IPv6-only networking on WLCG

Marian Babik[1], *Martin* Bly[2], *Tim* Chown[3], *Jiři* Chudoba[4], *Catalin* Condurache[2], *Thomas* Finnern[5], *Terry* Froy[6], *Costin* Grigoras[1], *Kashif* Hafeez[2], *Bruno* Hoeft[7], *David P.* Kelsey[2*], *Raul* Lopes[8], *Fernando* López Muñoz[9,10], *Edoardo* Martelli[1], *Raja* Nandakumar[2], *Kars* Ohrenberg[5], *Francesco* Prelz[11], *Duncan* Rand[12], and *Andrea* Sciabà[1]

[1]European Organization for Nuclear Research (CERN), CH-1211 Geneva 23, Switzerland

[2]STFC Rutherford Appleton Laboratory, Harwell Campus, Didcot, Oxfordshire OX11 0QX, United Kingdom

[3]JISC, Lumen House, Library Avenue, Harwell Campus, Didcot, Oxfordshire OX11 0SG, United Kingdom

[4]Institute of Physics, Academy of Sciences of the Czech Republic, Na Slovance 2 182 21 Prague 8, Czech Republic

[5]Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, D-22607 Hamburg, Germany

[6]Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom

[7]Karlsruher Institut für Technologie, Hermann-von-Helmholtz-Platz 1, D-76344 Eggenstein-Leopoldshafen, Germany

[8]College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge, UB8 3PH, United Kingdom

[9]Port d'Informació Científica, Campus UAB, Edifici D, E-08193 Bellaterra, Spain

[10]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain

[11]INFN, Sezione di Milano, via G. Celoria 16, I-20133 Milano, Italy

[12]Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

Abstract.

The use of IPv6 on the general internet continues to grow. The transition of WLCG central and storage services to dual-stack IPv4/IPv6 is progressing well, thus enabling the use of IPv6-only CPU resources as agreed by the WLCG Management Board and presented by us at earlier CHEP conferences.

During the last year, the HEPiX IPv6 working group has been chasing and supporting the transition to dual-stack services. We have also investigated and fixed the reasons for the use of IPv4 between two dual-stack endpoints when IPv6 should be preferred. We present the status of the transition and some tests that have been made of IPv6-only CPU showing the successful use of IPv6 protocols in accessing WLCG services.

The dual-stack deployment does however result in a networking environment which is much more complex than when using just IPv6. The group is investigating the removal of the IPv4 protocol in more places. We present the areas where this could be useful together with our future plans.

*e-mail: david.kelsey@stfc.ac.uk

1 Introduction

The HEPiX IPv6 Working Group [1] has been investigating the many issues related to the move of WLCG services to dual-stack IPv6/IPv4 networking, thus enabling the use of IPv6-only CPU resources as agreed by the WLCG Management Board and presented by us at CHEP2018 [Ref].

The dual-stack deployment does however result in a networking environment which is much more complex than when using just IPv6. Some services, e.g. the EOS storage system at CERN, are using IPv6-only for internal communication, where possible. Several Broadband/Mobile-phone companies, such as T-Mobile in the USA and BT/EE in the UK, now use IPv6-only networking with connectivity to the IPv4 legacy world enabled by the use of NAT64/DNS64/464XLAT. Large companies, such as Facebook, use IPv6-only networking within their internal networks, there being good management and performance reasons for this. Based on these examples of IPv6-only networking, we have therefore been investigating the future removal of the IPv4 protocol in more places within the WLCG infrastructure.

This paper presents the status of the WLCG transition to dual-stack services, together with our work and plans for moving to an IPv6-only networking environment for WLCG.

2 Status of the transition to dual-stack storage

2.1 Deployment at Tier-0 and Tier-1's

Efforts were made to investigate whether the WLCG software packages could be enabled to run in a dual-stack environment or even become protocol agnostic. The first software packages that were examined were data transfer software packages like FTS and SRM. After the examination some software packages were replaced like AFS with EOS, or CASTOR with DPM. Today the storage environment is dual/stack ready and at CERN the Tier-0 is IPv6 and IPv4 dual-stack enabled. The Tier-1 sites: CA-Triumf, DE-KIT, ES-PIC, FR-CCIN2P3, IT-INFN-CNAF, NDGF, NL-T1 (SARA-Matrix and NIKHEF), RRC-JINR-T1, TW-ASGC, UK-T1-RAL, US-T1-BNL, US-T1-FNAL are dual-stack deployed as shown in the following figure. 1. But even while the IPv6 redeaness deadline in April 2018 is long ago, there

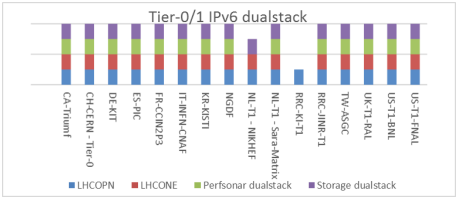

Figure 1. Tier-0/1 IPv4/6 dual-stack redyness incl. dual-stack perfsonar server deployment

is one part of the russian Tier-1 foderation RRC-KI-T1 still deployed with IPv4 only. The dual-stack perfsonar server is deployed at almost all sites except NL-T1-NIKHEF and RRC-KI-T1. The FTS server at FNAL is still running in IPv4 prefered mode. There were a long standing malfunctioned transfer issue to IPv4 only US-Tier-2 sites which is solved now. This last server will get deployed in dual-stack as soon as possible.

2.2 Deployment at Tier-2 sites

The deployment of IPv6 at Tier-2 sites is proceeding even after the official deadline expired at the end of 2018. It was decided not to give the deadline a formal extension, but just to

encourage all remaining sites to complete the IPv6 deployment "as soon as possible": the main motivations were that *a)* sites behind schedule were encountering objective difficulties and *b)* the most effective deadline would be imposed by the experiments themselves, if they wished, for example, to require IPv6 for production. This choice was confirmed by the steady progress observed during 2019, as it can be seen in figure 2.

The time evolution of the site status shows a steady increase of the number of sites that have deployed IPv6, until a more recent slowdown. This is consistent with the hypothesis that the remaining sites are those facing the biggest difficulties. A detailed analysis of the tickets shows that, in many cases, sites need to wait for the IPv6 deployment on site, which often depends on people different from the WLCG site staff. The fraction of the Tier-2 storage that is accessible via IPv6 is shown in table 1 for each experiment, and significant differences are apparent. Two experiments (ALICE and CMS) are very close to having all their Tier-2

Table 1. Fraction of Tier-2 storage available over IPv6

ALICE	ATLAS	CMS	LHCb	Global
85%	59%	89%	75%	73%

storage on IPv6, LHCb has little Tier-2 storage to begin with due to their particular computing model and ATLAS is getting better, but still far from the goal.

2.3 LHCOPN and LHCONE

The LHCOPN and LHCONE are both virtual private networks (VPN) serving the Large Hadron Colider Experiments. Both networks are from the end of 2016 anward dual-stack ready. LHCOPN is a CERN (Tier-0) centric star network mainley deployed for the distribuion of the raw detector data to the tier-1 sites. Since the majority of Tier-1 sites are dual-stack ready and even while the protocol IPv6 is prefered it is still not the situation that

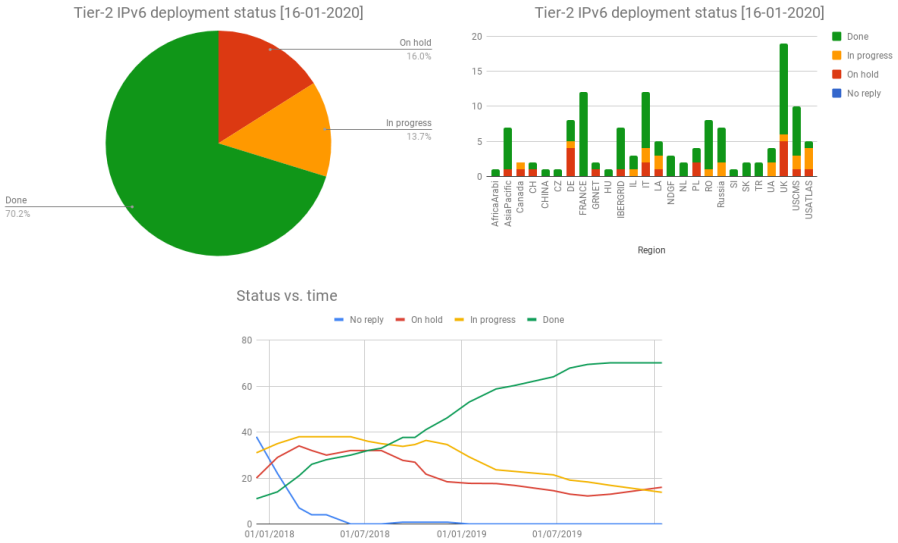

Figure 2. (left) Tier-2 deployment status by site globally, (right) by region, and (bottom) time evolution

IPv6 is the only transfer protocol, but a tendence towards IPv6 fle transfers are recognizible. LHCONE is a netwrork of close to 140 sites connected trough Virtual Routing and Forwarding implementations at 26 diffrent network service providers (NSP). All connected endsites deploying a Border Gateway Protocol (BGP) routing table and adverticing their own CIDR to the connecting NSP. The network itself is already since long IPv6 ready. The connected end sites are becomming more and more IPv6 ready. This is recogizable at the transfer protocol changes from IPv4 towards IPv6. Out of the high usage of the IPv4 transfer protocol it is still recogisable the the fraction of IPv4 only sites is quite substancial.

2.4 File transfers

Over the last 2 years we have been regularly tracking the fraction of WLCG file transfers that take place over IPv6. (needs more description).

The fraction of data transfers over FTS on IPv6 as a function of date is shown in figure 3.

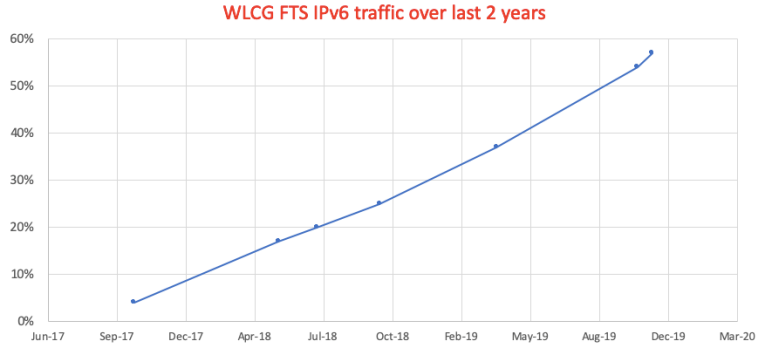

Figure 3. Percentage of FTS data transfers over IPv6

3 IPv6-only networking

A few years ago, RFC 6586 ([18]) reported on a survey on IPv6-only networking for mainstream applications (gaming, telephony, multimedia, etc.) and observed that *«it is possible to employ IPv6-only networking»* and that *«for large classes of applications there are no downsides or the downsides are negligible»*. This, along with the good working relations we established over the past years with the HEP software stack developers, encouraged us to test scenarios where the transient complication of running and managing two independent network stacks is eventually over and we are 'back' to just running IPv6.

3.1 Aims of moving to IPv6-only and issues to be tackled

A dual-stack IPv6/IPv4 setup includes many components and services that need to be deployed twice *and kept in sync*: firewall rules and access lists, address assignment services, routing rules, network monitoring, diagnostics and intrusion detection infrastructures - just to name a few. Removing this duplication is highly desirable both for better maintainability and cost-saving. However, this *requires* a technical solution to access any site and service that may remain accessible via IPv4 *only*. This trailing remainder of sites will be hopefully

shrinking but will likely exist for a very long time (see e.g. [10] and references therein). It is actually expected that after large blocks of public IPv4 addresses start being returned to the market, their market price will decrease and offset any economic drive connected to the IPv4 address shortage, *relieving* the pressure for migration.

The standard solution for accessing IPV4-only services from an IPv6-only network is the deployment of DNS64 (RFC6147) and stateful NAT64 (RFC6146 [18]) services. DNS64 maps names that are resolved to an IPv4 address only ('`A`' records) to a synthesized IPv6 address composed of a default prefix[1] plus the four bytes of the IPv4 address. Traffic towards the DNS64 prefix is then routed to a NAT64 service attached to the public IPv4 network. This will in turn map the source ports, translate the IP packets to IPv4 and convert any return traffic back to IPv6. While NAT64 and DNS64 services start being incorporated by several technology developers (especially in the 'carrier-grade' NAT appliance market), just a few open-source reference implementations exist for UNIX, with JOOL[2] apparently being the only one under active development.

While IPv6-only environments can present a few operational challenges that can be worked around[3], one-step (6→4) address translation techniques do and will fail whenever IPv4 literal addresses are explicitely handled, stored or signaled by network applications or protocols. We feel that the time is ripe to start identifying this class of applications and protocols and direct an early effort at cleaning them of *any* usage or reference to IPv4 literals. While two-step (4→6→4) address translation techniques such as XLAT464[4] are currently being added to network stacks especially at the request of telephony carriers that operate IPv6-only networks, we see this extra indirection as an (inefficient) workaround that just hides issues that should be fixed at the application level. Locating this issues as early as possible motivates the experimental operation of typical WLCG sites with IPv6-only networking, as described later (§3.3).

3.2 The case for an IPv6-only LHCOPN

The LHCOPN[5] is the private network that connects CERN, the WLCG Tier0, to the 14 Tier1s datacentre. It is made of multiple 10Gbps and 100Gbps and provide more than 1Tbps out of the Tier0.

Since the EOS storage service has been supporting IPv6, a large fraction of the LHCOPN data transfers have changed transmission protocol, moving from IPv4 to IPv6. Since June 2019, LHCOPN carries more IPv6 packets than IPv4[6].

It could be envisaged that in the near future, once all the Tier1s will have implemented dual-stack storage services, the LHCOPN could be turned into an IPv6 only network. There are some advantages that an IPv6 only LHCOPN could bring:

- Increased security: LHCOPN links connect directly into Tier1s data-centres, often bypassing border firewalls. Removing one protocol would decrease the attack surface;
- Simpler opeations: maintining one transmission protocol would simplify the operation of the networks and the resolution of problems.

[1] Usually `64:ff9b::`.

[2] `https://www.jool.mx`

[3] Some OS-specific network management tools, firewall appliances and network monitoring and diagnostic tools were found to be defective or immature, see RFC 6586 [18] for details.

[4] See RFC 6877 [18]. XLAT464 keeps a private IPv4 address assigned to devices connected to IPv6-only networks and performs an additional address translation at the device level.

[5] LHCOPN twiki: https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

[6] LHCOPN traffic comparison: https://twiki.cern.ch/twiki/bin/view/LHCOPN/LHCOPNEv4v6Traffic

The HEPiX IPv6 Working Group will encourage the LHCOPN community to move to IPv6 ONLY as soon as possible.

3.3 Testing of IPv6-only

An IPv6-only WLCG production cluster, composed of an ARC-CE head node, two worker nodes and three (SQUID-based) web cache nodes based has been in operation at Brunel University since March 2018. Given the value we place on early detection of IPv4-only code sections (especially non-address-translatable constructs such as the use of IPv4 literals in data structures and signaling - see above, §3.1), no transition techniques (e.g. NAT64/DNS64) were used for this infrastructure.

WLCG production jobs for three (out of four) major LHC experiments were routed to this IPv6-only cluster, with LHCB jobs running successfully in 2018, CMS jobs (submitted via a dedicated queue) running successfully in 2019, and ATLAS jobs, also handled by a special IPv6-only queue, requiring an in-depth, and still partly on-going, investigation of issues mainly within the Frontier[7] distributed database service.

This reality check does confirm that IPv4 is still *required* in part of the WLCG software base, with services failing in case IPv4 connectivity cannot be established. While the development time that has been spent in early troubleshooting and linting of these cases will definitely be rewarded as the transition progresses, we plan to complement this study with an assessment on how many of the residual issues aren't or cannot be covered by available address-translation techniques.

4 Conclusions and future plans

We have presented the status of the WLCG transition to the use of dual-stack IPv6/IPv4 services. The Tier-1 transition is very nearly complete and more than 70% of the Tier-2 storage is now available over IPv6. The transition will only be fully completed once we remove the complexity of dual-stack networking and the WLCG core infrastructure is based entirely on IPv6-only networking.

Insufficiently tested or immature code and the requirement that IPv6-based tools and infrastructures perform at least equally well as their IPv4 counterparts have been the opposite, conflicting poles of every IPv6 deployment effort so far. This continues to be true in the process of completing the transition. We therefore keep considering testing activities (and the consequent early detection of further application development needs), as a value that our group can provide. So we plan to increase the number of sites and stakeholders involved in testing IPv6-only scenarios. The aim is to stress-test existing networking software components that implement any needed transition protocol (especially NAT64 and DNS64, as their implementations under current maintenance are rare) and detect residual uses of IPv4 literals or IPv4-specific APIs in both applications and network protocols as early as possible.

Any use of IPv4 that cannot respond properly to a NAT64-mediated transaction [8] should be seen as an issue to be reported, tracked and eventually addressed by developers: we plan to deal with these just as we did with the lack of IPv6 support or the incorrect address selection strategies we were able to identify so far.

5 The References need updating

These are from the CHEP2018 paper!

[7] `http://frontier.cern.ch/`

[8] More complex and inefficient address translation solutions such as the deployment of 'customer'-side address translation for RFC 6877 ([18]) 424XLAT should be seen as options of last resort, see §3.1 above.

References

[1] S. Campana et al, J. Phys. Conf. Ser. **513**, 062026 (2014)

[2] M. Babik et al, J. Phys. Conf. Ser. **898**, 082033 (2017)

[3] A. J. Peters et al, J. Phys. Conf. Ser. **664(4)**, 042042 (2015)

[4] S. Bagnasco et al, J. Phys. Conf. Ser. **119(6)**, 062012 (2008)

[5] L. Bauerdick et al, J. Phys. Conf. Ser. **396** 042009 (2012)

[6] A. Grigora et al, J. Phys. Conf. Ser. **523**, 012010 (2014)

[7] I. Sfiligoi, J. Phys. Conf. Ser. **119(6)**, 062044 (2008)

[8] D. Thain et al, Concurrency - Practice and Experience, **7(2-4)**, 323 (2005)

[9] A. Tsaregorodtsev et al, J. Phys. Conf. Ser. **513**, 032096 (2014)

[10] M. Nikkhah and R. Guérin, IEEE/ACM Transactions on Networking, **24(4)**, 2291 (2016)

[11] A. Aimar et al, J. Phys. Conf. Ser. **898(9)**, 092033 (2017)

[12] Marian Babik, CERN, http://etf.cern.ch/docs/latest/

[13] A. Hanemann et al, In Boualem Benatallah, Fabio Casati, and Paolo Traverso, editors, *Service-Oriented Computing - ICSOC 2005*, pages 241–254, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[14] S. McKee et al, J. Phys. Conf. Ser. **664(5)** 052003 (2015)

[15] perfSONAR Consortium, http://psmad.grid.iu.edu/maddash-webui/

[16] CERN monitoring team, https://monit-grafana.cern.ch/

[17] A. A. Ayllon et al, J. Phys. Conf. Ser. **513(3)** 032081 (2014)

[18] All Internet Engineering Task Force Requests For Comments (RFC) documents are available from URLs such as http://www.ietf.org/rfc/rfcNNNN.txt where NNNN is the RFC number, for example `http://www.ietf.org/rfc/rfc2460.txt`