Towards IPv6-only on CERN's Worldwide Large Hadron Collider Computing Grid (WLCG)

Tim Chown (Jisc), tim.chown@jisc.ac.uk
TechEX 2024, Boston, 11 December 2024

Agenda

Topics:

- The WLCG - the LHCOPN and LHCONE networks
- The HEPiX IPv6 WG and IPv6 requirement
- Phases: analysis and enabling the storage and compute elements
- IPv6 in the 2024 WLCG Data Challenge
- IPv6 innovation - Scitags per-packet marking
- Towards IPv6-only

Co-authors: Nick Buraglio (ESnet), Tim Chown (Jisc), Dale Carder (ESnet), Bruno Hoeft (KIT), David Kelsey (UKRI-STFC), Edoardo Martelli (CERN), Carmen Misa Moreira (CERN), Francesco Prelz (INFN - Sezione di Milano), Andrea Sciabà (CERN).

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

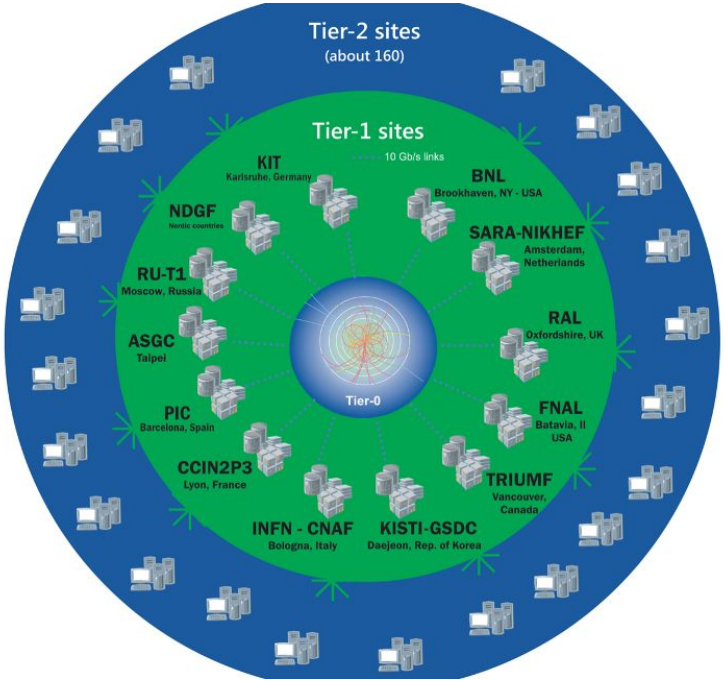
Sites in three tiers:

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

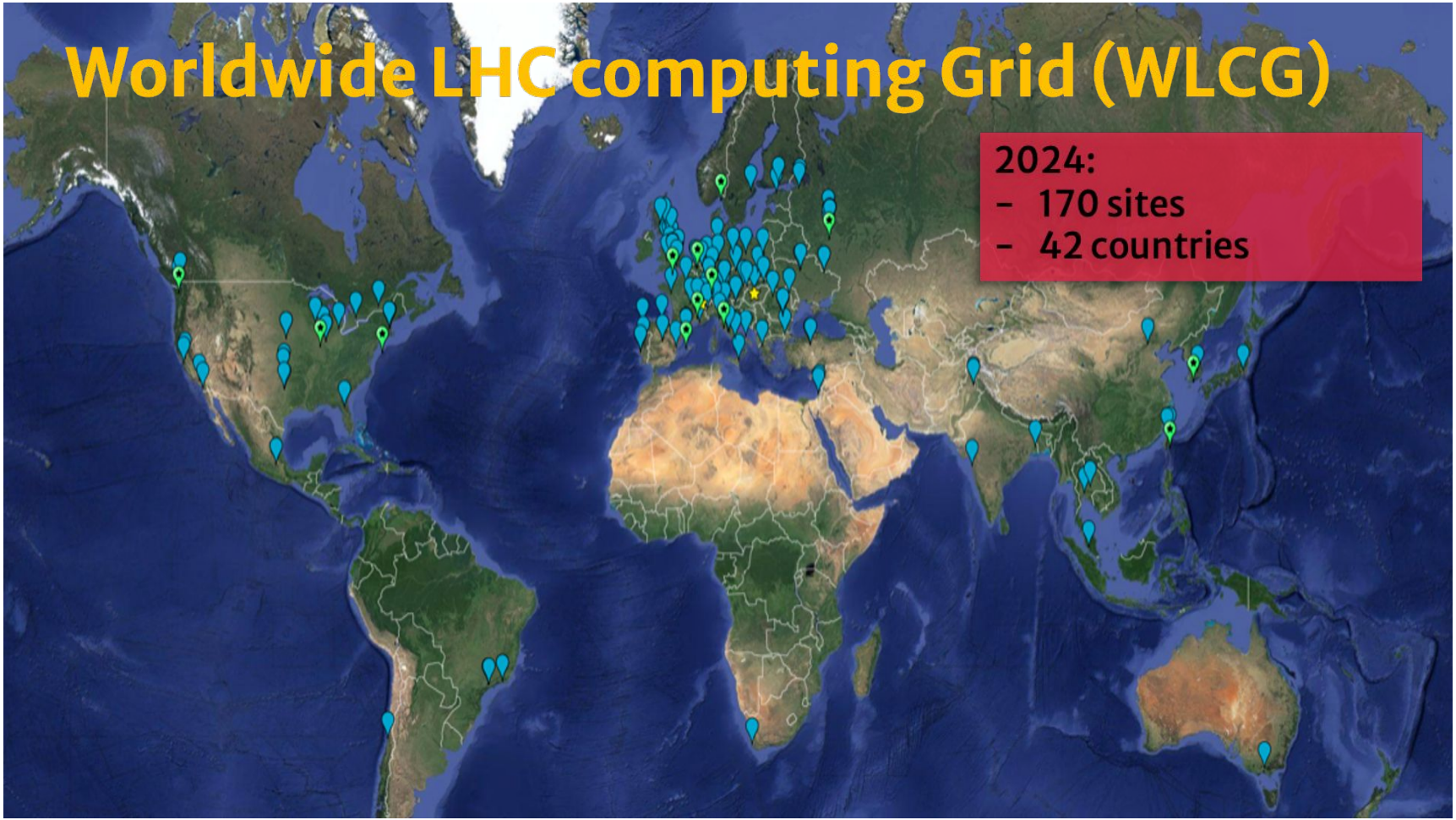
Two main networks used: LHCOPN (private optical) and LHCONE (L3VPN/VRF)

Transfers typically orchestrated via Rucio/FTS/XRootD

Very limited use of cloud or commercial networks



Worldwide LHC computing Grid (WLCG)



WLCG infrastructure administration

Network:

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

Campus network infrastructure

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

Particle physics storage and compute facility

- Local WLCG teams are the physics department or campus computing centre staff

The large number of different administrative teams supporting the WLCG makes coordinating change more challenging - each site needs to deploy IPv6 for its own WLCG facility

Why the WLCG's original interest in IPv6?

Triggered by the IANA statement on IPv4 exhaustion…. **13 years ago**!

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

WLCG sites were surveyed for IPv6 readiness

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

But also important to be able to scale the WLCG - to be able to expand sites and add new sites

And be performant - avoid, NATs, proxies, and unnecessary network middleboxes

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

More recent reasons to deploy IPv6

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

SciTags - accountability for traffic

- Per-flow marking with UDP ‘fireflies’ (IPv4 or IPv6)
- Per-packet marking - only supported by IPv6, using the Flow Label

The (slow) pace of IPv6 deployment in R&E networks

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 network IT teams and local WLCG teams for successful deployment

The HEPiX IPv6 WG

It was clear that expertise was needed to both assist and drive the adoption of IPv6

Hence the creation of the HEPiX IPv6 WG, chaired by Dave Kelsey (STFC)

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

Phase 1 (2011-2106):

- Analysis of work to be done
- Review of applications, middleware, system and network tools, security
- Creation and operation of a distributed test-bed

Initial plan to support IPv6-only clients (worker nodes and compute elements) drawn up in 2014

Led to Phase 2 goal of IPv6-enabling (dual-stack) all storage

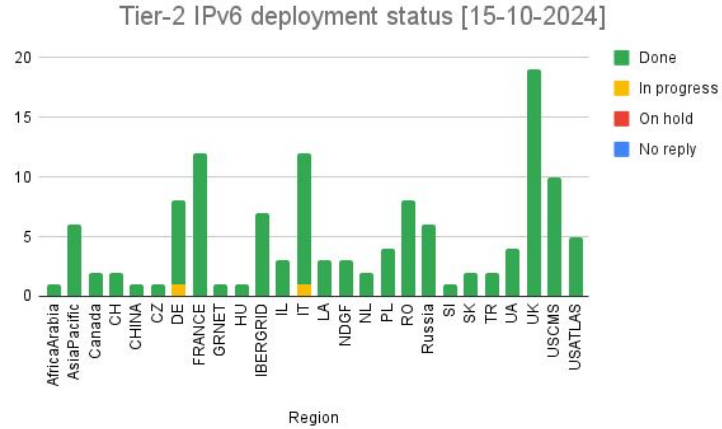
Phase 2: IPv6-enabling storage (2017-2023)

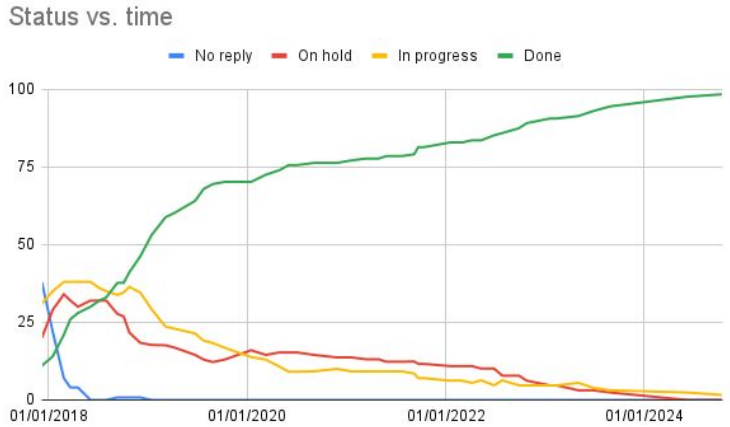
Ticket campaign: enabling IPv6 for Tier-2 storage - slow but steady

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

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

(checked on 15-10-2024)





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

Phase 3: *Towards* IPv6-only (2019-)

Running IPv6-only would 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 170 different organisations

Storage is IPv6-enabled, but to move to IPv6-only we need to support IPv6 in all the worker nodes (WNs) and compute elements (CEs)

- Allow efficient, direct WN/CE communication with storage. No NAT.

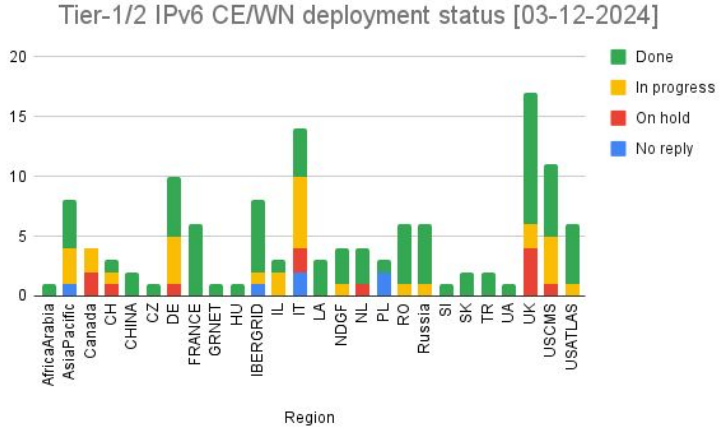
Led to a new ticketing campaign for WNs/CEs

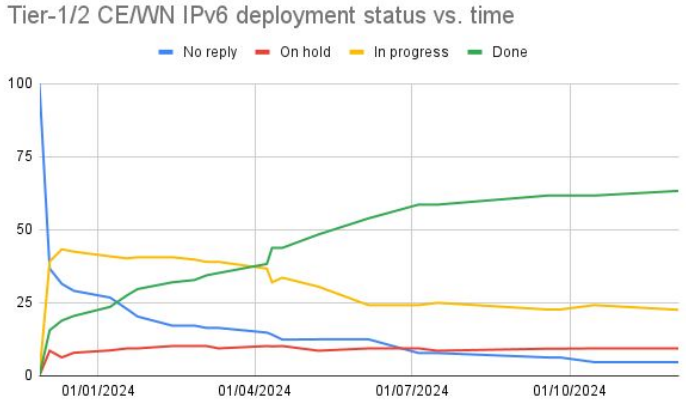
- Some already enabled, but a driven campaign was required to encourage others

IPv6 on WNs and CEs

Ticket campaign: enabling IPv6 for WNs/CEs, from Nov 2023, initial aggressive **June 2024** deadline

Status today: 63% of compute resource now IPv6-enabled (**dual-stack**), 23% “in progress”





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

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, etc.
- 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 CentOS 7 end-of-life
- Concern that WNs currently behind IPv4 NAT will become more “exposed”
- Lack of local expertise

Only 5% of sites have not responded to the campaign, 9% are “on hold”.

Also worth noting that some sites keen to go IPv6-only now (though not **yet** recommended), e.g., Brunel University is currently piloting an IPv6-only cluster.

IPv6 and 9000 MTU

Throughput can be significantly improved by using jumbo frames

WLCG proposal a few years ago was to support 9000 MTU on path

There's a reasonable level of adoption at Tier 1/Tier 2 sites. CERN is testing.

The larger (non-standard) MTU is a potential problem for IPv6, because IPv6 does not fragment on the path, so Path MTU Discovery (PMTUD) MUST work to allow 9000 MTU sites to talk to 1500 MTU sites

Issues seen where some sites over-filter ICMPv6, against RFC 4890 recommendations

The Fasterdata site recommends setting `net.ipv4.tcp_mtu_probing=1`

This issue has made some sites nervous of IPv6, and jumbo frames

IPv6 support in other required services and tools

This is required, and 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

Summary of obstacles to IPv6

Dual-stack has been 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:

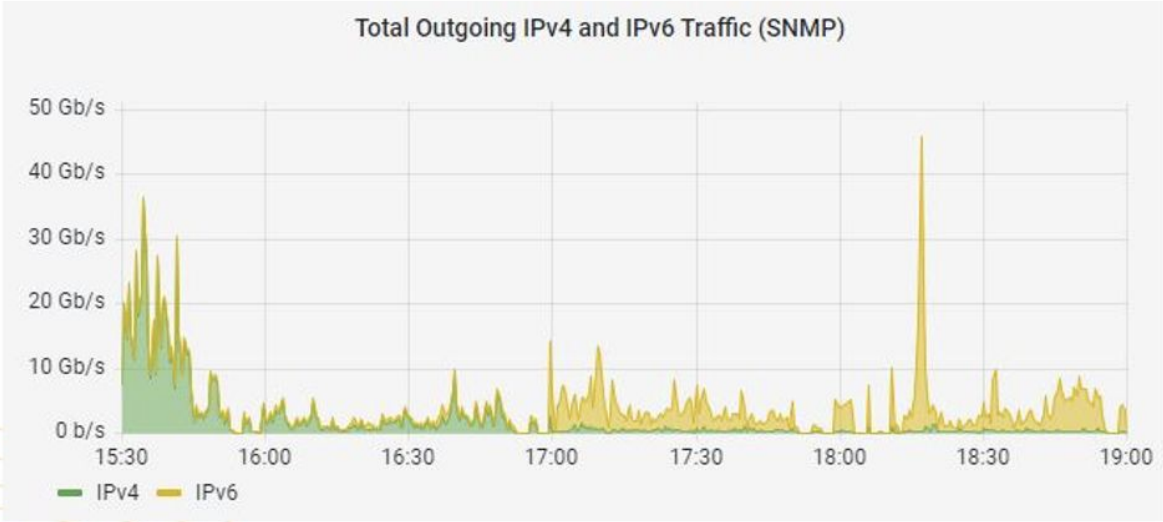
1. WLCG **sites** had not deployed IPv6 networking (~done)
2. Sites have IPv6 but Tier-2 has no **dual-stack storage** (~done)
3. Lack of IPv6 support on **compute resources** (new campaign, 63% there)
4. **IPv6 monitoring** not available or broken (still a challenge, to differentiate IPv4/IPv6)
5. **Service dual-stack but IPv4 is still being used** (a heavy focus recently)

Issue 5 is often ‘just’ a bad toggle default, but may be more subtle

One example is Java, as identified at Imperial College (UK Tier 2)...

Example: dCache/WebDAV transfers (Imperial College)

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



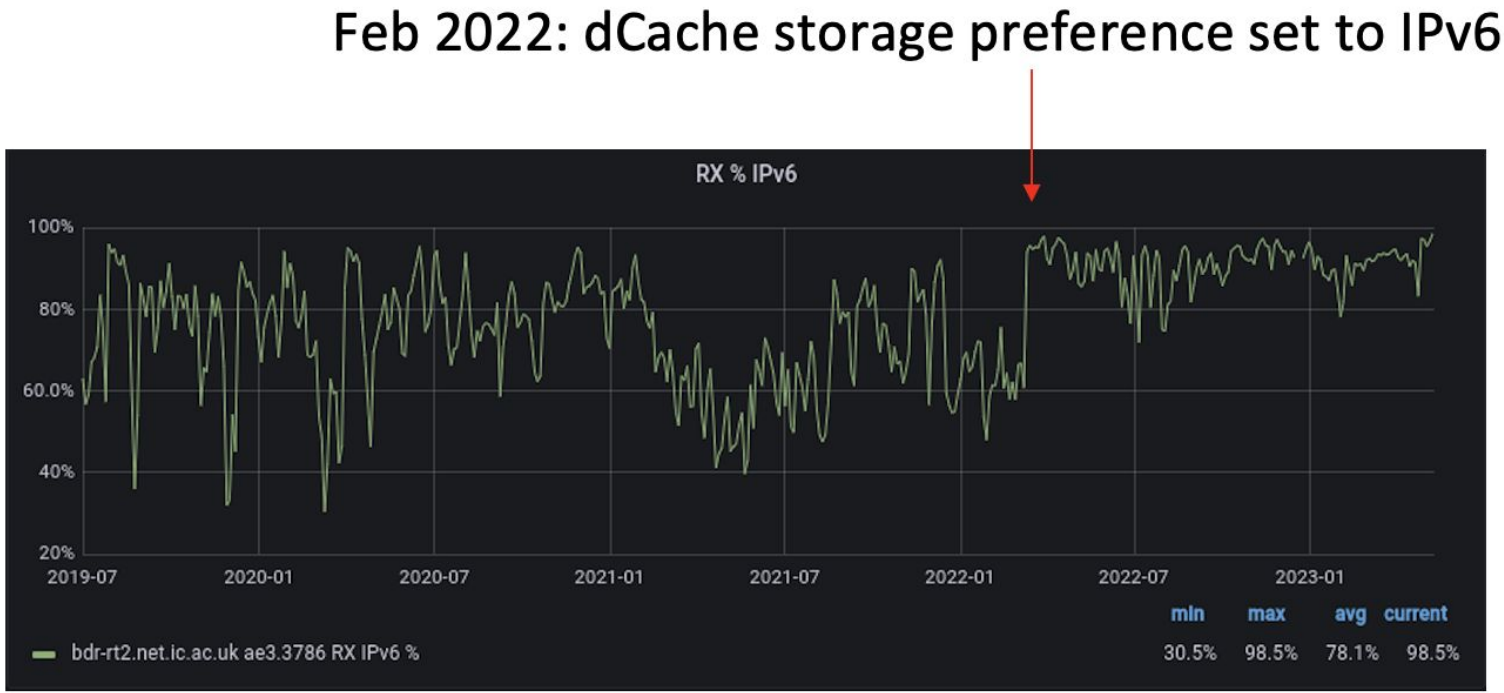
Green: IPv4; Yellow: IPv6

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



Monitoring traffic and network characteristics

The WLCG can draw on various sources of traffic information:

- Application oriented - e.g., FTS logs
- Router interface utilisation at site egress - sites were requested to expose this data to a CERN collector for DC24
- Netflow records - kept by sites for a short period of time

Allows reasonable investigation into use of IPv6 and causes of residual IPv4

We can also test with perfSONAR, an open source platform to measure latency, loss, path and throughput, see https://www.perfsonar.net/. Most WLCG sites have at least one perfSONAR server

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

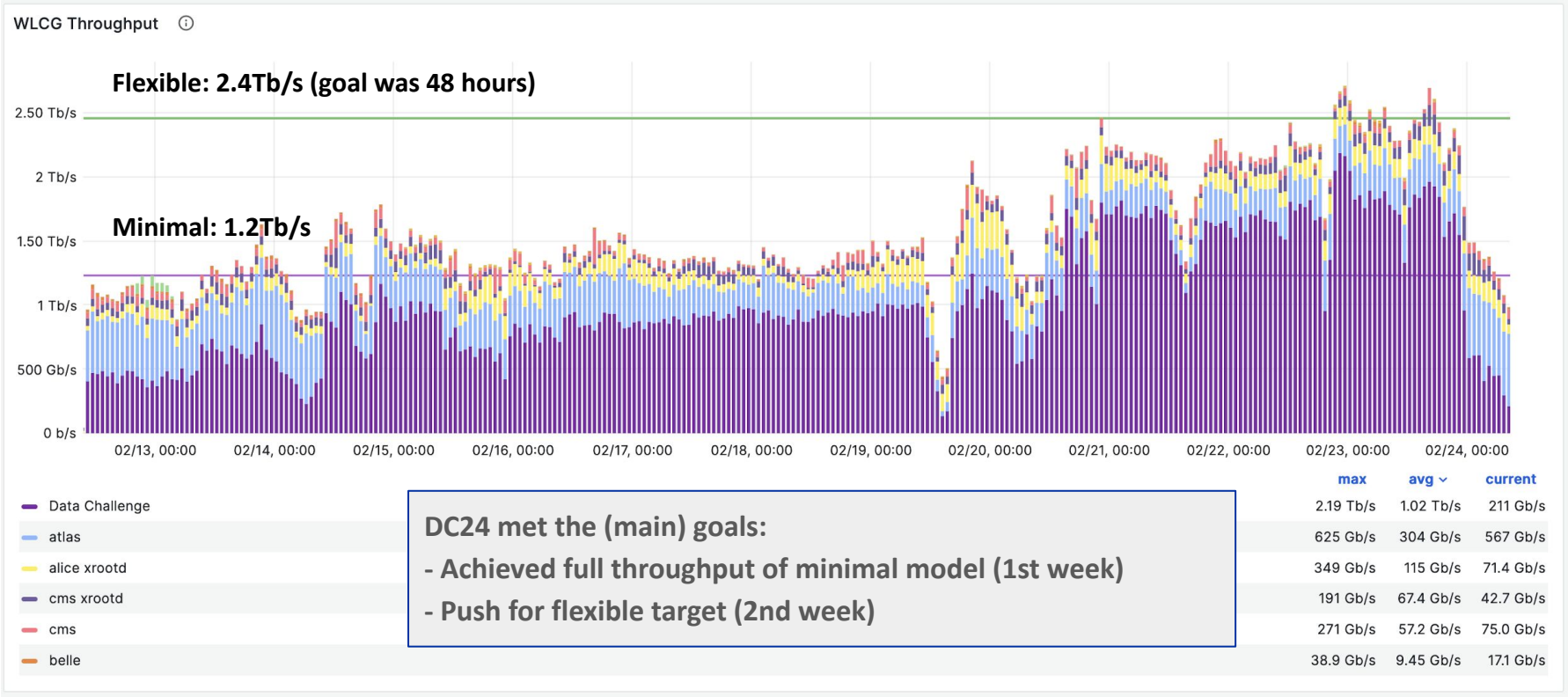
- Total target across all sites = 2,430 Gbps
- Expected requirement in 2029 = 9,620 Gbps

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)

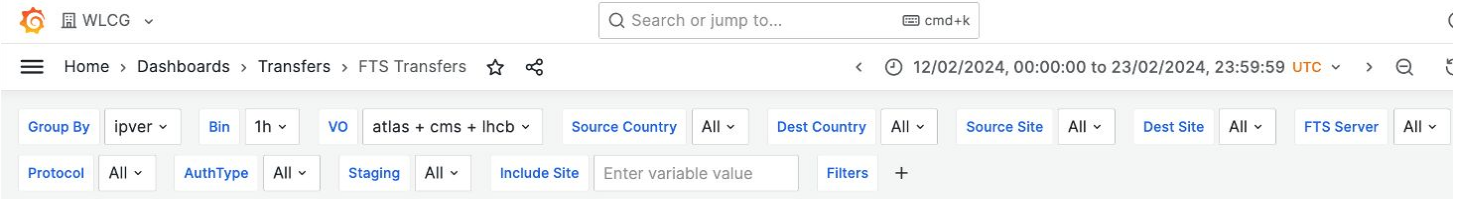
DC24 overall throughput by experiment

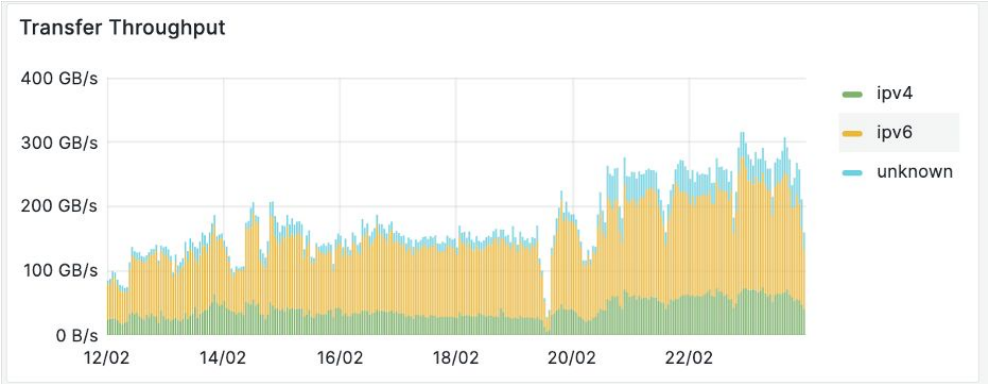


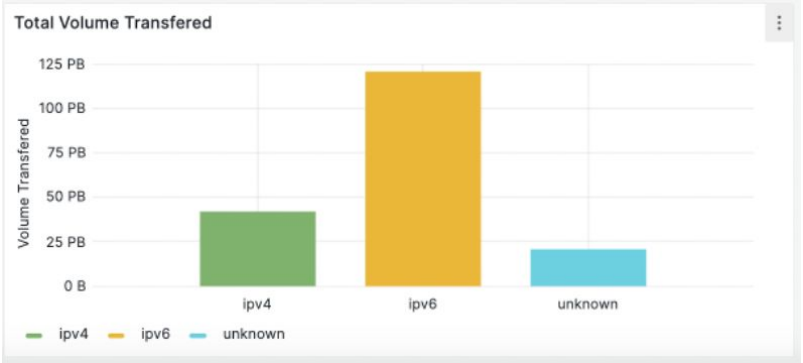
DC24 met the (main) goals:

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

DC24: Relative FTS use of IPv4 and IPv6







WLCG traffic, and thus IPv6, monitoring

Monitoring collected at CERN

Examples….

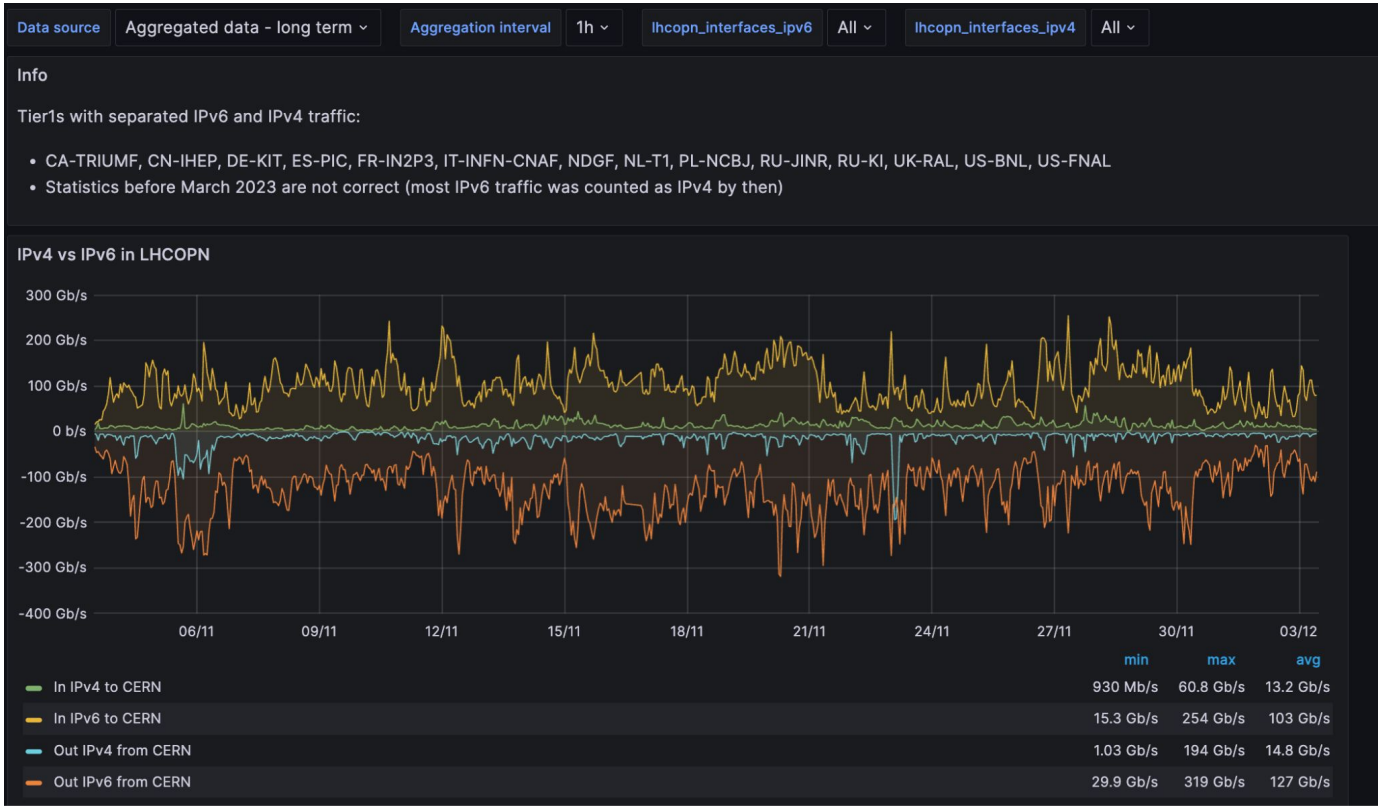
IPv4 and IPv6 traffic for each Tier 1:

- https://monit-grafana-open.cern.ch/d/000000523/home?orgId=16&viewPanel=1

Relative IPv4 vs IPv6 traffic:

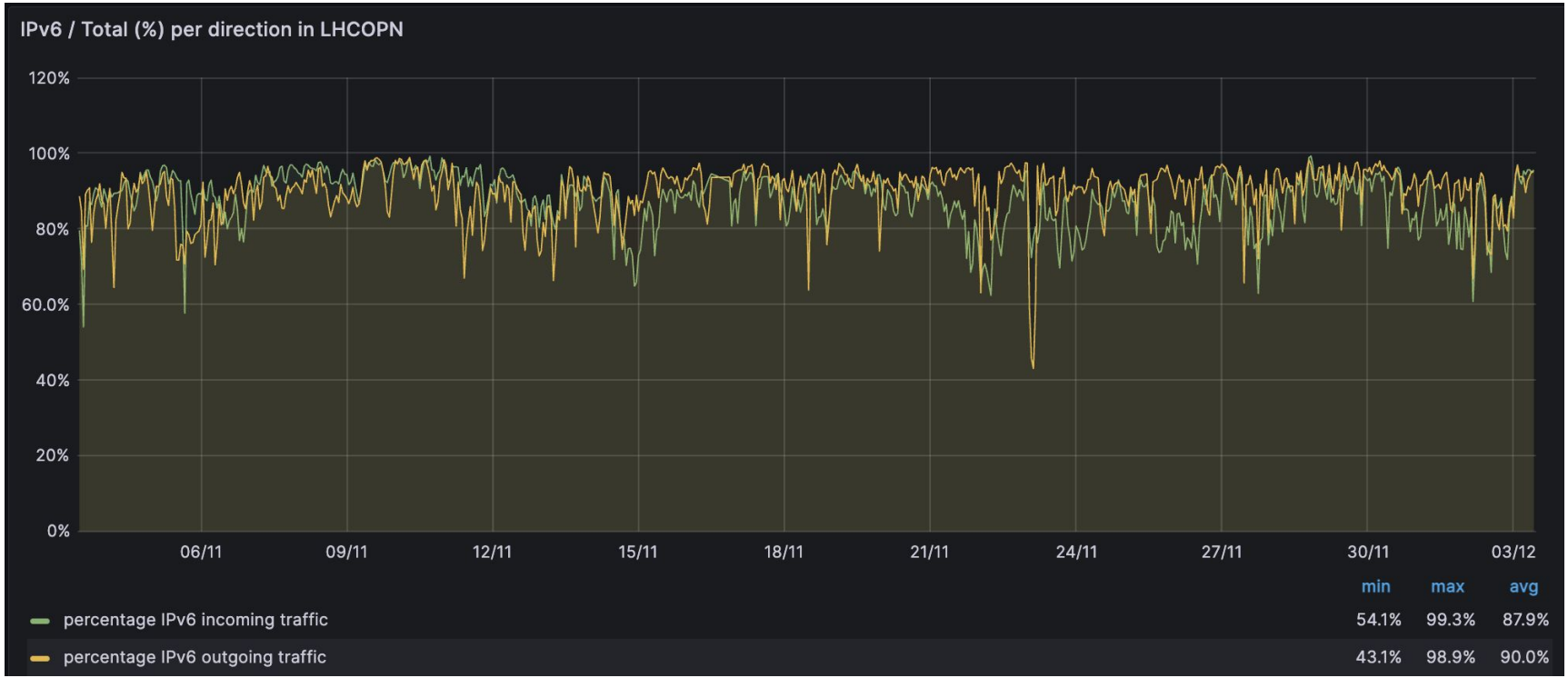
- https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4?orgId=16

Tier 1s: IPv4/IPv6 in past 30 days (~Nov 2024)



< all interfaces

% IPv6 in/out in past 30 days (~November 2024)



Diving into specific Tier1s

You can look at traffic per site, e.g., RAL (UK T1):
https://monit-grafana-open.cern.ch/d/dsY3tf5nk/uk-ral?orgId=16

Then pick the specific interfaces used, e.g., here CERN - RAL:

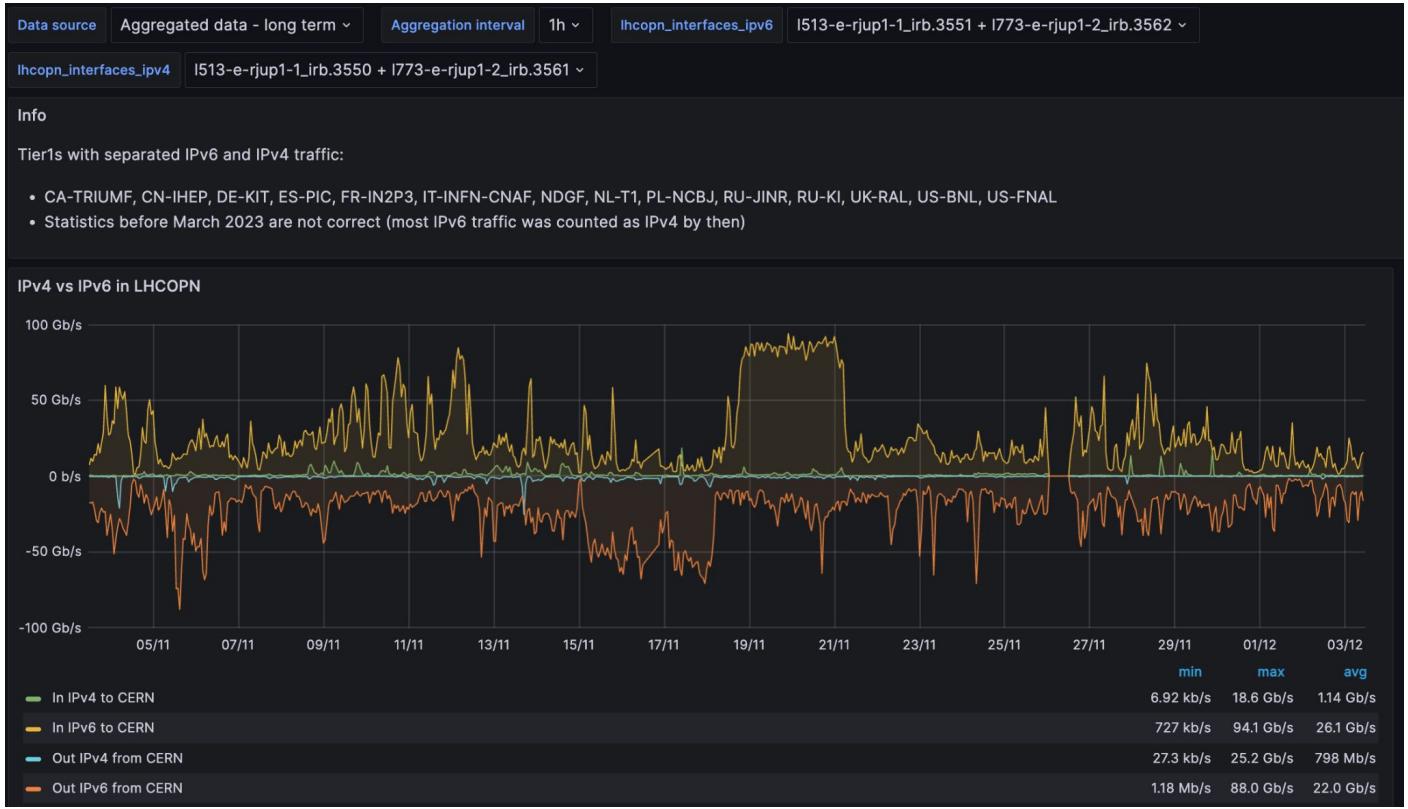
https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4?orgId=16&var-source=long_term&var-bin=1h&var-lhcopn_interfaces_ipv6=l513-e-rjup1-1_irb.3551&var-lhcopn_interfaces_ipv6=l773-e-rjup1-2_irb.3562&var-lhcopn_interfaces_ipv4=l513-e-rjup1-1_irb.3550&var-lhcopn_interfaces_ipv4=l773-e-rjup1-2_irb.3561

(can use this link and substitute in other interface names from slide 23)

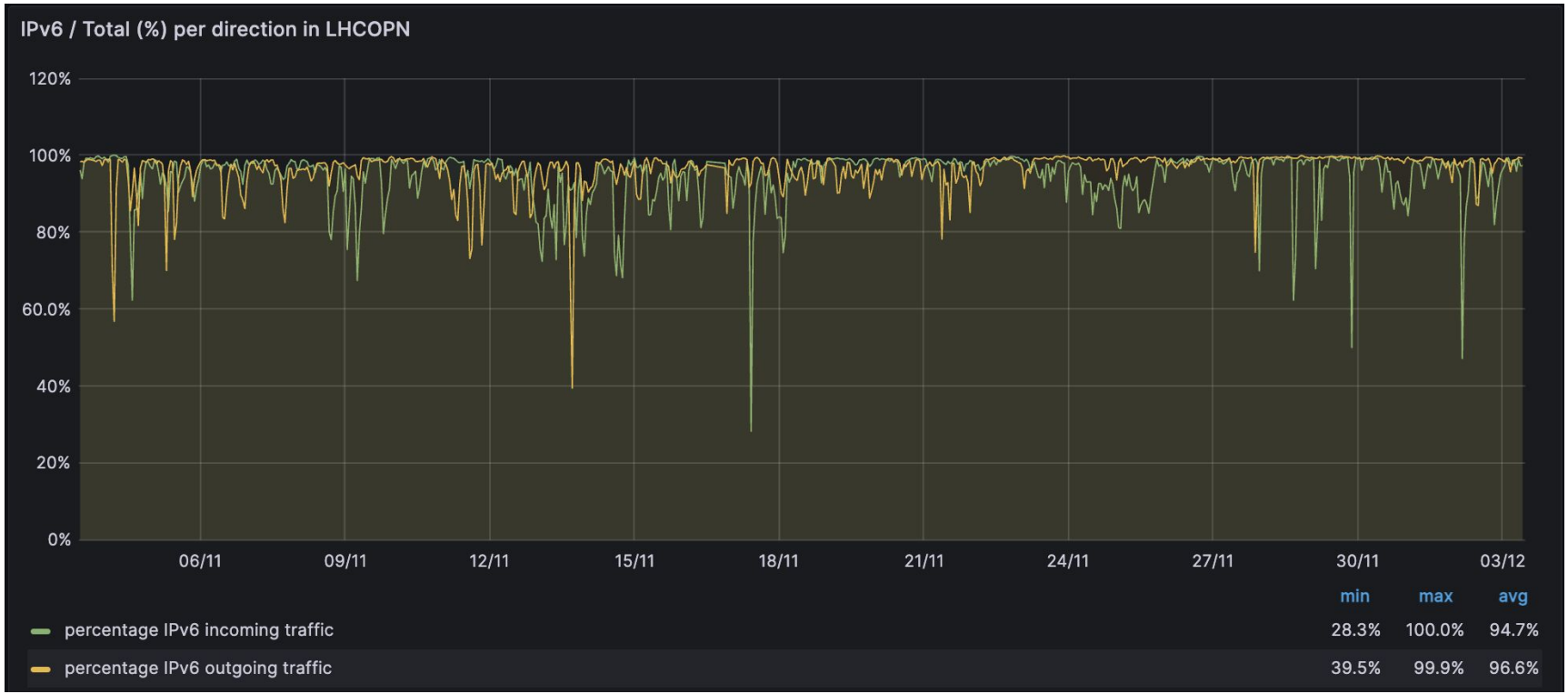
Example: CERN-RAL (with IPv4 & IPv6 interfaces/VLANs)

Example: CERN-RAL, total traffic and rates (~Nov 2024)

< RAL interfaces



Example: CERN-RAL, percentage IPv6 (~Nov 2024)



Back to the *towards IPv6-only* goal

What does the HEPiX IPv6 WG want to achieve?

- Removing complexity of dual-stack; simpler management and troubleshooting
- More performant operation, no NAT
- No longer chasing residual use of IPv4
- Benefiting from IPv6 per-packet Scitag marking

And by when?

- The timetable is still to be defined and agreed with Management Board
- Suggested target “end of Run 3”, or “before LHC Run 4” (starting in ~2030)

What do we mean by IPv6-only?

Many choices!

WLCG site services are IPv6-only (CE, SE, …)

WLCG Tier 2 is fully IPv6-only

Other WLCG central services (e.g., Rucio, FTS, etc) are IPv6-only

LHCOPN and/or LHCONE networks are IPv6-only

All WAN WLCG traffic is IPv6-only

…

Aside: SciTags

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

Plans for an IPv6-only WLCG

First steps:

- Any site can today have IPv6-only clients and fully function in WLCG
- We are gradually moving all WLCG services to be fully dual-stack
- We need more sites to test IPv6-only clients, worker nodes, etc

Ongoing plan:

- By end of Run 3 ***all*** WLCG services to support IPv6 (today ~75%)
- Continue to remove use of legacy IPv4 on LHCOPN (until end of Run 3)
- Turn off IPv4 peering on LHCOPN when possible
- Remove all IPv4 WAN traffic

Aside: “IPv6 Mostly” - removing IPv4

The WLCG is generally compute and storage, rather than user devices

A new IETF I-D has proposed an “IPv6 Mostly” model

- https://datatracker.ietf.org/doc/draft-ietf-v6ops-6mops

Hosts can negotiate turning off IPv4 via use of DHCPv4 Option 108

Supported on Android, macOS, iOS and soon Windows

Implemented on Imperial College London eduroam WiFi, a few thousand APs, working well as reported at the UK IPv6 Council annual meeting last month.

- 70,000 users and visitors seamlessly running IPv6-only (CLAT, NAT64) in a week (>75%)

Consider this on your campuses!

Summary

The 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

The large majority of data transfers use 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 (61% and rising) and WLCG services

End-game remains IPv6-only services; unofficial target is end of Run 3 / before Run 4

Any new research infrastructure should build with IPv6 from day one