Science & Technology Facilities Council

UK Research and Innovation

HEPiX

The deployment of IPv6 data storage on WLCG and UK GridPP

David Kelsey

(Head of Particle Physics Computing Group)

STFC Rutherford Appleton Laboratory
- UK Research and Innovation

Talk at UKNOF42, London, 15 Jan 2019

GridPP
UK Computing for Particle Physics

WLCG
Worldwide LHC Computing Grid

What is STFC?

Science and Technology Facilities Council, UK

- One of Europe's largest multi-disciplinary scientific research organisations

Science & Technology Facilities Council

UK Research and Innovation

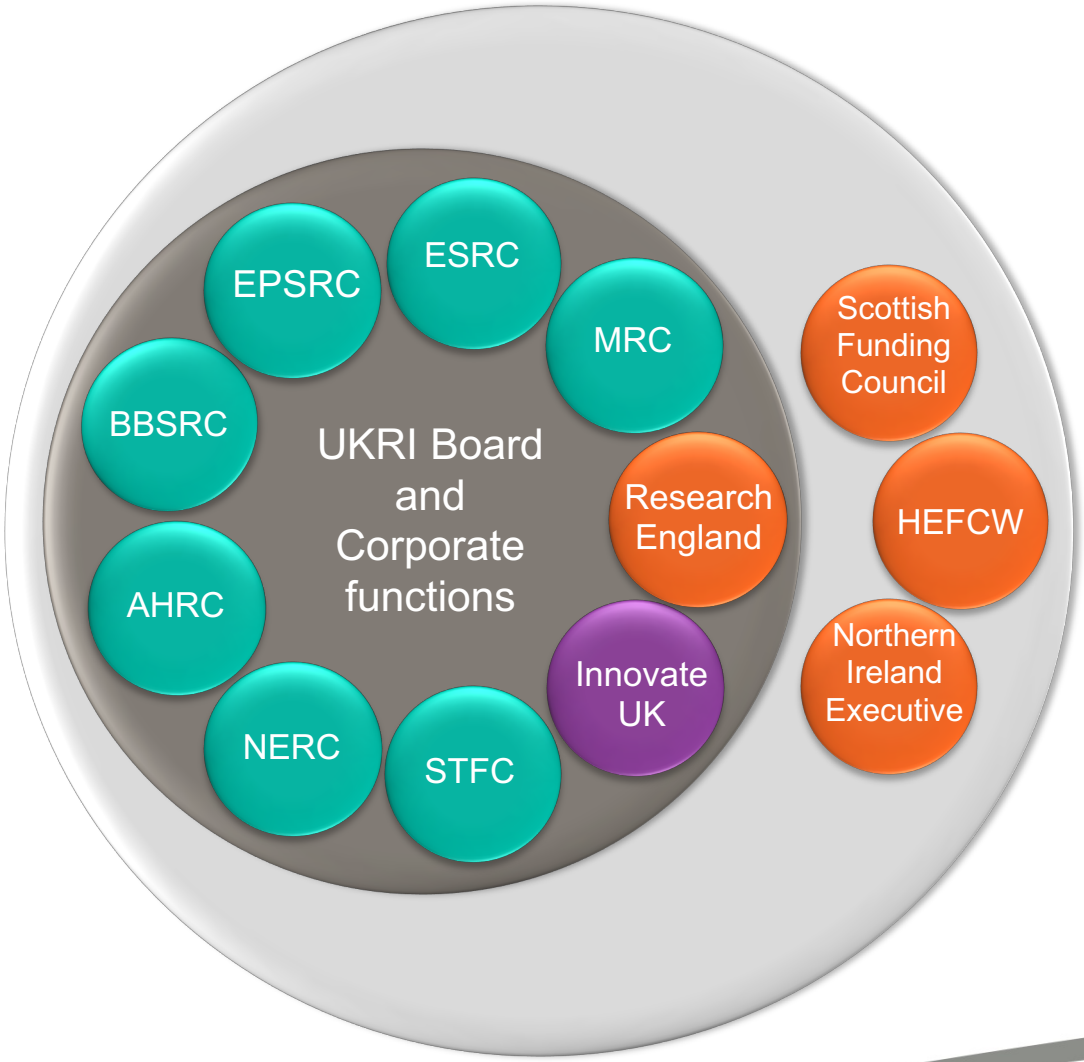
What we do (STFC)



World class research, innovation and skills

- Broad range of physical, life and computational sciences
- In-house scientists in particle and nuclear physics, and astronomy
- Access for 7,500 scientists to world-leading, large-scale facilities
- Science and Innovation Campuses at Daresbury and Harwell
- Globally-recognised capabilities and expertise in technology R&D
- Inspiring young people to undertake STEM

How we're funded



UK Research and Innovation starts 1 April 2018 as the new funding organisation for research and innovation in the UK.

It brings together the seven UK research councils including STFC, Innovate UK and a new organisation, Research England which will work closely with its partner organisations in the devolved administrations.

Where we are (STFC)

UK Astronomy Technology Centre
Edinburgh, Scotland

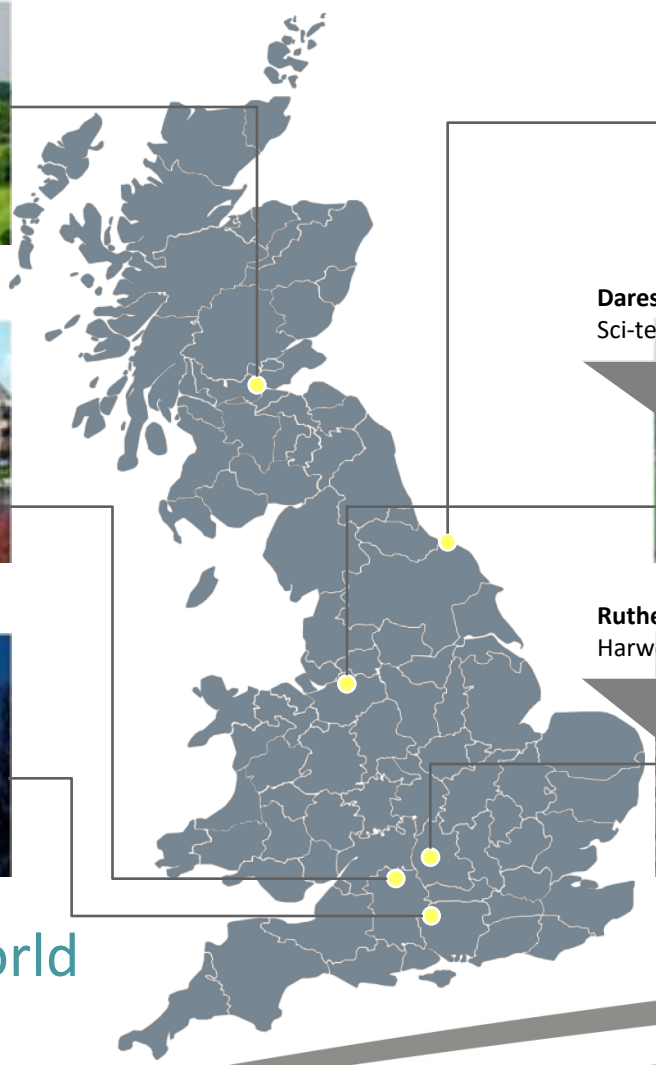
Boulby Underground Laboratory
North Yorkshire

Polaris House
Swindon, Wiltshire

Daresbury Laboratory
Sci-tech Daresbury Campus, Liverpool City Region

Chilbolton Observatory
Stockbridge, Hampshire

Rutherford Appleton Laboratory
Harwell Didcot, Oxfordshire



…and around the world (including CERN)

Science & Technology Facilities Council

UK Research and Innovation

Contents of talk

- CERN Large Hadron Collider, WLCG & GridPP
- HEP networking and data transfers
- Why use IPv6?
- Preparatory work during 2011-2016
- The transition 2016-2020
- Problems & lessons learned
- Summary

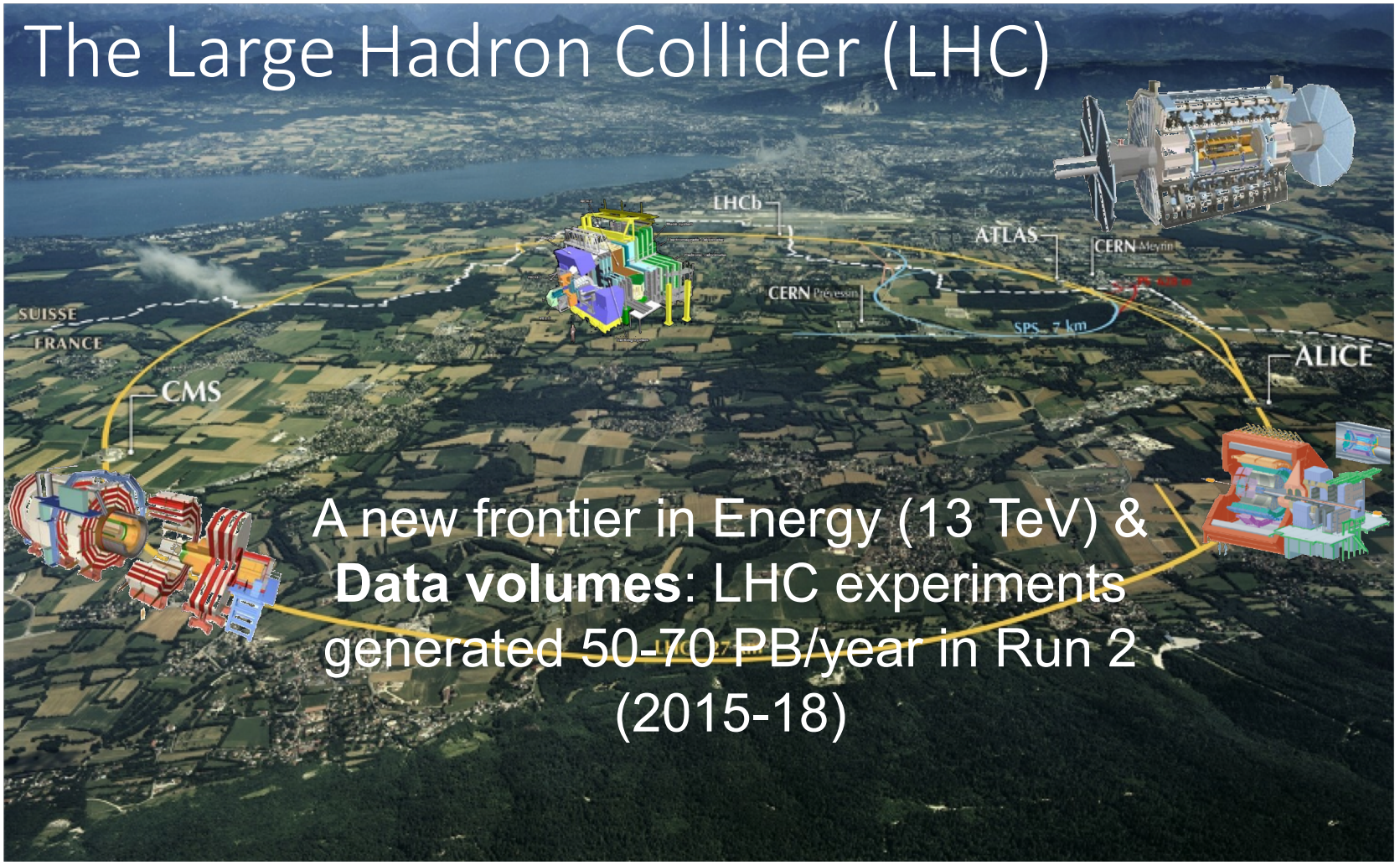
Acknowledgements: All my many colleagues in the HEPiX IPv6 WG, the WLCG IPv6 task force and experts in the Experiments and Sites

David Kelsey

- Experimental particle physicist – moved to IT
- Lead computing group in Particle Physics Dept, STFC-RAL
- Trust, security & identity coordination roles in WLCG, GridPP, EGI, EOSC-hub & AARC2
- Chair of the HEPiX IPv6 Working Group
  - HEPiX is a worldwide body of HEP IT specialists

Large Hadron Collider (LHC) at CERN, WLCG & UK GridPP

The Large Hadron Collider (LHC)



A new frontier in Energy (13 TeV) & **Data volumes**: LHC experiments generated 50-70 PB/year in Run 2 (2015-18)

Physics results (Run1) including…

In 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

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

Worldwide LHC Computing Grid (WLCG)

- 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 with three tiers:
  - Tier-0 at CERN (and Wigner in Hungary)
  - 13 Tier-1s (mainly national laboratories, incl RAL in UK)
  - >150 Tier-2s (generally university physics laboratories)

WLCG Tiers Hierarchy

- **Tier-0 (CERN and Hungary): data recording, reconstruction and distribution**
- **Tier-1: permanent storage, re-processing, analysis**
- **Tier-2: Simulation,**
- **end-user analysis**
- **~750k CPU cores**
- **~ 1 EB storage**
- **> 2 million jobs/day**
- **10-100 Gbps links**

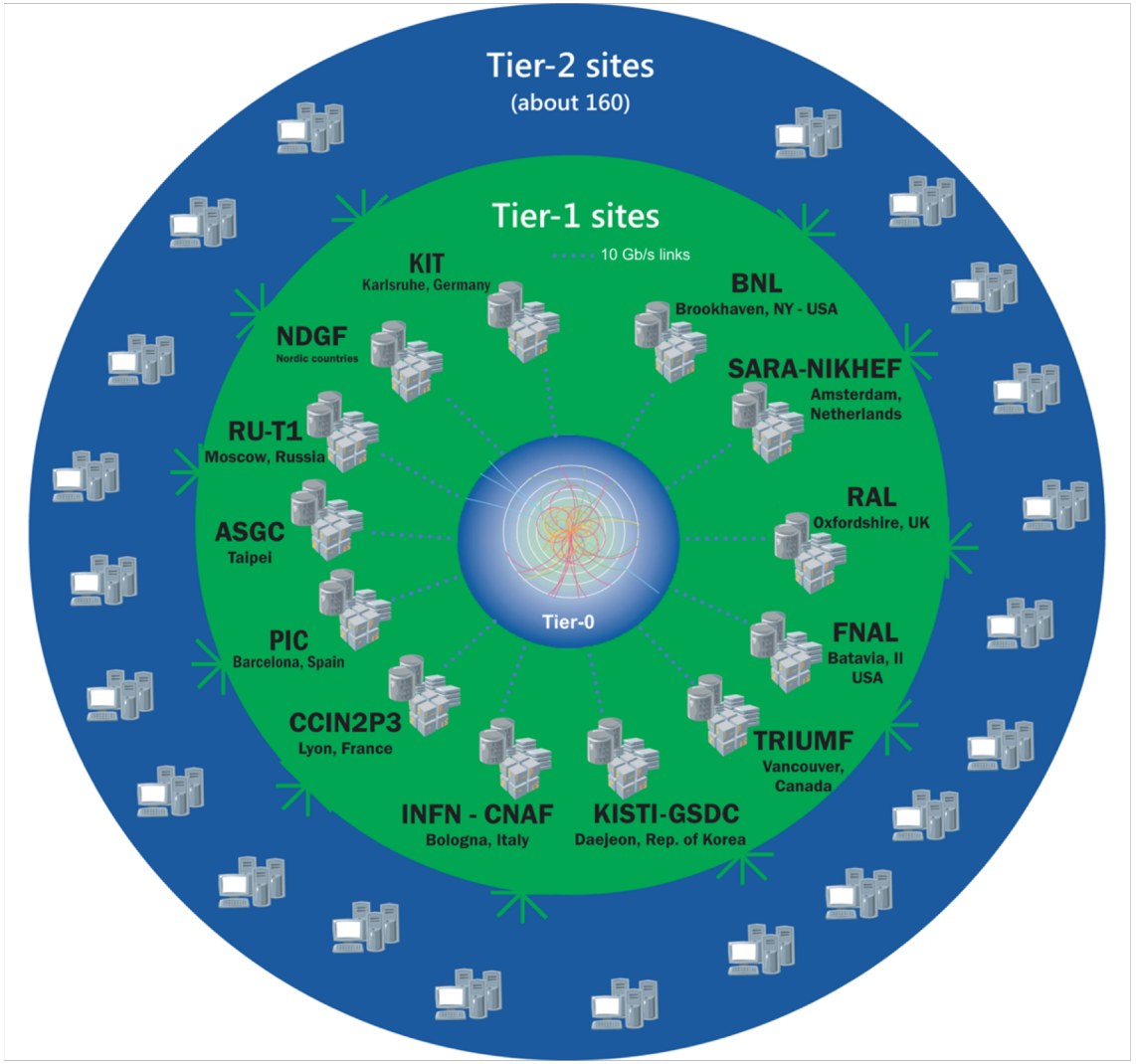
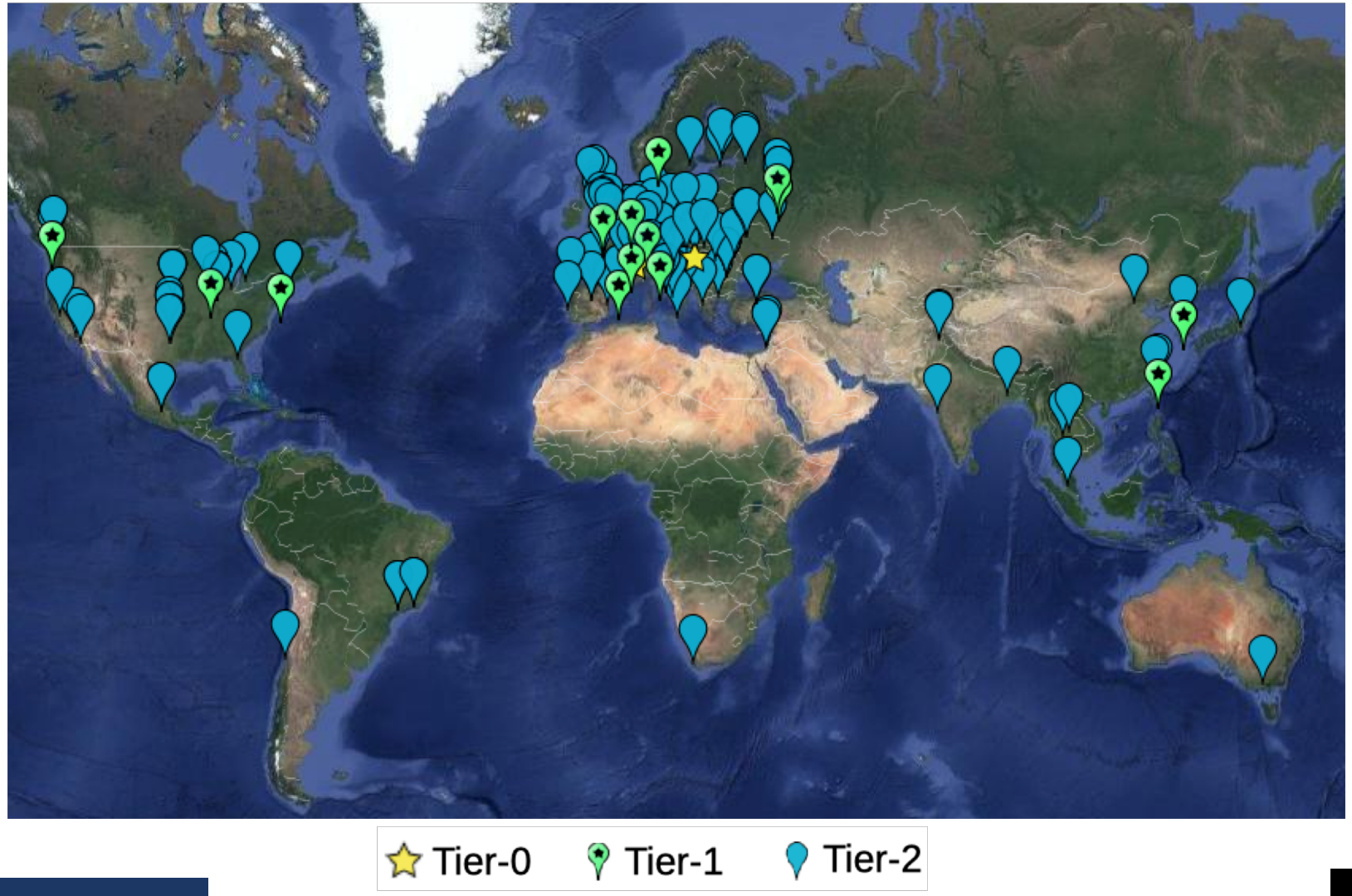


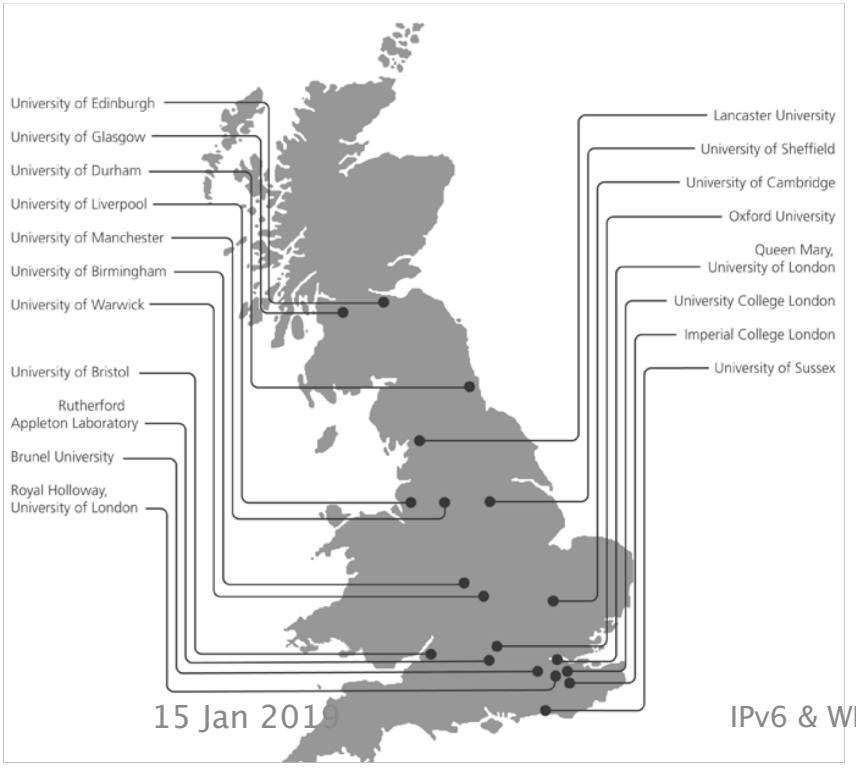
Image from 2014

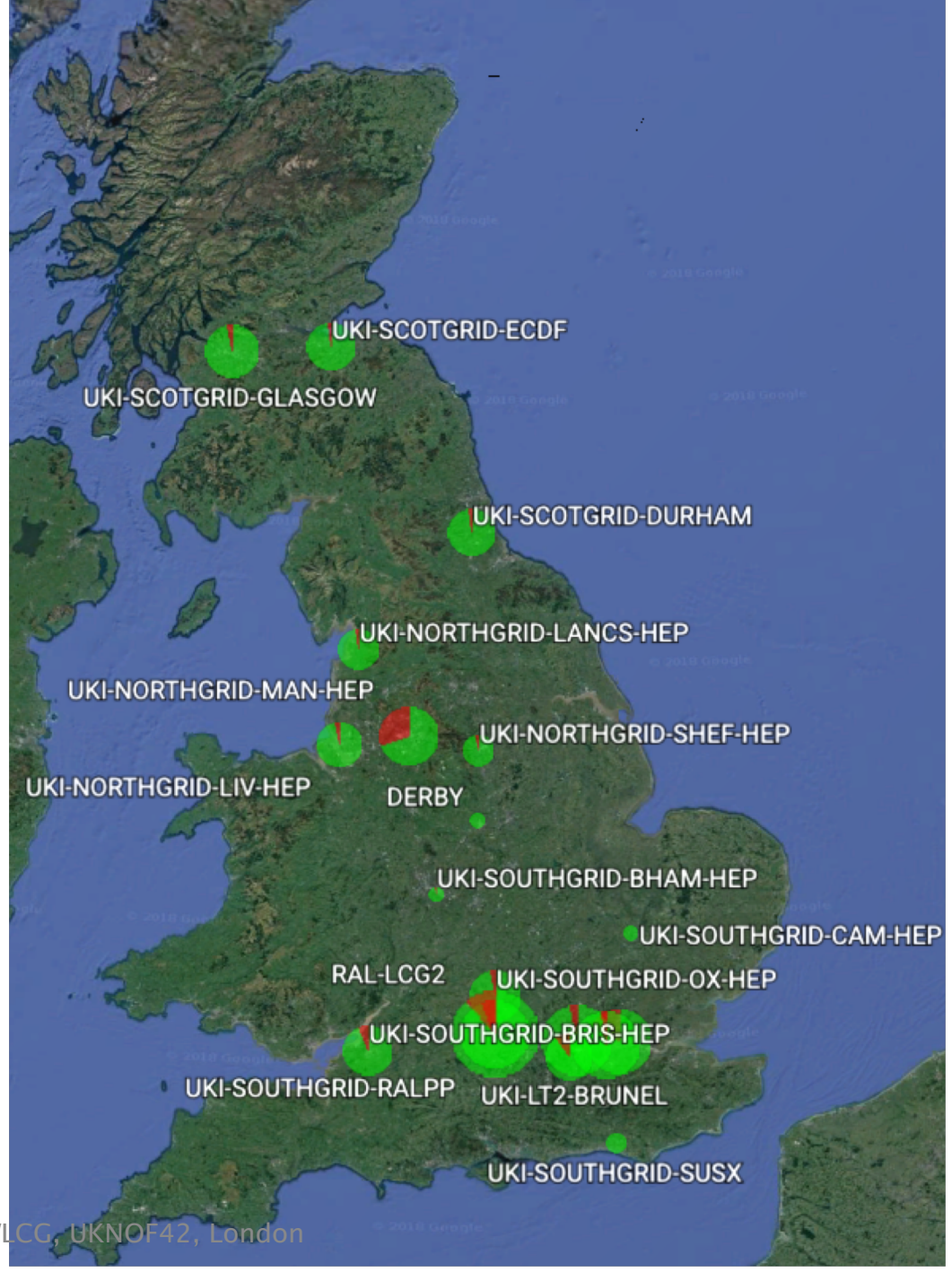
WLCG sites



GridPP
UK Computing for Particle Physics

GridPP is a collaboration of nineteen institutes providing data-intensive distributed computing resources for the UK High Energy Physics community and the UK contribution to the WLCG

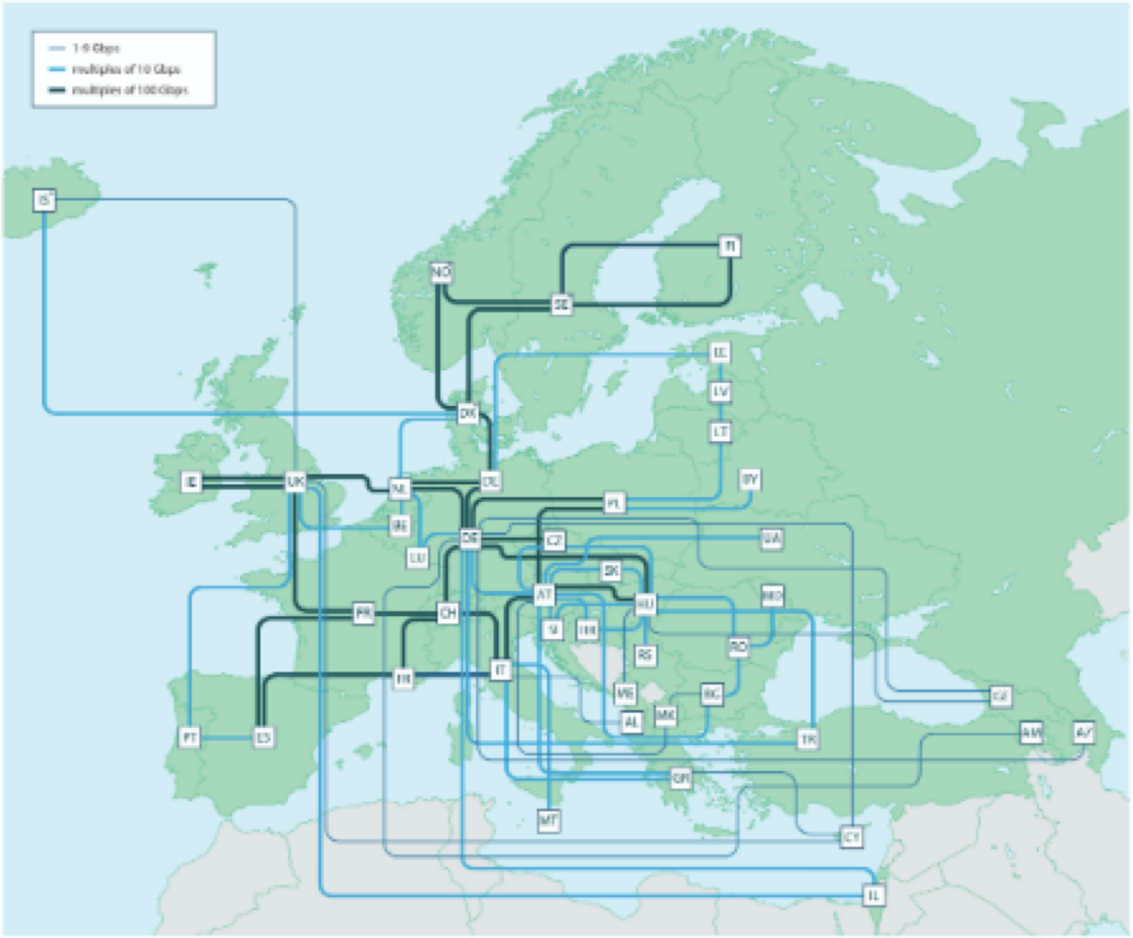




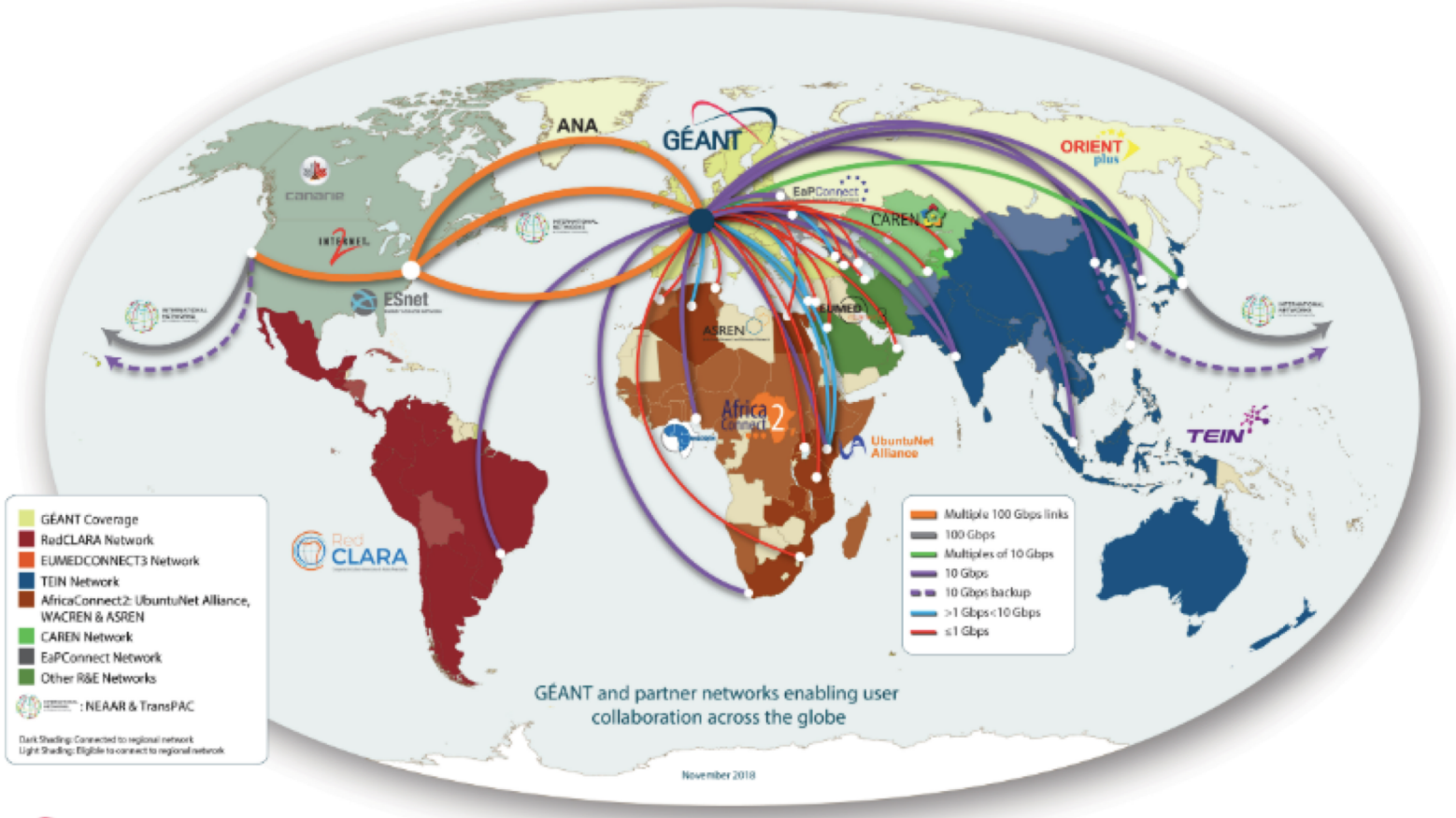
High Energy Physics Networking

GÉANT

GÉANT's pan-European **research and education network** interconnects Europe's National Research and Education Networks (NRENs).
Together we connect over **50 million users** at 10,000 institutions across Europe.



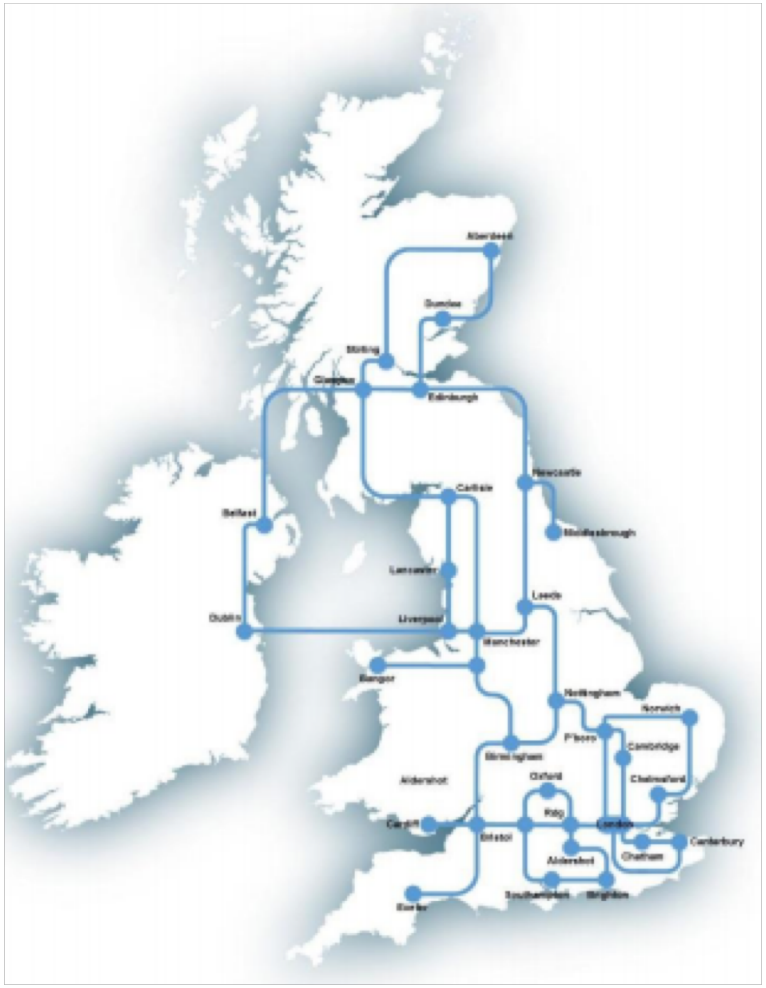
At the Heart of Global Research and Education Networking



GÉANT
Networks · Services · People
www.geant.org

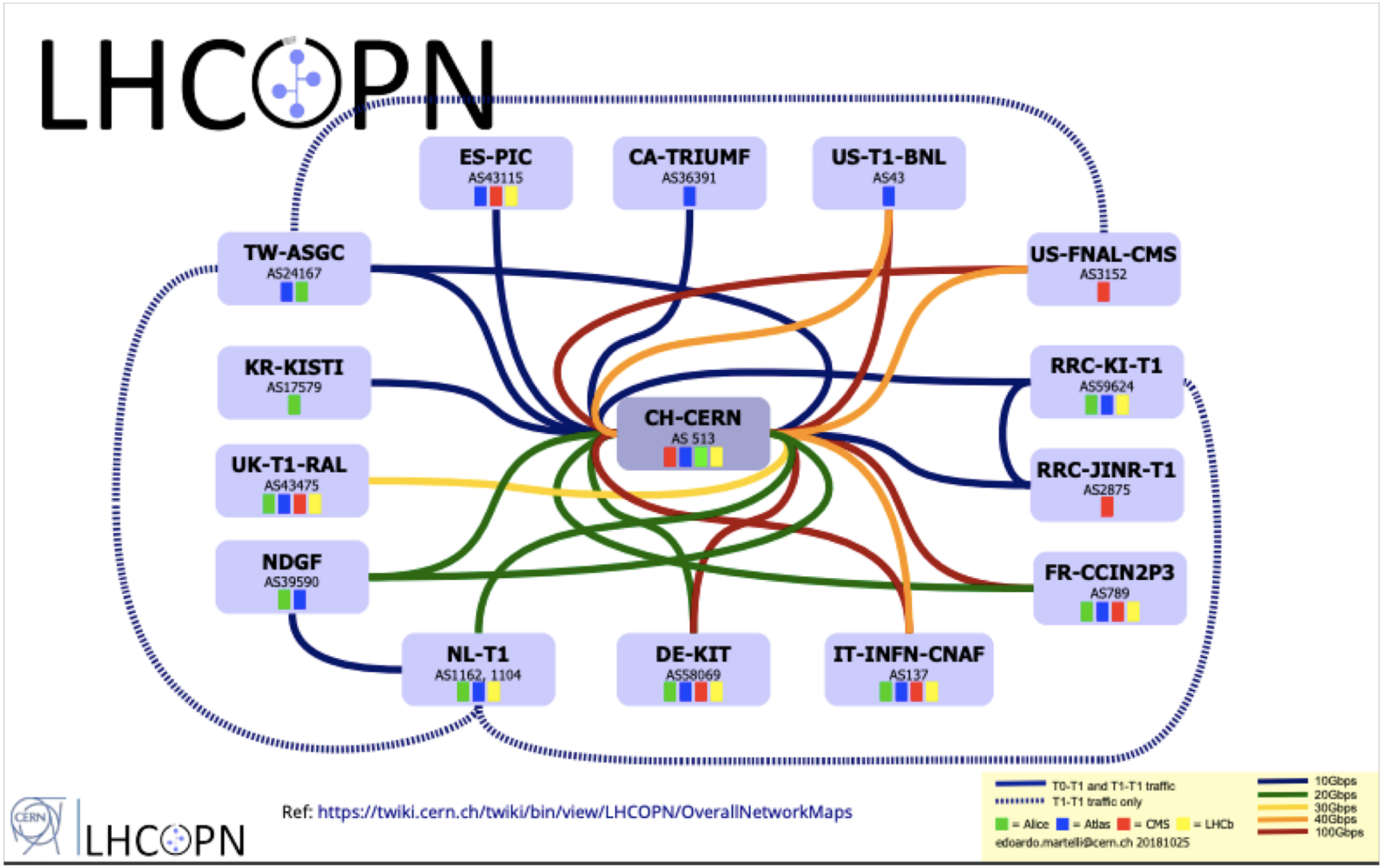
This document is produced as part of the GÉANT Specific Grant Agreement GN4-2 (No. 731122,, that has received funding from the European Union's 2020 research and innovation programme under the GÉANT2020 Framework Partnership Agreement (No.653965).
In addition to GN4-2, the following projects have received funding from the European Union. AfricaConnect2, CAREN2, AsiaConnect (DG DEVCO), EaPConnect and EUMEDCONNECT3 (DG NEAR).
The content of this document is the sole responsibility of GÉANT and can under no circumstances be regarded as reflecting the position of the European Union.
Funded by the European Union

UK – JANET network (Jisc)

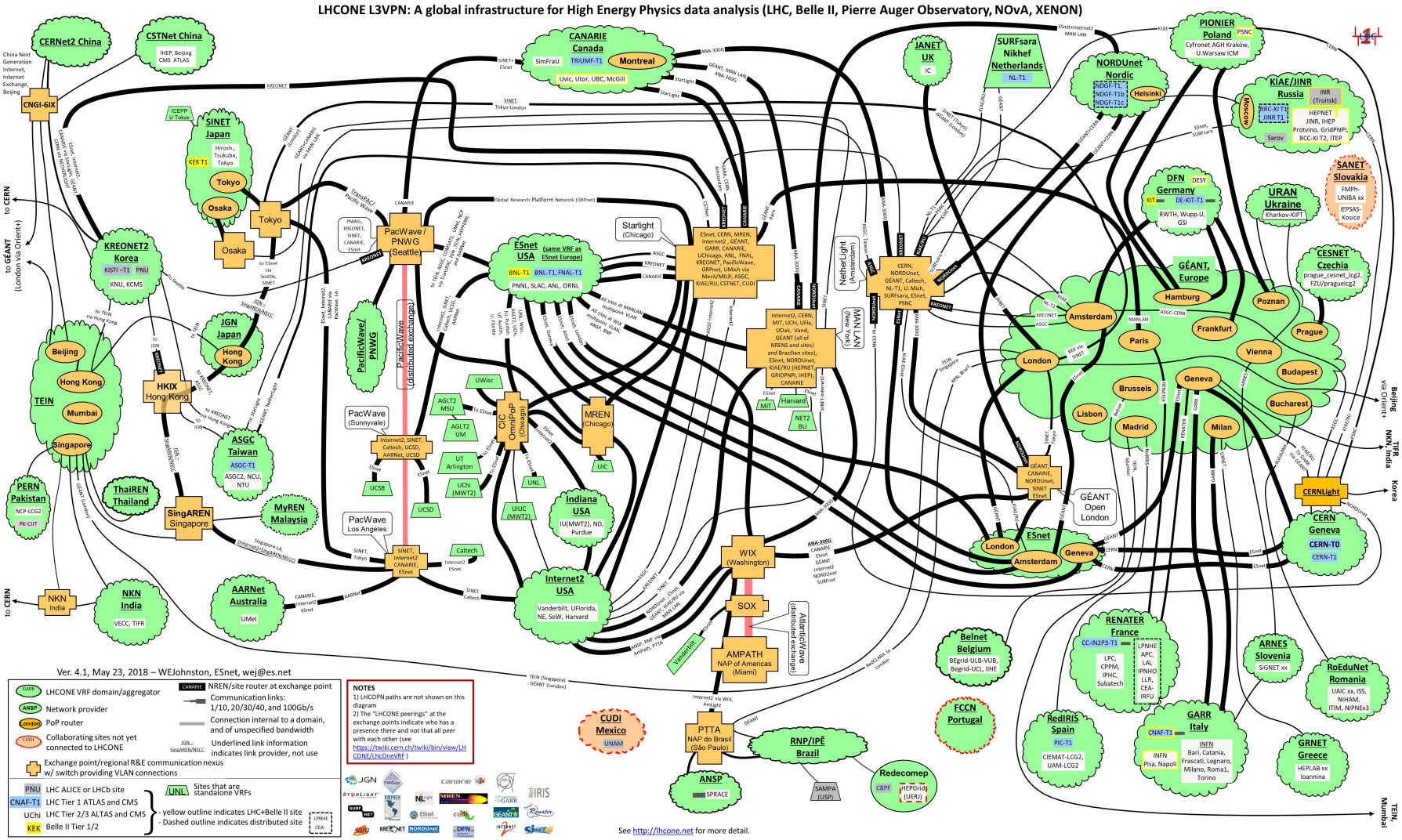


LHCOPN – optical private net

Connect Tier-0 and Tier-1s



LHCONE – L3VPN



WLCG Data Transfers

Data transfers in WLCG

- From Tier-0 to Tier-1s
- From Tier-1s to Tier-2s
- Requirements – Fast and reliable!
- Multiple protocols and implementations, but the standard approach is:

FTS3 and GridFTP

Bulk data transferred between storage clusters with the File Transfer Service (FTS3) using GridFTP

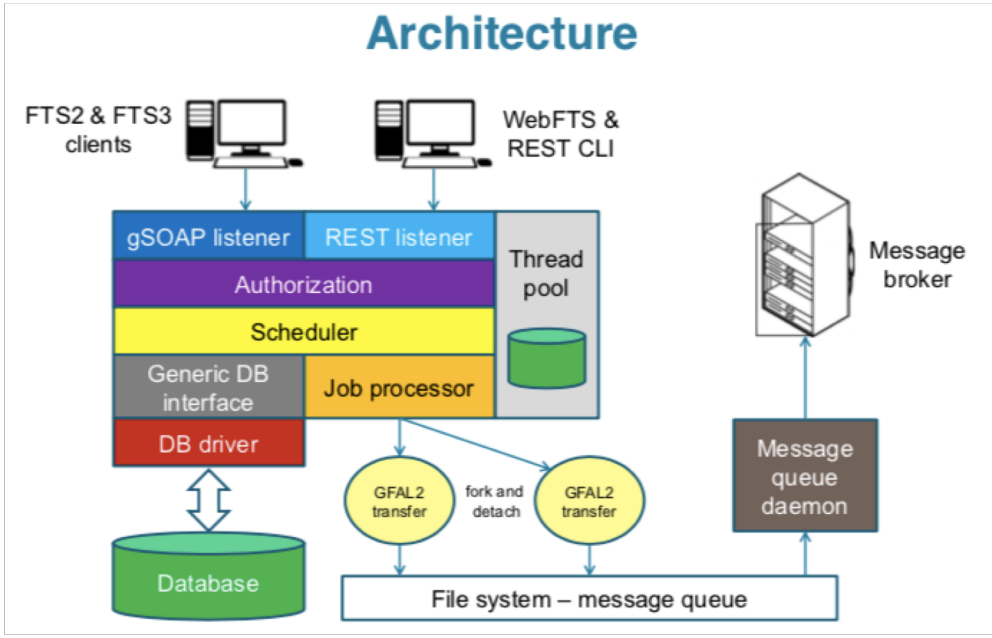
- Also data transfer from federated data storage using a HEP-specific protocol called XrootD
- direct access to data by an analysis job at one site from storage at another

Globus GridFTP

- High-performance, reliable, optimized for high-bandwidth WANs
- Based on FTP protocol
  - with extensions for high-performance operation and security
- Standardized through Open Grid Forum (OGF)
- Implementation provided by the Globus Alliance
- Performance
  - Parallel TCP streams, optimal TCP buffer
- - Non TCP protocol such as UDT (reliable UDP)
- Cluster-to-cluster data movement
- Multicasting, Overlay routing
- Multiple security options
  - Anonymous, password, SSH, GSI
- Support for reliable and re-startable transfers

File Transfer Service (FTS3)

- Powerful and reliable file transfer service
- Supports multiple protocols, standard API
- zero configuration
- web monitoring
- web interface
- use of federated IDs
- 3rd-party transfers:
  - source and destination can both be remote data centers



Why should WLCG use IPv6?

Why IPv6?

- Survey of 18 major HEP sites (Sep 2010) – IPv6 readiness
  - National NRENs ready, Universities and Labs not ready
  - Some reported lack of IPv4 address space, including CERN
- HEPiX meeting – Cornell, Ithaca NY – Nov 2010
  - Projected IANA IPv4 address exhaustion
  - Sep 2010 – memo from US Federal CIO to all Exec depts (incl DOE)
- Offers of opportunistic CPU resources which could be IPv6-only
  - Experiments want to be able to use them
- Recognition that much of our middleware, software and technology was not yet IPv6 capable
- HEPiX decided to create a working group (started April 2011)
  - No specific funding – but motivated, competent volunteers!

Preparatory work during 2011-2016

HEPiX IPv6 Working Group

- Phase 1 – full analysis of work to be done
  - Applications, system and network tools, operational security
  - Create and operate a distributed test-bed
    - No interference with WLCG production data analysis!
  - Propose timetable and plan for transition

2012

- CERN announces shortages of routable IPv4 addresses
  - explosion of virtualisation
- Active HEPiX IPv6 test-bed with ~ 12 sites
  - engagement of all 4 LHC experiments
- Testing regular data transfers across the testbed
- Testing dual-stack services (production) at Imperial College London
- Concluded not able to support IPv6-only clients until at least 2014

At CHEP2013 conference

- > 2 PB data transferred over IPv6 in last 6 months
- Success rate > 87%
- Very High!

GridFTP IPv6 data transfer mesh

2013-14 Data Management

- Testing the important data transfer protocols, technology and data storage/file systems
  - For IPv6-readiness
- GridFTP, 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!

2015

- At CHEP conference in April 2015
  - 75% of Tier-1 sites are IPv6-ready (but only 20% of Tier2)
  - 10% of sites now reporting lack of IPv4 addresses
- Most important IPv6-only use case
  - Sites, Clouds providing CPU (virtual machines)
  - Opportunistic resources may be IPv6-only
  - Need dual-stack federated storage services
  - And dual-stack central WLCG and Experiment services

The transition 2016-2020

2016

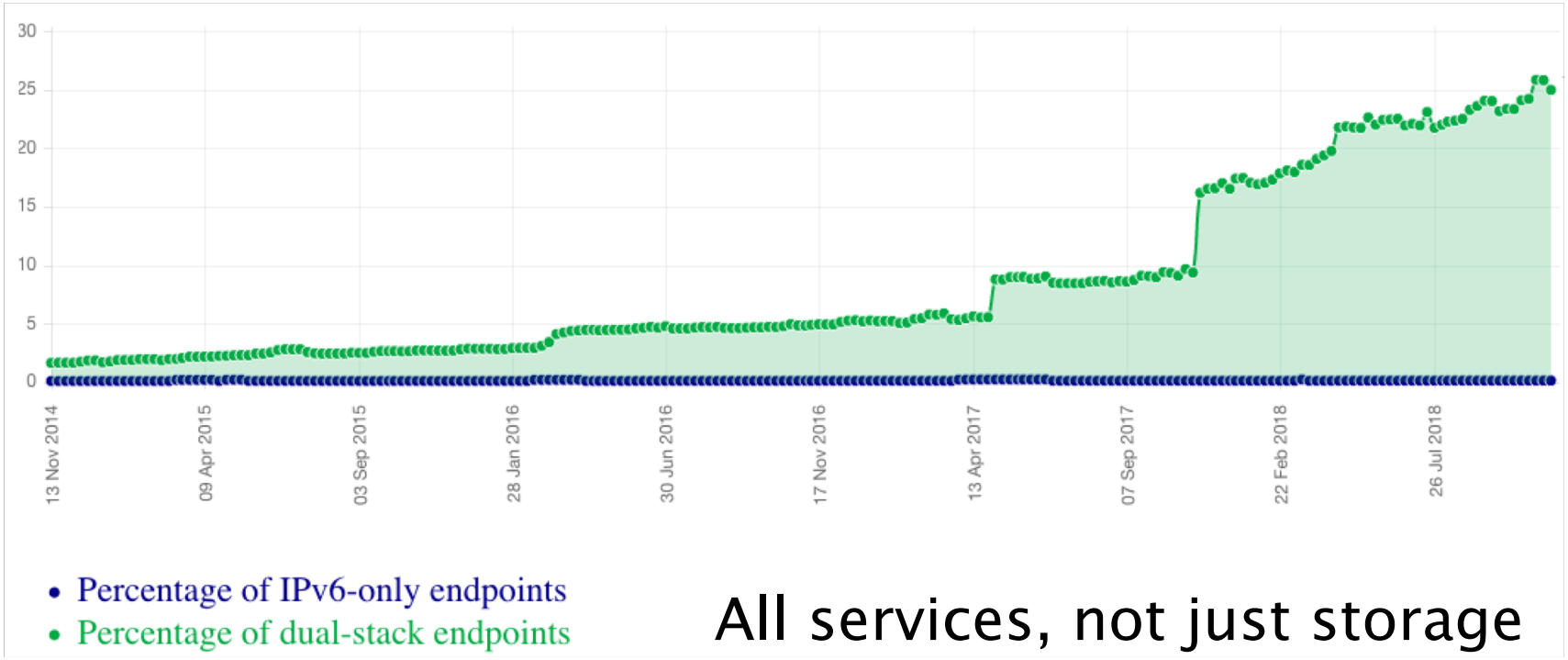
- Continue to push for
  - deployment of production dual-stack data services
  - LHCOPN (Tier0-Tier1 private network)
    - IPv6 peering everywhere
- perfSONAR – end to end network monitoring – dual-stack
- Move central services and central monitoring to IPv6
- Wrote guidance on IPv6 security for WLCG sites
- Deployment timetable approved by WLCG Management Board (Sep 2016)

WLCG – IPv6 deployment

Plan approved by WLCG Management Board

- April 2017 – support for IPv6-only CPU starts
  - Tier-1s to provide dual-stack storage (in testbed)
- April 2018
  - Tier-1 dual-stack storage in production mode
- By end of LHC Run 2 (end 2018)
  - A large number of Tier-2s to migrate storage to IPv6
  - All requested to do this

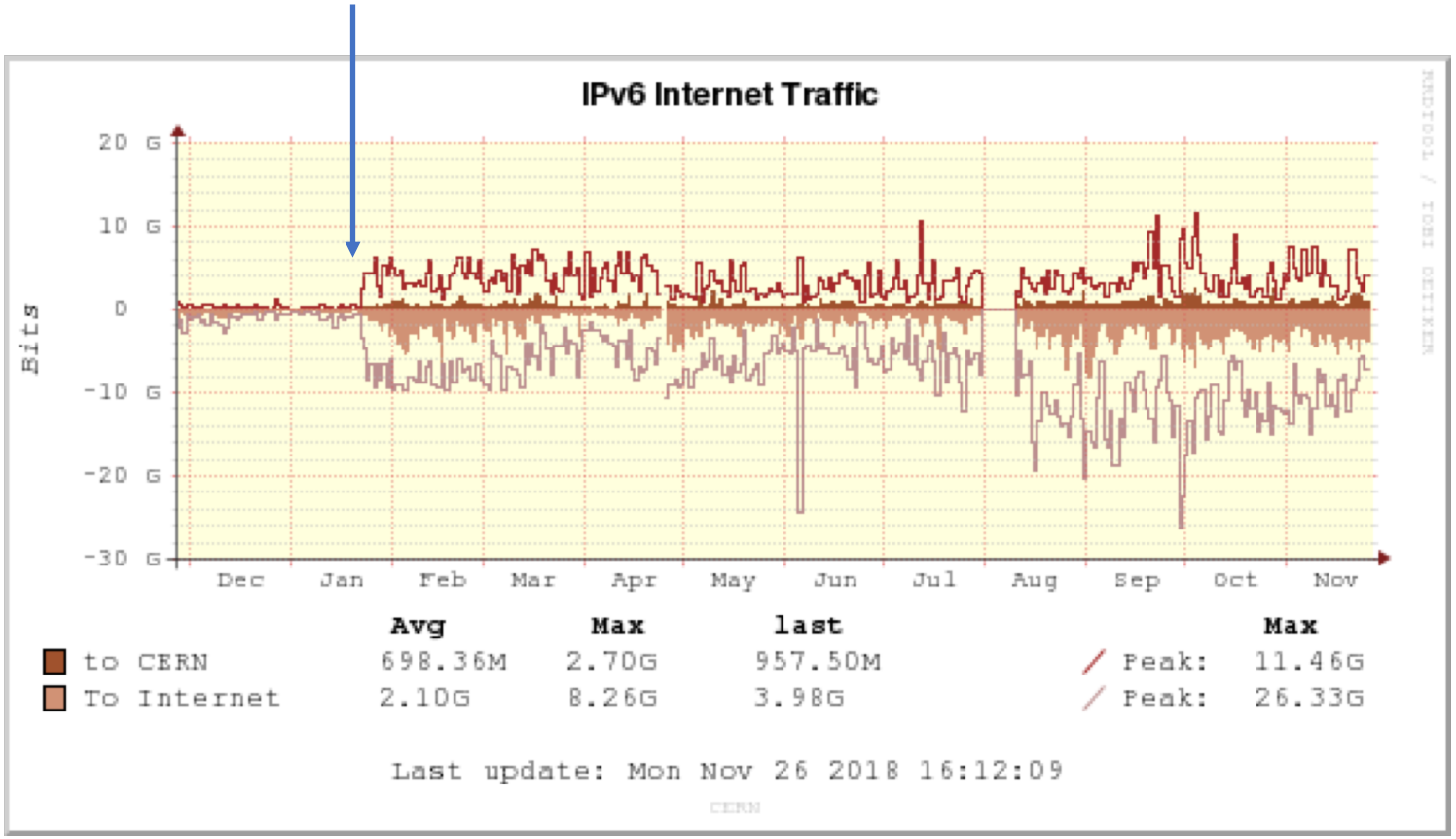
Growth of dual-stack hosts in the WLCG



Fraction of endpoints listed in the CERN central BDII (lcg-bdii.cern.ch) where the DNS returns a dual-stack IPv6-IPv4 (A+AAAA) resolution (green line) or an IPv6-only resolution (blue line). (http://orsone.mi.infn.it/~prelz/ipv6_bdii/).

Turning on IPv6 on CERN Tier-0 disk storage (EOS) in Jan 2018

Non-LHCOPN/non-LHCONE traffic



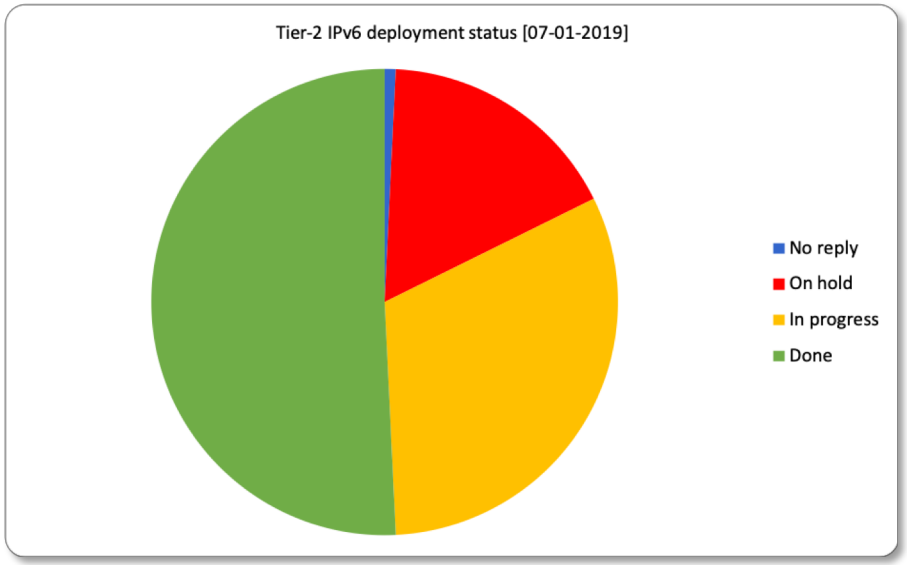
Tier-1 and Tier-2 transition tracking

- 11 Tier-1s are now IPv6 capable
- All have dual-stack storage except for 3
  - To be fixed in 1Q2019

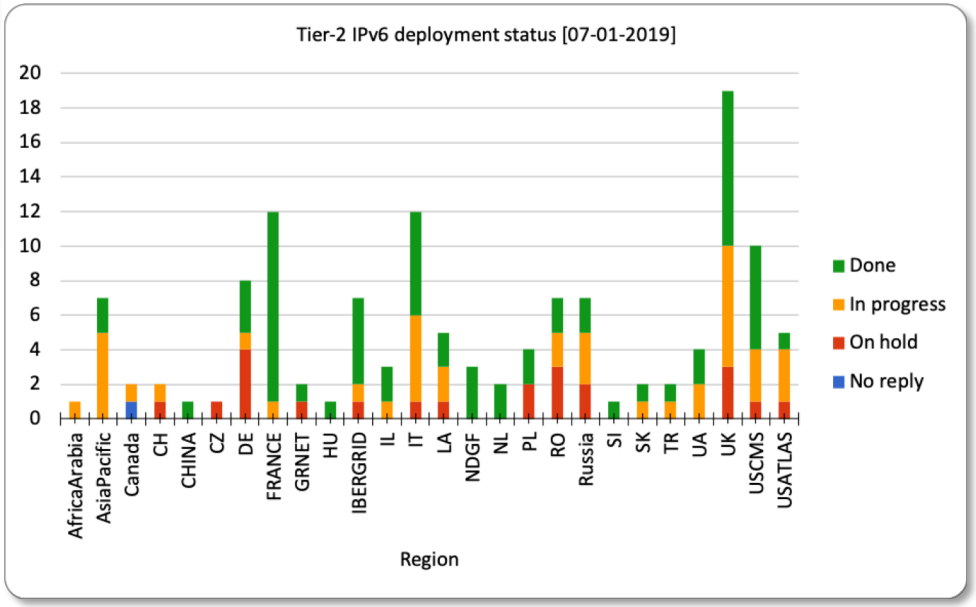
- 115 Tier-2 sites requested to deploy dual-stack perfSONAR and storage by end of Run 2 (end of 2018)
  - USA taking care of their sites
- Follow up with assistance, checking deployment etc
- Largest blocker:
  - Sites waiting for campus infrastructure to be IPv6-ready

Tier-2s: Current status - 130 Tier-2 sites (Jan 2019)

>50% of Tier-2s now with dual-stack perfSONAR and storage



https://twiki.cern.ch/twiki/bin/view/LCG/WlcgIpv6#WLCG_Tier_2_IPv6_deployment_stat

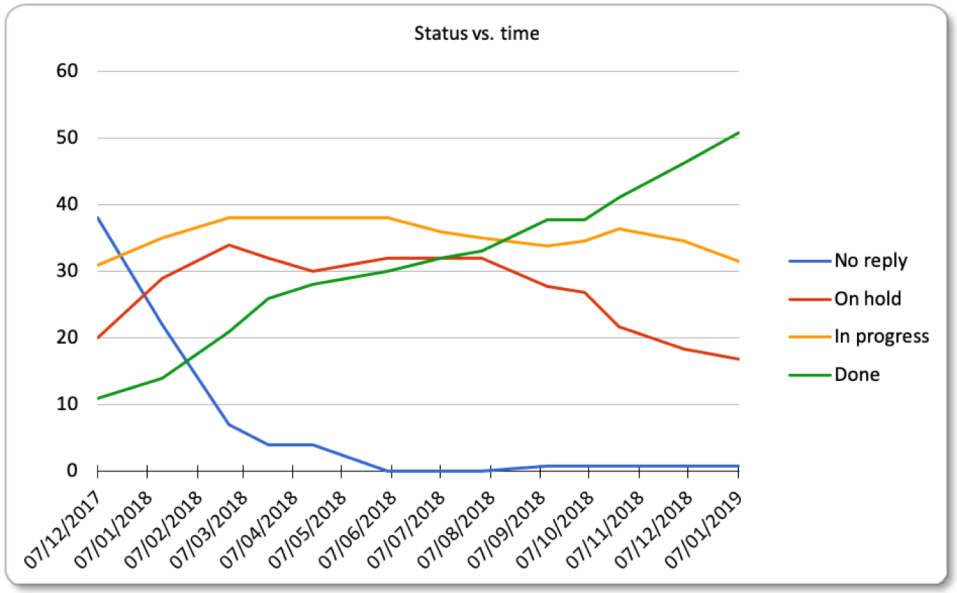


Storage accessible now over IPv6

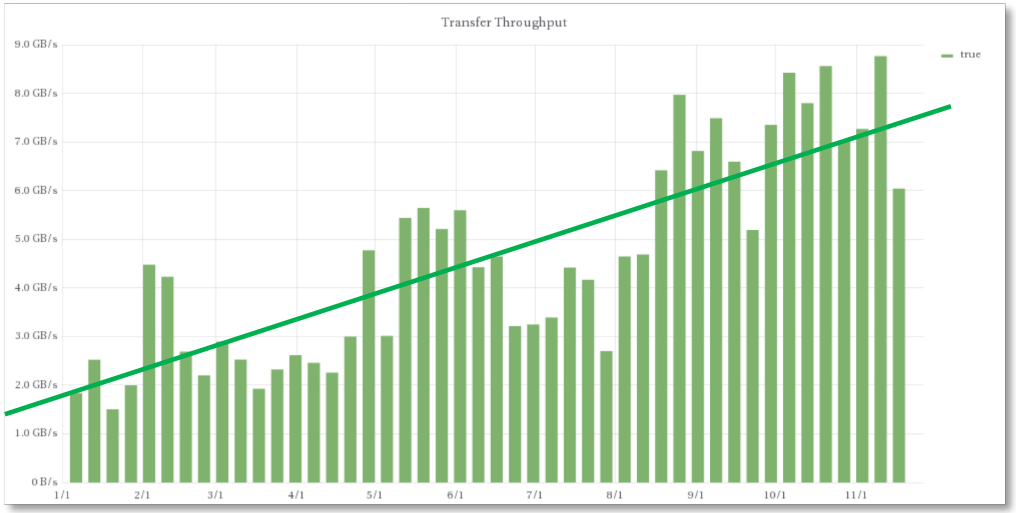
Experiment	Fraction of Tier-2 storage accessible via IPv6
ALICE	51%
ATLAS	37%
CMS	65%
LHCb	33%
Overall	49%

Country	Fraction of Tier-2 storage accessible via IPv6
UK	53%

Tracking IPv6 during 2018

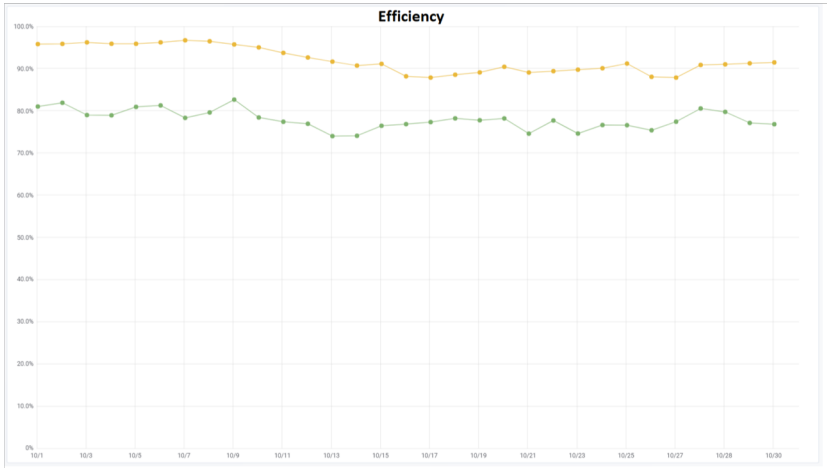


Transfer Throughput over IPv6

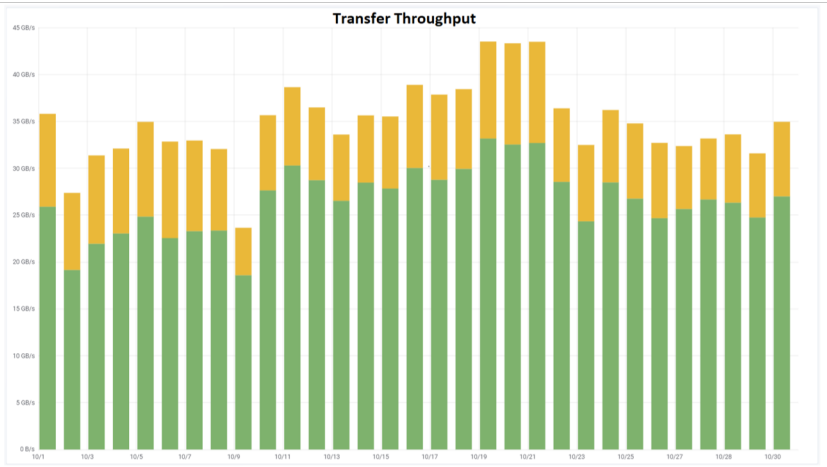


FTS3 file transfers (Oct 2018) ~24% IPv6

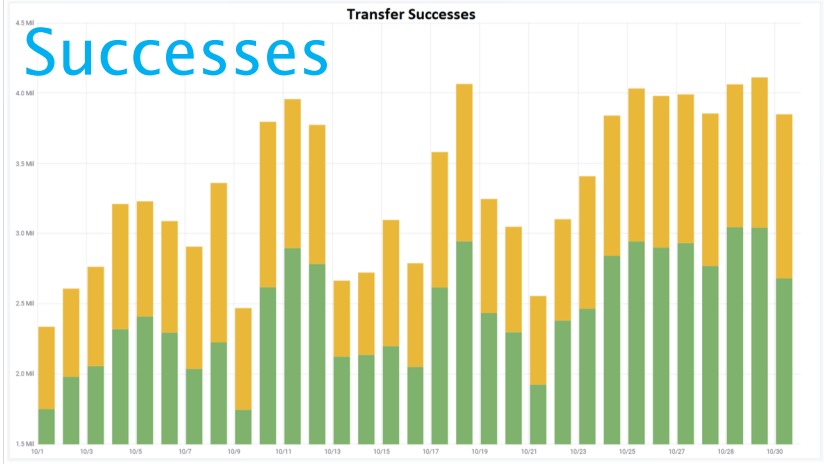
Efficiency

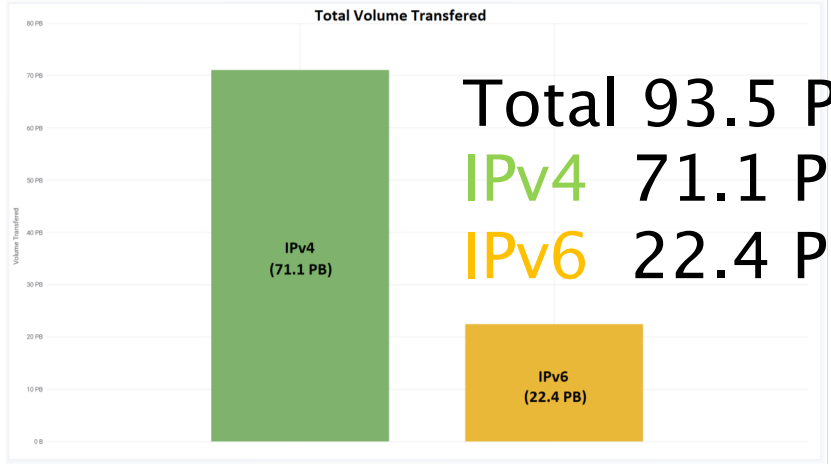


Throughput



Successes





Total 93.5 PB
IPv4 71.1 PB
IPv6 22.4 PB

End to end network monitoring

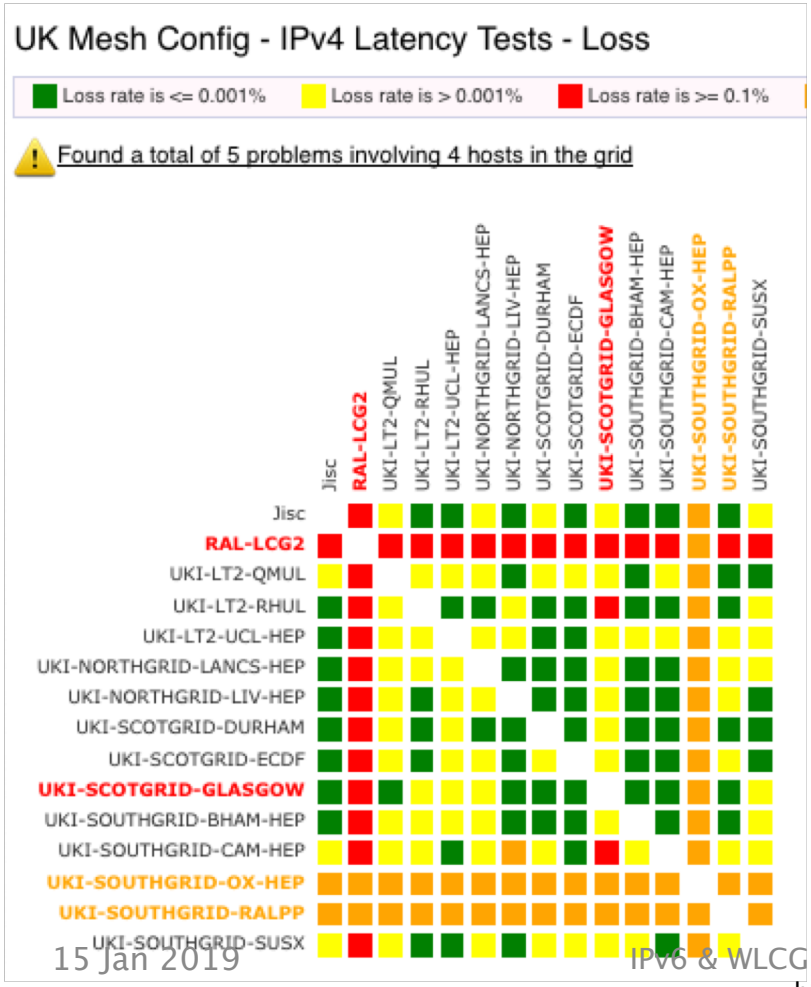
perfSONAR – network monitoring

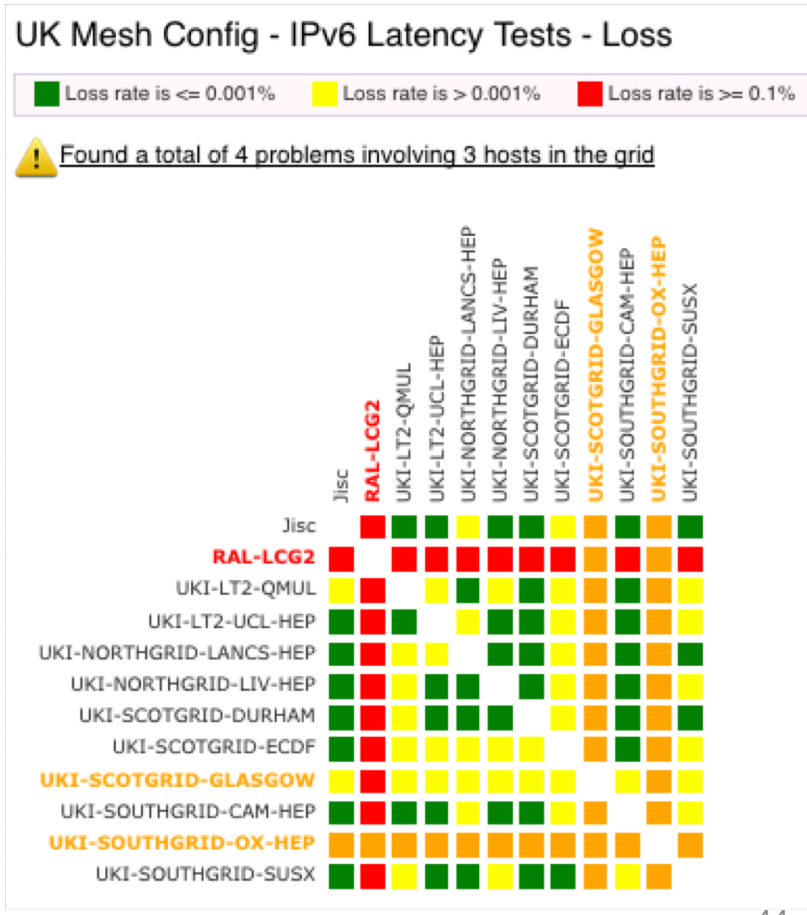
- Developed by ESnet, GEANT, Indiana University and Internet2
- perfSONAR, a widely-deployed test and measurement infrastructure
- used by science networks and facilities around the world
- to aid in network diagnosis
- allowing users to characterize and isolate problems
- measurements of network performance metrics over time as well as “on-demand” tests’
- perfSONAR is IPv6 compatible
- http://www.perfsonar.net/about/what-is-perfsonar/

- WLCG goals with perfSONAR
  - Find and isolate “network” problems; alerting in time
  - Characterize network use such as finding base-line performance
  - In the future: provide a source of network metrics for higher level services

perfSONAR dashboards

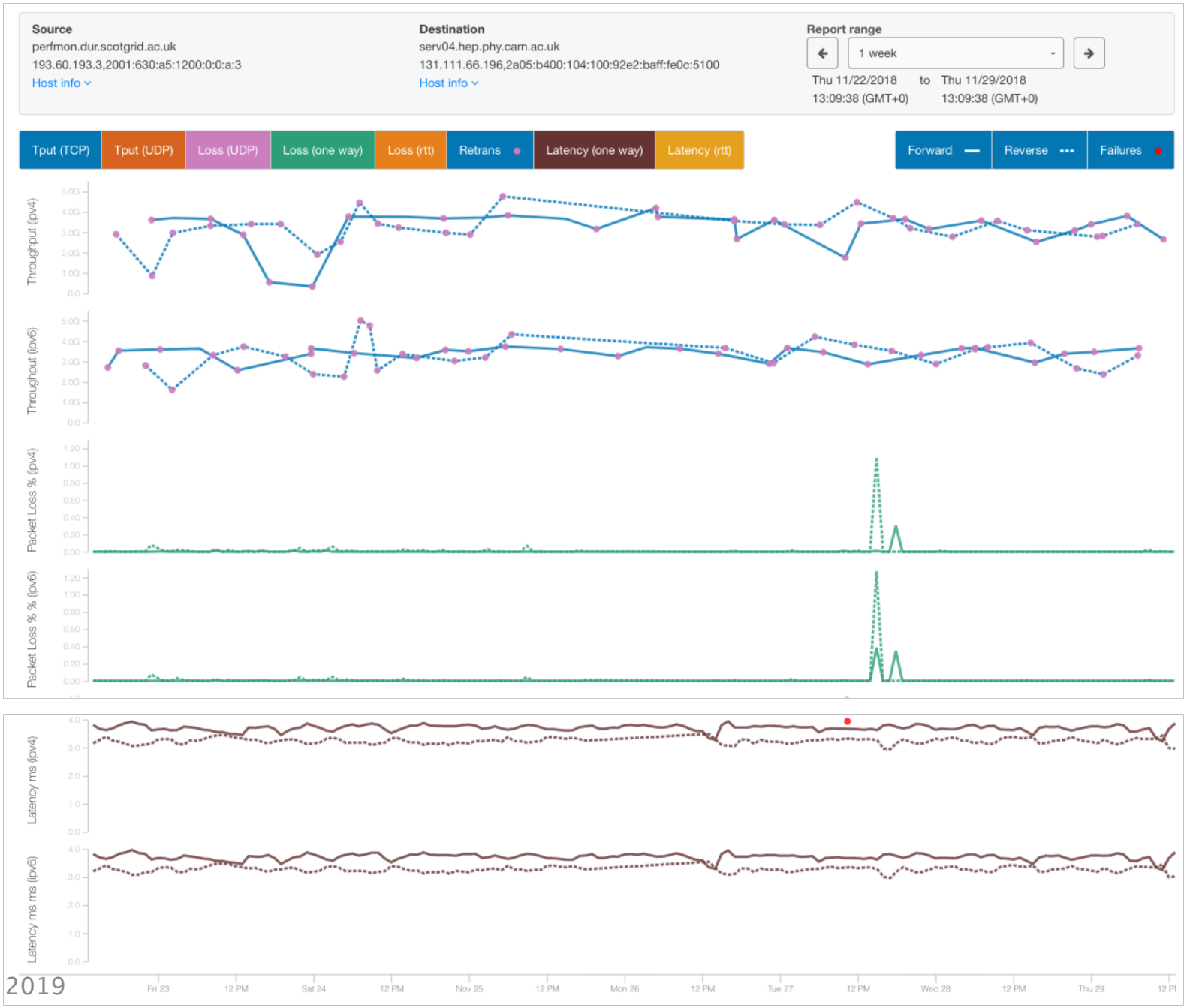
- WLCG has meshes for a variety of groupings e.g. the LHCOPN, CMS and ATLAS
- UK also runs one: throughput, latency, loss, traceroute
- Gives insight into network performance over IPv6 and IPv6 within UK





https://ps-dash.dev.ja.net/maddash-webui/index.cgi?dashboard=UK%20Mesh%20Config

Example perfSONAR results: Durham to Cambridge



IPv4 throughput

IPv6 throughput

IPv4 packet loss

IPv6 packet loss

IPv4 latency

IPv6 latency

GridPP Network Tests

- Jobs are sent to each site to read 1GB-3GB files from each site's Storage Element using various protocols
- The table shows average bandwidth computed from the times taken for each combination (including the local SE)
- Also test over IPv6
- Also recording the percentage of UK CPU and storage available over IPv6
- **UK Tier 2s:** 53% of disk storage available over IPv6 Jan 2019

https://pprc.qmul.ac.uk/~lloyd/gridpp/ukgrid.html

	Capacities				Network						
Site	CPU	Core	HS06	Disk	lcg	gfal4	gfal6	http4	http6	xroot4	xroot6
Brunel	352	5644	67445	1413		55.9	69.9	45.0	46.9	39.5	65.0
Imperial	716	5718	56664	4966	19.7	9.7	25.0	10.3	26.6	13.1	45.8
QMUL	361	4000	62351	5031	3.1	3.7				7.9	
RHUL	442	4624	48121	1460	28.8	30.0		14.5		11.9	
								43.8		46.6	
UCL	0	0	0	0							
Lancaster	420	3360	48384	3074		31.6		34.3		40.8	
Liverpool	173	1816	18466	1425	5.2	5.0		6.4		5.9	
					81.2			34.5		65.3	
Manchester	219	4297	46010	4545		32.0	41.8	25.4	33.7	40.6	41.1
								49.3		70.7	
Sheffield	100	800	10560	531	56.5	51.9		41.2		62.7	
Durham	592	4758	63758	423	16.8	18.1	18.0	12.8	21.1	23.8	17.0
Edinburgh	66	528	6811	2208		93.5		93.6		100.2	
Glasgow	629	5032	43980	3816	9.7	8.4		13.6		11.4	
								14.5		20.0	
Birmingham	152	1584	16996	0				87.3		92.6	
Bristol	82	1320	14744	726	37.2	29.9	6.3	31.5	3.0	34.5	34.7
Cambridge	78	528	6146	264	37.8	33.8		30.6		43.8	
								44.3		49.5	
Oxford	407	3256	33586	939	30.9	37.2		28.6		40.4	
RAL PPD	516	4648	46480	3424	12.0	18.0		13.4		13.8	
Sussex	71	568	5583	84	11.2	10.4		16.3		15.2	
CLOUD											
RAL Tier-1	2347	28168	281680	12819	10.1	12.2		13.3		12.1	

Capactities	CPU		Disk	
Tier-2 Totals:	5376	52481	596085	34329
IPv6 Totals:	1700	16936	196617	18256
IPv6 Percent:	32%	32%	33%	53%

18256/34329 = 53%

Not everything went smoothly!

Problems & lessons learned

- Many blocking issues outside of our own control
  - Both software and site networking teams
- Developers claim that software is fully IPv6-compliant!
- Software/protocols fixed-size storage for IP addresses
- Software/protocols assume single address (as in IPv4)
- Performance differences between IPv4 & IPv6
  - IPv6 must perform at least as well
- Have to understand cases where fraction of IPv6 is smaller than expected
  - Preference for IPv6 over IPv4 must be established
- Can be lots of development effort and testing is not easy when no other positive change re functionality
- Sys admins, operations staff, security team, developers
  - All need TRAINING and experience

Summary

- The WLCG needs to be ready to use IPv6-only CPU resources in LHC Run3 (2021)
- After years of work we are now making excellent progress!
- 2/3rds Tier-1 storage and half of Tier-2 storage is now accessible over IPv6
- The volume of data transferred over IPv6 has increased by a factor of ~3.5 over the last year, ~20-25% of bulk data transfers now go over IPv6
- ~50% of WLCG perfSONAR hosts now reporting 'IPv6-enabled'

- One side-effect is that this is expediting IPv6 adoption in ~170 research institutes worldwide