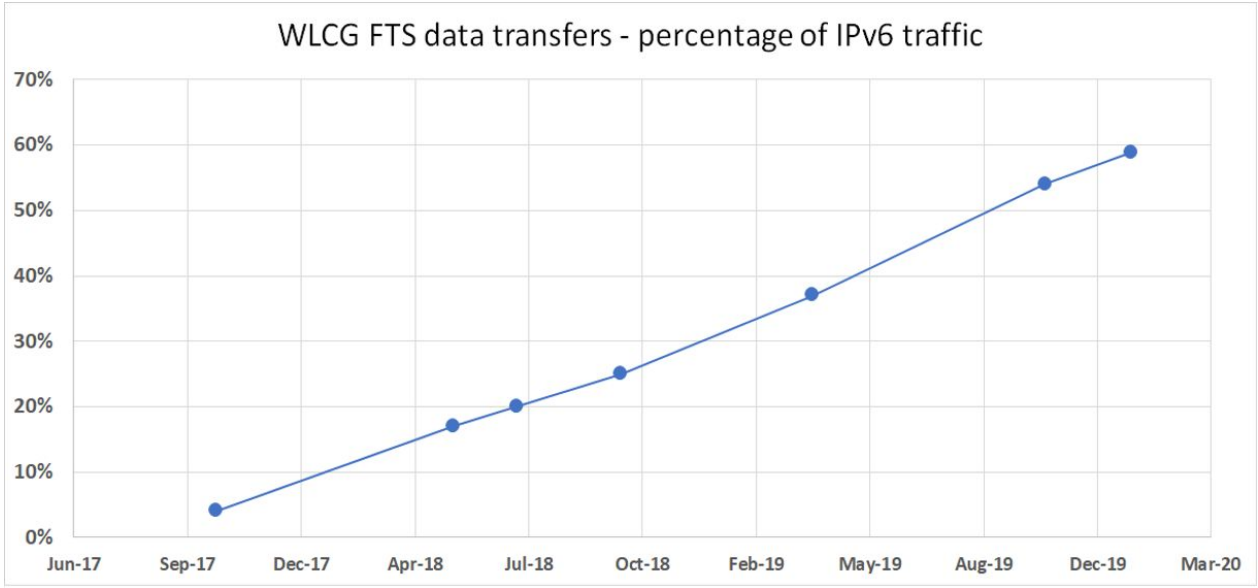




See https://twiki.cern.ch/twiki/bin/view/LCG/WlcgIpv6#WLCG_Tier_2_IPv6_storage_deploym

Percent of WLCG data transfers over IPv6

2017-2020, all experiments - as measured by FTS for GridFTP transfers

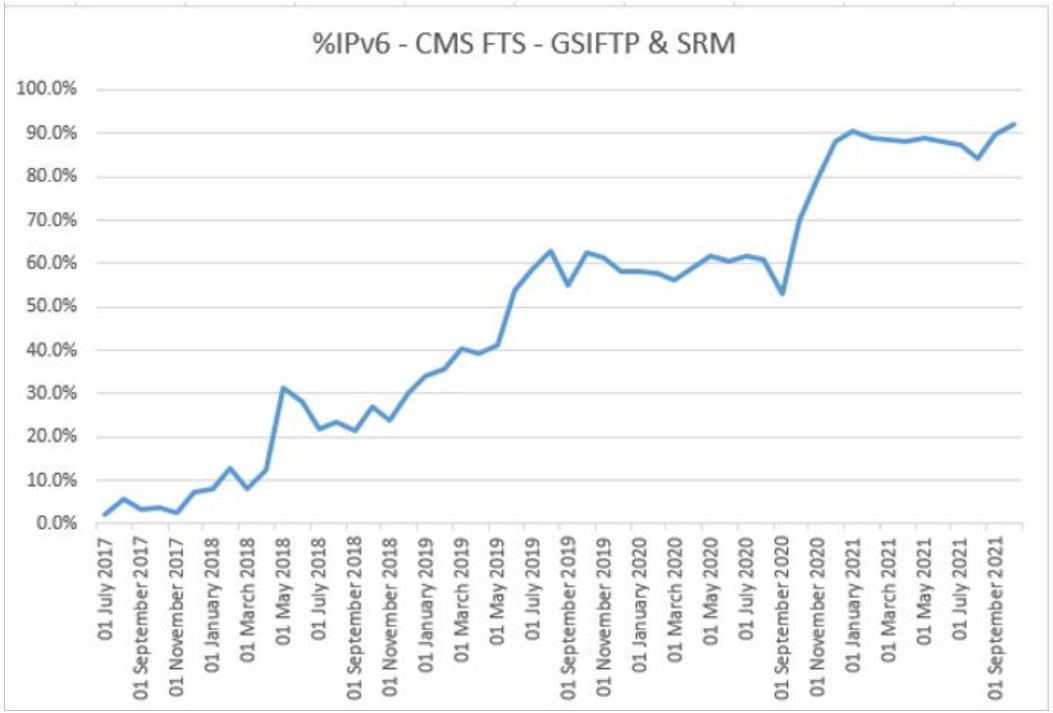


IPv6 works!

Experiments and physicists are happy

and unaware of the protocol used!

% of CMS experiment data transfers IPv6



Experiments no longer using GSIFTP & SRM

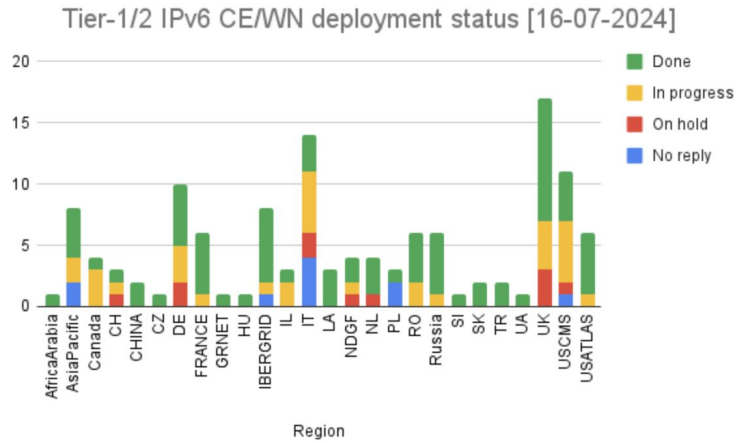
Reason for no plots in 2022

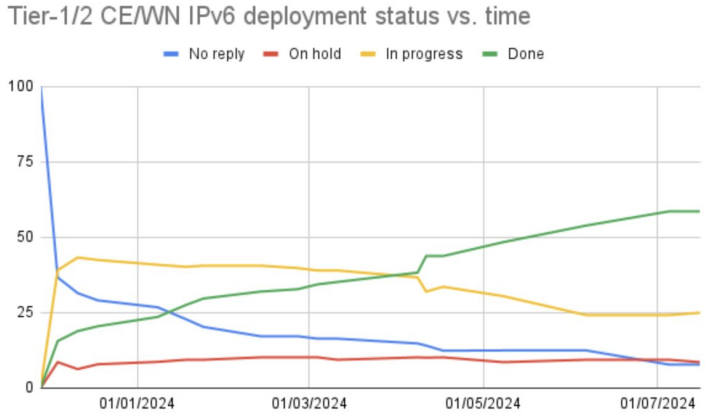
IPv6 WebDAV is not yet “visible” in our data monitoring!

IPv6 on compute elements (CE) and worker nodes (WN)

Campaign: enabling IPv6 for WNs/CEs, from Dec 2023, with an aggressive **June 2024** deadline

Current status shows 58% of compute resource now IPv6-enabled (**dual-stack**)





See https://twiki.cern.ch/twiki/bin/view/LCG/WlcgIpv6#WLCG_IPv6_CE_and_WN_deployment_s

Example: Imperial College London

UK WLCG Tier-2 site, on LHCONE

2x100G to Janet, one data centre, one campus

After its 100G upgrade the DC link soon filled (95.2Gbps) with CERN data, often 100% IPv6

LHCONE in green, other IP traffic in orange

Aside: Imperial is full dual-stack, looking to remove IPv4. Running IPv6 Mostly (with DHCPv4 Option 108) successfully over ~200 WiFi APs with eduroam, and planning to extend it to a few thousand APs later this summer

Issues reported with enabling IPv6 for WNs/CEs

Common examples:

- Delays where enabling IPv6 needs to be coupled to or depends on other changes such as OS updates, new hardware, internal routing changes, …
- Some sites have more complex or special configurations to consider with respect to NAT for IPv4 and global IPv6, e.g., needing to replace IPv4 NAT(s) with dual-stack router(s)
- Other priorities, like new WLCG auth tokens or handling the CentOS7 end-of-life
- Concern that WNs currently behind IPv4 NAT will become more “exposed”
- Lack of local expertise

Only 8% of sites have not responded to the campaign, 9% are “on hold”, the rest should be complete soon.

Also worth noting that some sites keen to go IPv6-only now (though not **yet** recommended)

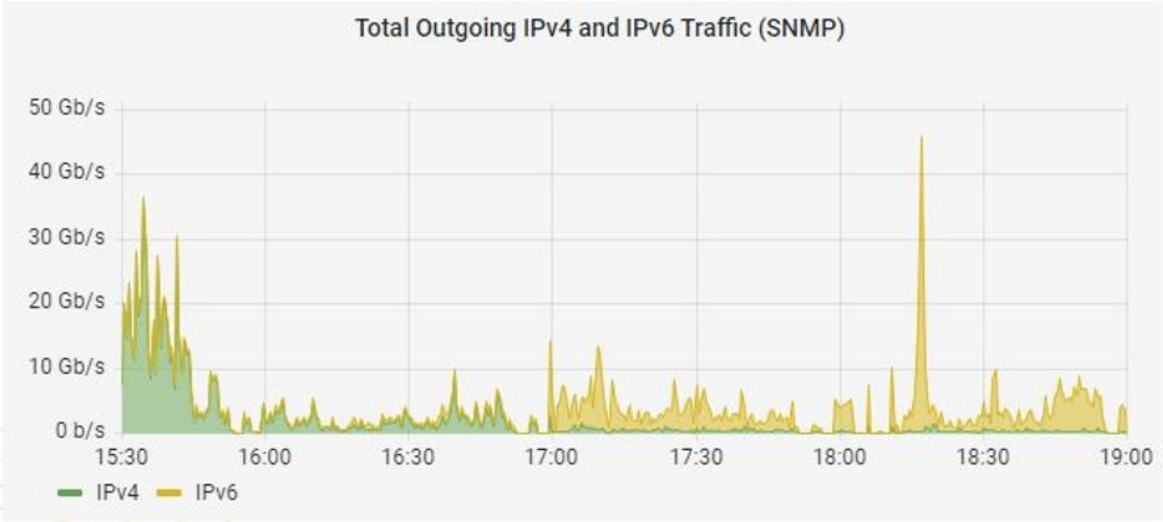
IPv6 support in other required services and tools

This is in a good position now. Examples include:

- Rucio - higher level data storage management - https://rucio.cern.ch/
- FTS - data movement orchestration - https://fts.web.cern.ch/fts/
  - Supports many third party transfer tools - GridFTP, XRootD, WebDAV/https, S3, …
- XRootD - third party data transfer tool - https://xrootd.slac.stanford.edu/
- HTCondor - high throughput cluster computing - https://htcondor.org/
- dCache - distributed cache - https://wlcg-ops.web.cern.ch/dcache
  - Interesting example of where a ‘prefer IPv6’ toggle needs to be set!
- CVMFS - CERN VM file system
  - Has a similar toggle - cvmfs_ipfamily_prefer=6
- Puppet - for configuration management

IPv4/IPv6 choice for dCache/WebDAV transfers

java.net.preferIPv6Addresses (default: false) **- Now set to "true"**



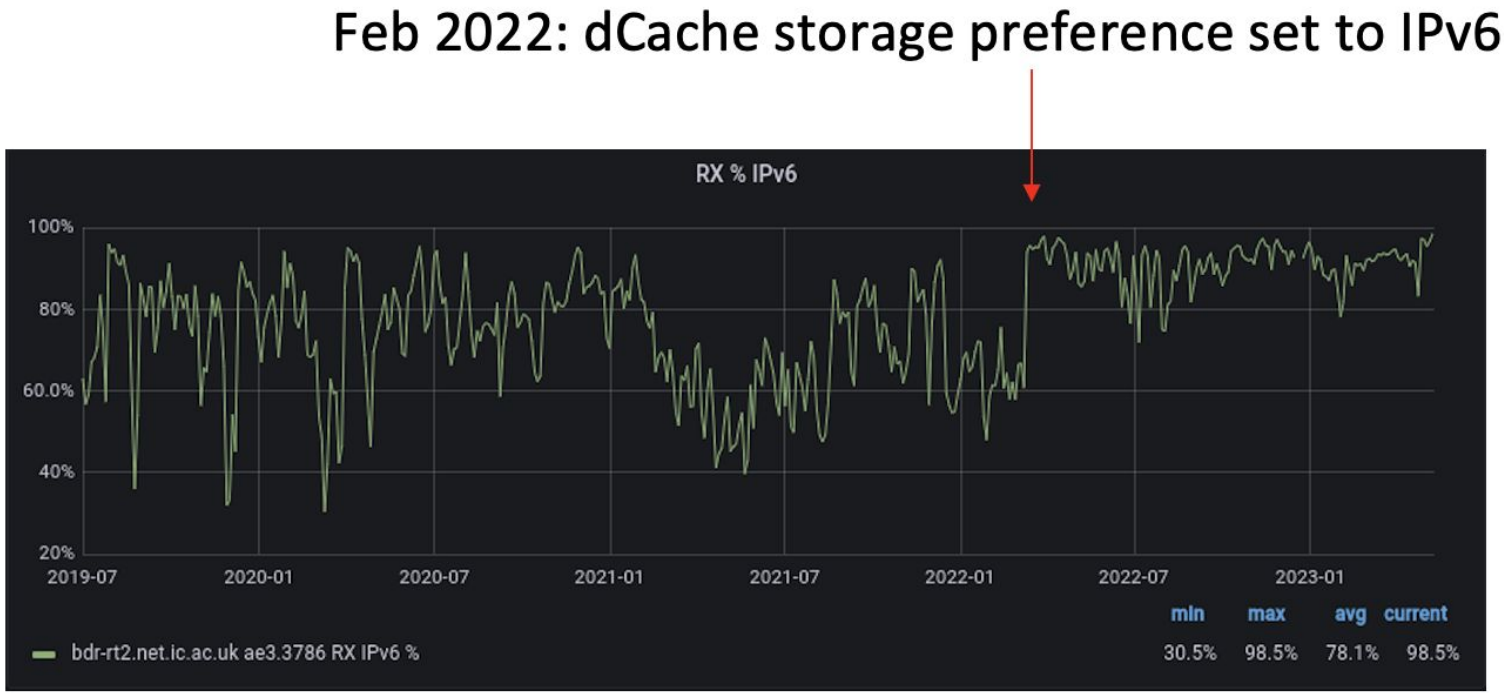
Green: IPv4; Yellow: IPv6

Default behaviour changed to prefer IPv6 at 17:00 local time on 14 Feb 2022

The fix works!

Then asked all sites to change the configuration

% IPv6 on LHCONE for Imperial College



Phase 3: IPv6-only

Consensus for IPv6-only to simplify operations: reduce complexity, streamline security

Positive examples of other worldwide infrastructures running IPv6-only, e.g., Facebook

- But that is one organisation, not a community of 140 different organisations

As we just saw, 98% of storage is IPv6-enabled, compute is at 58% and growing

The WLCG is working towards IPv6-only for the majority of WLCG services and clients but needs to consider the remaining IPv4-only clients (compute resources)

Timetable still to be defined and agreed with Management Board

Might start with data movement over Tier-0 to Tier-1 links (which have separate IPv4/IPv6 VLANs)

What have been the obstacles to IPv6?

Dual-stack is still considered an essential step on the journey to IPv6-only

Many quite detailed issues have been encountered

The higher-level challenges addressed by the HEPiX IPv6 WG include:

1. WLCG sites not yet deployed IPv6 networking (~done)
2. Sites have IPv6 but Tier-2 has no dual-stack storage (~done)
3. Lack of IPv6 support on compute resources (relatively new campaign)
4. IPv6 monitoring is not available or broken
5. Service is dual-stack but IPv4 being used (a heavy focus recently)

Issue 5 is often ‘just’ a bad toggle default, but may be more subtle - new examples still arise

WLCG Data Challenge 2024

Organised for two weeks in Feb 2024

Preparation exercise for LHC high luminosity phase starting around 2029

Plan - inject extra traffic at ~25% of HL level

- Total target across all sites = 2,430 Gbps
- Expected requirement in 2029 = 9,620 Gbps

Find bottlenecks - Backbones? Campuses? Storage? Elsewhere?

DC24 let us observe IPv6 usage and identify where IPv4 is still seen and why

- Looked at specific links, e.g., T0 CERN -> T1 KIT (DE)

Monitoring traffic and network characteristics

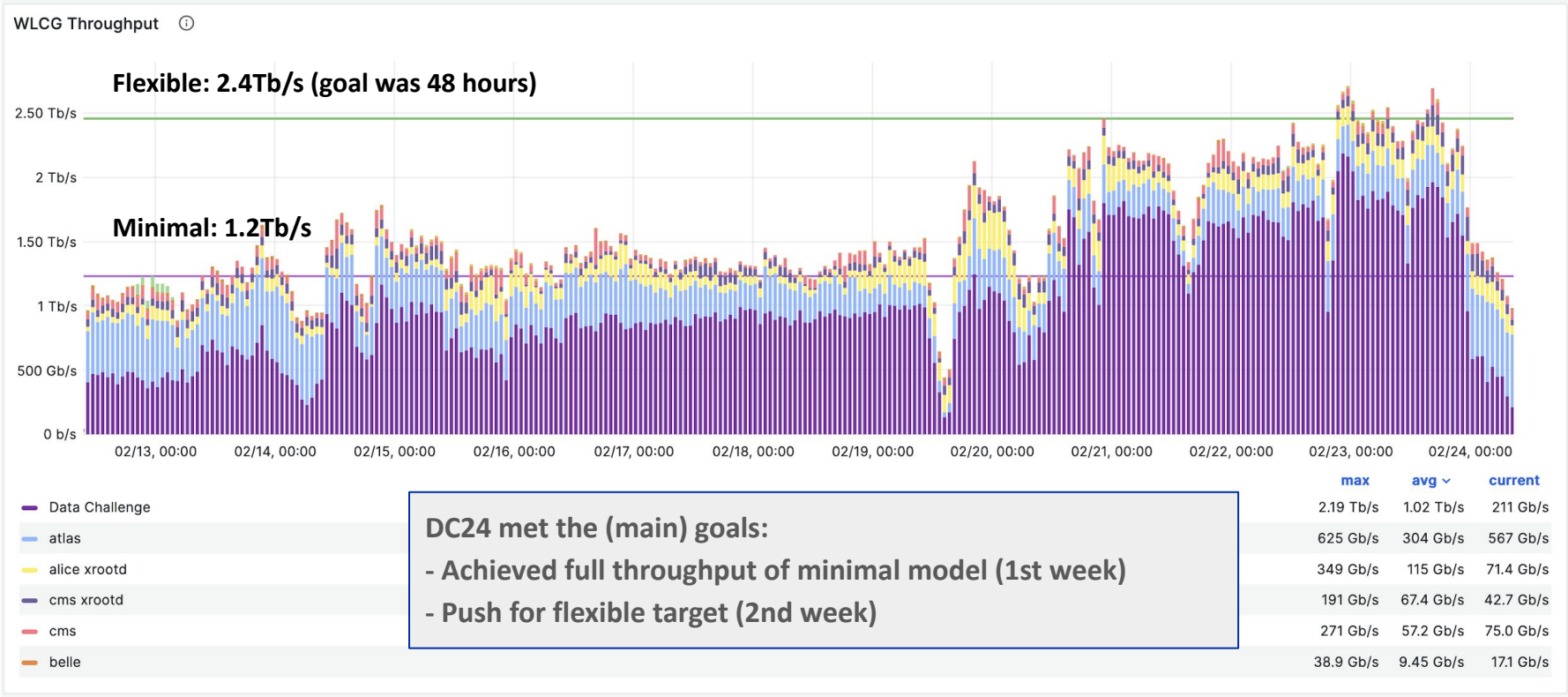
The WLCG can draw on various sources of traffic information:

- Application oriented - FTS logs
- Router interface utilisation at site egress - sites were requested to expose this data to a CERN collector for DC24
- perfSONAR - open source platform to measure latency, loss, path and throughput (most sites have at least one perfSONAR node) - https://www.perfsonar.net/
- Netflow records - kept by sites for a short period of time

Allows reasonable investigation into causes of IPv4 traffic

- Some may be intended, e.g., perfSONAR tests IPv4 and IPv6

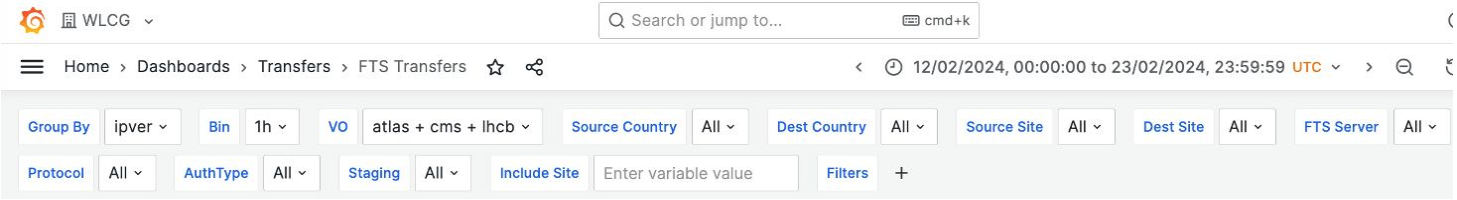
DC24 overall throughput by experiment

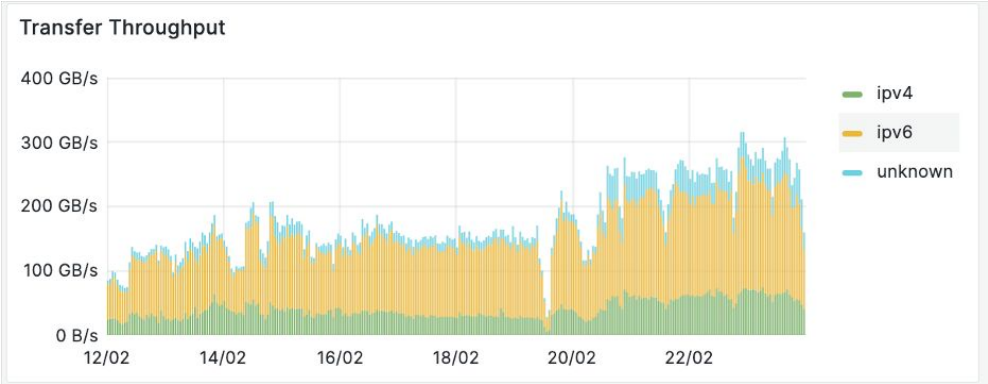


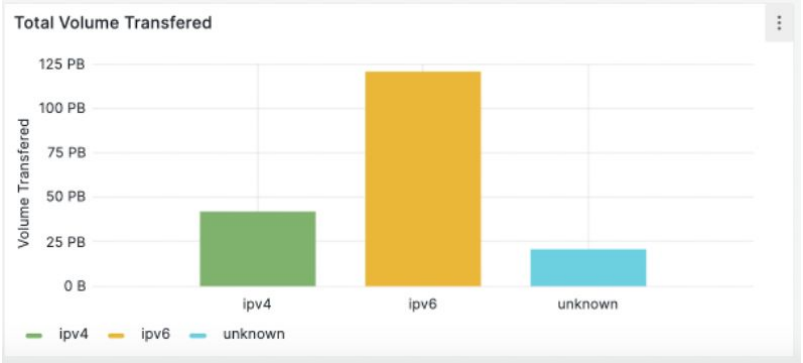
DC24 met the (main) goals:

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

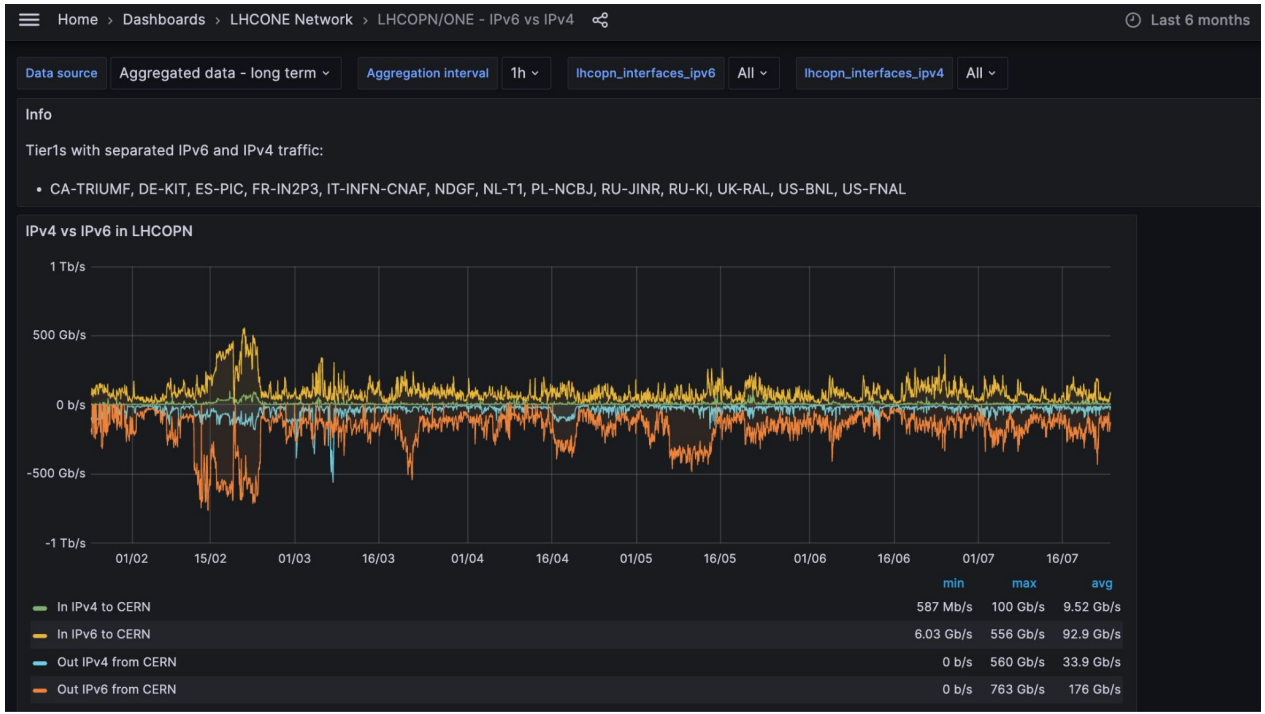
DC24: Relative FTS use of IPv4 and IPv6







LHCOPN IPv4 vs IPv6, volume, last 6 months



See https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4?orgId=16&from=now-6M&to=now

LHCOPN IPv4 vs IPv6, in/out %, last 6 months

See https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4?orgId=16&from=now-6M&to=now

DC24: IPv6 observations

There have been per-experiment debriefings; much of the focus has been on bottlenecks, in particular FTS tuning (concurrent flows) and auth token operation

IPv6 analysis looked at traffic levels and netflow data.

Examples of observations:

- Many OPN links were fully IPv6, e.g., PIC, RAL
- Two OPN links were all IPv4 one way and all IPv6 the other (CNAF, JINR)
- One had a surprisingly high level of IPv4 traffic
- It was noted that IPv6 transfers have a higher successful completion rate
- Useful lessons learnt at KIT from the more detailed netflow analysis there

The site egress traffic utilisation collection set up for DC24 was not IP version-specific - this will be addressed for DC26 (which will be 50% of LHC-HL traffic levels)

SciTags - marking WLCG traffic

An IPv6-specific capability and additional benefit for using IPv6

Defined by the WLCG Research Networking Technical Working Group (RNTWG)

Rationale is to allow NRENs or WLCG participants to identify and account for the experiment and activity associated with traffic seen on the networks

Uses IPv6 Flow Label - IETF draft - 20 bits: 9 for the experiment, 6 for activity, and 5 entropy bits

Written up as IETF ID: https://datatracker.ietf.org/doc/draft-cc-v6ops-wlcg-flow-label-marking/

There are also per-flow UDP “firefly” packets under test, which can be IPv4 or IPv6 - these were successfully demonstrated at some scale with XRootD support during DC24

See https://scitags.org

Other WLCG RNTWG activities

The WG studies a range of technologies that may be used to enhance the performance and capabilities of the WLCG. These are often drawn from IETF WGs and outputs.

Recent examples include:

- TCP-BBRv3
- Use of jumbo frames - and related IPv6 PMTUD operation
- Packet pacing - addressing microbursts and small buffer network devices
- SciTags

Testing is run both within labs and between WLCG sites worldwide.

perfSONAR can be used to vary CCA, MSS, TCP buffers, etc

WLCG is not (yet) using BBRv3 - **awaiting inclusion in production Linux distributions**

Other IETF IPv6 considerations

The WLCG, through the HEPiX IPv6 WG, tracks and has recently contributed to IETF WGs

Not everything is applicable though, e.g.:

- IPv6 Mostly - the WLCG generally manages servers not clients, with configuration by puppet/ansible/etc
- Happy Eyeballs - there is WebDAV/https traffic, but not browser traffic. Most applications have IP version preference toggles, and may or may not failover

Summary

IPv6 on WLCG is the flagship example of IPv6 in R&E networks, but it has taken 13 years

Tier-1 storage 100% IPv6, Tier-2 is 98% IPv6-enabled

- So the WLCG now effectively supports IPv6-only clients as per the original goal

Most data transfers use IPv6; LHCOPN/LHCONE is 90-95% IPv6

- (Annoying) challenge is hunting down use of IPv4 when both ends have IPv6 enabled

Obstacles to IPv6 continue to be addressed

- Current focus on IPv6 on WNs/CEs (58% and rising) and WLCG services

End-game remains IPv6-only services; IPv4 is legacy networking

Any new research infrastructure should build with IPv6 from day 1 - SKA is doing so