The production deployment of IPv6 on WLCG

J Bernier[1], S Campana[2], K Chadwick[3], J Chudoba[4], A Dewhurst[5], M Eliáš[4], S Fayer[6], T Finnern[7], C Grigoras[2], T Hartmann[8], B Hoeft[8], T Idiculla[5], D P Kelsey[5], F López Muñoz[9], E Macmahon [10], E Martelli[2], R Nandakumar[5], K Ohrenberg[7], F Prelz[11], D Rand[6], A Sciabà[2], U Tigerstedt[12], R Voicu[13], C J Walker[14] and T Wildish[15]

[1] IN2P3 Computing Centre, Boulevard du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France
[2] CERN, CH-1211 Genève 23, Switzerland
[3] Fermi National Accelerator Laboratory, Batavia, Il 60510, U.S.A.
[4] Institute of Physics, Academy of Sciences of the Czech Republic Na Slovance 2 182 21 Prague 8, Czech Republic
[5] STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxfordshire OX11 0QX, United Kingdom
[6] Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
[7] Deutsches Elektronen-Synchrotron, Notkestraße 85, D-22607 Hamburg, Germany
[8] Karlsruher Institut für Technologie, Hermann-von-Helmholtz-Platz 1, D-76344 Eggenstein-Leopoldshafen, Germany
[9] Port d'Informació Científica, Campus UAB, Edifici D, E-08193 Bellaterra, Spain
[10] The University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, United Kingdom
[11] INFN, Sezione di Milano, via G. Celoria 16, I-20133 Milano, Italy
[12] CSC Tieteen Tietotekniikan Keskus Oy, P.O. Box 405, FI-02101 Espoo
[13] California Institute of Technology, Pasadena, Ca 91125, U.S.A.
[14] Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
[15] Princeton University, Jadwin Hall, Princeton, NJ 08544, U.S.A.

E-mail: `david.kelsey@stfc.ac.uk, ipv6@hepix.org`

Abstract. The world is rapidly running out of IPv4 addresses; the number of IPv6 end systems connected to the internet is increasing; WLCG and the LHC experiments may soon have access to worker nodes and/or virtual machines (VMs) possessing only an IPv6 routable address. The HEPiX IPv6 Working Group (`http://hepix-ipv6.web.cern.ch/`) has been investigating, testing and planning for dual-stack services on WLCG for several years. Following feedback from our working group, many of the storage technologies in use on WLCG have recently been made IPv6-capable. The worldwide HEP computing community now needs to deploy dual-stack IPv6/IPv4 services on WLCG to allow such use of IPv6-only resources. This paper will present the IPv6 requirements, tests and plans of each of the four LHC experiments together with the tests performed both on the IPv6 test-bed and in targeted use of WLCG production services. This is primarily aimed at IPv6-only worker nodes or VMs accessing several different implementations of a global dual-stack federated storage service. The changes required to the operational infrastructure, including monitoring and security, will be addressed as will the implications for site management. The working group will present its deployment plan for dual-stack storage services, together with other essential central and monitoring services, to start during 2015.

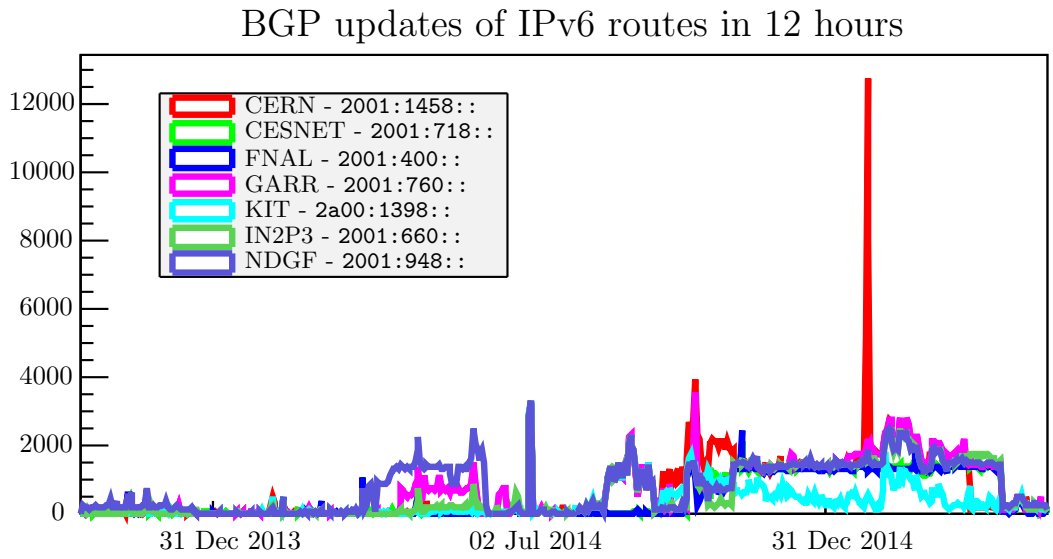

Figure 1. Number of routing topology updates involving the reference LIR of the working group testbed participants recorded by RIPEstat [3] every 12 hours since CHEP2013.

1. Introduction

The much-heralded exhaustion of the IPv4 networking address is with us, etc. etc.

2. Status and hurdles of the worldwide IPv4→IPv6 transition

We now offer a quick panorama of the IPv6 statistical trends since the last CHEP conference and identify two factors that may be currently limiting the IPv6 adoption rate.

2.1. Survey of available statistical data

An analysis of the available statistical data collected since 2013 by the Regional Internet Registries [4–8] and by major internet service providers [9,10] shows a steady, but still polynomial growth in the volume of IPv6 traffic from roughly 2% up to 6% of the total. Not everything has been progressing at a relaxed pace, though: we collected from RIPEstat ([3]) into Figure 1 the rate of BGP routing topology updates that affects the IPv6 `/32` prefixes that serve our one may wonder about the causes that prevent a more rapid adoption pace.

Transport and provider issues have to be excluded right away: most local registries, including all of the national research networks, have been providing IPv6 transport for 7 years or more ([2]). Performance issues also have to be ruled out: available studies, especially the ones carried out during the "world IPv6 launch" in 2012 (see [13]) show performance on the two stacks to be comparable within statistics. The reality of the IPv4 address depletion should also be widely perceived by now, as many regional registries are handling the *final* IPv4 assignment to local registries.

Two residual classes of factors may be quenching IPv6 adoption, one affecting network administrators and one affecting application developers:

(i) The IPv4/v6 difference in address allocation schemes, the pivot role of ICMPv6 Router Advertisements, the short-term need to implement measures to counter rogue Router Advertisements (see RFC6104, [12]) are all adding to the already sizeable initial investment of implementing monitoring and security tools for IPv6.
Also, existing IPv6 code in the Operating Systems and related tools often shows by inspection not

to have underwent full coverage testing: a phase of initial fault finding and patching is foreseen and feared.

(ii) Apart from the syntactical differences in IPv6 addresses (e.g. parsing '`defa`' in *default* as a hex digit), and the need to label, sort and pick IPv4 vs IPv6 addresses, a large *semantical* change is needed in applications supporting IPv6. Every network endpoint on the public IPv6 network has *at least* two IPv6 addresses assigned (global and link-local), and possibly more. Applications have therefore to *always* deal with multi-homed network endpoints, which means a complex $1 \rightarrow n$ change for many of them. The status of porting to IPv6 of many applications of interest for our community is good ([11]), but the related development effort cannot be underestimated.

3. The 2014 survey of IPv6 readiness at WLCG sites

In the summer of 2014 we performed a survey of all WLCG sites to determine the their readiness for IPv6 and also whether they foresee running out of IPv4 address space in the next years.

More words needed.

The results of the survey are shown in table X.

Table 1. Site IPv6-readiness

Type of Site	Answered	IPv6-ready	Ready soon	No IPv6 plans	Lack of IPv4
Tier 0/1	12	9	3	0	2
Tier 2	96	20	20	56	10

The conclusion of the survey is...

More words needed

4. LHC Experiment requirements and main use case

The shortage of available IPv4 addresses implies that there is a significant possibility that new large computing facilities will not be able to give IPv4 addresses to all of the machines in their network. The most likely consequence is that for these sites, worker nodes – which constitute the largest fraction of independent computing nodes will have purely IPv6 network addresses. Hence, the main use case for the LHC experiments is to enable jobs to run on these machines, access their software areas and input data and upload their outputs to various grid storages or services as needed.

The LHC experiments generally assume [LHCassumption reference] that the storage on different sites and supporting middleware [middleware reference] like the LFC will either be directly dual-stack, or support dual stack operation in some way, enabling seamless access to the storage as needed for either downloading or saving. For example, it is expected that dual-stack squid proxies will be needed for CVMFS and xrootd will soon be dual-stack, to handle storage technologies like Castor which will not be IPv4 only. The servers that the LHC experiments use to handle the grid infrastructure are / will also be dual-stack.

As an example of the above, we look at LHCb [LHCb reference] which is the experiment on the LHC, optimised for studying beauty and charm physics. LHCb uses the DIRAC [DIRAC reference] interware to manage is grid operations. The DIRAC software was coded to be able to handle both IPv4 and IPv6 addresses in late 2014, with the modifications being easy enough to make by non-expert programmers. While testing the processes on a dual-stack machine, it was found that there was a significant number of connections which were not going through to

the servers which was finally traced back to a missing enable_ipv6 option to compile python by an external provider causing errors in identifying IPv6 addresses. Using a version of the library, the problem with dropped connections went away and testing DIRAC will restart soon.

In general, testing of the grid middleware by the different experiments is going ahead as fast as possible given the manpower and time constraints of the LHC startup and the immediate issues with handling purely IPv6 worker nodes are expected to be sorted by sometime in 2016.

[LHCassumption reference] : WLCG pre-GDB - IPv6 Workshop (A. Dewhurst et al.), https://indico.cern.ch/event/313194/session/1/contribution/3/0/material/slides/0.pdf [middleware reference] : https://wiki.egi.eu/wiki/Middleware_products_verified_for_the_support_of_IPv6 , http://hepix-ipv6.web.cern.ch/wlcg-applications [LHCb reference] : LHCb Collab. (A. Alves et al.), JINST 3, S08005 (2008), http://dx.doi.org/10.1088/1748-0221/3/08/S08005 [DIRAC reference] : DIRAC: a community grid solution (A. Tsaregorodtsev et al.), 2008 J. Phys.: Conf. Ser. 119 062048, http://iopscience.iop.org/1742-6596/119/6/062048

5. Testbed operation: testing FTS3/dCache

5.1. The Transfer Testbed

The transfer testbed was upgraded in March 2015. Until then, it operated with gridFTP transfers between all sites, providing a low-level test of connectivity and functionality for the almost two years that it ran. Since March 2015 the testbed uses FTS3 [14] to initiate the transfers, moving up the middleware stack. Since FTS3 is used for the vast majority of experiment transfers in WLCG this provides an important full-stack test.

At the present time, the testbed consists of 7 storage elements at sites distributed around Europe. One is IPv6-only, the rest are all dual-stack. All the SEs are running dCache. Most are stable installations, but one (DESY) is rebuilt every morning with the latest patches from dCache, providing a valuable regression-test for both the dCache and IPv6 teams.

As before, each site serves as both a source and a destination, with each source sending a 1 GB file to each destination. The file-size is validated at the destination using **gfal-ls**, then the destination is cleaned with **gfal-rm** and the transfer duration is recorded. Then the cycle is repeated after a short delay, to avoid abusing the hardware/network with too much traffic. Physical file names are specified using the SRM protocol.

Two FTS3 servers are deployed for the testbed, one at Imperial College and one at KIT, though currently only the one at KIT is used.

Figure. 2 shows the transfers in the FTS3 testbed so far. Most sites transfer efficiently in both directions, but the effect of the firewall at KIT on inbound traffic can be clearly seen.

5.2. FTS3 server and dCache SE at KIT

For managing file transfers between sites, an FTS3 instance is setup at KIT. Furthermore, a storage element based on dCache 2.10 on Scientic Linux is created. Both instances are rolled out on physical machines.

5.2.1. FTS3 FTS3 supports IPv6 in its baseline version. The service has to be bind to IPv6 locally in the FTS3 conguration (IP=::) and to be enabled explicitely for gfal2 (IPV6=true). The host is available in the DNS as default dual-host, with IP4 and IPv6 announced in the A and AAAA records, and also with IPv4-only or IPv6-only names with an -ipv4 or -ipv6 appendix, respectively. Thus, all aliases have to included in the host certicate.

File transfers are successfully brokered by the FTS3 instance via IPv4 and IPv6 between the sites. Since most FTS3 instances in production use a separated database instance for performance and failsafe reasons, moving the database to a dedicated machine was tested as well. For the SQL db backends supported by FTS3, IPv6 support had been implemented in MySQL v5.5.3 and MariaDB v5.5.35, which are not available in the baseline SL6 repositories.

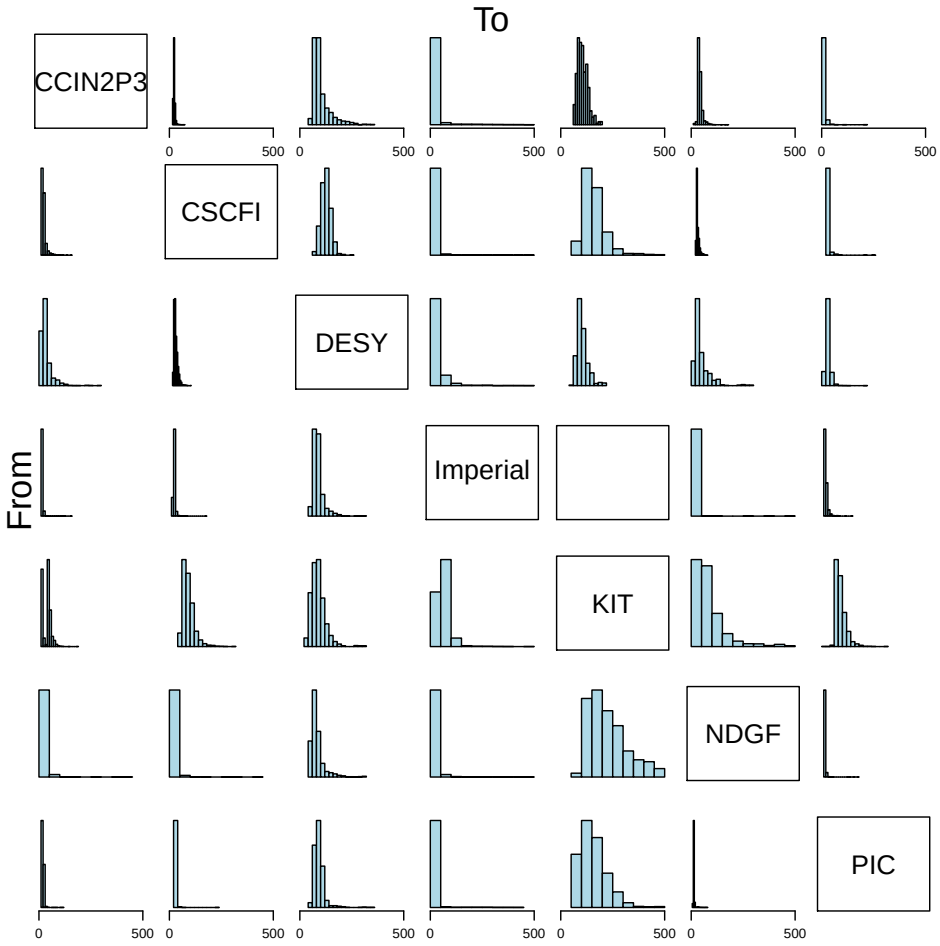

Figure 2. The FTS3 transfer testbed. Rows show transfers from the named site, columns show transfers to the destination. The horizontal axis for all plots is fixed at 500 seconds, i.e. transfers that proceed at less than 2 MB/sec will overflow.

MariaDB was installed on a dedicated host with version 5.5.42. After binding mysql locally to IPv6 as well ([mysqld] bind-address = ::), the database could be connected remotely with an IPv6-ready mysql client. For the FTS3 service to connect to the remote database via IPv6, the address has to be escaped explicitly, i.e., encapsulating the IP as [IP6]:PORT/fts3 and may depend on the version of the database library used by the FTS3 service.

5.2.2. dCache based SE A dCache instance is setup on a dedicated host. The main hurdle for file transfer access was a reverse lookup by the instance when receiving a file request. After explicitely setting the dual-stack, IPv4-only, and IPv6-only host names in the dCache congfiguration (srm.net.local-hosts=hostname-ipv4,ipv6) and the general hosts file, the storage element is accessible via IPv4 and IPv6 as well.

6. IPv6 readiness of storage technology for WLCG

6.1. Storage technology

For the FTS3-based testbed SRM [17] was selected as more production-like than the old gridftp-based one.

All sites participating selected dCache, as it had matured into the only full-stack storage system to fully support dualstack and IPv6-only setups.

6.1.1. Protocol: SRM SRM is built on SOAP which in turn is built on HTTP. It's designed to be protocol-independent as it only sends data in one stream and only processes TURLs and SURLs. It's however only used to set up the transfer, and the real transfer is handled by other protocols.

6.1.2. Protocol: HTTP Transfer over HTTP or HTTPS work without problems for reads, but the implementations for writing are differing between different software leading to limited usability. It is used in production for reads at some sites, most notably NDGF-T1.

6.1.3. Protocol: GSIFTP GSIFTP or GridFTP [15] is FTP with an extra layer on top for multistreaming and autentication. It's currently widely used with SRM, but has issues with IPv6 since it sends the IP address of where to connect to instead of using hostnames. All current servers and clients break the GridFTP-2.0 document [16] in the same way to support "Delayed Passive", a method for redirecting writes from the client directly to the storage without a multihost storage environment like dCache having to proxy the write.

6.1.4. Protocol: XROOTD XROOTD by SLAC is a general purpose random-IO protocol for data access. It supports IPv6 from release 4.0.0.

6.1.5. Software: dCache dCache is a Java-based software for building distributed storage solutions. It supports many different access protocols and authentication methods.

6.1.6. Software: DPM DPM or Disk Pool Manager was the software used by many sites for the previous testbed. It

7. Status of LHCOPNv6/LHCONEv6

8. LHCONE/LHCOPN

Based on the actions initiated at Grid Deployment Board in November and December 2014, to request tier-1s to join the HEPiX-IPv6 working group and to encourage sites moving their production endpoints to dual stack even if this requires concessions of the quotation of their site relatability and site availability. The proposal of the LHC experiment Atlas was to

- request that all Tier-1s provide, besides an IPv6 peering to LHCOPN, a dual stack PerfSONAR machine by first of April 2015
- request that Tier-2s provide, besides an IPv6 peering to their LHCONE connection, a dual stack PerfSONAR machine by August 2015.

At the last LHC[OPN/ONE] meeting a proposal was put forward to the effect that LHCOPN connecting Tier-1 sites to CERN would get IPv6-ready by 1. April 2015 and LHCONE connecting Tier-[123] sites would become IPv6 ready by August 2015. No objections to this proposal were presented. The url:
`http://maddash.aglt2.org/maddash-webui/index.cgi?dashboard=Dual-Stack%20Mesh%20Config`
shows the status of the tier-[12] sites connecting a dual stack perfsonar and implies that there are still some LHC tier-1 sites more than two month after the agreed deadline not offering ipv6 cidr via their LHCONE peering.

9. IPv6 perfSONAR measurements

The WLCG has adopted the perfSONAR toolkit [18] for the monitoring of its network infrastructure and this project is being coordinated by the WLCG Network and Transfer Metrics group [19]. The WLCG perfSONAR configuration system operates around groups of sites with a common purpose and these are known as meshes. For example, there are meshes for each WLCG country group (e.g. UK, DE, FR etc.) or experiment such as USATLAS or USCMS or network groupings such as LHCONE and LHCOPN. Until recently testing between members of the groups was configured using JSON files held on a web-server at CERN specifying the group members and test parameters. A site administrator configuring a perfSONAR host for a mesh needed to add the URL of the JSON file corresponding to that mesh into a configuration file on the perfSONAR host. This worked but required effort from all site administrators involved. The mesh configuration system has recently undergone development. This is described in detail in [19], but briefly, the system has evolved from the use of the manually configured JSON files to a more automated system which is significantly easier to configure. Each perfSONAR host now has a so-called auto-mesh URL e.g.
`https://myosg.grid.iu.edu/pfmesh/mine/hostname/psum01.aglt2.org`
containing configuration details for the meshes that the perfSONAR host has been added to. The meshes are maintained using a mesh-configuration GUI provided by the US Open Sciences Grid (OSG) using data collected from both the GOCDB and OSG Information Management System (OIM). The perfSONAR toolkit has the ability to monitor both IPv4 and IPv6 network connectivity. Consequently, in addition to the meshes mentioned above, we have added a mesh containing perfSONAR hosts known to have both IPv4 and IPv6 connectivity, i.e. a dual-stack mesh. The mesh tests throughput and latency between hosts over both IPv4 and IPv6. Results are available from the web sites of the relevant perfSONAR hosts, for example the perfSONAR bandwidth host at WLCG site UKI-SOUTHGRID-OX-HEP at the University of Oxford:
`http://t2ps-bandwidth.physics.ox.ac.uk/toolkit/`

10. Outlook and future plans

References

[1] `http://hepix-ipv6.web.cern.ch`

[2] The status of IPv6 readiness of the local Internet Registries that refer to RIPE is monitored in the "IPv6 RIPEness" pages (`https://ipv6ripeness.ripe.net/` at the time of writing).

[3] RIPE provides access to a comprehensive set of statistical data via `https://stat.ripe.net/data`. This data set is also used for the online diagnostics and statistical tools (e.g. BGPlay).

[4] Specific IPv6 trend graphs for the RIPE LIRs can be accessed at `https://labs.ripe.net/statistics/?tags=ipv6`.

[5] A collection of trend plots collected by ARIN can be found here: `https://www.arin.net/knowledge/statistics/`.

[6] A somewhat out-of-date set of IPv6 statistics data gathered by APNIC can be found here: `https://labs.apnic.net/ipv6-measurement/`.

[7] The LACNIC IPv6 portal is at: `http://portalipv6.lacnic.net/en/`.

[8] A collection of AFRINIC IPv6 resources can be found here: `http://www.afrinic.net/en/services/statistics/ipv6-resources`.

[9] The historical statistics of IPv6 traffic on Akamai usare are recorded at `http://www.akamai.com/ipv6/` at the time of writing.

[10] The historical statistics of IPv6 client traffic on Google are recorded at `http://www.google.com/intl/en/ipv6/statistics.html` at the time of writing.

[11] Our working group tracks the IPv6 readiness of applications of interest to WLCG at `http://hepix-ipv6.web.cern.ch/wlcg-applications`.

[12] 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`

[13] Plonka D, Barford P, Assessing performance of Internet services on IPv6, 2013 IEEE Symposium on Computers and Communications (ISCC), doi:10.1109/ISCC.2013.6755050

[14] A A Ayllon and M Salichos and M K Simon and O Keeble *FTS3: New Data Movement Service For WLCG*, J. Phys.: Conf. Ser. 2014, **513** 3 032081, doi:10.1088/1742-6596/513/3/032081, http://dx.doi.org/10.1088/1742-6596/513/3/032081,

[15] https://www.ogf.org/documents/GFD.20.pdf

[16] https://www.ogf.org/documents/GFD.21.pdf

[17] http://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html

[18] B Tierney, J Metzger, J Boote, A Brown, M Zekauskas J Zurawski, M Swany, M Grigoriev, perfSONAR: Instantiating a Global Network Measurement Framework, 4th Workshop on Real Overlays and Distributed Systems (ROADS09) Co-located with the 22nd ACM Symposium on Operating Systems Principles (SOSP), January 1, 2009.

[19] S McKee, M Babik et al. Integrating network and transfer metrics to optimize transfer efficiency and experiment workflows, CHEP2015, Journal of Physics: Conference Series.