Towards an IPv6-only WLCG: more successes in reducing IPv4

David Kelsey

RAL, STFC, UK Research and Innovation

(on behalf of the HEPiX IPv6 Working Group)

CHEP2024, Krakow, Poland, 24 Oct 2024

On behalf of all members of the HEPiX IPv6 working group - (many thanks all!)

M Babik (CERN), M Bly (RAL), N Buraglio (ESnet), T Chown (Jisc), D Christidis (CERN/ATLAS), J Chudoba (FZU Prague), P Demar (FNAL), J Flix (PIC), C Grigoras (CERN/ALICE), B Hoeft (KIT), H Ito (BNL), D P Kelsey (RAL), E Martelli (CERN), S McKee (U Michigan), C Misa Moreira (CERN), R Nandakumar (RAL/LHCb), K Ohrenberg (DESY), F Prelz (INFN), D Rand (Imperial), A Sciabà (CERN/CMS), T Skirvin (FNAL), C J Walker (Jisc)

- Special thanks to underlined co-authors for provision of some slides
- Many more in the past, and members join/leave from time to time
- ***many thanks** also to WLCG operations, WLCG sites, LHC experiments, networking teams, monitoring groups, storage developers…*

Outline

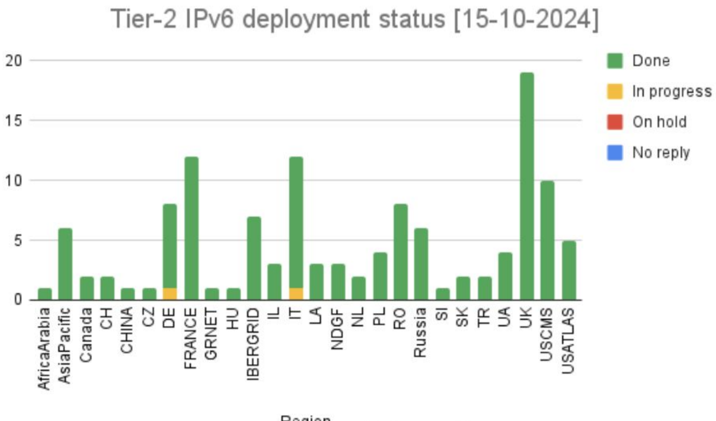
- The HEPiX IPv6 working group - reminder
  - Drivers for IPv6
  - IPv6/IPv4 dual-stack storage
- Dual-stack CPU & worker nodes campaign
- Observations during WLCG Data Challenge (DC24)
- Plans for IPv6-only WLCG
- Summary

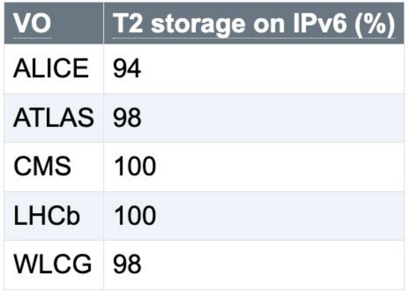
HEPiX IPv6 working group - History and drivers for use of IPv6

- Phase 1 - 2011-2016 - analysis, investigations, testbed, fix storage
- Phase 2 - 2017-2023 - deploy dual-stack storage on WLCG
- Phase 3 - 2019-onwards - plan for IPv6-only
- Sites running out of routable IPv4 addresses (avoid NAT)
  - Use IPv6 addresses for external public networking
- To be ready to support use of IPv6-only CPU clients
- There are other drivers for IPv6:
  - scitags.org – packet marking (in header of IPv6 packets)
    - Research Networking Technical Working Group (RNTWG)
  - USA Federal Government – directive on “IPv6-only” (Nov 2020)

Dual-stack WLCG Storage (Tier2s)

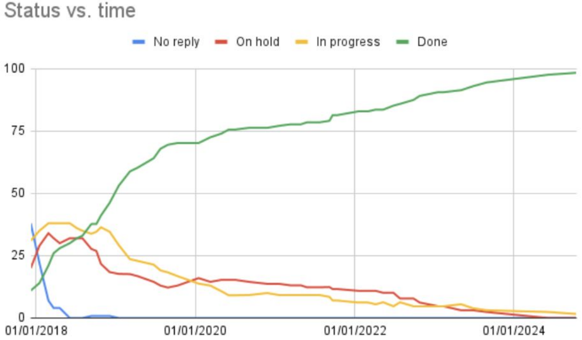
- Campaign “IPv6 on storage services” started in 2017
- Goal to allow IPv6-only WNs
- Main reason for delay - the institute networking
- Today, almost all WLCG sites have dual-stack IPv6/IPv4 Storage





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

(checked on 15-10-2024)

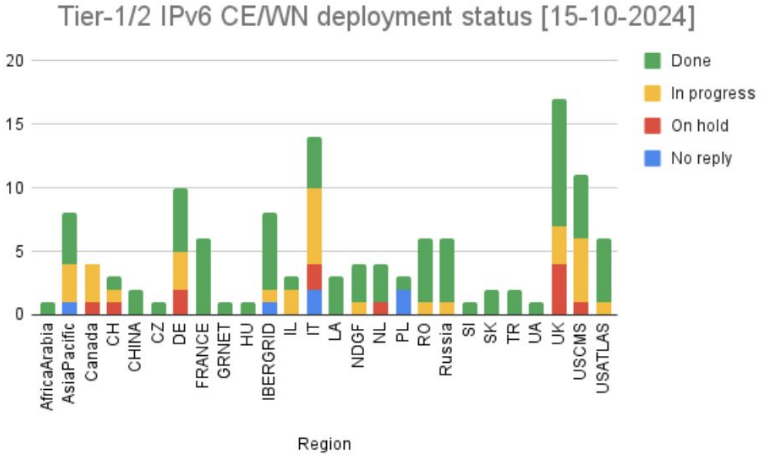


Dual-stack CPU and WN campaign

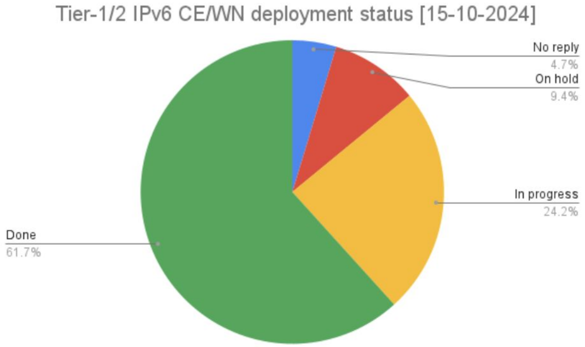
WLCG CPU - GGUS ticket campaign

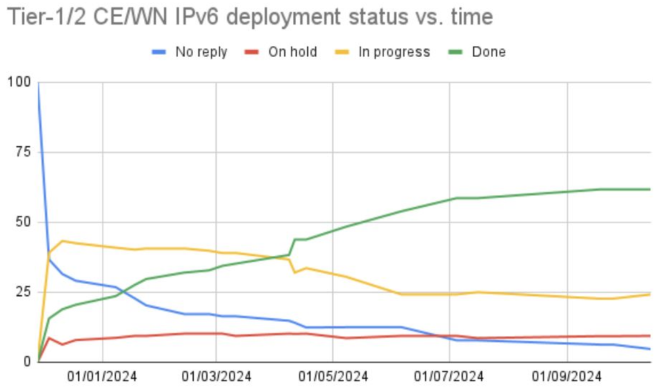
- Eliminate a large remaining source of IPv4 traffic
  - Data transfers between WNs and remote storage systems
- Approved by WLCG MB in October 2023
- Launched on 28 November 2023 on all WLCG sites
- **“Please deploy dual-stack connectivity (IPv4+IPv6) on your computing services (computing elements and worker nodes) as soon as possible and by 30 June 2024 at the latest”**
- Provide estimates for timescale and details on the necessary steps
- If cannot meet the deadline, then explain why

CPU Current status


Tier-1/2 IPv6 CE/WN deployment status [15-10-2024]

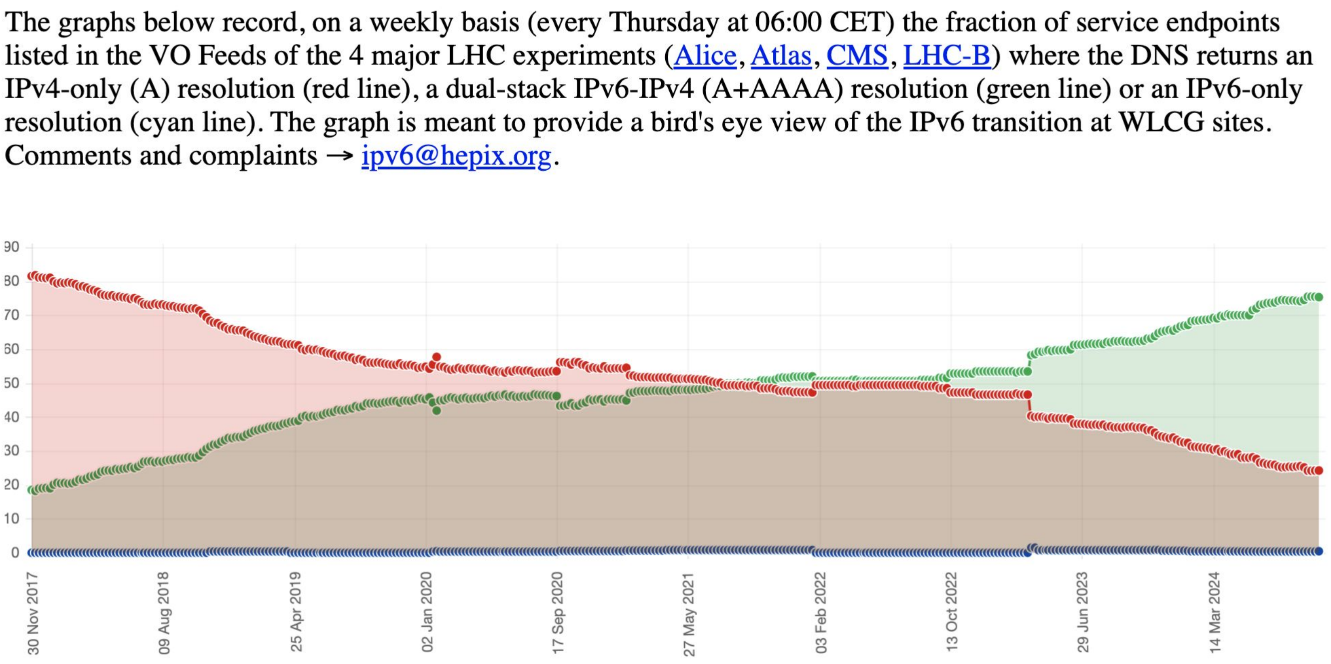
62% done - Status always visible from a twiki page


Tier-1/2 IPv6 CE/WN deployment status [15-10-2024]


Tier-1/2 CE/WN IPv6 deployment status vs. time

All WLCG services - “VOfeeds”

https://orsone.mi.infn.it/~prelz/ipv6_vofeed/



The graphs below record, on a weekly basis (every Thursday at 06:00 CET) the fraction of service endpoints listed in the VO Feeds of the 4 major LHC experiments (Alice, Atlas, CMS, LHC-B) where the DNS returns an IPv4-only (A) resolution (red line), a dual-stack IPv6-IPv4 (A+AAAA) resolution (green line) or an IPv6-only resolution (cyan line). The graph is meant to provide a bird's eye view of the IPv6 transition at WLCG sites. Comments and complaints → ipv6@hepix.org.

~75% dual stack

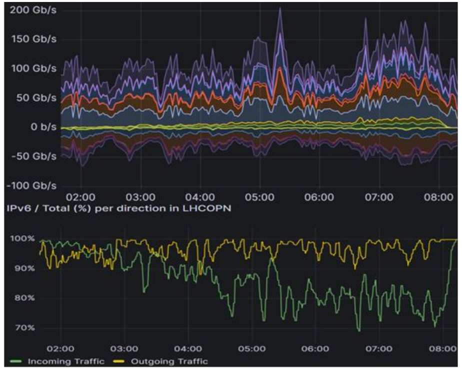
~25% IPv4

Observations during WLCG Data Challenge (DC24)

During WLCG DC24 - IPv6 sub-project

- Work to study the LHCOPN link between CERN and KIT
- Understand when and why IPv4 is being used
- Early on - large IPv4 transfer seen to ALICE at CERN
  - Failed transfers on IPv6 failing over to use of IPv4
- Later - some transfers from KIT to NL-T1
  - All end-points were dual-stack but NL-T1 preferred IPv4 to avoid some observed problems with many concurrent IPv6 streams
- Then see next slide
  - Plot of XRootD file transfers from CERN
  - Squid at KIT - all would work if IPv6-only but often fails back to IPv4
- Lots of detailed investigations - and STILL ongoing (see later slides)

XRootD file transfer from CERN



```
2024-02-20 06:50:17.012   22.500 TCP    128.142.56.61  59332   192.108.47.90  1094    2.7 M   4.1 G   1.5 G  1499    1
2024-02-20 06:02:38.012   16.000 TCP   128.142.57.111  40594   192.108.47.89  1094    2.7 M   4.1 G   2.1 G  1499    1
2024-01-31 09:33:31.833   11.653 TCP   128.142.63.105  43670   192.108.46.89  1094    2.8 M   4.2 G   2.9 G  1498    1
```

Summary: total flows: 597053, total bytes: 33.0 TeraByte

1625 Server at CERN

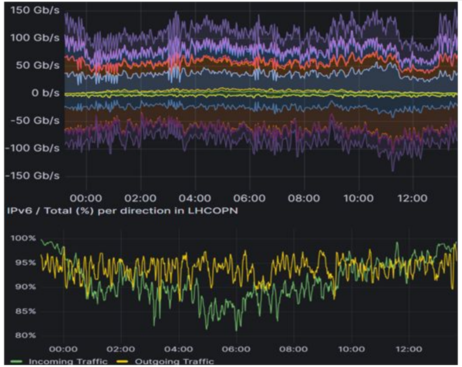
Only 16 Server at KIT

- cvmfs-sq4.gridka.de. dual-stack
- cvmfs-sq1.gridka.de. dual-stack
- cvmfs-sq3.gridka.de. dual-stack
- cvmfs-sq5.gridka.de. dual-stack
- cvmfs-sq6.gridka.de. dual-stack
- cvmfs-sq2.gridka.de. dual-stack
- frontier-sq1.gridka.de. dual-stack
- fw-nat-inside-outside.gridka.de.

8 Storage Server at DE-KIT (XRootD Port – 1094):

- f01-032-114-e.gridka.de.
- f01-124-110-e.gridka.de. dual-stack
- f01-124-159-e.gridka.de. dual-stack
- f01-124-160-e.gridka.de. dual-stack
- f01-124-161-e.gridka.de. dual-stack
- f01-125-159-e.gridka.de. dual-stack
- f01-125-160-e.gridka.de. dual-stack
- f01-125-161-e.gridka.de. dual-stack

Green line - CERN to KIT %IPv6 70 to 80%



```
2024-02-20 23:26:09.012    0.500 TCP   128.142.249.74  38908  192.108.68.144  1094      12    2266   36256   188    1
2024-02-21 02:39:33.262    0.250 TCP   128.142.240.76  55700   192.108.46.89  1094      10     706   22592    70    1
```

Summary: total flows: 1460049, total bytes: 43.1 TeraByte

2426 Server at CERN

Only 25 Server at KIT

Squid service

Port 3401

cvmfs-sq4.gridka.de.
cvmfs-sq1.gridka.de.
cvmfs-sq3.gridka.de.
cvmfs-sq5.gridka.de.
cvmfs-sq6.gridka.de.
cvmfs-sq2.gridka.de.
frontier-sq1.gridka.de.
fw-nat-inside-outside.gridka.de.

XRootD Port 1094

f01-124-109-e.gridka.de.
f01-124-110-e.gridka.de.
f01-124-112-e.gridka.de.
f01-124-155-e.gridka.de.
f01-124-159-e.gridka.de.
f01-124-160-e.gridka.de.
f01-124-161-e.gridka.de.
f01-125-109-e.gridka.de.
f01-125-110-e.gridka.de.
f01-125-155-e.gridka.de.
f01-125-159-e.gridka.de.
f01-125-160-e.gridka.de.
f01-125-161-e.gridka.de.
f01-117-137-e.gridka.de.
f01-152-140-e.gridka.de.
f01-152-191-e.gridka.de.
f01-152-192-e.gridka.de.

Green line - and again

Plans for IPv6-only WLCG

IPv6-only on WLCG (CHEP2019)

https://doi.org/10.1051/epjconf/202024507045

- The end point of the transition from IPv4 is an IPv6-only WLCG core network - agreed by WLCG MB
- To simplify operations
  - Dual-stack infrastructure is the most complex
  - Reduced complexity reduces chance of making security errors
- Large infrastructures (e.g. Facebook, Microsoft,...) use IPv6-only internally
- The goal we are still working towards
  - “IPv6-only” for the majority of WLCG services and clients
- Timetable still to be defined - but aiming for “before LHC Run 4”

What do we mean by IPv6-only?

Choices (one or more of):

- 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

What does the IPv6 working group wish to achieve:

- All WLCG services (site and central) are IPv6-only
- Removes complexity of dual-stack
- No longer have to chase use of IPv4 by dual-stack endpoints
- All WLCG WAN traffic is IPv6-only

Plans for 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 be fully dual-stack (today ~75%)
- Continue removing use of legacy IPv4 on LHCOPN (until end of Run 3)
- Turn-off IPv4 peering on LHCOPN when possible
- Remove all WAN traffic over IPv4

Working group observations/questions:

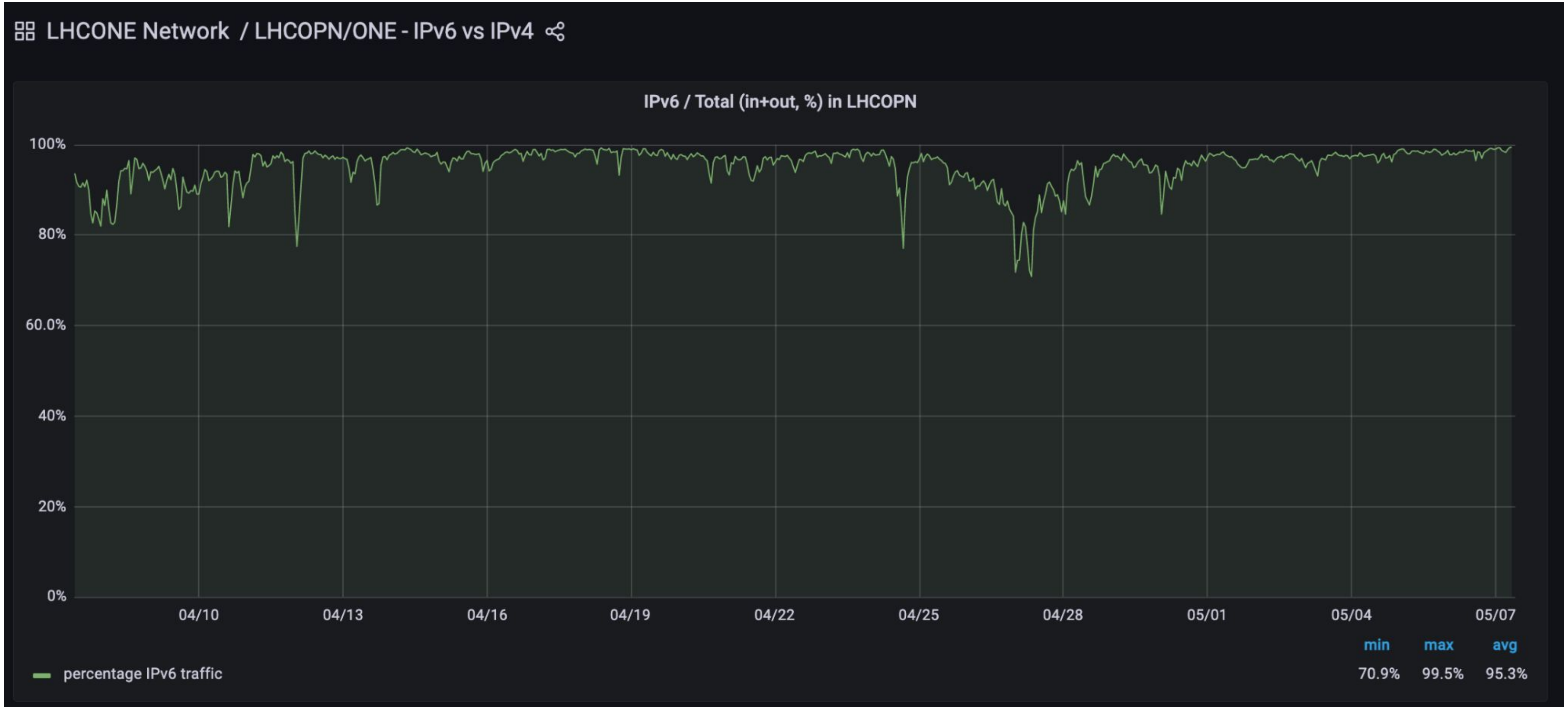
- When should perfSONAR stop performing IPv4 tests?
- Can we add “IPv4 versus IPv6” traffic split in the WLCG Site egress monitoring network I/O (for DC24) (every minute)?

Some plots: IPv6 and IPv4 traffic on LHCOPN (5 to 9 Oct 2024) (and compare with CHEP2023)

Will skip these if no time to show

LHCOPN - %IPv6 traffic - shown at CHEP2023

7 April to 7 May 2023 - shows drops in %IPv6



min	max	avg
70.9%	99.5%	95.3%

100%

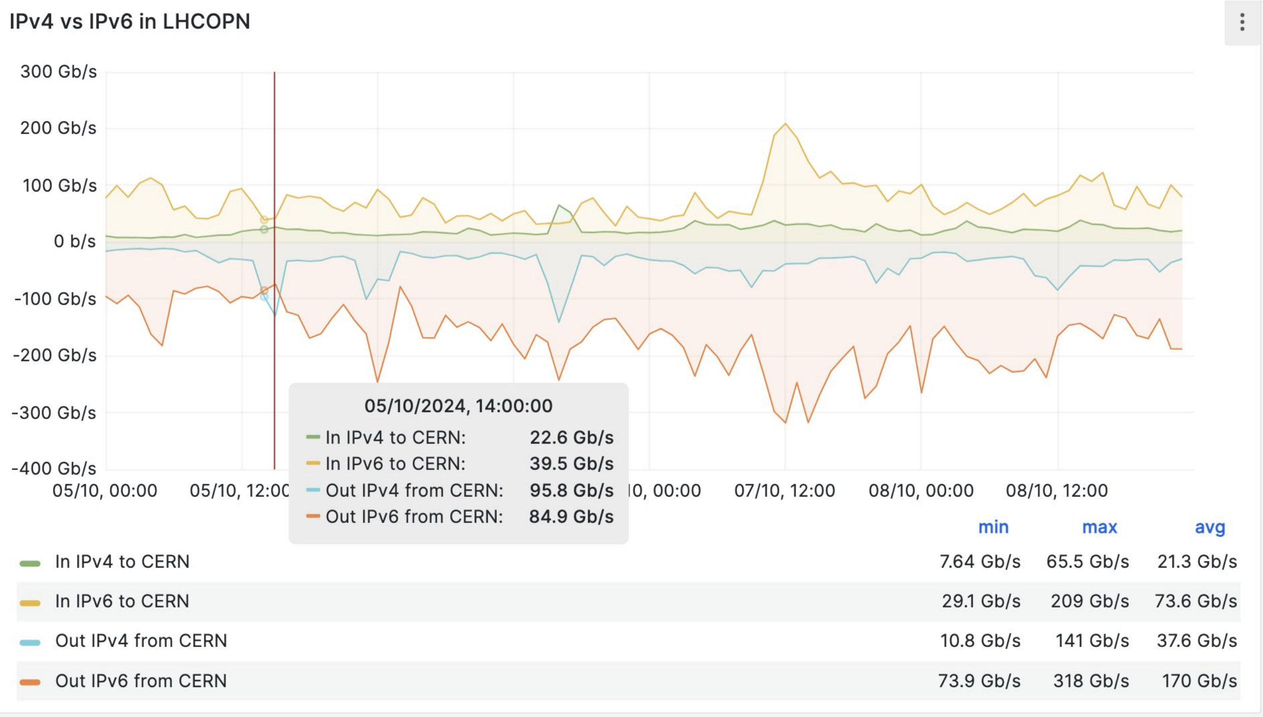
Max
99.5%

Avg
95.3%

Min
70.9%

LHCOPN total traffic, split IPv4 & IPv6 (as seen at CERN)

https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4?orgId=16&from=1728079200000&to=1728424799000


IPv4 vs IPv6 in LHCOPN

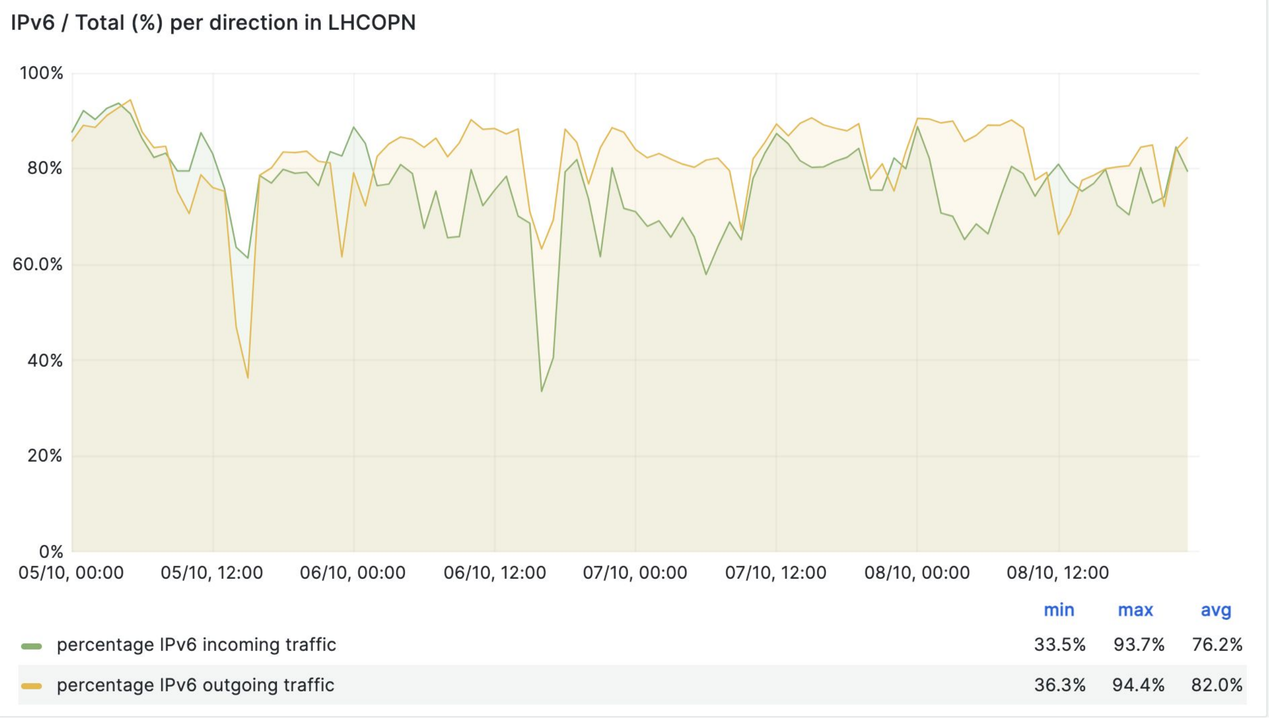
05/10/2024, 14:00:00

In IPv4 to CERN:	22.6 Gb/s
In IPv6 to CERN:	39.5 Gb/s
Out IPv4 from CERN:	95.8 Gb/s
Out IPv6 from CERN:	84.9 Gb/s

	min	max	avg
In IPv4 to CERN	7.64 Gb/s	65.5 Gb/s	21.3 Gb/s
In IPv6 to CERN	29.1 Gb/s	209 Gb/s	73.6 Gb/s
Out IPv4 from CERN	10.8 Gb/s	141 Gb/s	37.6 Gb/s
Out IPv6 from CERN	73.9 Gb/s	318 Gb/s	170 Gb/s

- 5 to 9 Oct 2024
- IPv6 Out of CERN
  - Avg 170 Gbps
- IPv4 Out of CERN
  - Avg 37.6 Gbps
- BUT
  - Large IPv4 peaks, e.g.
  - 5/10 @ 14:00
  - Out 95.8 Gbps

%IPv6 traffic - generally high - but large drops down to ~40%

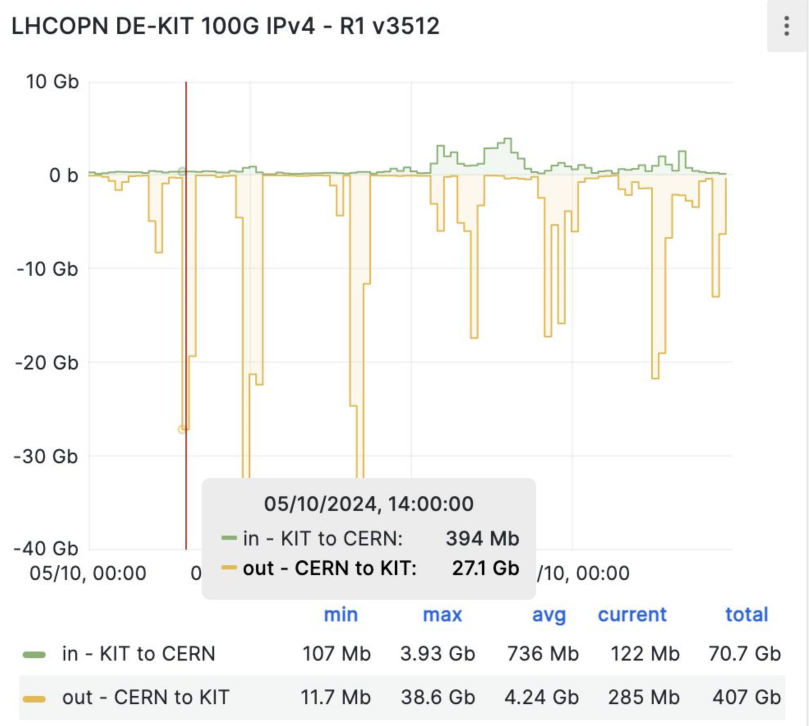

IPv6 / Total (%) per direction in LHCOPN

	min	max	avg
🟩 percentage IPv6 incoming traffic	33.5%	93.7%	76.2%
🟨 percentage IPv6 outgoing traffic	36.3%	94.4%	82.0%

- %IPv6
- In - avg 76.2%
  - Min 33.5%
- Out - avg 82.0%
  - Min 36.3%

LHCOPN traffic (CERN- KIT) German Tier1 - large IPv4 peaks

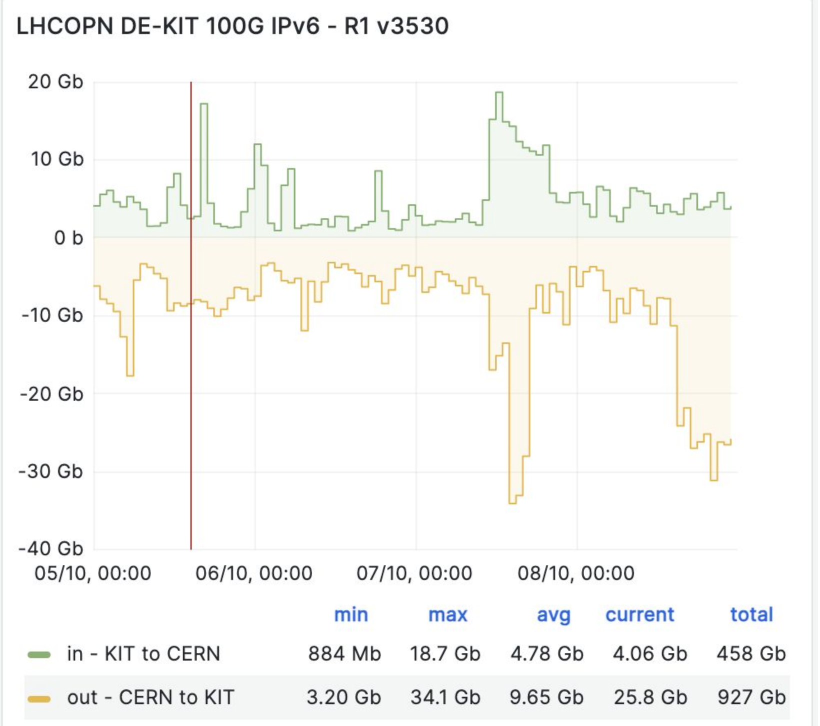
IPv4 plot



LHCOPN DE-KIT 100G IPv4 - R1 v3512

	min	max	avg	current	total
in - KIT to CERN	107 Mb	3.93 Gb	736 Mb	122 Mb	70.7 Gb
out - CERN to KIT	11.7 Mb	38.6 Gb	4.24 Gb	285 Mb	407 Gb

IPv6 plot



LHCOPN DE-KIT 100G IPv6 - R1 v3530

	min	max	avg	current	total
in - KIT to CERN	884 Mb	18.7 Gb	4.78 Gb	4.06 Gb	458 Gb
out - CERN to KIT	3.20 Gb	34.1 Gb	9.65 Gb	25.8 Gb	927 Gb

What are these large peaks of IPv4?

- Not easy
- Need access to Netflow data
- Study IP addresses and Port numbers
    - Aim to identify LHC Experiment
    - Source and Destination address
    - Type of data transfer
- Work in progress
    - But some evidence of Frontier/CVMFS/Squid, etc….

Summary

- WLCG already supports use of IPv6-only clients
- Dual-stack Storage campaign finished
  - Most WLCG data transfers use IPv6
- Campaign for dual-stack CPU and WN’s - well underway
- Observed use of legacy IPv4 during DC24 and afterwards
- We continue to chase use of legacy IPv4 and try to fix
- Aim to complete move to IPv6-only before start of HL-LHC Run 4

- ***Message to WLCG sites and LHC experiments:***
  - ***Deploy dual-stack on all services & clients and prefer use of IPv6***

HEPiX

Questions, Discussion?

Backup slides

The HEPiX IPv6 Working Group

- In 2010-11
  - some HEPiX sites running out of IPv4 addresses
  - IANA projecting imminent IPv4 address exhaustion
  - Moving to support IPv6 would not be fast - better start now!
- Phase 1 - 2011-2016 - full analysis, investigations, ran a testbed
  - lots of work by storage developers to be IPv6-capable
- Phase 2 - 2017-2023 - deploy dual-stack storage on WLCG
- Phase 3 - 2019-onwards - plan for IPv6-only
  - investigate and fix reasons for obstacles to deployment of IPv6
  - Deploy dual-stack CPU and worker nodes (2023-onwards)

https://www.hepix.org/e10227/e10327/e10326/

https://indico.cern.ch/category/3538/ (meetings)

“Obstacles” to IPv6

There are many reasons stopping the full use of IPv6/IPv4

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

The Obstacles that we have been addressing:

1. **WLCG Sites not yet deployed IPv6 networking** ~done
2. **Sites have IPv6 but Tier-2 has no dual-stack storage** ~done
3. **IPv6 monitoring not available or broken**
    - Monitoring is essential
4. **Service is dual-stack but IPv4 still being used**
    - We continue to chase these problems

Obstacles to IPv6 - being addressed

5. **Non-storage services not yet dual-stack**
   a. ~75% of all WLCG services are dual-stack today, we need 100%
6. **WLCG client CPU (worker nodes, VMs, containers) some IPv4-only**
   a. GGUS ticket campaign well underway
7. **Services/clients outside of WLCG Tier-1/Tier-2 not yet addressed**
   a. Tier-3, Public/Commercial Clouds, Analysis facilities, Experiment portals…
8. **Use of new or evolving technologies** not yet tested or tracked
   a. New CPU architectures (GPU, non-x86, …), container orchestration, …
9. **Staffing issues can be an obstacle**
   a. Lack of effort, lack of IPv6 training/knowledge, pressure of other work