Science and Technology Facilities Council

GridPP/WLCG & LHC IPv6

David Kelsey (UKRI/STFC-RAL)

UK IPv6 Council Meeting, London

28 November 2022

GridPP
UK Computing for Particle Physics

WLCG
Worldwide LHC Computing Grid

HEPiX

Overview

GridPP/WLCG & LHC IPv6

- Explanation of the title
  - GridPP: name of the STFC-funded IT Infrastructure for UK Particle Physics
  - WLCG: the Worldwide LHC Computing Grid (CERN)
  - LHC: The CERN Large Hadron Collider

- This talk
  - Introduction
  - IPv6 on WLCG (history) – the HEPiX working group
  - Dual-stack IPv4/IPv6 deployment
  - Monitoring
  - Why still IPv4 traffic?
  - IPv6-only WLCG
  - Summary

Science and Technology Facilities Council

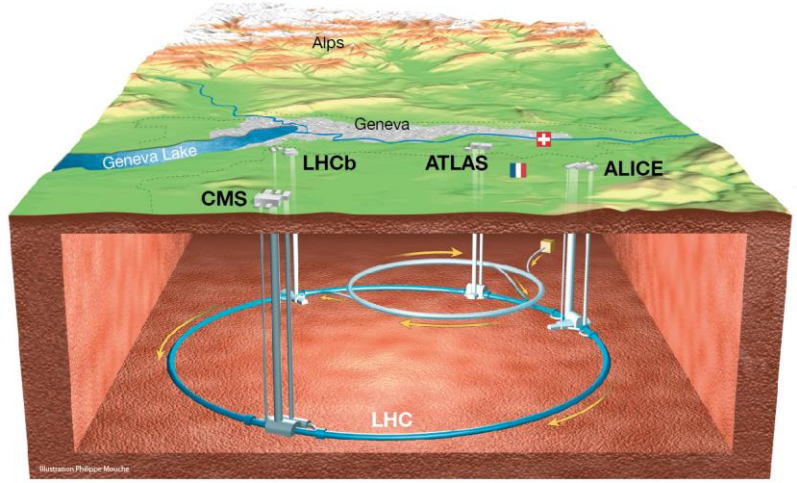
Introduction to …

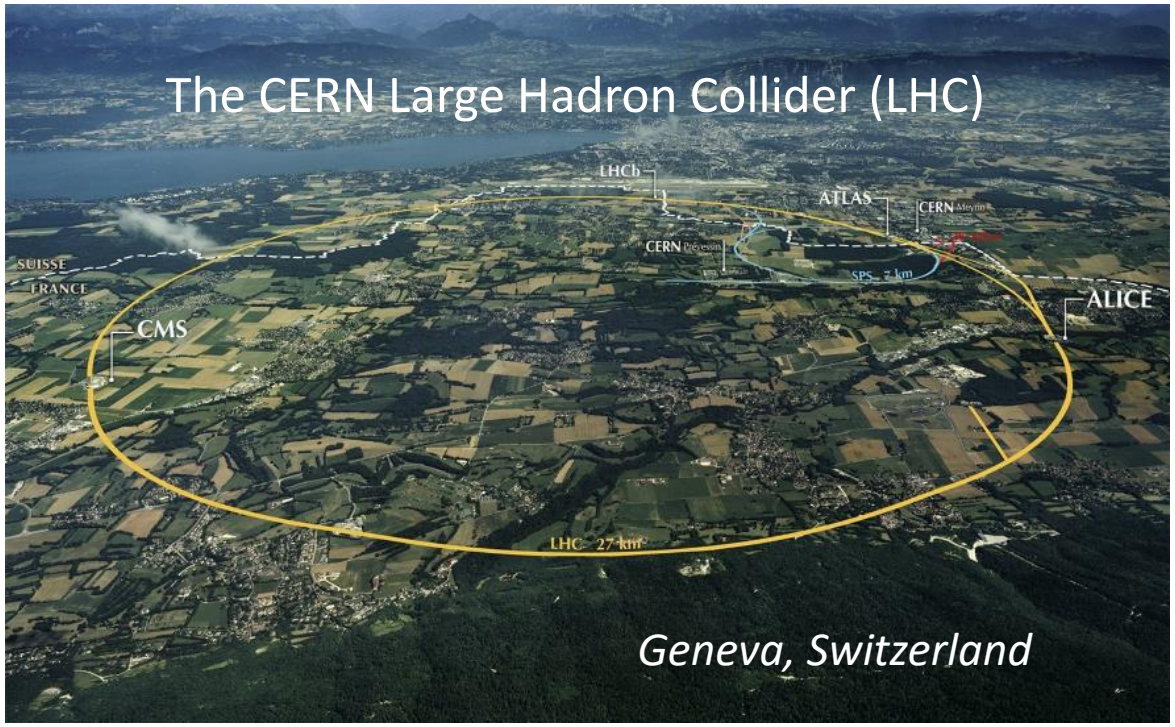
Large Hadron Collider (LHC) at CERN; WLCG/UK GridPP; HEP Networks

Science and Technology Facilities Council

Who is David Kelsey?

- Experimental particle physicist -> IT
- Lead computing group in UKRI STFC-RAL Particle Physics Dept
- Trust, security & identity roles in WLCG, GridPP, EGI, EOSC
- But also in networking …
- Chair of the HEPiX IPv6 Working Group
  - HEPiX is a worldwide body of High Energy Physics IT specialists





Science and Technology Facilities Council

The CERN Large Hadron Collider (LHC)

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

4 July 2012

Higgs boson-like particle discovery claimed at LHC

COMMENTS (1665)

By Paul Rincon

Science editor, BBC News website, Geneva

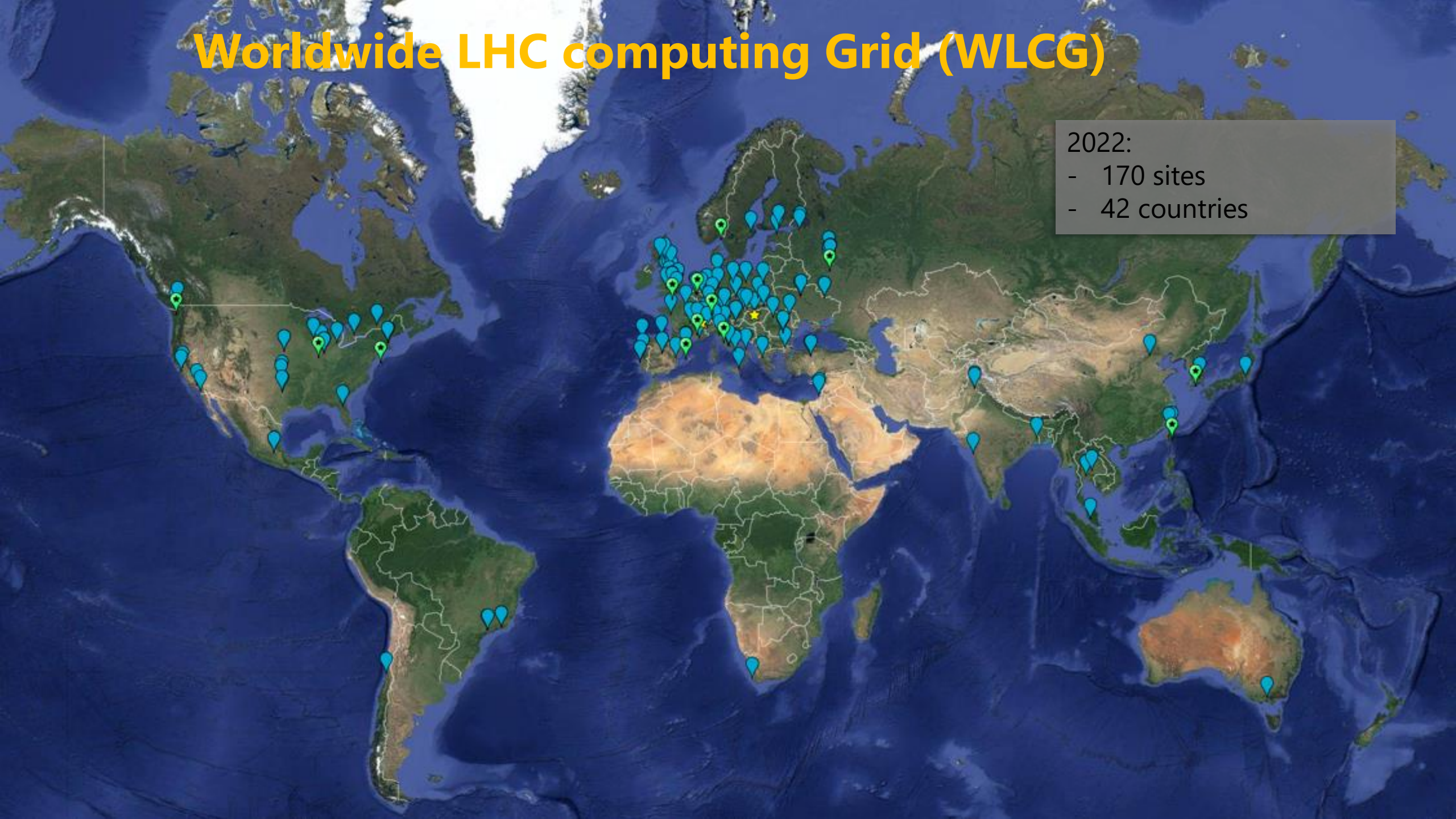
The moment when Cern director Rolf Heuer confirmed the Higgs results

Cern scientists reporting from the Large Hadron Collider (LHC) have claimed the discovery of a new particle consistent with the Higgs boson.

Relat

How do you get from this to this?

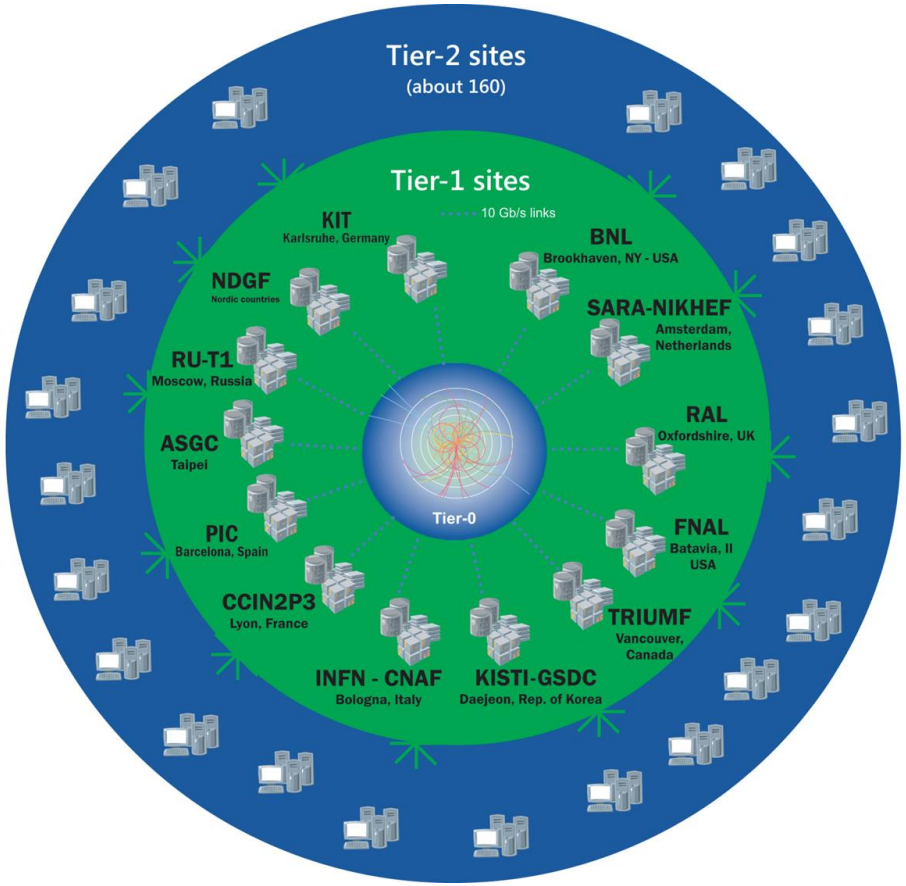
Worldwide LHC computing Grid (WLCG)



Worldwide LHC Computing Grid (WLCG)

Computing for the CERN Large Hadron Collider

- The WLCG is a global collaboration
  - more than 170 computing centres in 42 countries
- Its mission is to **store, distribute** and **analyse** the data generated by the LHC experiments
- Sites hierarchically arranged in three tiers:
  - Tier-0 at CERN
  - 14 Tier-1s (mainly national laboratories) incl. **RAL**
  - ~160 Tier-2s (university physics departments)
- > 1M CPU cores (> 2M jobs per day)
- > 1 EB of data storage



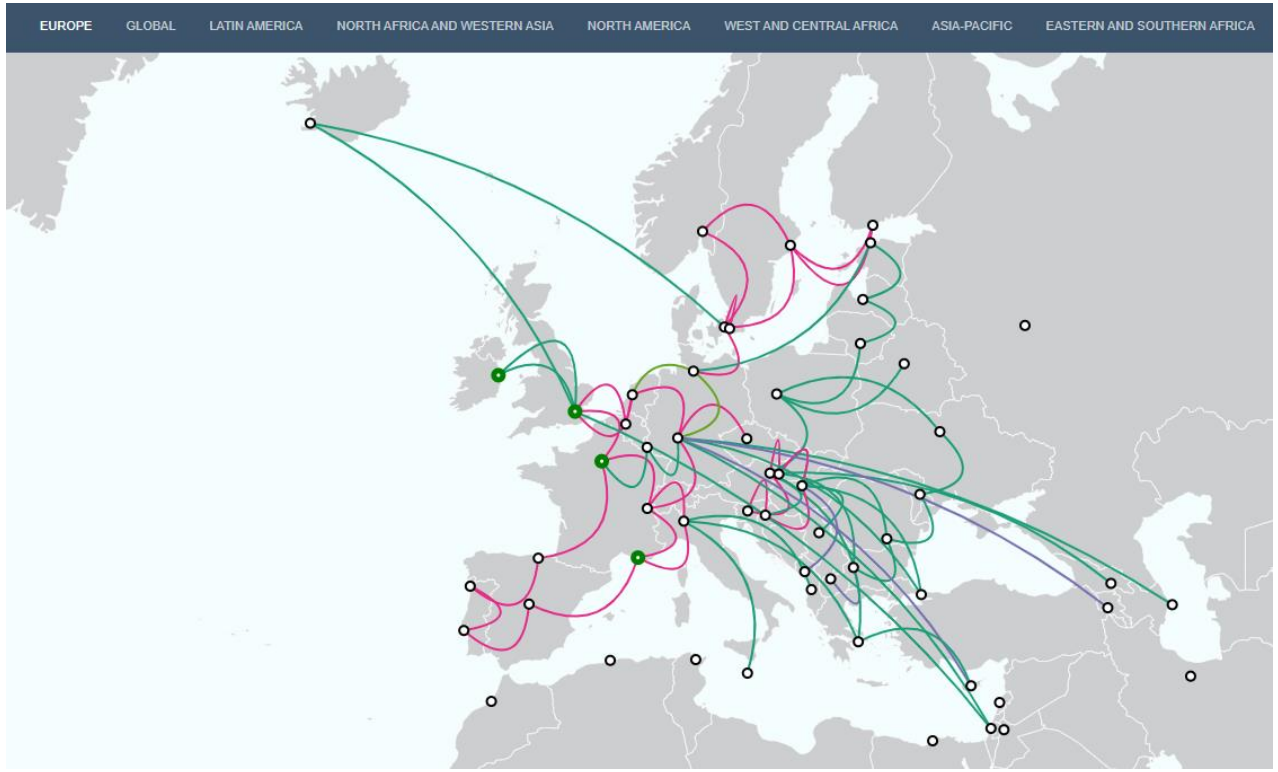
Science and Technology Facilities Council

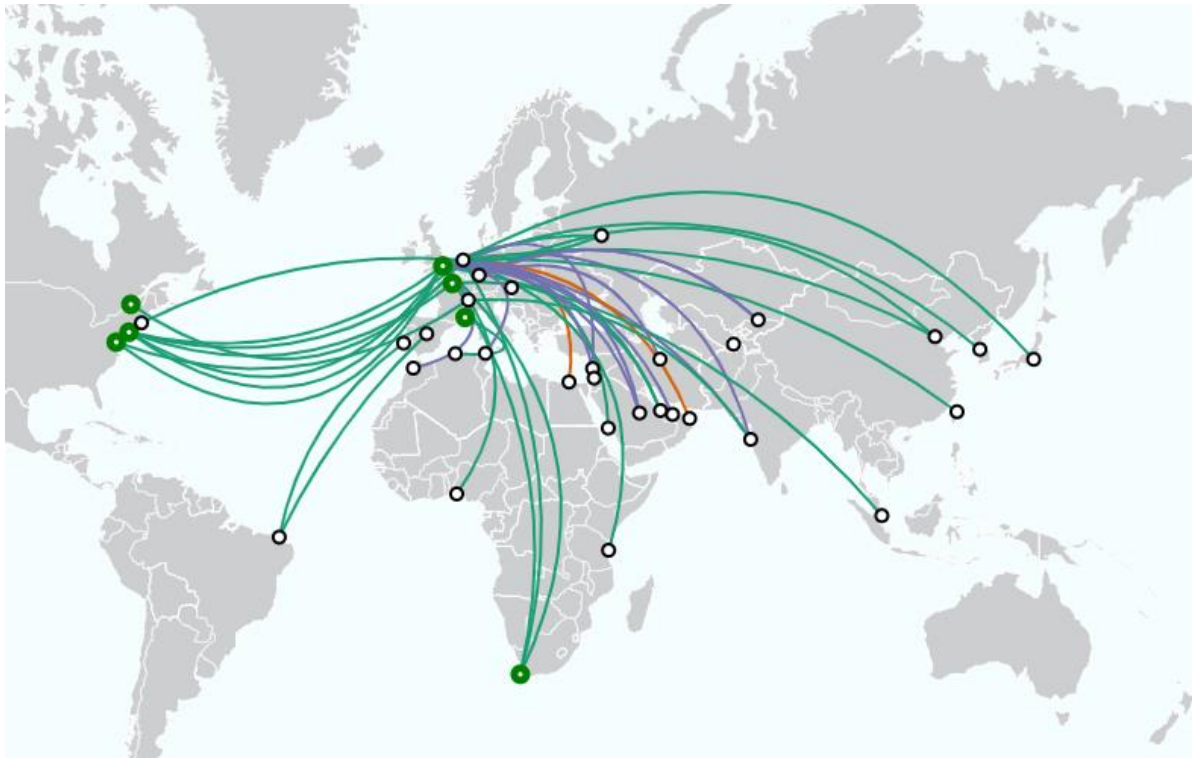
GridPP

STFC-funded IT infrastructure for Particle Physics

- A collaboration of UK institutes providing data-intensive distributed computing resources for the UK High Energy Physics community and the
- The UK contribution to the WLCG

HEP Networking – uses GÉANT and UK Janet (Jisc)

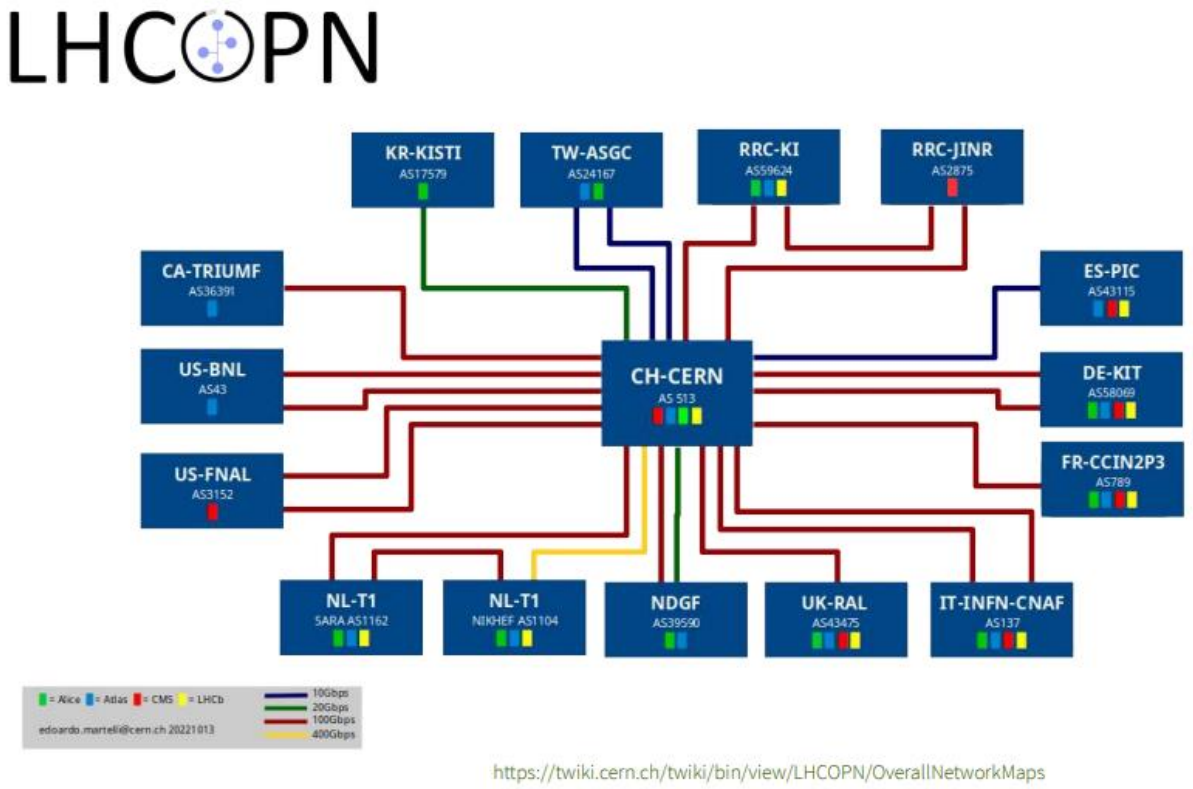




And also, ESnet & Internet2 (USA), NORDUnet and many other national networks

Science and Technology Facilities Council

LHCOPN – Optical Private Network



Numbers

- 14 Tier1s + 1 Tier0
- 12 countries in 3 continents
- Dual stack IPv4-IPv6
- 1.9 Tbps to the Tier0

And also, LHCONE – virtual network

Science and Technology Facilities Council

Science and Technology Facilities Council

IPv6 on WLCG & GridPP – a short history

Science and Technology Facilities Council

Why IPv6?

2010-11

- In 1990s – HEP/NASA/ESA global DECnet moved from Phase IV to DECnet/OSI Phase V
- Survey of HEPiX Community (Sep 2010) – “IPv6 readiness”
  - National NRENs are ready; Universities and Labs are not ready
  - Some lack of IPv4 address space, including CERN (WLCG wish to avoid NAT)
- IANA projecting IPv4 address exhaustion
- Sep 2010 – memo from US Federal CIO to all depts including Department of Energy (HEP national labs) - Deploy dual-stack
- Offers of opportunistic CPU resources could arrive and be IPv6-only
- Our middleware, software, technology and tools are not yet IPv6 capable
  - This will take lots of time to fix - so start a working group in April 2011!

Science and Technology Facilities Council

HEPiX IPv6 Working Group

2011-16 Phase 1

- full analysis of work to be done
  - Applications, middleware, system and network tools
  - Operational security
  - Created and operated a distributed test-bed
  - Aim for a timetable and plan for transition
    - Initial plan for support of IPv6-only clients was 2014
- Test the important data transfer protocols, technology and data storage/file systems - for IPv6-readiness

Science and Technology Facilities Council

IPv6 deployment

Phase 1 (continued)

- Storage and Data transfers DPM, dCache, xRootD, OpenAFS, FTS, CASTOR, …
  - Found ***many*** problems needing work
    - Worked closely with developer community
- Concluded IPv6 support will be much later than 2014!
- perfSONAR – end to end network monitoring – made dual-stack capable
- Wrote guidance on IPv6 security for WLCG sites
- **Challenges –** A Collaboration, not a single Management Domain
  - At many sites the Institute networking team decides when/if to deploy IPv6

Science and Technology Facilities Council

IPv6 Deployment on WLCG (Phase 2)

2017 onwards (as approved by WLCG Management Board)

- All Tier1 storage services in IPv4/IPv6 dual-stack mode from April 2018
- Tier-2 storage services
  - Aim for large number of dual-stack Tier-2s **by end 2018**

Monitoring is essential

- Network end to end performance – perfSONAR
- Track **number** of sites deploying IPv6 over time
- Track each **data transfer** – for IPv6 versus IPv4

Science and Technology Facilities Council

Science and Technology Facilities Council

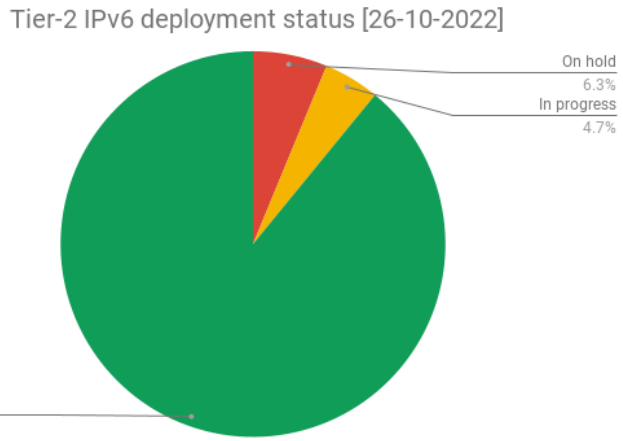
Now: Dual-stack IPv4/IPv6 deployment

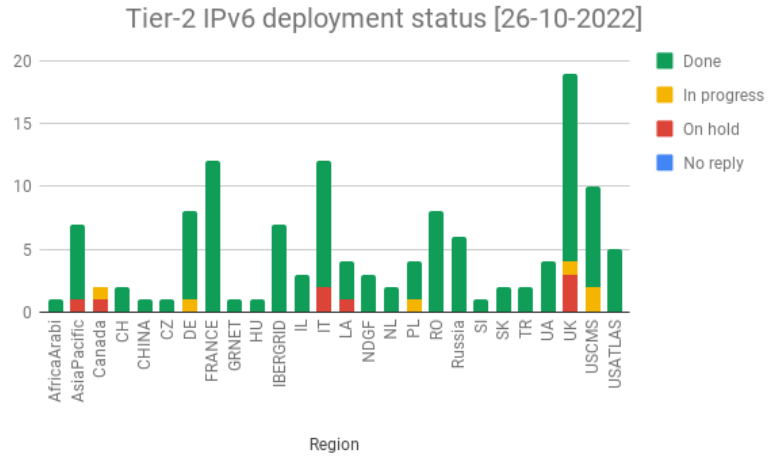
Tier1 storage has been IPv6-capable for several years

Science and Technology Facilities Council

IPv4/IPv6 deployment at WLCG Tier-2 sites

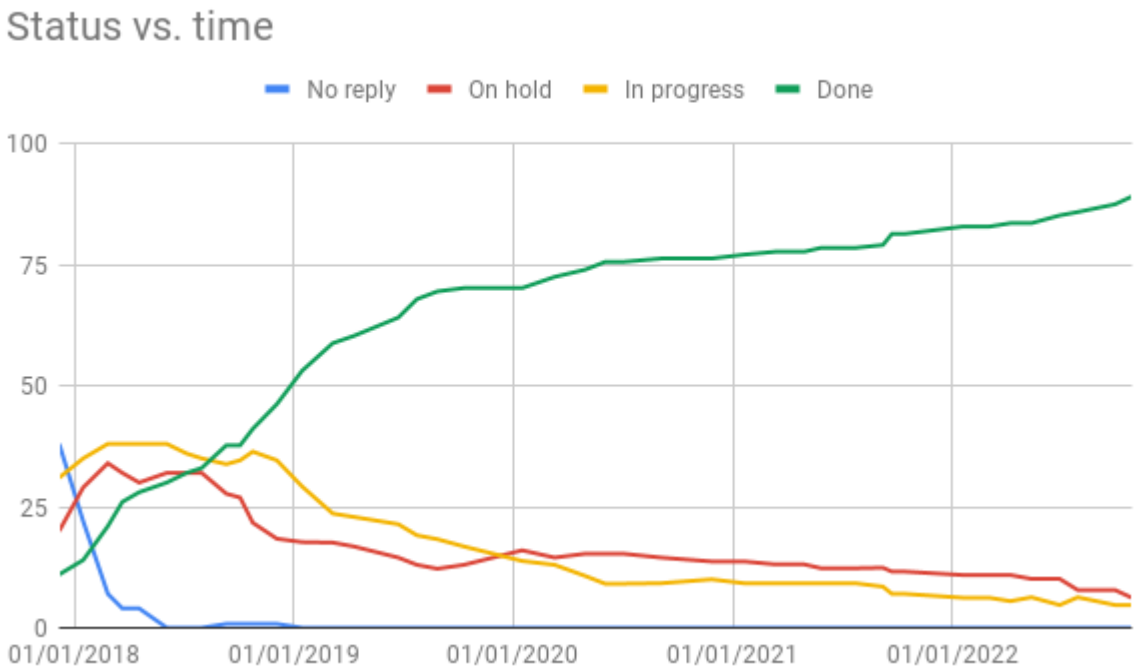
- The deployment campaign was launched in November 2017
- Steady progress
  - ~89% of Tier-2s have dual stack storage
  - 91% of storage

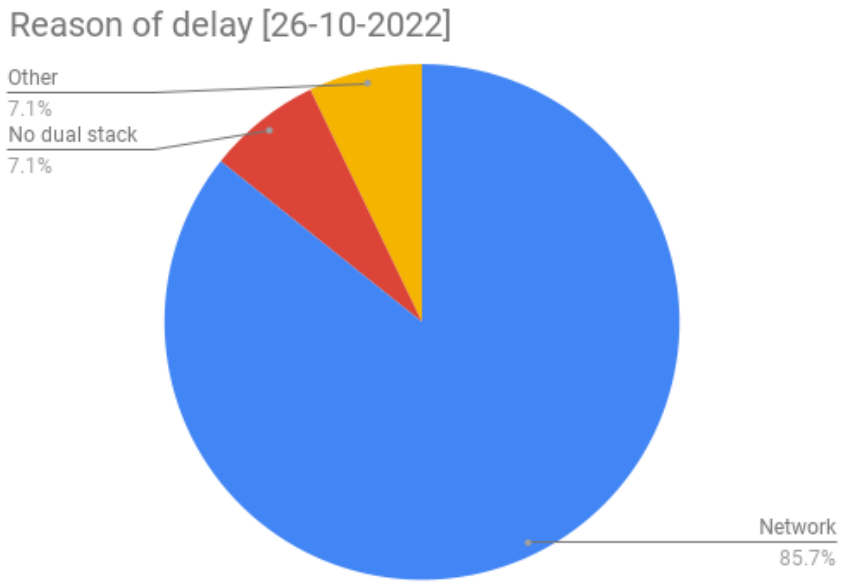




Experiment	Fraction of T2 storage accessible via IPv6
ALICE	89%
ATLAS	89%
CMS	94%
LHCb	79%
Overall	91%

Tier-2 evolution of dual-stack





Science and Technology Facilities Council

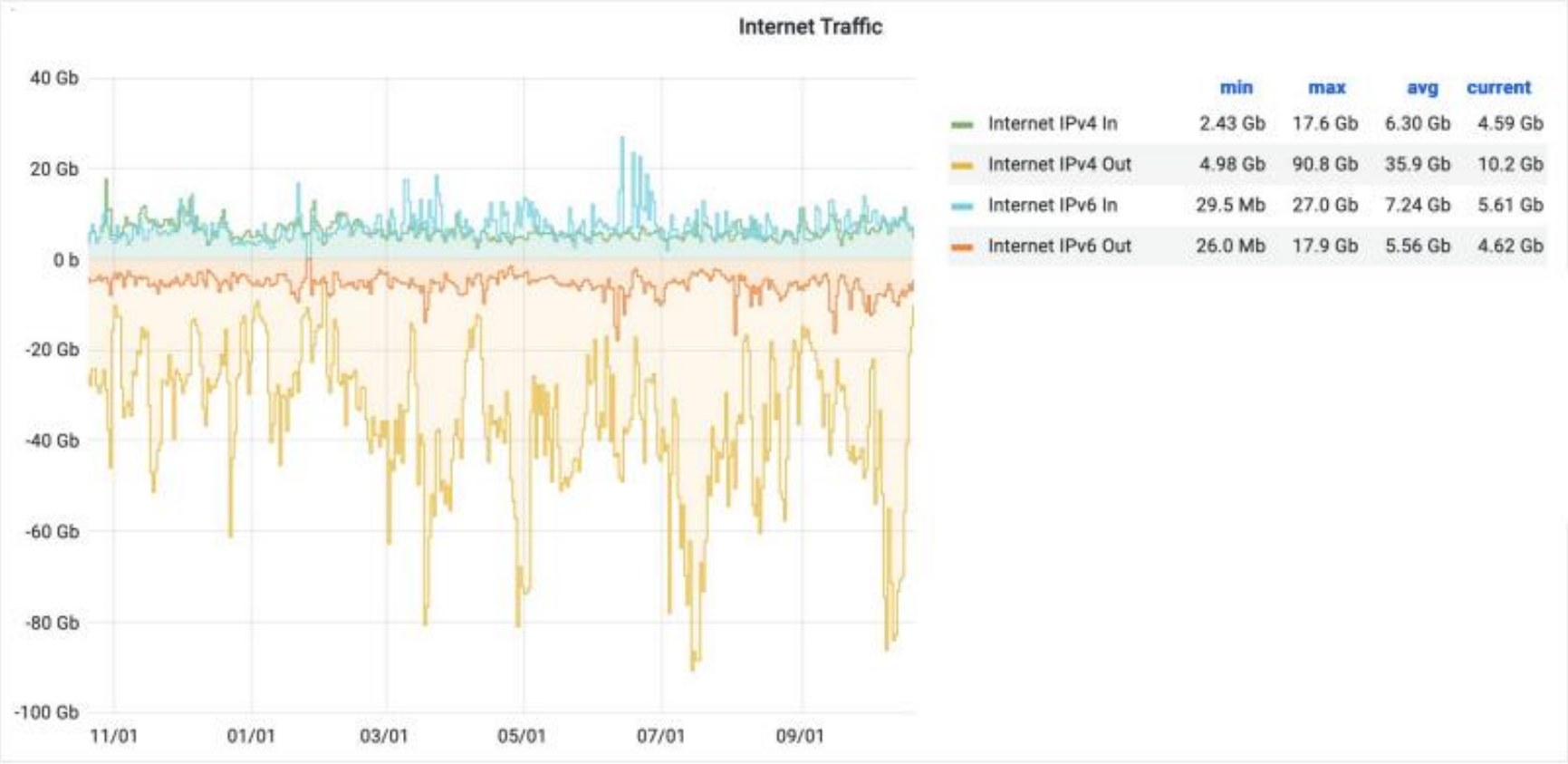
Science and Technology Facilities Council

Monitoring the network and individual data transfers

Science and Technology Facilities Council

CERN Internet Access (not just WLCG) – last 12 months

Incoming traffic: slightly more IPv6 than IPv4
Outgoing traffic: IPv4 7x more than IPv6

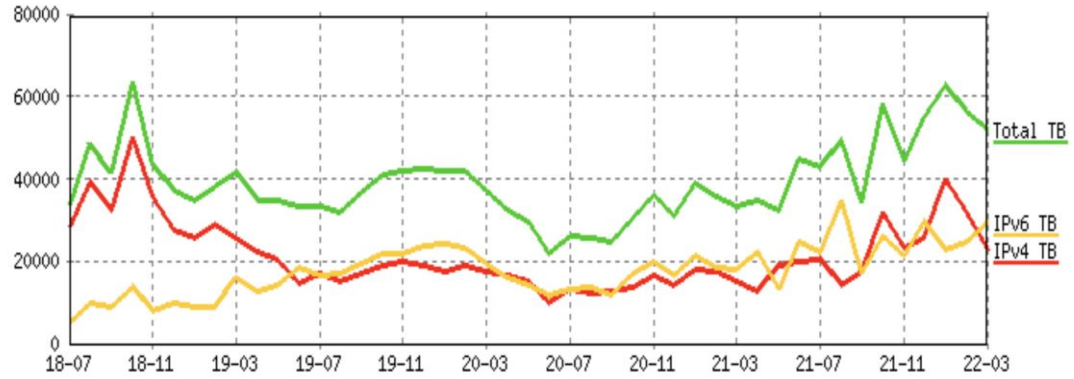


Science and Technology Facilities Council

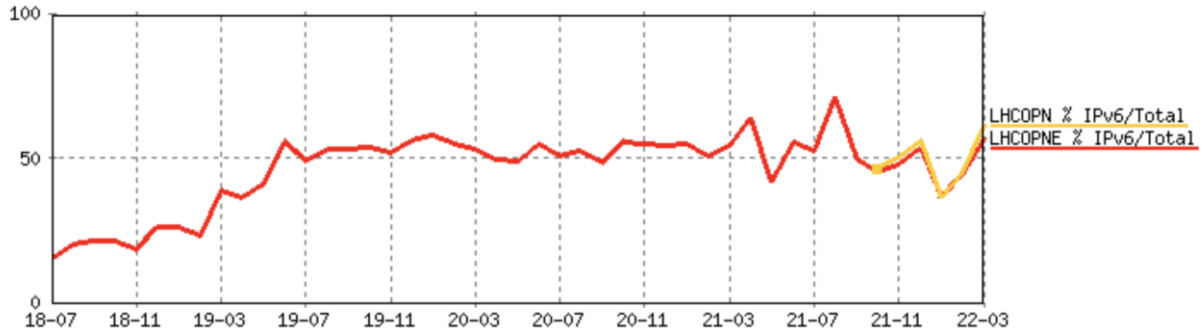
IPv6 traffic on LHCOPN/LHCONE at CERN

LHCOPN and LHCONE IPv4 and IPv6 traffic volumes seen at CERN Tier0

LHCOPN+LHCONE IPv4 and IPv6 traffic volumes month by month



Percentage of IPv6 traffic over the total



Facilities Council

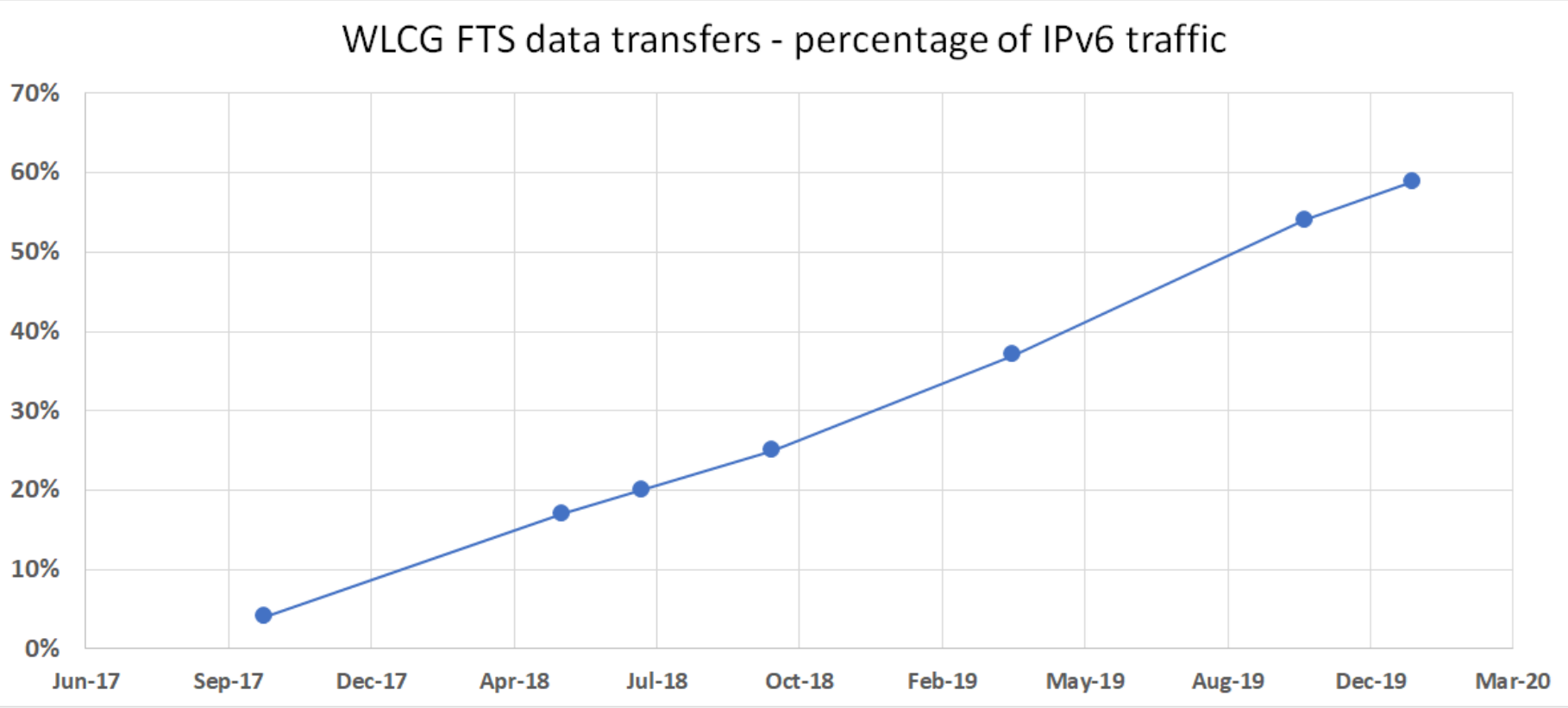
IPv6 traffic on LHCOPN/ONE as seen at CERN

- ~40-70% of all traffic is IPv6
- from June 2019 onwards

- Problems with data from April 2022 – awaiting fix from vendor

% of WLCG Data Transfers over IPv6

2017-2020 all experiments – all File Transfer Service (FTS) servers



IPv6 works!

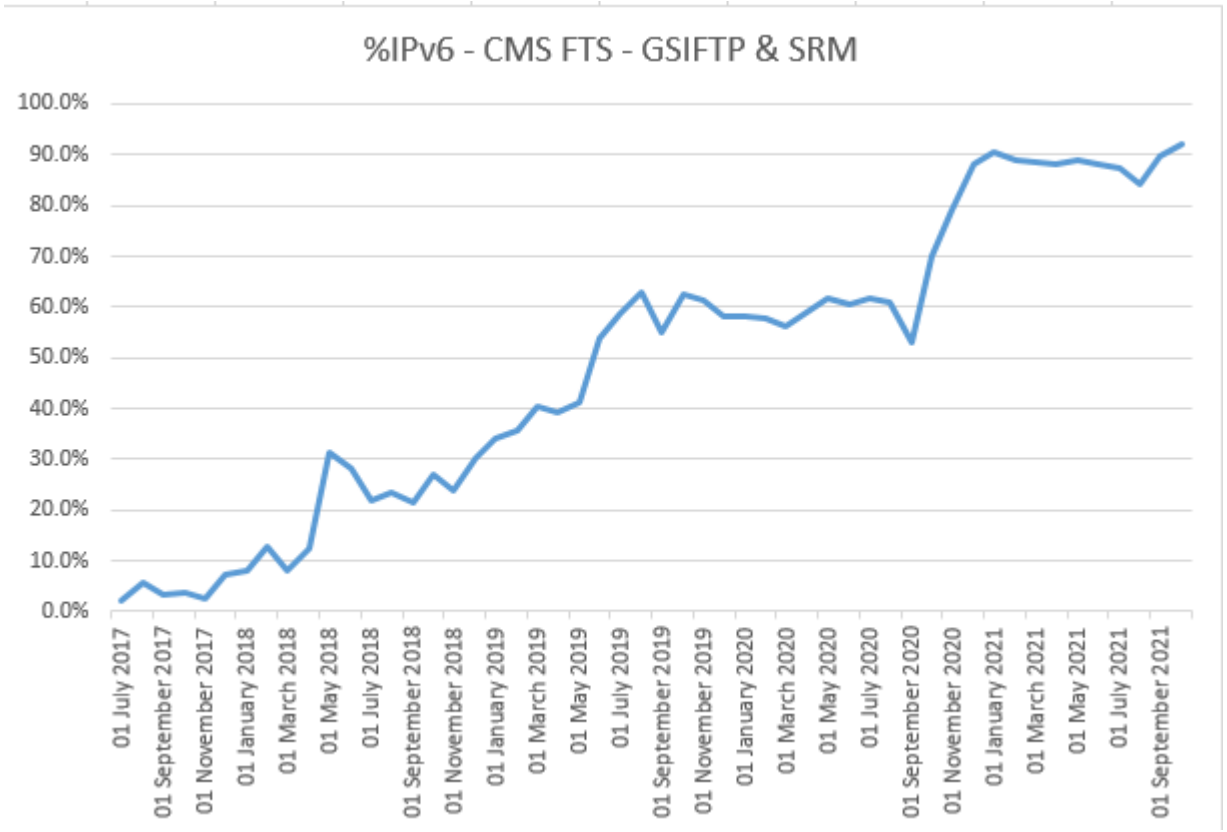
Experiments and physicists are happy

and unaware of the protocol used!

Science and Technology Facilities Council

% of CMS Experiment Data Transfers IPv6

2017-2021



Experiments **no longer** using GSIFTP & SRM. Moving to HTTP/WebDAV

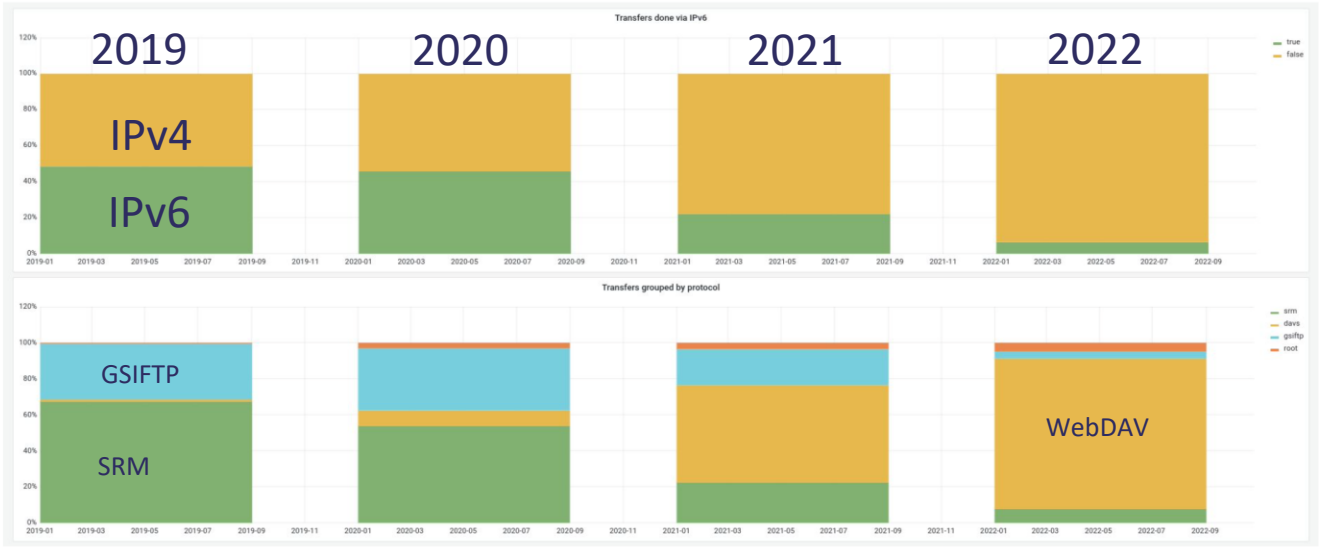
Reason for no plots in 2022

IPv6 HTTP/WebDAV is not yet "visible" in our data monitoring!

Science and Technology Facilities Council

File Transfer Service monitoring & IPv6

II. FTS & IPv6 Monitoring – the bad



Charts plotted using the FTS Aggregated data

IPv4 numbers are wrong!

WebDAV monitor is not yet capable of splitting IPv6 from IPv4

Currently the monitoring assumes IPv4 when the IP version is “unknown”

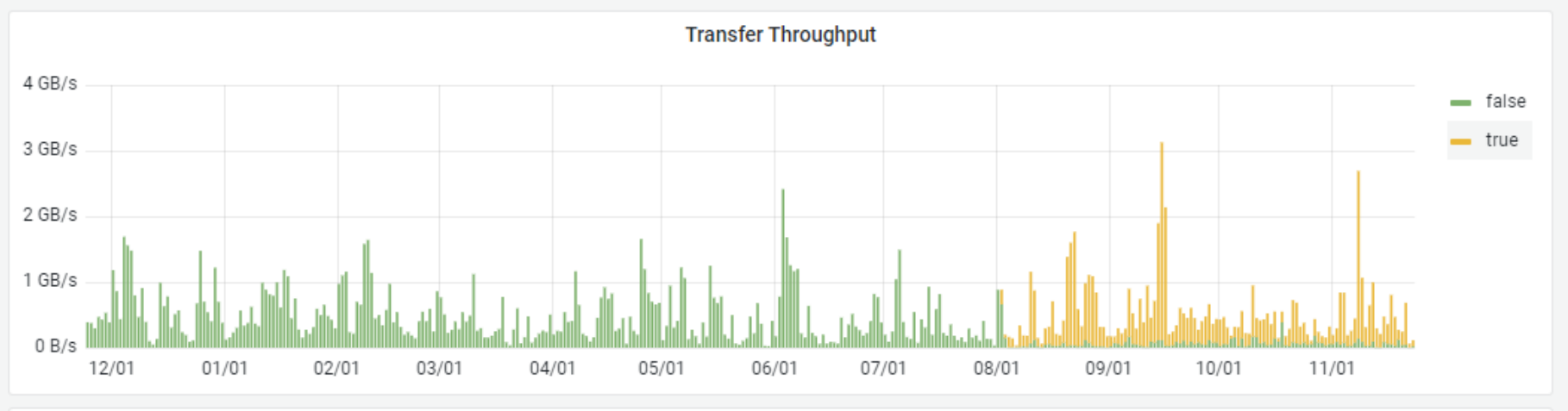
Conclusion – the amount of IPv6 traffic in 2022 is UNKNOWN

Science and Technology Facilities Council

Instrumentation of FTS transfers over HTTP/WebDAV – has started!

Some FTS servers are now able to distinguish IPv6 from IPv4

ATLAS & CMS HTTP transfers into CERN (last year) – IPv6 showing from August 2022 onwards



IPv6 is yellow

Science and Technology Facilities Council

Imperial London - LHCONE - 100 Gbps on IPv6

https://shapingthefutureofjanet.jiscinvolve.org/wp/uncategorized/100gbps-of-cern-data-over-ipv6-on-the-janet-network/

Figure 1 — Imperial monitoring shows the two-hour period where the 100G link was filled and where 100% of the LHCONE traffic was IPv6.

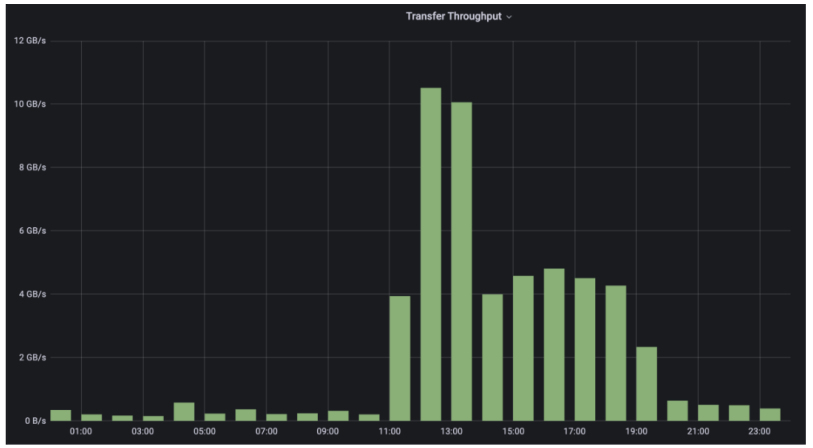


Figure 2 — The traffic levels seen in the network view correspond to those seen by the WLCG File Transfer Service (FTS) visualization tools.

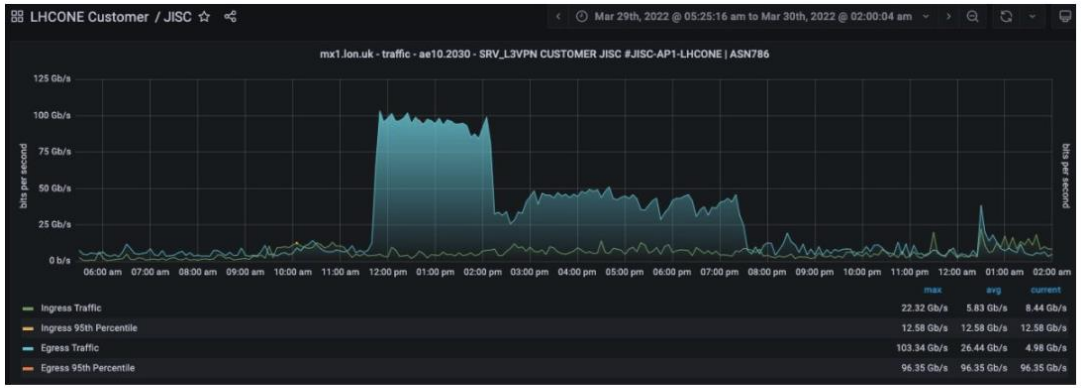


Figure 3 — It was also interesting to see this traffic reflected in the monitoring platform for the GÉANT pan-European research and education backbone network.

Why do we still see IPv4 traffic?

But why do transfers still use IPv4?

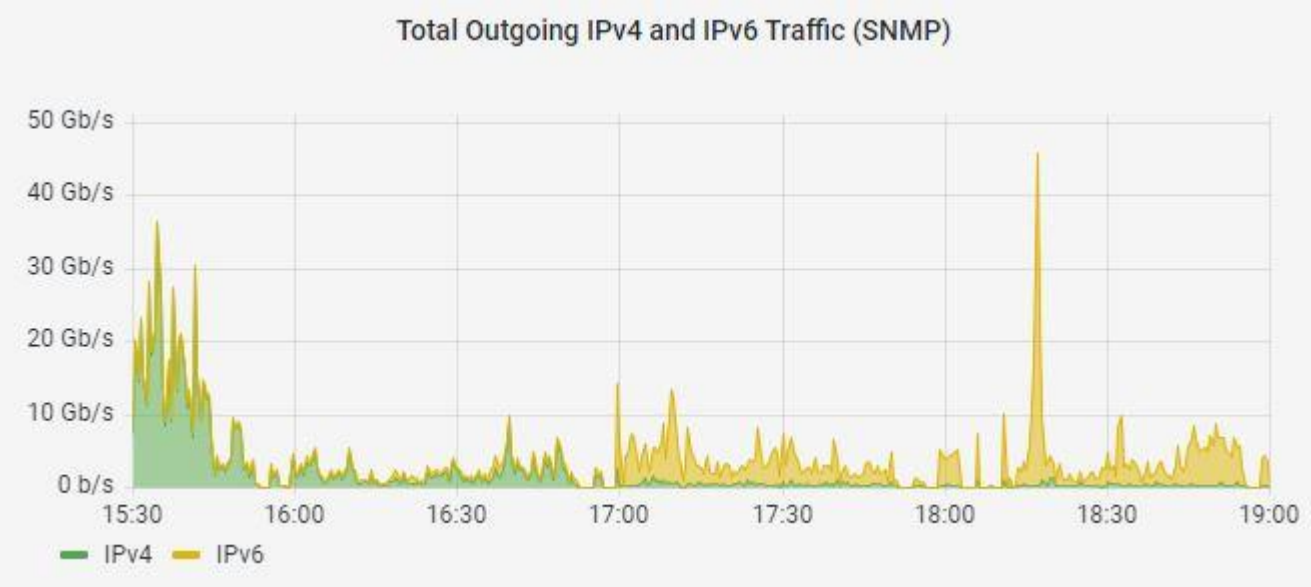
The working group's main activity in 2022!

- Tier 0/1s are dual-stack, but IPv4 often still used for transfers
- ***Possible reasons***:
  - Site/experiment issues - some storage end-points not yet dual-stack
  - Old software stacks (legacy deployments) are still deployed
  - Bad configuration or transfer request prefers IPv4
  - "happy eyeballs" (RFC6555/8305)
  - Worker nodes are still IPv4-only
- IPv6 WG is analyzing Tier-1/2 IPv4 transfers (case by case)
- One Example: Data transfers into USA/ATLAS Great Lakes Tier 2 (AGTL2)
  - Found to use IPv4 even when both ends dual-stack
  - Solved: A default Java configuration is set to prefer IPv4 (see next slide)

Science and Technology Facilities Council

IPv4/6 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!

Now chasing all sites to change the configuration

Science and Technology Facilities Council

Analysis of job logs – an example

- HTCondor jobs for one LHC experiment
  - Check for dual-stacked end points
- 1st iteration: 6th to 27th of June 2022
  - 17813 different data exchange events (3850 dual-stacked)
  - 76.8% (2956) dual-stacked (type = production)
  - 0.1% (3) dual-stacked (type = analysis)
- **Inform the experiment** of the IPv4-only analysis end-points
  - They slowly move to dual-stack (~1/3rd slots/collectors are now set to prefer IPv6)
- 2nd iteration: 1st to 14th of September 2022
  - 16072 different data exchange events (5347 dual-stacked)
  - 74.8% (4127) dual-stacked (type = production)
  - 22.1% (1220) dual-stacked (type = analysis) (**An improvement**)

Science and Technology Facilities Council

Phase 3 - IPv6-only

WLCG - from dual-stack to IPv6-only (CHEP2019) https://doi.org/10.1051/epjconf/202024507045

- Planning for an IPv6-only WLCG
- To simplify operations
  - Dual-stack infrastructure is the most complex
  - Dual-stack is less secure
- The goal we are working towards
  - IPv6-only for the majority of WLCG services and clients
  - With ongoing support for IPv4-only clients where needed
- Timetable still to be defined and agreed with Management Board

Science and Technology Facilities Council

Other drivers for IPv6-only

- Still need to be ready for use of (opportunistic) IPv6-only CPU
- BUT there are other drivers for IPv6-only:
  - lack of public IPv4 addresses in Data Centres
    - avoid use of NAT
  - SCITAG – packet marking (in header of IPv6 packets)
  - USA Federal Government – directive on IPv6-only
  - (see next 3 slides)

Science and Technology Facilities Council

Refresh - **US Government IPv6 Mandate**

FY23	• All new federal systems to be IPv6 enabled at deployment. • **20%** of all networked federal systems **IPv6-only**
FY24	• **50%** of all networked federal systems **IPv6-only**
FY25	• **80%** of all networked federal systems **IPv6-only** • Identify, plan, schedule retirement/replacement of remaining networked systems that cannot be converted to IPv6-only

- US National Labs (T1s) are included; **but** university-run T2s are not subject to the mandate

Science and Technology Facilities Council

Fermilab

3 11/2/22 Phil DeMar / Nick Buraglio I October 2022 HEPiX WG Meeting

Research Networking Technical WG

https://www.scitags.org/

Scitags Initiative

- Scientific Network Tags (scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

- Enable tracking and correlation of our transfers with Research and Education Network Providers (R&Es) network flow monitoring
- Experiments can better understand how their network flows perform along the path
  - Improve visibility into how network flows perform (per activity) within R&E segments
  - Get insights into how experiment is using the networks, get additional data from R&Es on behaviour of our transfers (traffic, paths, etc.)
- Sites can get visibility into how different network flows perform
  - Network monitoring per flow (with experiment/activity information)
    - E.g. RTT, retransmits, segment size, congestion window, etc. all per flow

4

Science and Technology Facilities Council

Scitags Technical Spec

Technical Spec

The detailed technical specifications are maintained on a Google doc

- The spec covers both Flow Labeling via UDP Fireflies and Packet Marking via the use of the IPv6 Flow Label.
  - **Fireflies** are UDP packets in Syslog format with a defined, versioned JSON schema.
    - Packets are intended to be sent to the same destination (port 10514) as the flow they are labeling and these packets are intended to be world readable.
    - Packets can also be sent to specific regional or global collectors.
    - Use of syslog format makes it easy to send to Logstash or similar receivers.
    - Two optional fields were added recently: **usage** and **netlink**
      - **Usage** reports on bytes sent/received as seen by the storage
      - **Netlink** adds TCP/IP low level metrics (RTT, Cwnd, Buffers, Busy time, etc.)
  - **Packet marking** uses the 20 bit flow label field in IPv6 packets.
    - To meet the spirit of RFC6437, we use 5 of the bits for entropy, 6 bits (64) for activity and 9 bits (512) for science domain (experiment).
- The document also covers methods for communicating owner/activity and other services and frameworks that may be needed for implementation.

8

Science and Technology Facilities Council

More information

Some papers from the HEPiX IPv6 working group

a) *"IPv6 Security"*

- M Babik et al 2017 J. Phys.: Conf. Ser. 898 102008
- http://dx.doi.org/10.1088/1742-6596/898/10/102008

b) *"IPv6 in production: its deployment and usage in WLCG"*

- M Babik et al, EPJ Web of Conferences 214, 08010 (2019)
- http://dx.doi.org/10.1051/epjconf/201921408010

c) *"IPv6-only networking on WLCG"*

- M Babik et al EPJ Web of Conferences 245, 07045 (2020)
- http://dx.doi.org/10.1051/epjconf/202024507045

Science and Technology Facilities Council

Summary

- WLCG is supporting use of IPv6-only clients (CPU)
- Tier-1s all have production storage accessible over IPv6
- Tier-2s ~90% sites are done
- Monitoring data transfers and configuration is essential
  - But currently broken – to be fixed
- Why do two dual-stack endpoints use IPv4 between them?
  - A priority for 2022/23
- Phase 3 – we are planning for move to IPv6-only services
  - Dual-stack is NOT the desired end-point!
- ***message to new research communities - build on IPv6 from start***

Science and Technology Facilities Council

Science and
Technology
Facilities Council

Questions?

UKRI Science and Technology Facilities Council

Thank you

Science and Technology Facilities Council @STFC_matters Science and Technology Facilities Council