WLCG IPv6 deployment strategy

HEPiX IPv6 Working group

June 21, 2016
Version 0.1

Executive Summary

This document describes the Worldwide LHC Computing Grid's (WLCG) strategy to allow sites to provide IPv6 resources to the LHC experiments. In summary:

- Sites can provide IPv6-only CPU resources from April 2017 onwards if necessary;
- Sites can provide IPv6-only interfaces to their CPU resources, if necessary;
- The VO infrastructure (e.g. central services provided by VOs) must provide an equal quality of service to both IPv4 and IPv6 resources;
- Sites should allow dual stack access to their storage resources, to allow remote access from IPv6-only resources.

Contents

1 Introduction

There are various motivations for WLCG sites to migrate services to IPv6. The most obvious is the exhaustion of the IPv4 address space, which is already putting constraints on some countries and institutions. 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.

Significant effort is also being put in by the WLCG community to investigate commercial cloud providers to see if they can provide resources (normally CPU) more cost effectively than traditional Grid sites. Some commercial providers charge more for machines with IPv4 connectivity over those with IPv6 only connectivity[1]. The commercial sectors adoption of IPv6 is significantly ahead of the WLCG. The rapid growth in the percentage of internet traffic going over IPv6 is expected to continue and large companies such as Apple now mandate that software should be validated on machines with IPv6 only connectivity[2].

The eventual goal with IPv6 deployment is for the entire internet to migrate to IPv6 only. However as IPv6-only machines are unable to talk to IPv4-only machines, during the migration, certain machines will need to have both IPv4 and IPv6 addresses (dual stack machines). One possible upgrade path would be for all sites to upgrade to dual stack before allowing any site to switch off IPv4. This is unworkable as some sites are already under pressure to migrate while others have not started thinking about it. As there is an additional overhead in running dual stack machines and the fact that the complete migration will take several years, where possible machines should be migrated directly to IPv6.

This document describes the required steps to allow existing and future sites and any opportunistic resource that may become available to provide IPv6-only CPU resources. In order to provide CPU resources, a site also needs to provide other services such as CEs, Squid caching proxies etc.; for this reason, when this document refers to IPv6-only CPU resources it means not only the WN but all related services can be IPv6-only.

In order not to penalise sites that choose to deploy IPv6-only CPU resources, central services need to not only work with IPv6 but should provide the same level of service (e.g. resilience and performance). Ideally the setup would be identical and in cases where this is not possible the differences should be clearly documented. Each of the WLCG VOs operate some central services which are usually hosted at CERN. CERN is already able to make these services dual stack. The WLCG also operates several central services, which should be made dual stack.

2 Site requirements within the current computing model

In the current computing model followed by the WLCG VOs a typical site provides CPU and Disk resources. Data can be transferred to the site from many other sites. In general jobs are still sent to the data therefore jobs running on a sites CPU resources normally only access the local storage. Some sites only provide CPU resources in which case they normally make use of a nearby site's storage.

For this reason, it should be possible for sites to directly upgrade their CPU resources to IPv6 as long as they upgrade their storage to dual stack first. For VOs to take advantage of [opportunistic] CPU only resources it will be necessary to have a nearby dual stack storage for them to connect to.

The WLCG VOs, to different extents, all make use of federated XrootD access. In the case of LHCb this is purely as a failover incase data access at the local sites fails. For ATLAS (FAX) and CMS (AAA), as well as making using of the failover mechanism, a small number of jobs make use of the XrootD federation to access files remotely. This is normally to take advantage of idle CPU at sites which lack the relevant data but have good connectivity to the site that does. ALICE uses a fully federated storage model.

The XrootD redirection mechanism has been designed to not direct requests from an IPv4-only source to an IPv6-only destination or vice a versa. Therefore if there are two copies of a file one on a dual stack storage and one on a IPv4-only storage a job on an IPv6-only machine should be able to access the file.

The original WLCG computing models used Tier 1 sites to distribute data to the Tier 2 sites. While this model has changed, Tier 1s are still extremely important as they have good connectivity, and provide access to a lot of data. They also still run many services that are required by VOs or other sites. It is therefore critical that the Tier 1s follow CERNs lead and provide dual stack access to their services.

2.1 Site Services

2.1.1 CPU

A site that provides CPU resources to the WLCG is likely to have deployed the following:

- Computer Element (CE): VOs submit their jobs to site CEs. The role of a CE is to convert this job submission over the Grid into something that the local batch system understands.
- Worker Nodes (WN): These provide the job slots where jobs run. Outbound connectivity is normally assumed.
- Squid proxies: These are used to cache requests from jobs to CVMFS and Frontier.
- Accounting: The number and usage statistics of jobs run is reported on a monthly basis to APEL.
- Information Provider: Site information is provided by the BDII, while usage of this service is dropping, it is still necessary for some functions.

There are some CPU resources (HPC, commercial clouds) that have different setups. However these normally place constraints on the VOs that would actually make it easier to be IPv6 compliant (e.g. no outbound connectivity). WN normally make up the majority of machines (and hence IP addresses) run by a site.

There are two likely upgrade paths. Some sites may wish to deploy all their CPU resources IPv6-only in one go. In this case all services that speak to any of these resources will need to be dual stack. Alternatively a site may wish to upgrade more slowly, they are likely to make their services dual stack and then migrate their WN to IPv6-only. When they are happy with the migration, they can drop the IPv4 support from the remaining services. In this case only the services that speak to WN will need to be made dual stack.

From April 2017 sites will be allowed to deploy IPv6-only CPU resources (and all related services). It is expected that the first few sites to migrate will choose a gradual upgrade which will hopefully avoid problems that could significantly affect the sites availability.

2.1.2 Disk

WLCG sites deploy a range of storage solutions. In general data can either be accessed directly from the storage node or via gateway (sometimes known as a doors or proxy) machines. Sites can use a variety of transfers protocols internally however the LHC VOs rely on the XrootD and GridFTP protocols, both of which have been shown to be IPv6 compliant. There is also a push within the WLCG to use http and this is also IPv6 compliant. Both dCache and DPM, the most popular storage services run by WLCG sites are already being run by a small number of sites as dual stack in production. Other storage service have been shown to be IPv6 complaint. For storage services which aren't IPv6 compliant (e.g. Castor) it is still possible to provide dual stack access via an XrootD/GridFTP dual stack gateway service.

All sites are encouraged to upgrade their storage to dual stack. Even if a site does not intend to migrate to IPv6 soon, if it provides external access to its services via dual stack gateways these will help the VOs with data access.

2.1.3 Tier 1 requirements

In addition to CPU and Disk resources, Tier 1s also provide Tape backed storage. Tape backed data is not used by jobs running on the Grid and there is no requirement at this time to make this service dual stack. Having said that most Tier 1 use dCache which provides a common interface to access both disk and tape back files so it is expected that tape backed service will become dual stack at the same time the disk is.

Tier 1s will be required to provide dual stack access to their storage with the following requirements:

- At least 1Gb/s and 90% availability by April 2017.
- At least 10Gb/s and 95% availability by April 2018.

Even if there are Tier 1s, that haven't started to think about IPv6, it should be possible to fulfil the April 2017 goal with a testbed setup which should be easily achievable. Any central services a Tier 1 provides will also need to be made dual stack with similar availability as for the storage.

2.2 Shared services

2.2.1 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[4]. 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[1] 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.

2.2.2 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.

2.2.3 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

[1] There is a bug in the handling of HTTP caching headers, whose resolution is expected for July.

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.

2.2.4 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.

2.2.5 Frontier Service

ATLAS and CMS both use the Frontier Service[5] 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.

2.2.6 Other Services

There are several other services such as certificate authorities, software repositories, the GOCDB/OIM, GGUS 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.

3 Experiment plans

The WLCG VO plans to allow IPv6 only CPU sites are detailed below. In general the motivation for VOs to support IPv6 is to be able to take advantage of any opportunistic resources that maybe IPv6 only. The VOs agree that sites should be able to migrate to IPv6 if it gives performance, cost or operational advantages. The VOs recommend all sites to upgrade their storage to dual stack and would expect a large number to have by the end of Run II. During Run II, the VOs still expect good site availability and reliability and where possible sites should retain their IPv4 connectivity until the end of Run II, even in a degraded form, as a precaution (e.g. don't hand back IPv4 addresses or completely decommission NATs if not necessary).

3.1 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.

3.2 ATLAS

The ATLAS workload management system is called PanDA [3]. 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.

3.3 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.

3.4 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 dips) which are then executed. If input data is needed, they contact the relevant storages using the sites SRM[2] 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.

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. There is one outstanding issue with the gLite software which has problems submitting to dual-stack cream CEs which needs to be fixed [6].

[2] 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.

4 Conclusion

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

[1] https://www.mythic-beasts.com/servers/virtual

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

[3] https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA

[4] http://cernvm-monitor.cern.ch/cvmfs-monitor/atlas.cern.ch/

[5] http://frontier.cern.ch/

[6] https://ggus.eu/index.php?mode=ticket_info&ticket_id=120586