Deployment of IPv6 only CPU resources at WLCG sites

M Batik[1], J Chudoba[2], A Dewhurst[3], A Sciaba[1], S Fayer[9], T Finnern[7], T Froy[6], C Grigoras[1], K Hafeez[3], B Hoeft[4], T Idiculla[3], D Kelsey[3], E Martelli[1], F Munoz[10], R Nandakumar[3], K Ohrenberg[6], F Prelz[5], D Rand[9], U Tigerstedt[11], D Traynor[6] and R Voicu[8]

[1] CERN, CH-1211 Genève 23, Switzerland
[2] Institute of Physics, Academy of Sciences of the Czech Republic Na Slovance 2 182 21 Prague 8, Czech Republic
[3] STFC - Rutherford Appleton Lab. UK
[4] Karlsruher Institut für Technologie, Hermann-von-Helmholtz-Platz 1, D-76344 Eggenstein-Leopoldshafen, Germany
[5] INFN, Sezione di Milano, via G. Celoria 16, I-20133 Milano, Italy
[6] Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
[7] Deutsches Elektronen-Synchrotron, Notkestraße 85, D-22607 Hamburg, Germany
[8] California Institute of Technology, Pasadena, Ca 91125, U.S.A.
[9] Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
[10] Port dInformació Científica (PIC), Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain
[11] CSC Tieteen Tietotekniikan Keskus Oy, P.O. Box 405, FI-02101 Espoo, Finland

E-mail: `alastair.dewhurst@cern.ch, ipv6@hepix.org`

Abstract. The fraction of internet traffic carried over IPv6 continues to grow rapidly. IPv6 support from network hardware vendors and carriers is pervasive and becoming mature. A network infrastructure upgrade often offers sites an excellent window of opportunity to configure and enable IPv6.

There is a significant overhead when setting up and maintaining dual stack machines, so where possible sites would like to upgrade their services directly to IPv6 only. In doing so, they are also expediting the transition process towards its desired completion. While the LHC experiments accept there is a need to move to IPv6, it is currently not directly affecting their work. Sites are unwilling to upgrade if they will be unable to run LHC experiment workflows. This has resulted in a very slow uptake of IPv6 from WLCG sites.

For several years the HEPiX IPv6 Working Group has been testing a range of WLCG services to ensure they are IPv6 compliant. Several sites are now running many of their services as dual stack. The working group, driven by the requirements of the LHC VOs to be able to use IPv6-only opportunistic resources, continues to encourage wider deployment of dual-stack services to make the use of such IPv6-only clients viable.

This paper will present the HEPiX plan and progress so far to allow sites to deploy IPv6 only CPU resources. This will include making experiment central services dual stack as well as a number of storage services. The monitoring, accounting and information services that are used by jobs also needs to be upgraded. Finally the VO testing that has taken place on hosts connected via IPv6 only will be reported.

1. Introduction

The fraction of internet traffic carried over IPv6 continues to grow rapidly. Over one eighth of queries to google traffic go via IPv6[**?**]. Apple recently announced[0] that all Apps produced for their products must be able to work over IPv6 only networks. Large cloud providers such as Amazon[**?**] and Microsoft[**?**] provision dual stack machines, while some smaller cloud providers[0] will offer cheaper VMs if they are IPv6 only. Within the HEP community there are over 10 sites that have deployed dual stack storage, while others have expressed a desire to deploy IPv6 only WN. Not only does IPv6 provide a solution to the limited number of IPv4 address it also offers potential benefits. e.g., for security conscious sites, it is possible to assign every job that runs on their batch system a unique IPV6 address, this means that if suspicious behaviour is detected it becomes significant easier to trace the source.

2. IPv6 only CPU resources

Despite the advantages of IPv6 and the ever increasing deployment across the world, deployment at WLCG sites has remained slow. The following reasons were identified for this.

- No appetite from the LHC VOs. The ability to access data and run analysis jobs had not been a problem, so from their point of view, why change?
- There is an initial cost (primarily manpower) to setup IPv6 at a site as well as a small ongoing overhead running dual stack services.
- IPv4 address exhaustion was not affecting several of the larger WLCG sites (Tier 1s) who often lead when it comes to adopting new technologies.

The WLCG is expected to evolve under the assumption of flat cash funding for computing resources and it is therefore important that sites are not hindered in their procurement by unnecessary restrictions from the WLCG VOs. Hardware procurements often have a significant lead time and will often be in production for several years. Even if a site does not intend to switch to IPv6 any time soon, they may well be making procurement decisions now which will influence their decision to migrate some time in the next 5 years.

The HEPiX WG came to the conclusion that the only way to ensure that IPv6 adoption did not become a problem was to make a strategic decision to mandate the LHC VOs as well as all Tier 1 sites to provide a minimum level of IPv6 support. This would allow any other site to provide IPv6 support with confidence that support would be available in the event of problems.

In order to provide an incentive for sites to move it was also decided that any agreement must allow sites to completely migrate some services to IPv6. Sites traditionally provide both Storage and CPU. The current LHC Computing models allow transfers between any two sites and therefore for IPv6 only storage to be supported, all sites would need to provide dual stack storage. For CPU, they traditionally only talk to their internal sites services as well as a handful of VO boxes for the VO running the jobs. It was therefore decided that it was easier to allow CPU resources to be migrated completely to IPv6.

In July 2016 the HEPiX Working group submitted a proposal to the WLCG Management Board setting out a plan to allow sites, if they so choose, to deploy their CPU resources as IPv6 only.

3. IPv6 Peering and PerfSonar work

In the past the HEPiX IPv6 working group has requested that sites offer IPv6 peering over the LHCOPN.

- All Tier 1s to offer IPv6 peering to LHCOPN and provide dual stack PerfSONAR machine by April 2015.

- All Tier 2s to offer IPv6 peering to LHCONE and provide dual stack PerfSONAR machine by August 2015.

As this was a request from a working group rather than a mandate from the WLCG take up was not full. In the month since the WLCG agreement was approved, two additional Tier 1s (BNL and Triumf) are now peering over the LHCOPN.

4. Software validation

The HEPiX IPv6 working group has also invested a significant amount of time in validating software as IPV6 ready. Key storage software and protocols work: dCache, DPM, StoRM XrootD 4, GridFTP, http

5. Central service migration

5.1. CVMFS

All the WLCG VOs as well as many others distribute their software across the Grid using CVMFS. The software is uploaded to a Stratum-0 server (located at CERN for the WLCG VOs) which then mirrors the data to several Stratum-1 servers[0]. Jobs will access the VO software from a cache on the local disk; if the file is not available, it will be looked for in the site Squid server, which in turn, will contact a Stratum-1 if needed. Squid 3.x is IPv6 compliant and is being used in production by some sites. It is essential that the Stratum-1 service at CERN is upgraded to dual stack by April 2017. When possible the Tier 1 should upgrade their service to dual stack and all Tier 1s should be upgraded by April 2018 at the very latest.

5.2. FTS

ATLAS, CMS and LHCb all use the FTS service extensive for data movement around the Grid. Jobs do not contact the FTS service directly so it is not necessary for the FTS service to be dual stack. All VOs are encouraging sites to make their storage dual stack. Transfers via two dual stack service should go via IPv6, however it is the FTS server which initiates the negotiation and sends a PASV (on IPv4) or an EPSV (on IPv6) to the destination and sends the IP (for the corresponding protocol) and port to the source. Therefore all FTS services should be upgraded to allow transfers between dual stack sites to go over IPv6.

Currently the FTS service at CERN is dual stack. There are IPv4 only FTS services at RAL, BNL and Fermilab that are used by the LHC VOs. While it is possible to work around this all FTS services should be upgraded to dual stack when possible and by April 2018 at the very latest.

5.3. PerfSonar

PerfSonar instances are required at all WLCG sites to implement the network monitoring infrastructure. All Tier-1s were requested to provide a dual stack perfSonar instance and GGUS tickets have now been submitted to those that have not. PerfSonar is a very good way of checking that the migration to IPv6 hasn't caused any network/routing problems. All sites are requested to provide a dual stack PerfSonar instance by April 2018 at the latest. While it is not essential for all Tier 2s to migrate, it would be concerning if they are unable to provide a PerfSonar instance by this time. Any site unable to provide a PerfSonar instance by April 2018 will be requested to provide a clear description of their IPv6 plans.

5.4. ETF test infrastructure

A separate IPv6-only ETF test infrastructure will need to be set up to monitor IPv6-ready sites. This must be done by April 2017. This will be run in parallel to the production ETF test infrastructure. This service will provide sites with low level monitoring to help them identify problems with their IPv6 migration and not used for official availability metrics unless the site is providing some resources on IPv6 only. From April 2018 the official ETF infrastructure will be migrated to dual stack. From this point on production work going over IPv6 should be considered entirely normal. This will hopefully encourage sites to investigate IPv6 before April 2018.

5.5. Frontier Service

ATLAS and CMS both use the Frontier Service[0] to access conditions data across the Grid. The Frontier service has three components:

- Frontier client: This software is run by ATLAS and CMS jobs. It converts a conditions database query into an HTTP request. The Frontier Client was made IPV6 compliant in January 2016.
- Squid proxy: Sites are expected to deploy squid servers to cache the conditions data requests.
- Frontier Launchpad: This converts the HTTP requests back into database queries which are then submitted to the conditions database.

5.6. Other Services

There are several other services such as certificate authorities, software repositories, the GOCDB/OIM, GGUS, VOMS and the BDii. These are not used directly by jobs but are needed when configuring the site. These services should be made dual stack when possible and ideally by April 2018 (although some services might not fall under the WLCG banner). It will depend heavily on the site setup as to whether the lack of IPv6 connectivity will cause problems. Problems will have to be followed up by the HEPiX working group as they appear.

6. VO migration

The following section details the ongoing work from the LHC VOs to upgrade their services to allow jobs to run on IPv6 only CPU resources.

6.1. ALICE

Unlike the other LHC VOs, ALICE uses fully federated storage, any site can access the storage element of another site if needed (reading, writing and data transfers). Therefore in order to ensure all job types can run on IPv6 only CPU all data needs to be accessible over IPv6. Some data is stored on multiple sites and therefore it does not necessarily mean all sites will need to be dual stack. To support IPv6, the site storage elements need to run xrootd v.4. The central ALICE Grid services have been tested to run on IPv6 and are running in dual stack mode for over a year. For sites supporting ALICE the current situation is:

- One third of the sites are still running SEs with xrootd v.3.
- 5% of the SEs are running in dual stack mode, while the remaining are IPv4.

6.2. ATLAS

The ATLAS workload management system is called PanDA [0]. Pilot factories generate pilot jobs which are sent directly to CEs at sites. Once these pilots are started by the batch system, they will contact a central Panda Server to pull in a job (done via http). They will also contact the Rucio Server for File lookup (done via http) and the local storage. Some ATLAS jobs access Conditions data using the Frontier service. At the end of the job the pilot will write the output files to a local SE. Every 30 minutes while the job is running the pilot will report to the Panda server (via http). It will also contact the Panda Server at the end of the job. ATLAS jobs running on IPv6 WN will need access to the following resources:

- The production panda server nodes.
- The Rucio Authentication nodes.
- The Rucio Production nodes.
- The Frontier servers at CERN, IN2P3, RAL and Triumf.

The pilot factories that submit jobs to CEs have been made dual stack. ATLAS also use the ARC Control Tower (aCT) to submit jobs primarily to NorduGrid but potentially any sites running an ARC CE. This will also need to be made dual stack. ATLAS are working on making all these services dual stack by April 2017.

6.3. CMS

The job submission middleware, glideinWMS, is used to launch HTCondor worker nodes and its major components (frontend and factory). These have been validated as IPv6 compliant. Some glidein factories are already deployed in dual stack. HTCondor itself is fully IPv6-compliant, but the collectors and schedds still need to be all dual-stack in production in order to support IPv6-only worker nodes.

The central services hub, cmsweb.cern.ch, has been validated for dual stack operation. The CMS-specific job management systems (WMAgent for production and CRAB3 for analysis) have not yet been fully tested on IPv6, but they are expected to work with little effort needed. In any case, they do not need to be in dual stack for the foreseeable future.

The data management system, PhEDEx, uses the Oracle client for communication between local site agents and the central service. Tests have not yet been done, but Oracle 12c fully supports IPv6.

Concerning AAA, the CMS storage federation, only a very small fraction of the data is accessible using xrootd or GridFTP via IPv6. The global and regional redirectors are only partly on dual stack.

CMS plans to immediately start upgrading all services to dual-stack. Upgrades will be coordinated to minimise operational disruption and will be completed by the end of Run II. For services contacted by worker nodes (like HTCondor) these will be given priority and will aim to be done by April 2017.

At the time of writing, only eleven CMS sites expose IPv6 addresses for their services. This is proven not to create any problem, either in the ETF tests or for real production or analysis jobs. This should be taken into account by sites that need to evaluate the risks of deploying IPv6 in production. Having said that, CMS strongly recommends sites not to switch off their IPv4 networking until the end of Run II, as a risk mitigation measure.

6.4. LHCb

LHCb uses the DIRAC framework to submit jobs to the grid. DIRAC officially supports IPv6 and some other VOs, who use DIRAC, are already using a dual stack service in production. LHCb submits generic pilot jobs to CEs as needed. When these pilots start on a WN, they contact the LHCb DIRAC central services for available tasks (via the dips protocol) which are then executed. If input data is needed, they contact the relevant storages using the sites SRM[1] to access the data. Production jobs typically retrieve / download the data to the worker node, as they know exactly how much data is needed. User jobs stream data from the storage directly.

[1] The job is given a list of locations of the input files by DIRAC. It currently contacts the site SRMs in turn to retrieve the data. This will in future be updated to bypass the SRM and construct the file location automatically using the information available.

Once the job is done, it will upload the output to a storage location. If the default preferred location is not available, all other possible locations (available for LHCb) are tried in turn until successful and a request is set in the central services of LHCb to transfer the file to the preferred location when possible. If no location is available, the job ends up in status "failed", and could be resubmitted depending on the conditions.

LHCb jobs running on an IPv6 only WN will need access to the following resources :

- LHCb's DIRAC central services
- Storage services supporting LHCb
- Optionally, one of six VO-boxes at LHCb Tier-1 sites

Currently there is one Tier-1 storage and one Tier-2D storage that support LHCb in a dual-stack configuration. The LHCb central services are being moved to dual-stack machines.

7. Conclusions

The LHC VOs are committed to being able to work on the Grid over IPv6. Much work still remains to be done to make this a reality. The HEPiX IPv6 working group has validated that all essential software is IPv6 compliant. Software developers should consider IPv6 compliance a standard requirement and the emphasis should be on them to test this. All the VOs have analysed their workflows on the Grid and have provided a list of services which they will need to make dual stack. While exact time lines have not been agreed the amount of work required is sufficiently small that it should be achievable by April 2017 without significantly disrupting normal WLCG operations.

From April 2017 sites will be allowed to deploy IPv6 only CPU resources. Sites wishing to deploy IPv6 only CPU must deploy dual stack storage if they provide it. All sites are encouraged to upgrade their storage to dual stack. From the contact the HEPiX IPv6 working group has with sites, we believe that there are at most one or two sites that wish to urgently upgrade making up less than 2% of the pledged WLCG CPU resources. Any site wishing to upgrade should be in contact with the HEPiX IPv6 working group to ensure that the inevitable teething problems are resolved promptly. By April 2018 it should be possible to deploy IPv6 only CPU resources with relative ease and by the end of Run II enough sites should have upgraded their storage to dual stack to allow almost complete data availability via federated XrootD over IPv6.

References

https://www.mythic-beasts.com/servers/virtual

https://developer.apple.com/news/?id=05042016a

https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA

http://cernvm-monitor.cern.ch/cvmfs-monitor/atlas.cern.ch/

http://frontier.cern.ch/