Overcoming obstacles to IPv6 on WLCG

Bruno Hoeft[1], Jiri Chudoba[2], David Kelsey[3], Edoardo Martelli[4]

HEPiX-Spring 2023, Taipei, 29 March 2023

1 - Karlsruhe Institute of Technology (KIT), Germany
2 - Institute of Physics of the Czech Academy of Sciences (FZU), Czech Republic
3 - UKRI Science and Technology Facilities Council (STFC UK), United Kingdom
4 - Conseil Européen pour la Recherche Nucléaire (CERN), Switzerland

On behalf of all co-authors in the HEPiX IPv6 working group

Active in HEPiX IPv6 Working Group – last 12 months

- M Babik (CERN), M Bly (RAL), N Buraglio (ESnet), T Chown (Jisc), D Christidis (U Texas/ATLAS), J Chudoba (FZU Prague), C Condurache (EGI.eu), 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), E Simmonds (FNAL), T Skirvin (FNAL)

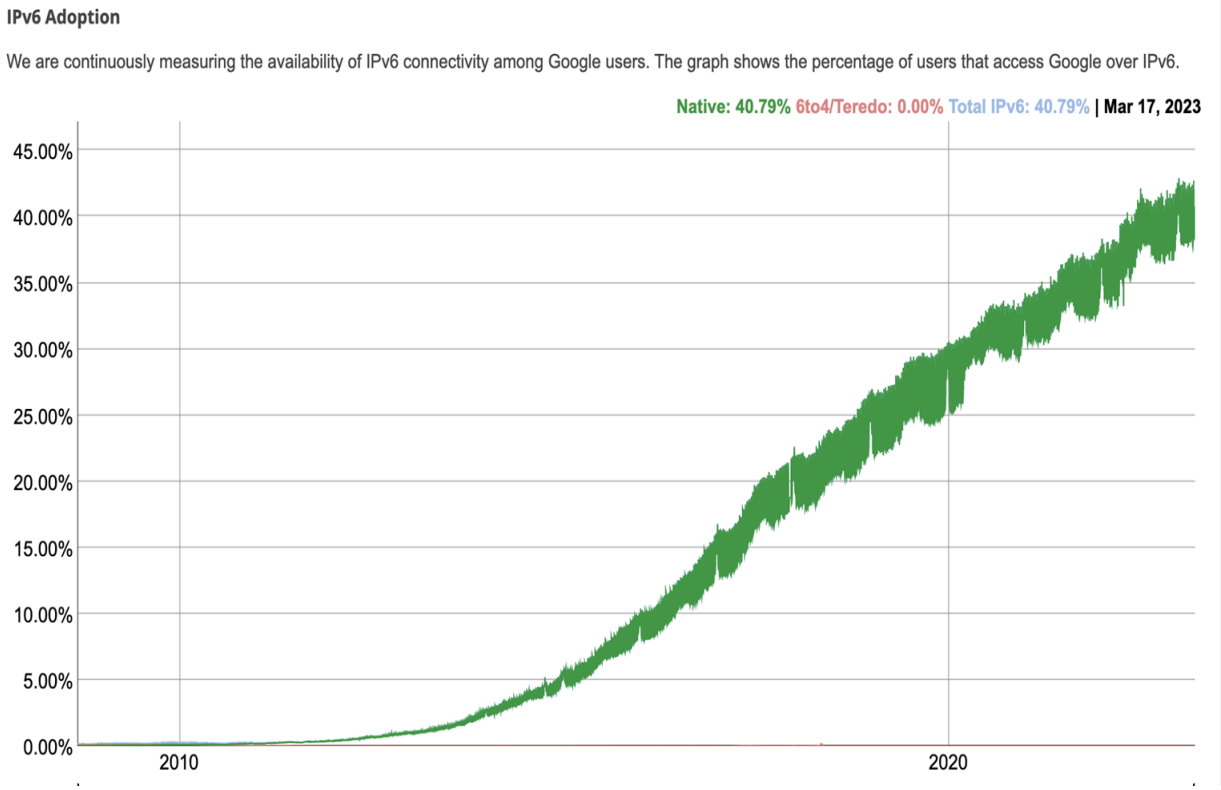
- Many more in the past, and others join from time to time
- *and thanks also to WLCG operations, WLCG sites, LHC experiments, networking teams, monitoring groups, storage developers…*

Outline

- Global IPv6 traffic
- The HEPiX IPv6 working group
- Deployment of IPv6/IPv4 dual-stack storage
- IPv6 monitoring
- Plans for IPv6-only WLCG
- Obstacles to IPv6 on WLCG
- Overcoming those obstacles
- Summary

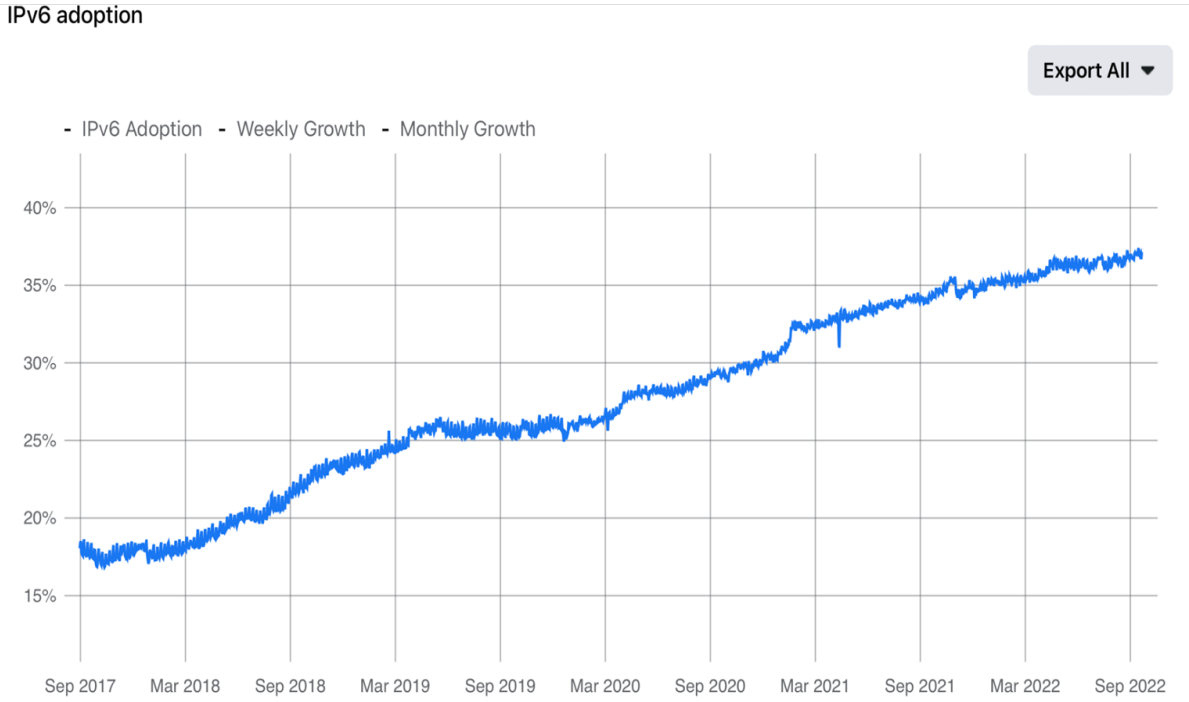
Global IPv6 traffic

General IPv6 traffic continues to grow (Google and Facebook)



Google

Facebook



(and also) IPv6 users -6lab.cisco.com

Taiwan

Display Users Data

50%

2018

France

Display Users Data

70%

2017

The HEPiX IPv6 working group

The HEPiX IPv6 Working Group

- Started in April 2011
  - 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 - 2017-2020 - deploy dual-stack storage on WLCG
  - in production
- Phase 3 - 2021-onwards - plan for IPv6-only
  - investigate reasons for data transfers over IPv4

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

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

Drivers for use of IPv6

- 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
- BUT there are other drivers for IPv6 too
  - scitags.org – packet marking (in header of IPv6 packets)
    - Research Networking Technical Working Group (RNTWG)
  - USA Federal Government – directive on “IPv6-only” (Nov 2020)
  - multiONE (several LHCONE for different communities)
    - uses the scitags marks in header flow label for policy based routing

US Government **IPv6** Mandate
(from Phil Demar and Nick Buraglio)

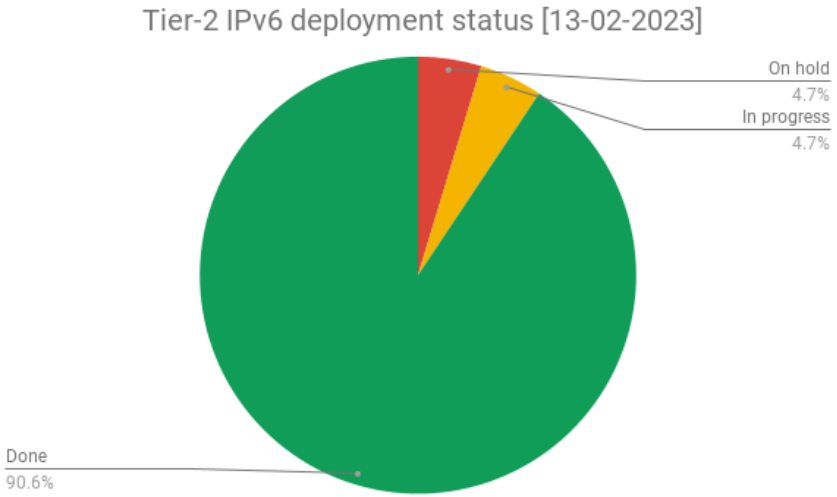
FY23	• All new federal systems to be IPv6 enabled at deployment. • **20%** of all networked federal systems **IPv6-only**
FY24	• **50%** of all networked federal systems **IPv6-only**
FY25	• **80%** of all networked federal systems **IPv6-only** • Identify, plan, schedule retirement/replacement of remaining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

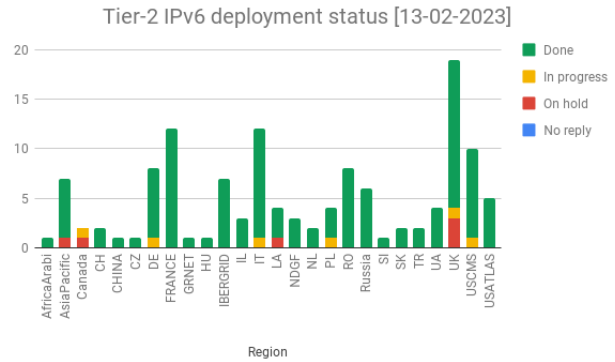
- US National Labs (T1s) are included; **but** university-run T2s are not subject to the mandate

Deployment of IPv6/IPv4 dual-stack storage

IPv6/IPv4 deployment at WLCG Tier-2 sites

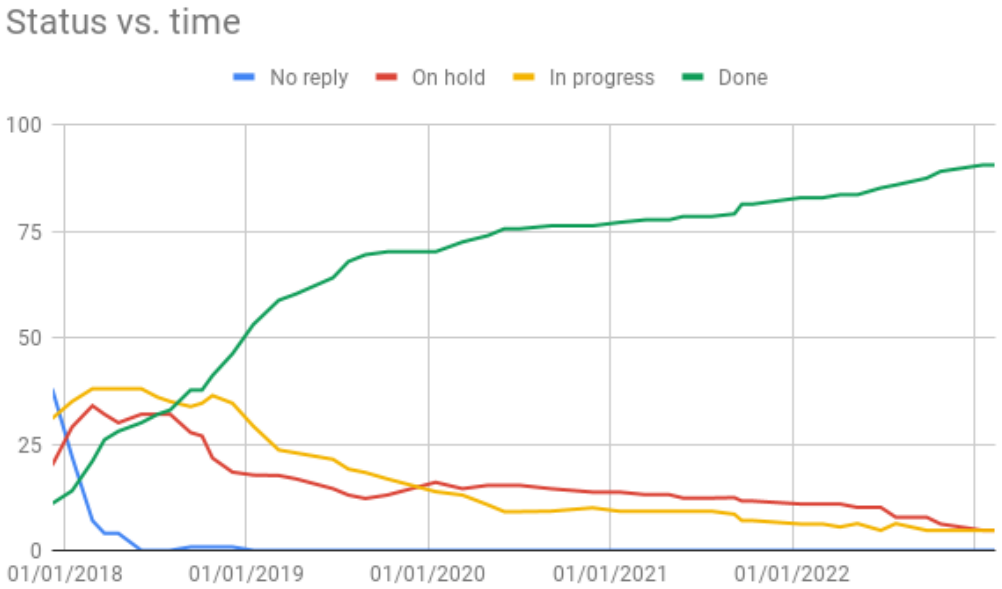
- The deployment campaign was launched in November 2017
- Steady progress (status)
  - ~91% of Tier-2s have dual stack storage
  - 91% of storage

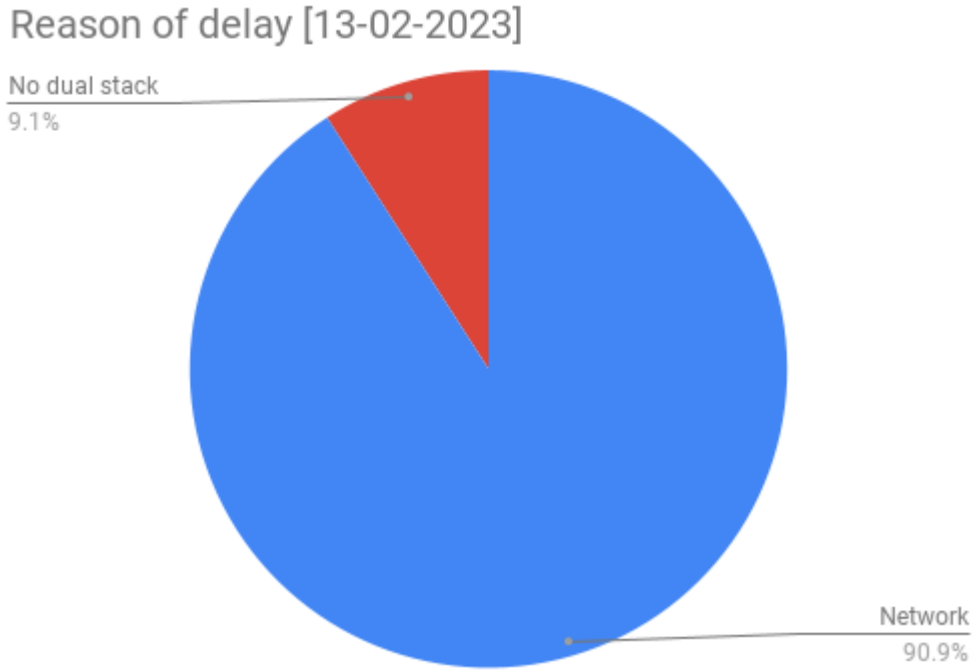




Experiment	Fraction of T2 storage accessible via IPv6
ALICE	90%
ATLAS	89%
CMS	96%
LHCb	79%
Overall	91%

Tier-2 evolution of dual-stack





IPv6 monitoring

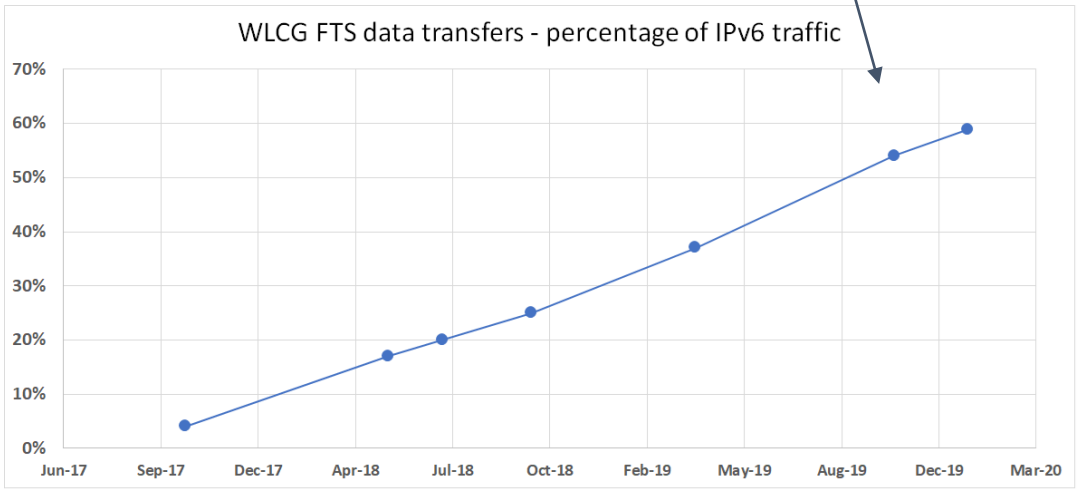
Importance of monitoring

- Must monitor
  - deployment of IPv6-capable services
  - fraction of data transfers taking place over IPv6
- Monitoring implementations used for IPv6
  - perfSONAR
  - ETF - experiment test framework
  - FTS (File Transfer Service)
  - Network utilisation and traffic plots
    - e.g. IPv6 versus IPv4 on LHCOPN/LHCONE as seen at the CERN routers

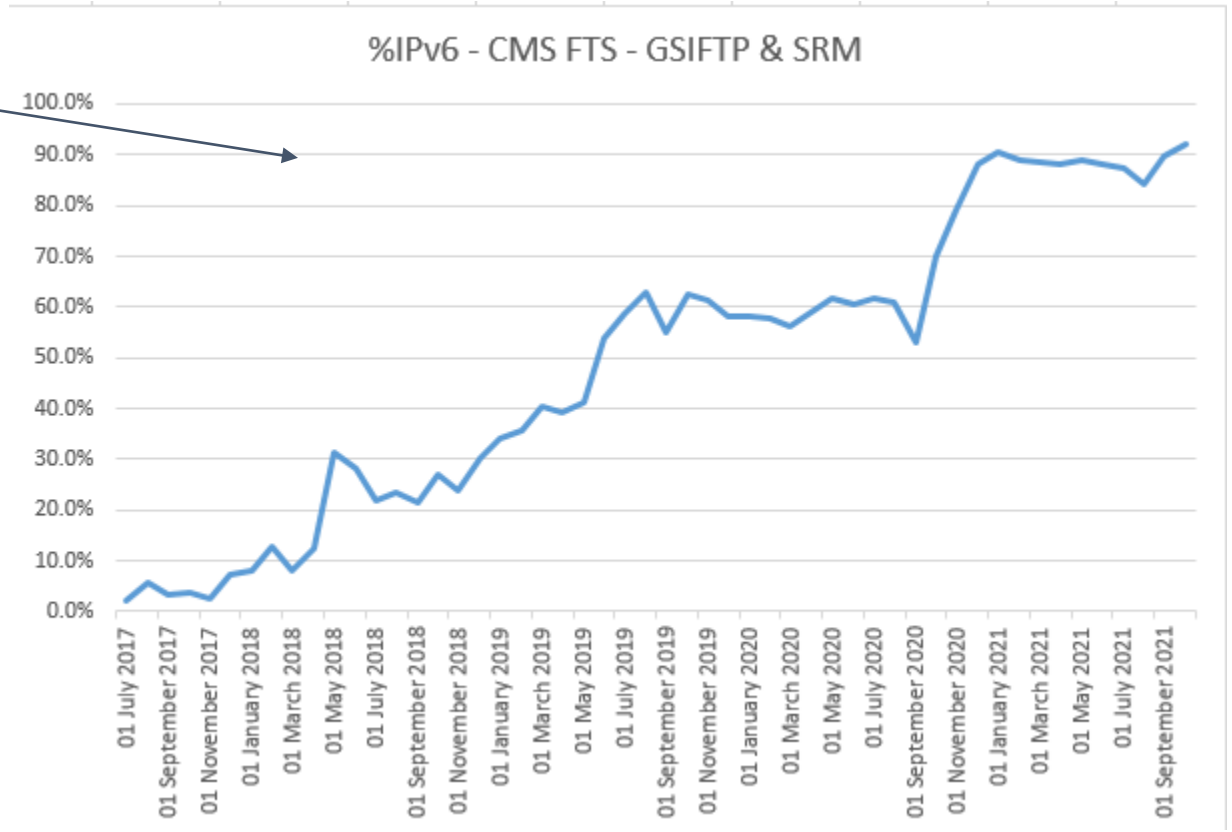
- But in recent years some existing **monitoring has ceased to work**

% of WLCG FTS data traffic over IPv6

- Some FTS protocols, e.g. DAVS still not able to monitor IPv6 traffic
- DAVS excluded from this plot
  - and explains end date here
- GSIFTP & SRM are OK



All protocols & all experiments



Broken FTS IPv6 monitoring

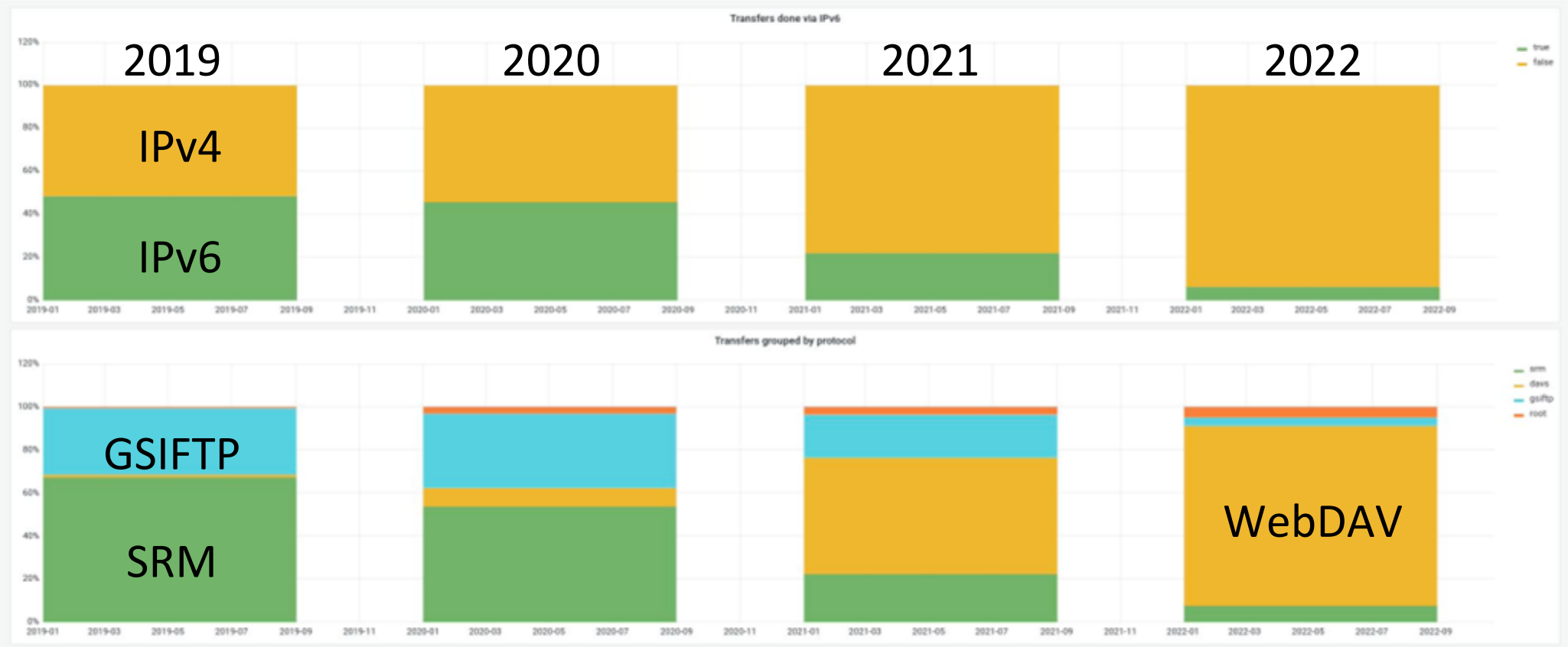
- Experiments **no longer** using GSIFTP & SRM
  - Moving to HTTP/WebDAV
- Reason for no plots in 2022
- IPv6 HTTP/WebDAV is not fully “visible” in our FTS monitoring!

In the next slide

- IPv4 numbers are wrong!
- WebDAV monitor is being made capable of splitting IPv6 from IPv4
- In past, wrongly assumed “IPv4” when the IP version is “unknown”
- Conclusion – the **amount of IPv6 traffic in 2022 is UNKNOWN**
- **This is now being fixed - but will take time to deploy**

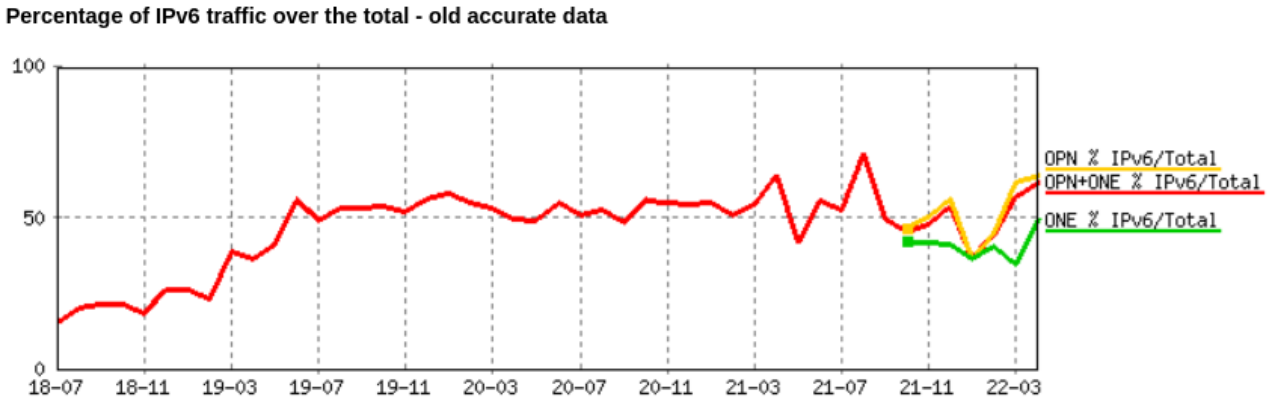
II. FTS & IPv6 Monitoring – the bad

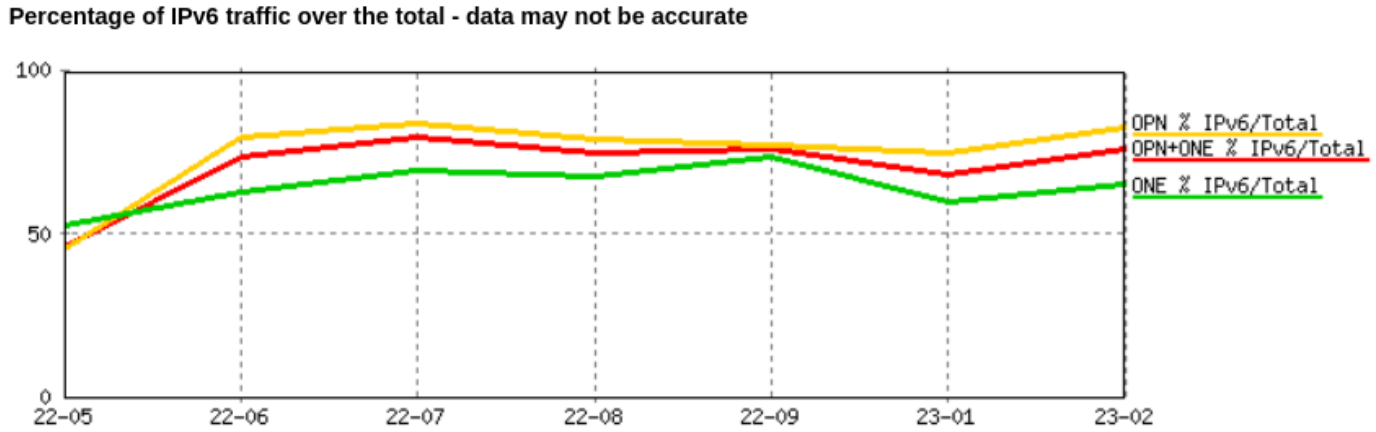
"IPv4" includes "unknown"



Charts plotted using the FTS Aggregated data

IPv6 traffic on LHCOPN/LHCONE at CERN



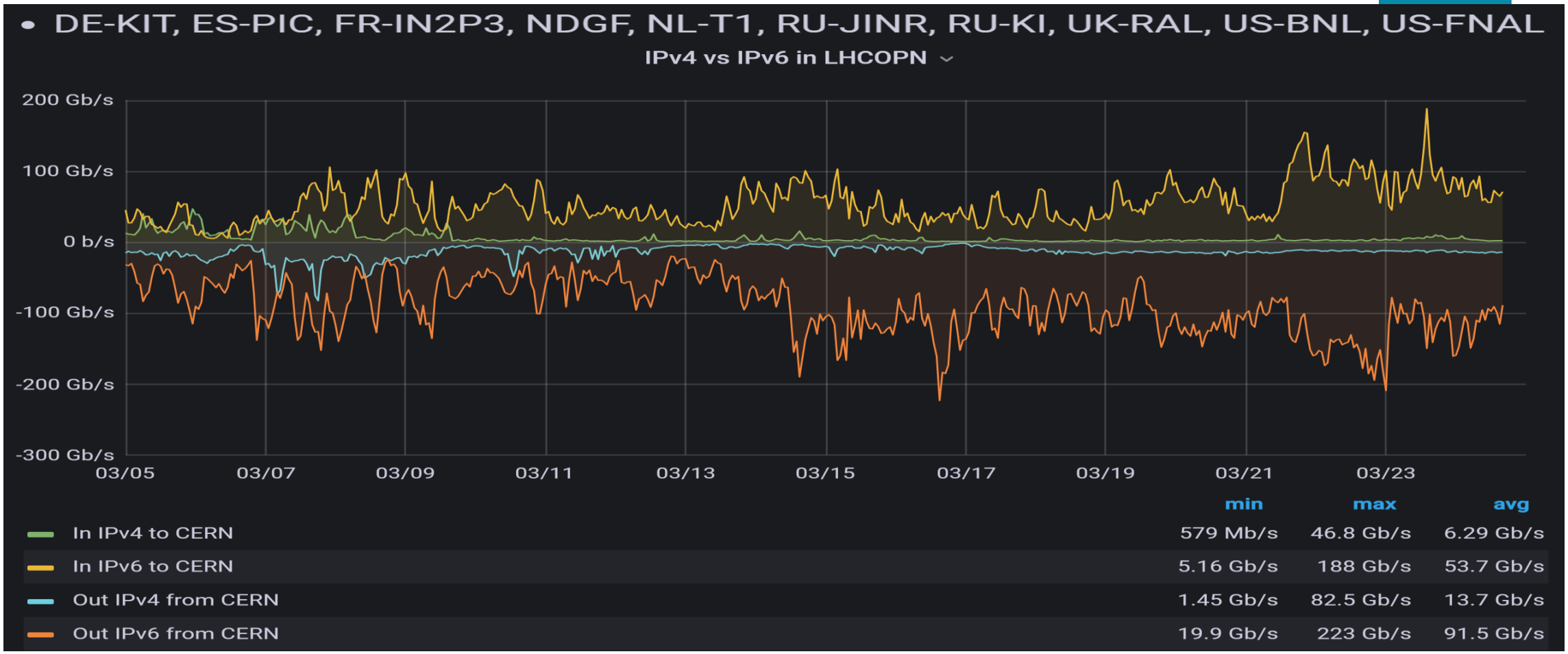


LINK to these plots

IPv6 traffic on LHCOPN/ONE as seen at CERN

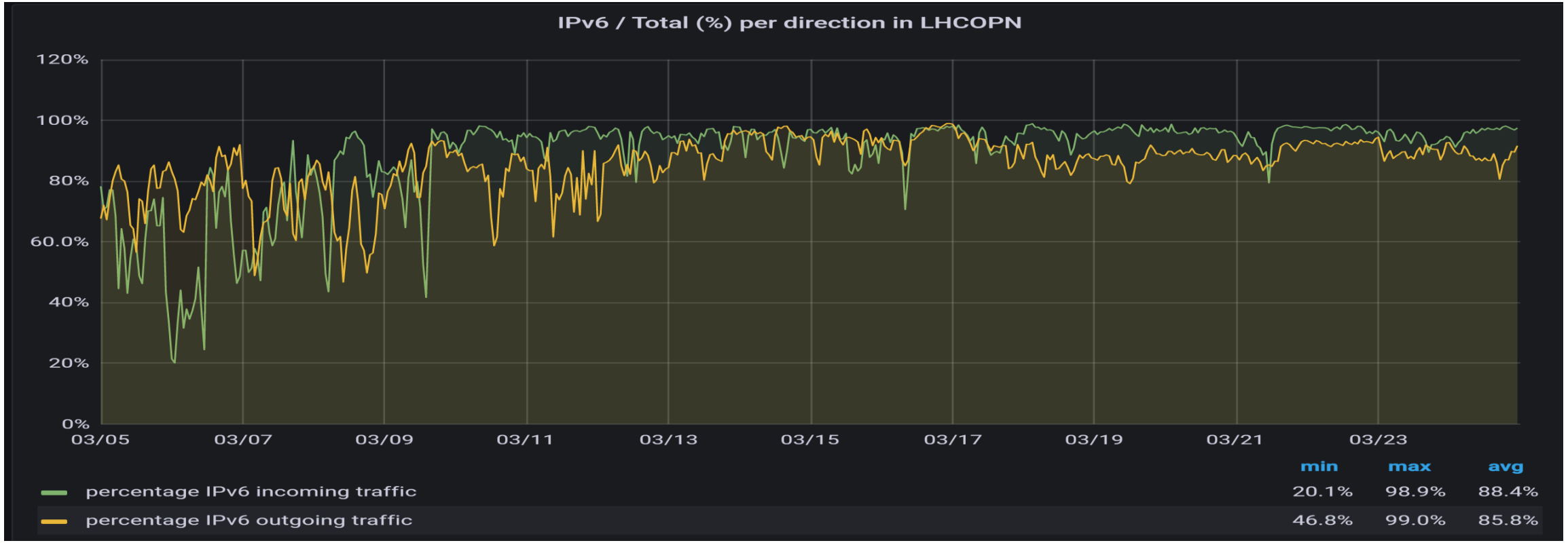
- *Problems with data from April 2022 onwards*
- May not be accurate from April 22
- Ratio IPv6/IPv4 may be correct
- Required fix from vendor
- Next slides fixed (after 3 Mar 23)

“Fixed” plots from CERN (>5 Mar 23)



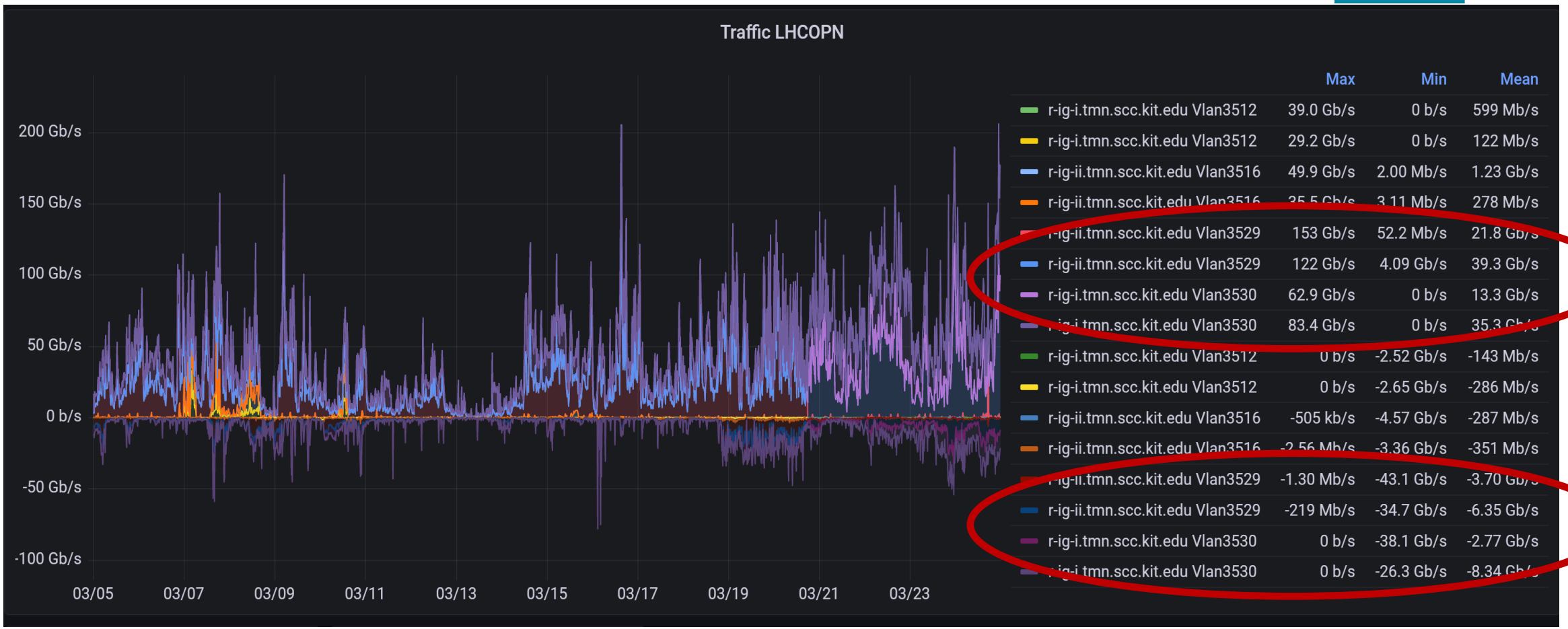
https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4

“Fixed” plots (2) from CERN (>3 Mar 23)



Average rates more than 85% IPv6 in each direction

LHCOPN at KIT



Plans for IPv6-only WLCG

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

- The end point of the transition from IPv4 is an IPv6-only WLCG core network
- To simplify operations
  - Dual-stack infrastructure is the most complex
  - Dual-stack has more security threat vectors
- Large infrastructures (e.g. Facebook) use IPv6-only internally
- The plan - 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 to be defined

Obstacles to IPv6 on WLCG

Why: IPv4 transfers on LHCOPN?

- Tier-1s are dual-stack, but IPv4 often still used for transfers
  - Site/experiment issues
    - Old software stacks (legacy deployments)
  - both ends dual-stack but configuration prefers IPv4
  - transfers are to/from WN's - and the WN's are IPv4-only
- IPv6 WG has been analyzing Tier-1 top-talkers over IPv4
  - understand reasons for IPv4 and request fixes to problems
- encourage all sites to deploy CPU as dual-stack or IPv6-only
- encourage all sites and all experiments to "prefer" IPv6

The obstacles to IPv6

Obstacles to IPv6 on WLCG	Within IPv6 WG or WLCG MB control?	Status	Possible actions
Institute/Site not yet providing IPv6 network support	No	>90% WLCG sites do support IPv6 networking	WG, MB and experiments can continue to “encourage”
Site provides IPv6 network support but still no dual-stack storage	Yes - MB mandate	>90% sites offering at least one dual-stack storage service	Experiments can request deployment of dual-stack storage. MB could “demand”
Non-storage services not IPv6 capable	Partial - no MB mandate	Not yet tracked. We have evidence suggesting ~60% of these services are dual-stack	Sites and experiments continue controlling and negotiating constraints that prevent IPv6. WG can continue to encourage

WG = HEPiX IPv6 Working Group MB = WLCG Management Board

The obstacles to IPv6 (2)

Obstacles to IPv6 on WLCG	Within IPv6 WG or WLCG MB control?	Status	Possible actions
WLCG clients (CPU) not IPv6 capable	Partial - no MB mandate	Not yet tracked. WG encourages sites. Some sites have fully deployed dual-stack worker nodes	Sites and experiments continue controlling and negotiating constraints that prevent IPv6. WG can continue to encourage
Monitoring is not available to track use of IPv6 by data transfers	No - but MB supports the need for this	Partially available. FTS now handles "IPversion"; Storage must provide "IPversion" on file close	WG to continue to encourage. Storage developers to provide IP version in logs and new implementations to be deployed by sites
Service is IPv6 capable but chooses not to use IPv6 (configuration error or deliberate choice).	Partial	A WG priority during 2023. We are actively seeking, tracking and chasing.	Experiments either control the configuration or can request configuration changes. WG to request.

Should we ask sites to move straight to IPv6-only CPU?

Overcoming those obstacles

Also - see talk at ISGC 2023 conference

- **IPv4 to IPv6 Worker Node migration in WLCG by Bruno Hoeft (KIT)**

Why is IPv4 used between dual-stack endpoints?

HTCondor job logs During summer of 2022

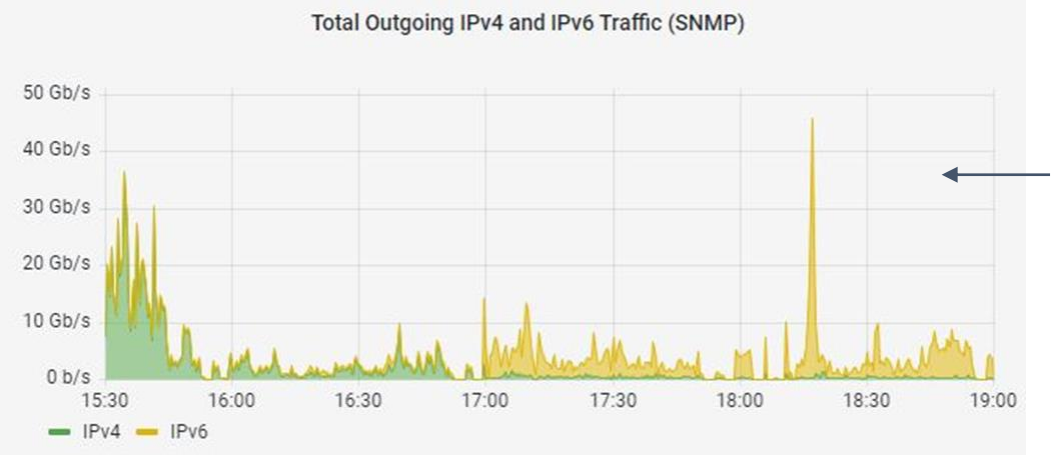
- study in June (17.8K jobs) for one LHC experiment
  - many HTCondor Scheduler Daemons not dual-stack (close to zero for “analysis” jobs)
- experiment was informed
  - they configure IPv6 preference on Schedds
- second study in September (16K jobs) showed a significant improvement
  - “analysis” now has 19% of the dual-stacked Schedds

Study of the **IPv4 and IPv6 Top-talkers** (each month) on LHCOPN/LHCONE (at CERN)

- Found many different sites and endpoints using IPv4
- Not easy to find protocols/services
- 2 improvements will help
  - weekly basis rather than monthly
  - log port numbers and not just IP addresses

Work continues on these studies

Some other obstacles fixed



Data transfers into USA/ATLAS Great Lakes Tier 2 (AGTL2)

Found to use IPv4 even when both ends dual-stack (dCache/WebDAV)

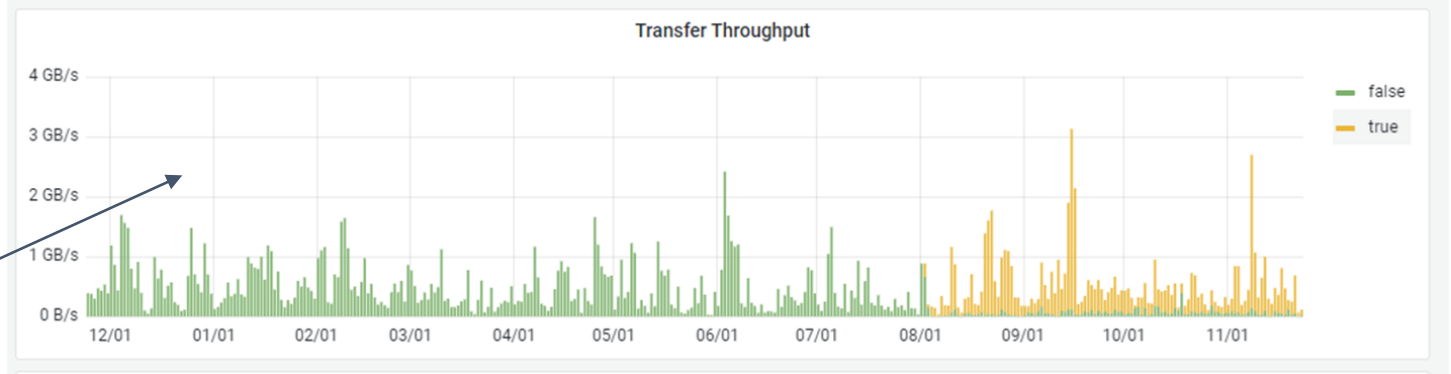
java.net.preferIPv6Addresses (default: false) **- Now set to "true"**

Fixed at 17:00 on 14 Feb 2022 (confirmed in the plot!)

This fix is essential for all dCache instances - fixed in v7.2.11

IPv6 is yellow

Some FTS monitoring now able to distinguish IPv6 from IPv4



ATLAS & CMS HTTP transfers into CERN (last year) – IPv6

showing from August 2022 onwards

Summary

Summary

- WLCG is ready to support use of IPv6-only clients
- Tier-1s all have production storage accessible over IPv6
- Tier-2s >90% sites are IPv6 capable
- Monitoring data transfers is essential - broken and being fixed
- We have investigated obstacles to IPv6 in WLCG (seeking fixes)
  - e.g. Why do two dual-stack endpoints use IPv4 between them?
- 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***

Questions, Discussion?

Backup slides

The CERN Large Hadron Collider (LHC)

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

4 July 2012

How do you get from this to this?

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

CERN

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.

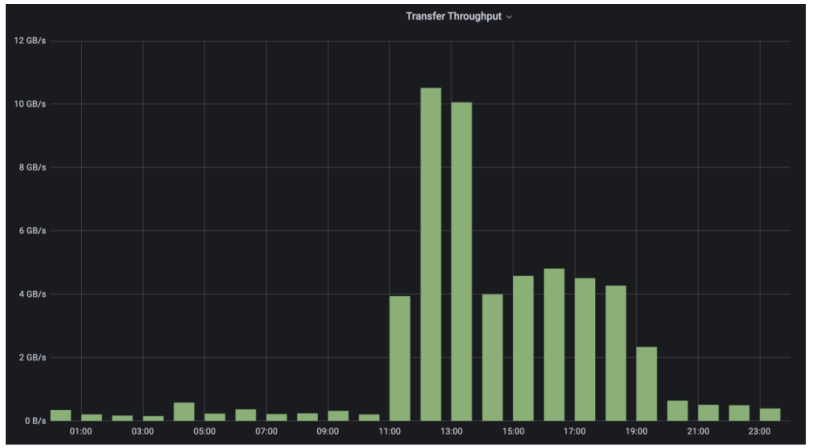


Figure 2 — The traffic levels seen in the network view correspond to those seen by the WLCG File Transfer Service (FTS) visualization tools.

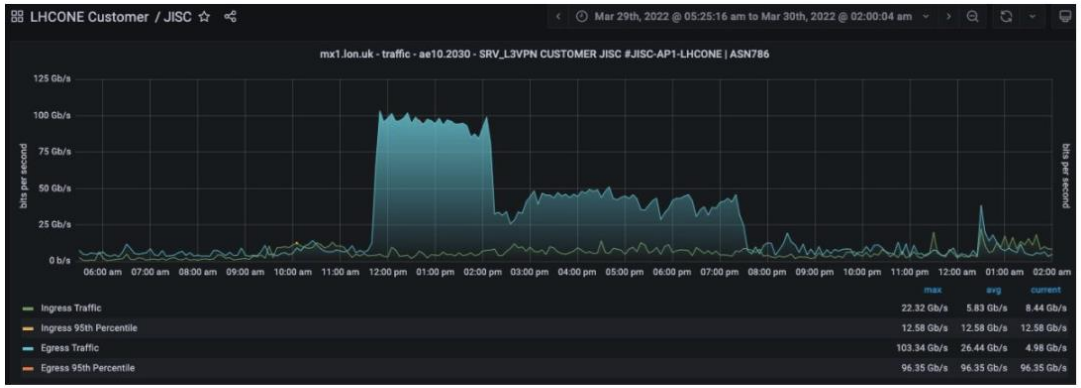


Figure 3 — It was also interesting to see this traffic reflected in the monitoring platform for the GÉANT pan-European research and education backbone network.

Messages to WLCG sites

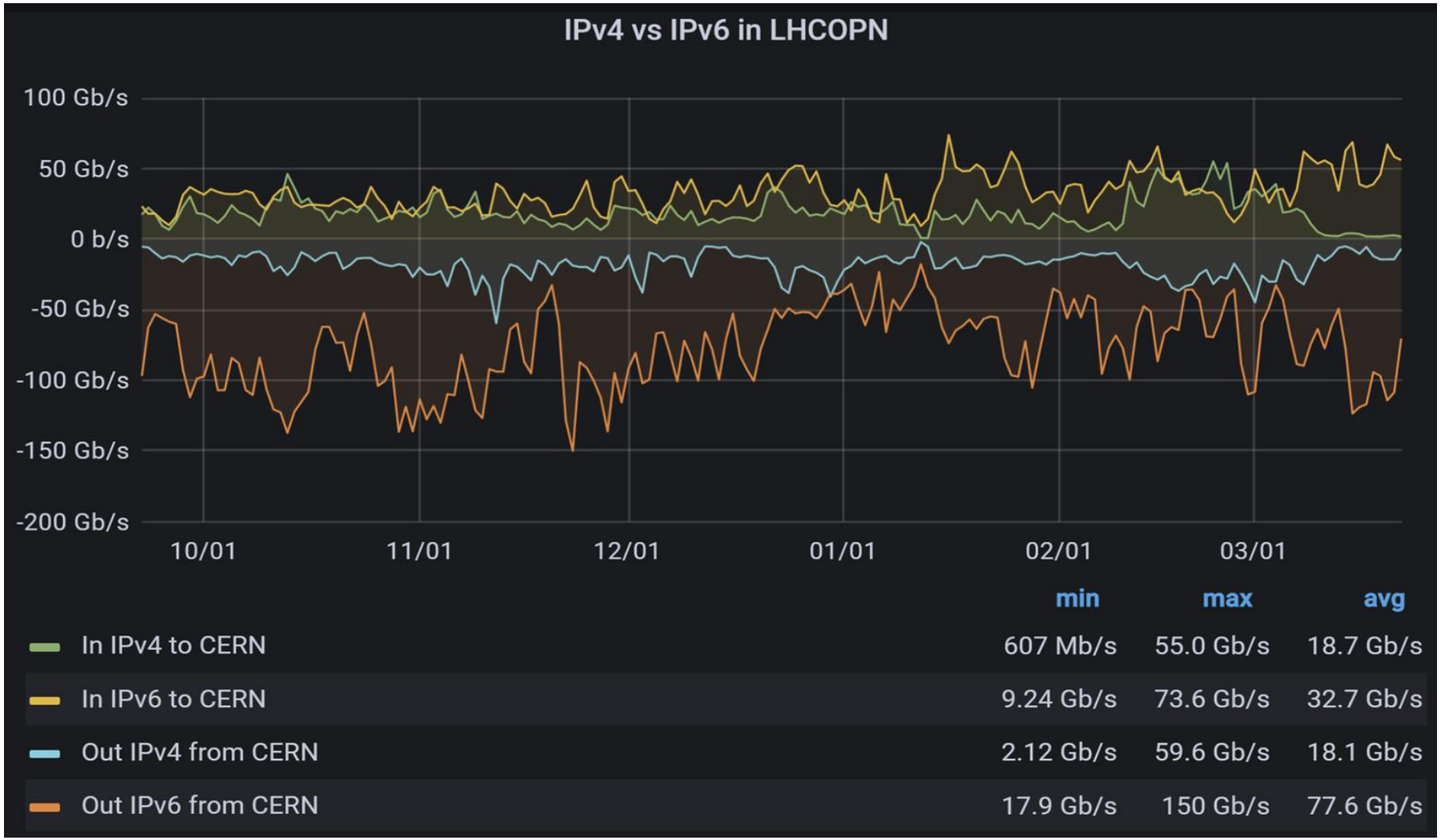
WLCG MB statements (July 2021)

- Deployment of dual stack storage remains the priority
  - this is a prerequisite to fully supporting IPv6-only WNs
- All sites and regions should plan accordingly and as soon as possible
- The final goal is IPv6-only (timetable to be agreed later)

Encouragement: IPv6 WG to all sites and experiments

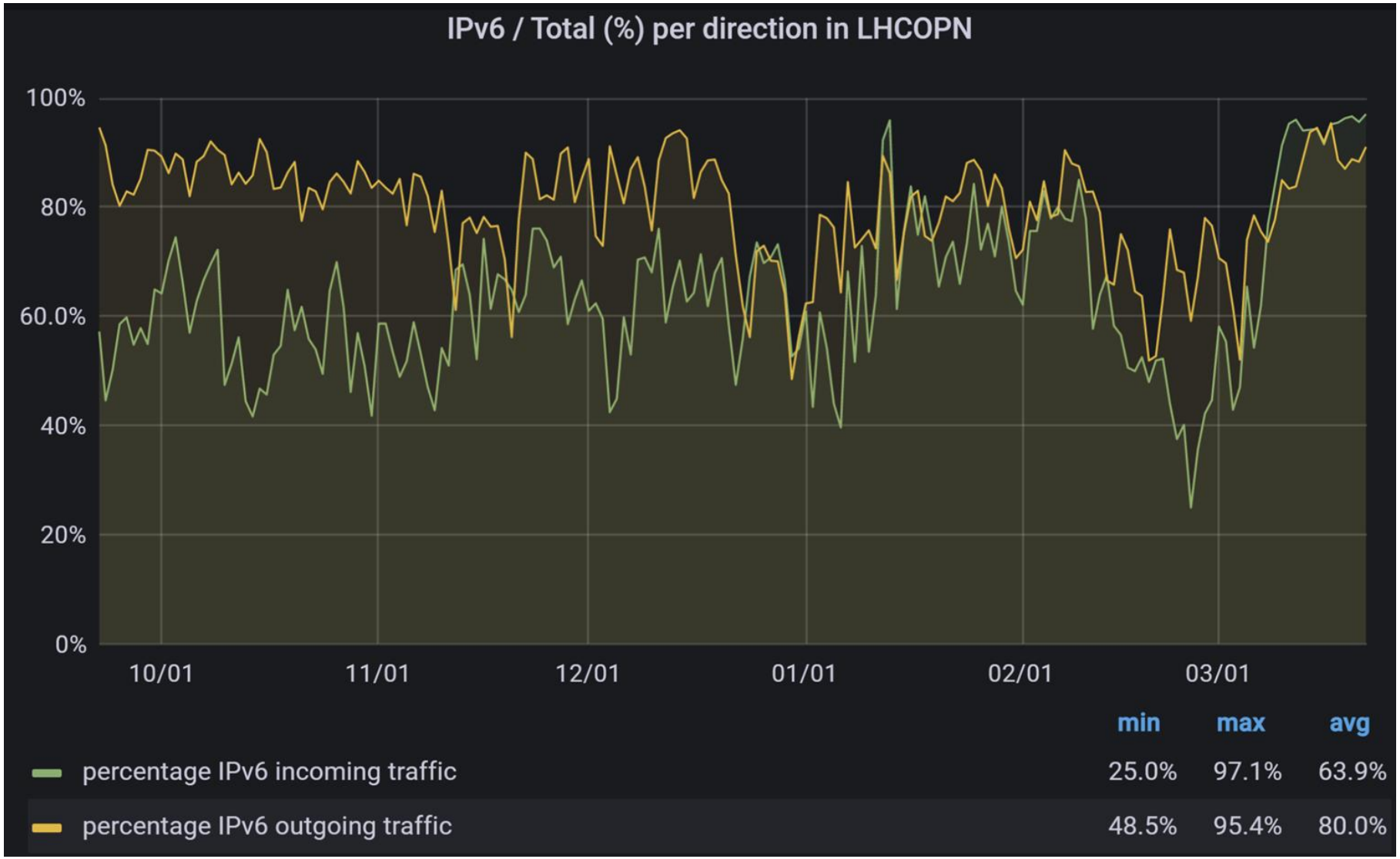
- deploy all WN, VM, containers, local services as dual-stack
- Configure to enable and "prefer" IPv6 transfers

New “fixed” plots from CERN



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

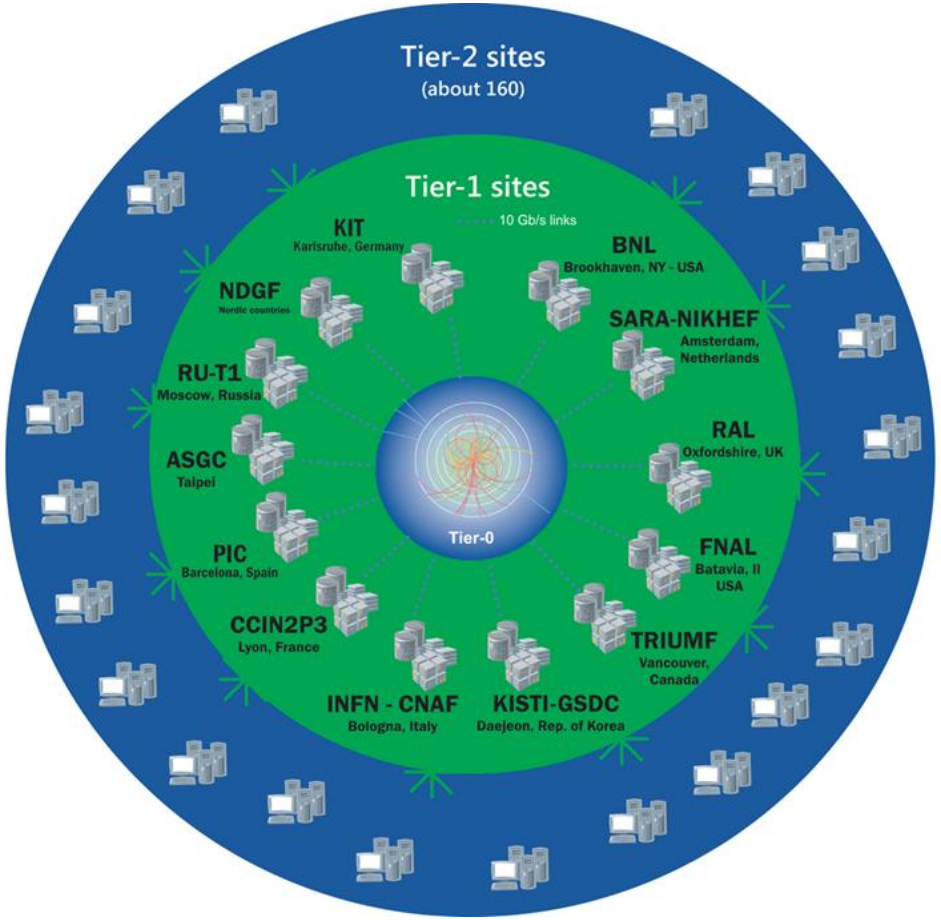
New “fixed” plots from CERN (2)



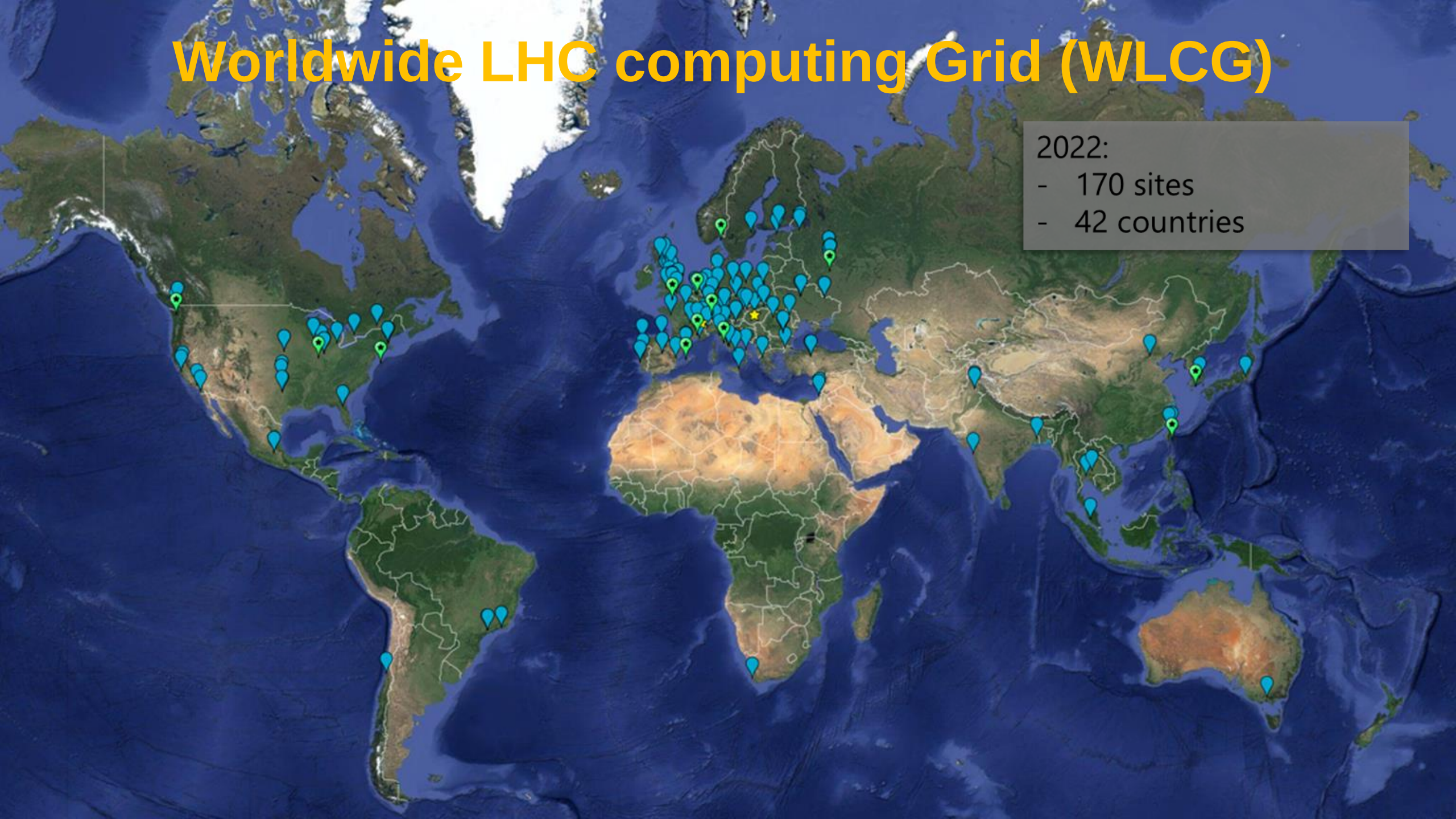
Worldwide LHC Computing Grid (WLCG)

Computing for the CERN Large Hadron Collider

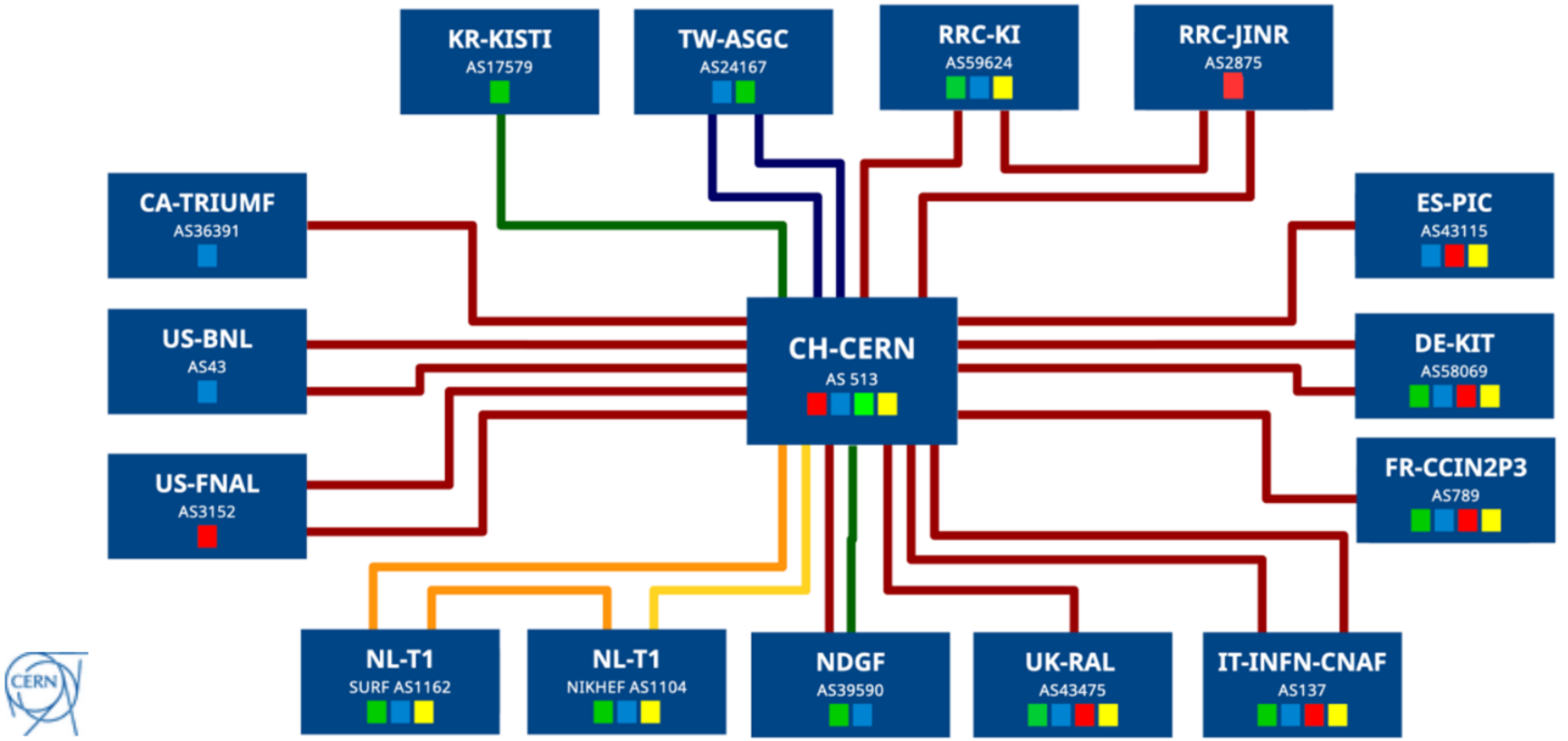
- The WLCG is a global collaboration
  - >170 computing centres, >40 countries
- Its mission is to **store, distribute** and **analyse** the data generated by the LHC experiments
- Sites hierarchically arranged in three tiers:
  - Tier-0 at CERN
  - 14 Tier-1s (national laboratories)
  - ~160 Tier-2s (universities)
- > 1M CPU cores (> 2M jobs per day)
- > 1EB of data storage



Worldwide LHC computing Grid (WLCG)



LHCOPN - Optical Private Network

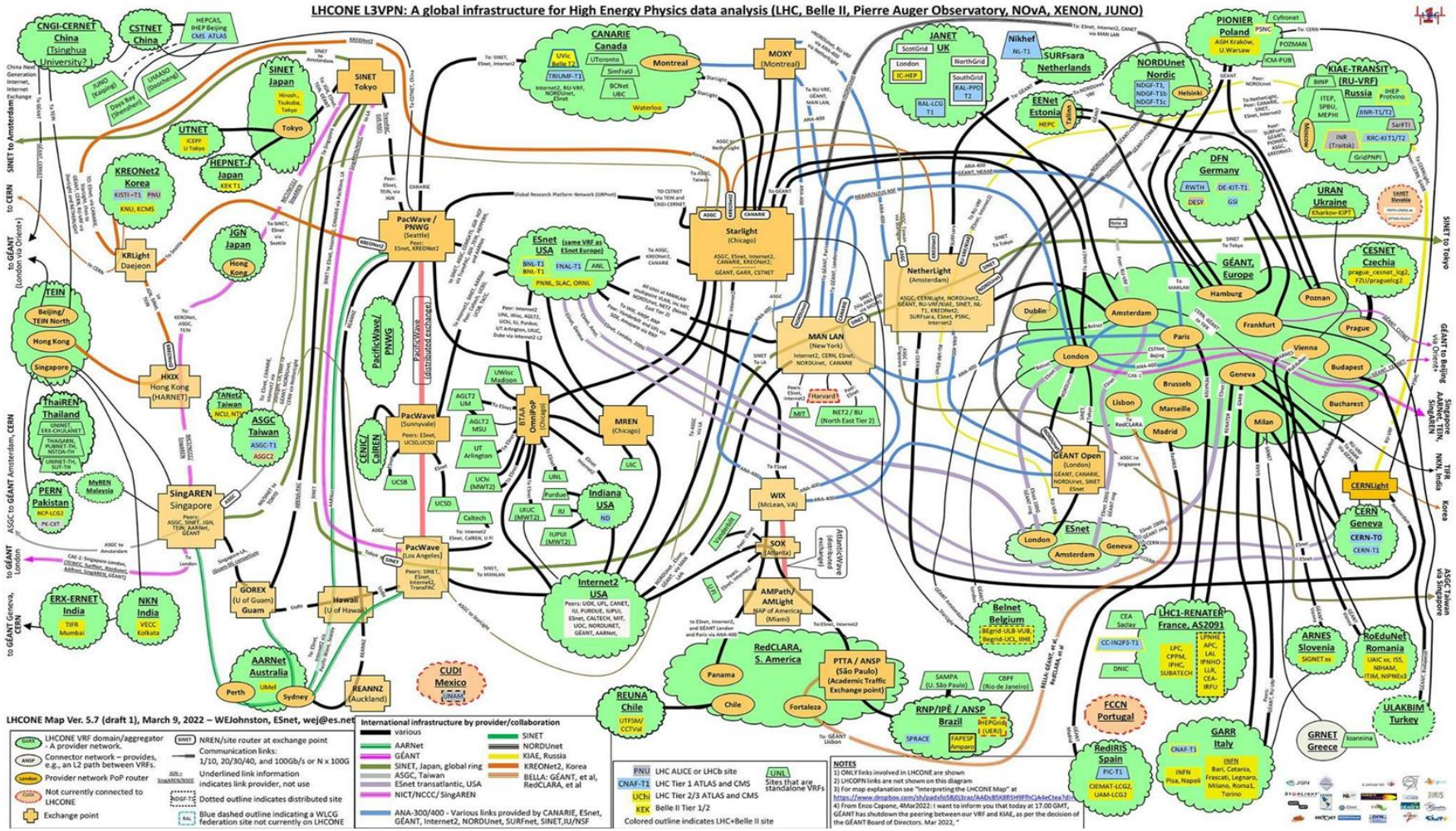


Numbers

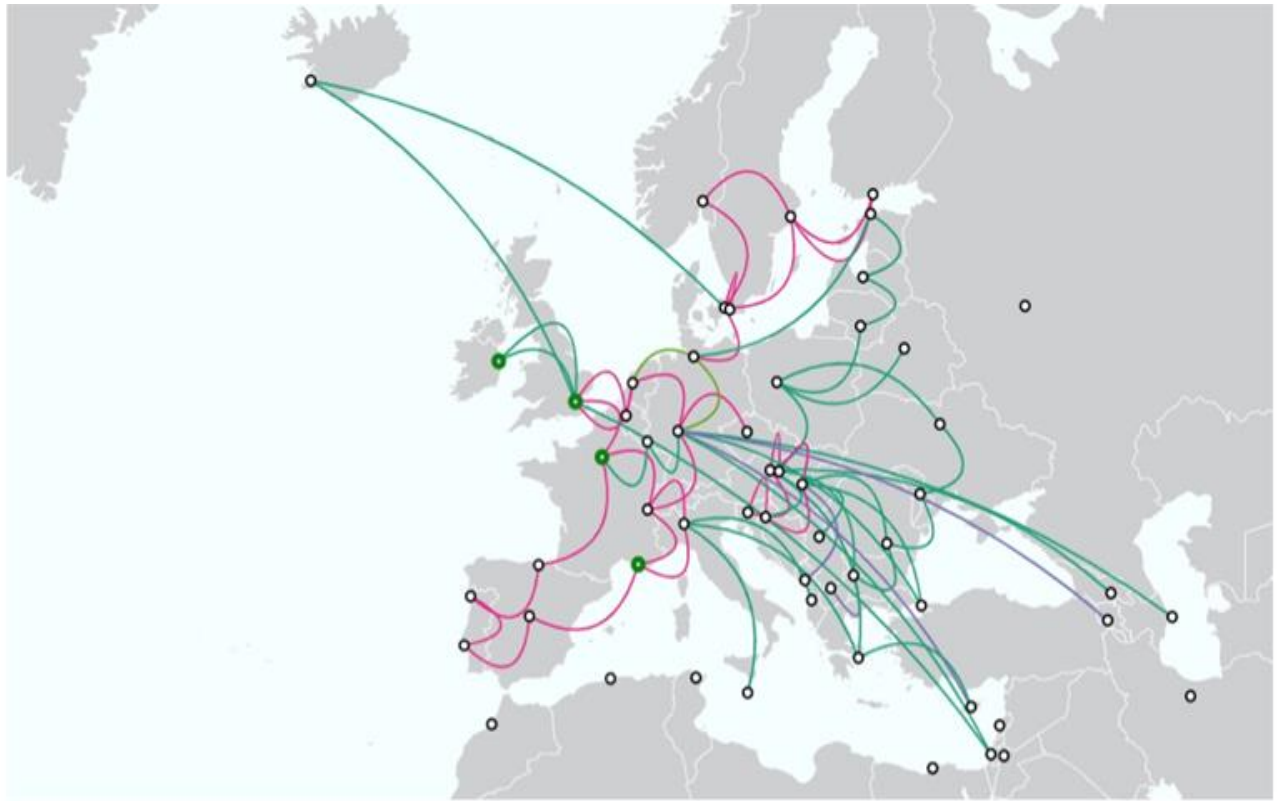
- 14 Tier1s + 1 Tier0
- 12 countries in 3 continents
- Dual stack IPv4-IPv6
- 1.9 Tbps to the Tier0

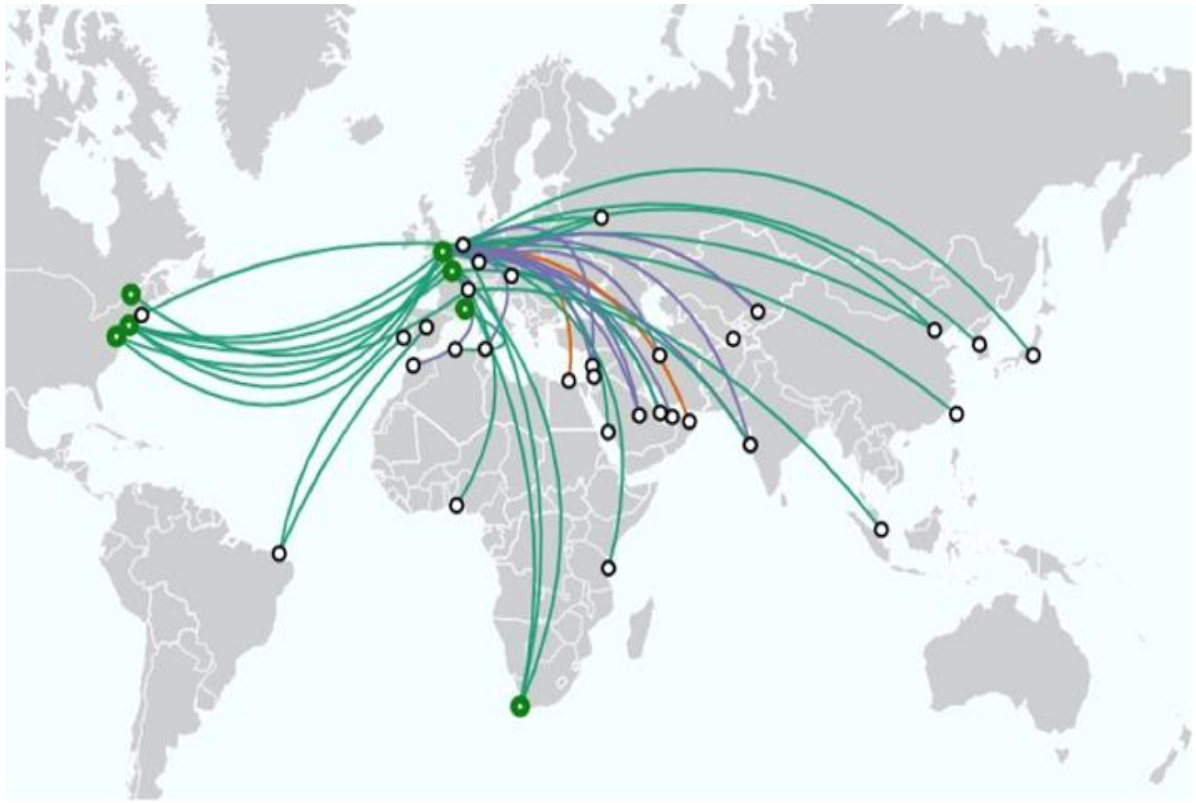
https://twiki.cern.ch/twiki/bin/view/LHCOPN/OverallNetworkMaps

LHCONE map


LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)
LHCONE Map Ver. 5.7 (draft 1), March 9, 2022 – WEJohnston, ESnet, wej@es.net
International infrastructure by provider/collaboration
various
GÉANT
SINET, Japan, global ring
ASGC, Taiwan
ESnet transatlantic, USA
NICT/NCCC/SingAREN
ANA-300/400 - Various links provided by CANARIE, ESnet, GÉANT, Internet2, NORDUnet, SURFnet, SINET,IU/NSF
SINET
AARNet
NORDUnet
KIAE, Russia
KREONet2, Korea
BELLA: GÉANT, et al, RedCLARA, et al
LHCONE VRF domain/aggregator - A provider network.
Connector network – provides, e.g., an L2 path between VRFs.
Provider network PoP router
Not currently connected to LHCONE
Exchange point
NREN/site router at exchange point
Communication links: 1/10, 20/30/40, and 100Gb/s or N x 100G
Underlined link information indicates link provider, not use
Dotted outline indicates distributed site
Blue dashed outline indicating a WLCG federation site not currently on LHCONE
PNU LHC ALICE or LHCb site
CNAF-T1 LHC Tier 1 ATLAS and CMS
UCh LHC Tier 2/3 ATLAS and CMS
KEK Belle II Tier 1/2
Colored outline indicates LHC+Belle II site
UNL Sites that are standalone VRFs
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
4) From Enzo Capone, 4Mar2022: I want to inform you that today at 17.00 GMT, GÉANT has shutdown the peering between our VRF and KIAE, as per the decision of the GÉANT Board of Directors. Mar 2022, "
CNGI-CERNET China (Tsinghua University)
CSTNET China
HEPCAS, IHEP Beijing
CMS ATLAS
JUNO (Kaijing)
LHAASO (Shaozhang)
Diya Bay (Shenzhen)
SINET Japan
SINET Tokyo
UTNET
HEPNET-J Japan
KEK T1
KREONet2 Korea
KISTI-T1
PNU
KNU, KCMS
KRLight Daejeon
IGN Japan
Hong Kong
TEIN
Beijing/ TEIN North
Hong Kong
Singapore
HKIX Hong Kong (HARNET)
ThaiREN Thailand
PERN Pakistan
PK-CIT
MyREN Malaysia
SingAREN Singapore
ERX-ERNET India
TIFR Mumbai
NKN India
VECC Kolkata
TANet2 Taiwan
NCU, NTU
ASGC Taiwan
ASGC-T1
ASGC2
GOREX (U of Guam) Guam
Hawaii (U of Hawaii)
AARNet Australia
UMel
Perth
Sydney
REANNZ (Auckland)
CUDI Mexico
UNAM
CANARIE Canada
UVic Belle T2
UToronto
TRIUMF-T1
SimFraU
BCNet
UBC
Waterloo
Montreal
MOXY (Montreal)
PacWave / PNWG (Seattle)
Peer: ESnet, KREONet2
PacificWave/ PNWG
PacificWave distributed exchange
PacWave (Sunnyvale)
Peers: ESnet, UCSD,UCSC
CENIC/ CalREN
UCSB
UCSD
Caltech
PacWave (Los Angeles)
Peers: SINET, ESnet, Internet2, TransPAC
ESnet USA (same VRF as ESnet Europe)
BNL-T1
FNAL-T1
ANL
PNNL, SLAC, ORNL
Starlight (Chicago)
ASGC, ESnet, Internet2, CANARIE, KREONet2, GÉANT, GARR, CSTNET
MAN LAN (New York)
Internet2, CERN, ESnet, NORDUnet, CANARIE
Harvard
MIT
NET2 / BU (North East Tier 2)
OMniPoP (Chicago)
MREN (Chicago)
UWisc Madison
AGLT2 UM
AGLT2 MSU
UT Arlington
UChi (MWT2)
UNL
Purdue
IU
UIUC (MWT2)
IUPUI (MWT2)
Indiana USA
ND
UIC
WIX (McLean, VA)
SOX (Atlanta)
AtlanticWave (distributed exchange)
AMPath/ AMLight NAP of Americas (Miami)
Internet2 USA
Peers: UOK, UFL, CANET, NJ, PURDUE, IUPUI, ESnet, CALTECH, MIT, UOC, NORDUNET, GÉANT, AARNet,
RedCLARA, S. America
Panama
Chile
Fortaleza
PTTA / ANSP (São Paulo) (Academic Traffic Exchange point)
SAMPA (U. São Paulo)
CBPF (Rio de Janeiro)
RNP/IPÊ / ANSP Brazil
SPRACE
IHEPGrid (UERJ)
FAPESP Amparo
REUNA Chile
UTFSM/ CCTVal
JANET UK
London
IC-HEP
NorthGrid
SouthGrid
RAL-PPD T2
RAL-LCG T1
Nikhef
NL-T1
SURFsara Netherlands
EENet Estonia
HEPC
NORDUnet Nordic
NDGF-T1, NDGF-T1b, NDGF-T1c
Helsinki
PIONIER Poland
PSNC
Cyfronet
AGH Krakow, U.Warsaw
POZMAN
ICM-PUB
KIAE-TRANSIT (RU-VRF) Russia
BINP
ITEP, SPBU, MEPHI
IHEP Protvino
BNR-T1/T2
SarFTI
INR (Troitsk)
RRC-KI T1/T2
GridPNPI
Moscow
DFN Germany
RWTH
DE-KIT-T1
DESY
GSI
URAN Ukraine
Kharkov-KIPT
SANET Slovakia
NetherLight (Amsterdam)
ASGC, CERNLight, NORDUnet, GÉANT, RU-VRF/KIAE, SINET, NL-T1, KREONet2, SURFsara, ESnet, PSNC, Internet2
GÉANT, Europe
Amsterdam
Dublin
Paris
London
Brussels
Lisbon
Marseille
RedCLARA
Madrid
Hamburg
Frankfurt
Geneva
Milan
Poznan
Prague
Vienna
Budapest
Bucharest
CESNET Czechia
prague_cesnet_lcg2
FZU/praguelcg2
GÉANT Open (London)
GÉANT, CANARIE, NORDUnet, SINET, ESnet
ESnet
London
Amsterdam
Geneva
CERNLight
CERN Geneva
CERN-T0
CERN-T1
Belnet Belgium
Belgrid-ULB-VUB, Belgrid-UCL-IIHE
LHC1-RENATER France, AS2091
CEA Saclay
CC-IN2P3-T1
DNIC
LPC, CPPM, IPHC, SUBATECH
LPNHE APC, LAL IPNHO LLR, CEA-IRFU
ARNES Slovenia
SIGNET xx
RoEduNet Romania
UAIC xx, ISS, NIHAM, ITIM, NIPNEx3
ULAKBIM Turkey
FCCN Portugal
RediRIS Spain
PIC-T1
CIEMAT-LCG2, UAM-LCG2
GARR Italy
CNAF-T1
INFN Pisa, Napoli
INFN Bari, Catania, Frascati, Legnaro, Milano, Roma1, Torino
GRNET Greece
Ioannina

HEP Networking - uses GÉANT





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