WLCG and IPv6 - the HEPiX IPv6 working group

K. Chadwick[1], S. Campana[5], G. Chen[2], J. Chudoba[3], P. Clarke[4], M. Elias[3], A. Elwell[5], S. Fayer[6], Q. Fazhi[2], T. Finnern[7], L. Goossens[5], C. Grigoras[5], B. Hoeft[8], D. Kelsey [9], T. Kouba[3], F. Lopez Muñoz[10], E. Martelli[5], M. Mitchell[11], A. Nairz[5], K. Ohrenberg[7], A. Pfeiffer[5], F. Prelz[12], D. Rand[6], M. Reale[13], S. Rozsa[14], A. Sciabà[5], R. Voicu[14], T. Wildish[15]

[1] Fermi National Accelerator Laboratory, Batavia, Il 60510, U.S.A.
[2] Institute of High Energy Physics, 19B YuquanLu, Shijingshan District, 100049 Beijing, China
[3] Institute of Physics, Academy of Sciences of the Czech Republic Na Slovance 2 182 21 Prague 8, Czech Republic
[4] The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
[5] CERN, CH-1211 Genève 23, Switzerland
[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] STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxfordshire OX11 0QX, United Kingdom
[10] Port d'Informació Científica, Campus UAB, Edifici D, E-08193 Bellaterra, Spain
[11] University of Glasgow, Kelvin Building, University Avenue, Glasgow, G12 8QQ, United Kingdom
[12] INFN, Sezione di Milano, via G. Celoria 16, I-20133 Milano, Italy
[13] Consortium GARR, Via dei Tizii 6, I-00185 Roma, Italy
[14] California Institute of Technology, Pasadena, Ca 91125, U.S.A.
[15] Princeton University, Jadwin Hall, Princeton, NJ 08544, U.S.A.

E-mail: ipv6@hepix.org

Abstract. The HEPiX (http://www.hepix.org) IPv6 Working Group has been investigating the many issues which feed into the decision on the timetable for the use of IPv6 networking protocols in HEP Computing, in particular in WLCG. RIPE NCC, the European Regional Internet Registry, ran out of IPv4 addresses in September 2012. The North and South America RIRs are expected to run out in 2014. In recent months it has become more clear that some WLCG sites, including CERN, are running short of IPv4 address space, now without the possibility of applying for more. This has increased the urgency for the switch-on of dual-stack IPv4/IPv6 on all outward facing WLCG services to allow for the eventual support of IPv6-only clients. The activities of the group include the analysis and testing of the readiness for IPv6 and the performance of many required components, including the applications, middleware, management and monitoring tools essential for HEP computing. Many WLCG Tier 1 and Tier 2 sites are participants in the group's distributed IPv6 testbed and the major LHC experiment collaborations are fully engaged in the testing. We have worked closely with similar activities elsewhere, such as EGI and EMI. We are constructing a group web/wiki which will contain useful information for sites on the IPv6 readiness of the various software components. This includes advice on IPv6 configuration and deployment issues for sites (http://hepix-ipv6.web.cern.ch/knowledge-base).
This paper will describe the work done by the HEPiX IPv6 working group since CHEP2012. This will include detailed reports on the testing of various WLCG services on IPv6 including

data management, data transfer, workload management and system/network monitoring. It will also present the up to date list of those applications and services which function correctly in a dual-stack environment together with those that still have open issues. The plan for more testing on the production infrastructure with a dual-stack IPv4/IPv6 setup and the work required before the support of IPv6-only clients is realised will be described.

1. Introduction

The much-heralded exhaustion of the IPv4 networking address space has recently started, but while the backbone networks are fully ready to support IPv6 there is still a well-known lack of communities and institutes using IPv6. Many of the individual components of distributed applications have already been reported to be IPv6-ready, but to ensure a smooth transition by the HEP community a full-systems analysis is required. This is a time consuming and complicated endeavour. The HEPiX IPv6 Working Group (ref) has been investigating the many issues feeding into the transition to the use of IPv6 on the Worldwide Large Hadron Collider Grid (WLCG). Its activities include the analysis and testing of the readiness for IPv6 and performance of the many different components essential for HEP computing and planning for the impact on operations and security. The work of the group to date is presented in this paper together with its future plans.

2. IPv6: the general problem

3. IPv6: the general problem

The intention of the designers of the IPv6 was protocol was to make it full of appealing features, in order to push its adoption widely and quickly. The IPv6 specifications (RFC 1883) were set back in the 1995, when the Internet community realized that the classfull allocation policies of the time were causing a quick depletion of the address space.

Unfortunately for IPv6, the IPv4 problem was quickly fixed with the adoption of the classless allocations (CIDR, RFC 1519) and by the invention of Address and Port translation techniques (NAT, RFC 1631). These events, together with the fact that the IPv6 advantages were far less appealing than the cost of deploying it, put the protocol in a limbo where it stayed for almost twenty years, until IPv4 addresses became scarce again.

At the end of the first decade of the XXI century, Regional Internet Registries started warning the Internet community that IPv4 addresses were soon be exhausted and urged everyone to adopt IPv6. IPv6 was quite quickly deployed on the Internet backbones, but not where it would have brought the most of its benefits, at the client and content side. Plagued by the chicken and the egg problem (no users if no content, no content if no users), in 2012 finally some of the biggest content providers made the bold move to make their services available over IPv6. One year after, still the IPv6 global traffic counts as a negligible fraction of the total Internet traffic.

4. IPv6 at CERN

Many academic institutions, which joined the Internet when it was in a very early stage, are still enjoying the large allocations that were given in those days; thus they are lacking of any urge to move to IPv6.

This was the situation at CERN till 2011, when Server Virtualization started being used. The virtualization technique proved to be very effective and its adoption at CERN has grown exponentially. In 2012 , when the plan for the services to be run in the upcoming remote data centre in Wigner (Budapest, HU) was finalized, it became clear that something like 250,000 public IP addresses would be needed in the near future. At the same time, RIPE, the European Internet Registry, was announcing the adoption of a new conservative allocation policy that

would grant no more than 1024 IPv4 addresses to any requester. It could have been an impasse for the IT deployment plans, but luckily CERN had been testing with IPv6 since 1998 and in 2011 the management of the IT department approved the project to deploy IPv6 in the CERN campus and datacentres, when its need had yet to be proven.

For large enterprises like CERN, deploying IPv6 is not as simple as configuring a dozen of routers to be dual stack. The Network Management System and the Network database had to be made IPv6 aware, all the IPv6 information generated, all the basic network services configured (DNS, NTP, DHCPv6..). At the same time the network security had to be kept at the same level as always. After two years, the deployment is almost completed. Right in time to tackle the IPv4 exhaustion problem that most likely will hit CERN in 2015 when the Wigner datacentre will reach its full capacity. Many applications still cannot make use of IPv6, thus is very premature to deploy IPv6 only Virtual Servers. The strategy will be to deploy an hybrid solution where servers get a private IPv4 address and a public IPv6 one. The private IPv4 address will allow legacy applications to work within the CERN domain, while the public IPv6 address will allow world-wide reachability.

Hopefully the availability of LHC data over IPv6 will push IPv6 adoption in the large WLCG community.

5. The HEPiX IPv6 working group

6. The IPv6 testbed

7. IPv6 testing and results

7.1. Testing scenarios

In order to prepare the WLCG infrastructure for IPv6, the best approach is to identify beforehand the relevant use cases, starting from the simplest ones, to take into account the likely constraints from sites and finally to define realistic scenarios to be used for testing.

It is reasonable to assume that all central services must work in dual-stack mode, to be compatible with both IPv4 and IPv6 clients. Site services are strongly encouraged to be run in dual stack mode as well, lest they incur in limitations (for example in joining storage federations). On the other hand, clients (including users, software agents and jobs running on worker nodes) should be able to exclusively use either protocol: users may connect from arbitrary nodes (e.g. their laptops) while compute nodes at some sites might have only IPv6 public addresses.

To summarise, any testing should comply with these requirements:

- all services must be tested in dual stack
- all user clients must be tested in IPv4 and dual stack
- all batch nodes must be tested in IPv4, IPv6 and dual stack (not all configurations might be possible for a given site)

The use cases to test are of increasingly complexity and should be addressed in this order:

- basic job submission, either direct (via CREAM client, Condor-G, etc.) or via workload management services (EMI WMS, glideinWMS, PanDA, etc.)
- basic data transfer from/to a user node to/from a storage element
- third party data transfer (e.g. via FTS)
- production data transfer (e.g. via PhEDEX, DIRAC, etc.)
- conditions data access (e.g. via Frontier/squid)
- experiment software access (e.g. via CVMFS)
- experiment workflow, running a complete production/analysis task
- information system query (e.g. via BDII)
- job monitoring (e.g. via MonALISA, experiment dashboards, etc.)

In all cases, all relevant client/service combinations in terms of network protocols should be addressed. For simplicity, "auxiliary" services (ARGUS, VOMS, MYProxy, etc.) running only IPv4 may be used.

7.2. Point-to-point testing

We used the PhEDEx LifeCycle agent [2] to drive transfers between pairs of sites, using gridftp with the IPv6 connectivity flags. Filesizes were checked at the destination, and any failures recorded. Files were transferred in both directions between each site pair.

Initially, we simply tested connectivity and basic functionality. We also tested under specific conditions, e.g. to compare throughput and error rates with IPv6 vs. IPv4 connectivity. This was useful for debugging issues with firewalls etc.

Since March 2013 the transfer testbed has been running continuously, with more sites joining over time. Finally we have 11 sites transferring 1 GB files between each other. With this many sites, we have had to introduce a delay between successive transfers, to reduce load on the servers.

To date, we have transferred over 2 PB of data between the 11 sites over the 6 months since the testbed started continuous operations. This is 7% of the rate that CMS achieve in daily operations, so not an insignificant amount. The overall success rate for transfers is 87%, which is very high considering that the testbed was operated at-risk, with errors only detected when someone decided to look for them. There were periods when a site or site-pair had errors lasting for a week or more, and the testbed was left running to debug them. So we can conclude that gridftp transfers over IPv6 are in fact very reliable, given adequate hardware to run on.

Figure 1 shows the transfer results for the full mesh of sites. Transfers from a site are shown along the rows, transfers to a site are shown in the columns. All plots are scaled to an x-axis of 500 seconds (which corresponds to a transfer rate of 2 MB/sec), and only successful transfers are shown.

We can see that, in general, transfers were fast, the graphs mostly peak to the left. Some sites (IHEP, Chicago) have long tails for transfers out, though transfers to them are more successful.

7.3. Testing with PhEDEx transfers

We also tested with PhEDEx ([3]), the CMS data-placement system. We used two WLCG/GridPP sites (Imperial College London and Glasgow) with IPv6-enabled DPM storage elements, and transfers via a dual-stack FTS3 server at Imperial College. Transfers were throttled to limit the load on the servers, and have been running smoothly for nearly two months at the time of writing. Figure 2 shows the accumulated volume of data transfered from mid-August until early October 2013, by which point over 120 TB of data had been transferred with very few errors. This tells us that PhEDEx can indeed operate to CMS production standards with IPv6-enabled services.

7.4. Dual-stack hosts at Imperial College London

The WLCG/GridPP site within the HEP group at Imperial College London (UKI-LT2-IC-HEP) has configured a subset of its hosts to be dual-stack. The site currently runs dual-stack DNS, SSH, NFS, EMI 2 and EMI 3 CREAM CEs, EMI 2 Worker Nodes, ARC CE and dCache (headnode, SRM component only) services. Additionally, all BDII services including top and site BDIIs run in dual-stack mode. The Puppet configuration system and the local OS install system have been left IPv4-only. This set up was achieved in a number of stages. Initially, the local campus network team enabled stateless address autoconfiguration (SLAAC) on the subnet routers servicing the grid hosts. All hosts acquired an IPv6 address via autoconfiguration. No subsequent problems were observed and no hosts required IPv6 to be turned off in order to continue normal operations. Next, AAAA records were added to core services such as mail and

Figure 1. Transfer performance for the IPv6 testbed continuous transfers. A 1 GB file is transferred between each pair of sites, then deleted, then transferred again, continuously. The plots show the distribution of transfer duration times per site pair. The source site is named in the row, the destination site is named in the column. So the top-right plot shows transfers from Caltech to PIC, the bottom-left shows transfers from PIC to Caltech. The x-axis is in seconds, from 0 to 500 for each plot. The number inset in each plot shows the approximate number of transfers between that site pair in that direction.

LDAP. The DNS servers had static IPv6 addresses added. The IPv6 DNS server hostnames were added into resolve.conf on all hosts. AAAA and PTR records were added to the worker nodes and these were made to point at the SLAAC addresses. On relevant service hosts the IPv6 firewall was configured as appropriate. On hosts running a BDII service the IPv6 option was enabled explicitly (by setting BDII_IPV6_SUPPORT=yes in /etc/sysconfig/bdii).

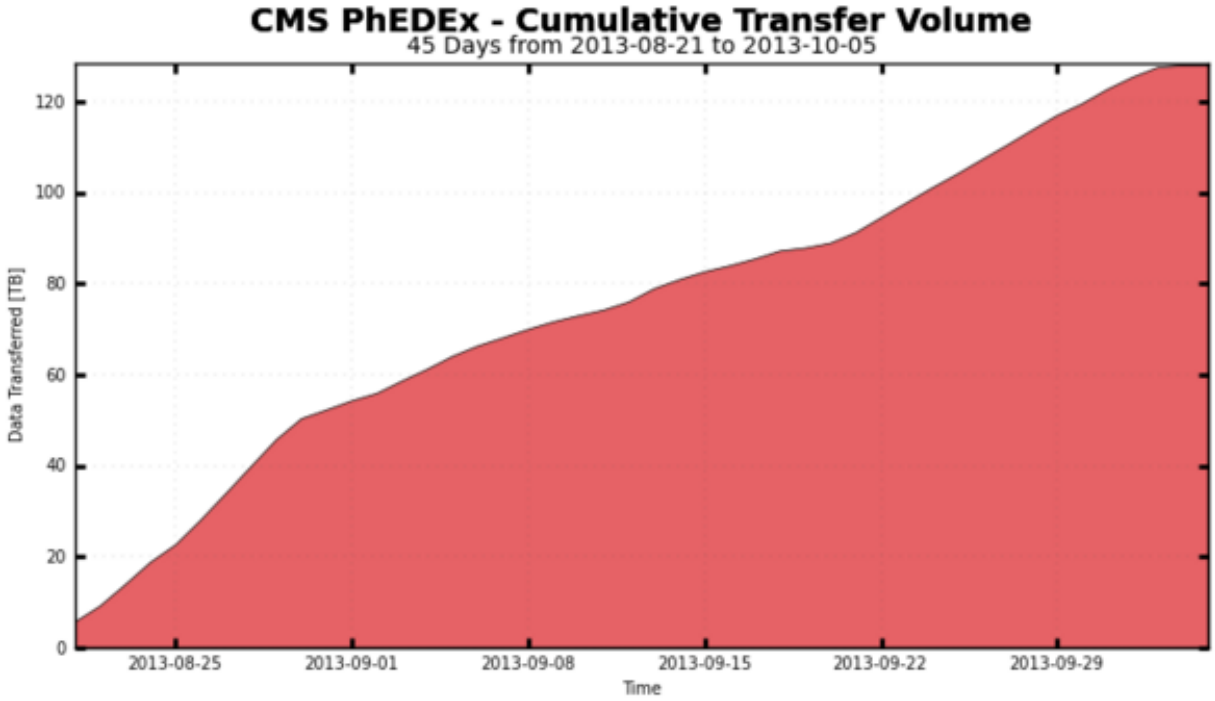

Figure 2. Cumulative data-transfer between Imperial College and Glasgow using PhEDEx on the IPv6 testbed.

8. Software and tools survey

9. Outlook and future plans

10. Conclusions

References

[1] S. Hogg, E. Vyncke, *"IPv6 security"*, Cisco Press 2008 ISBN-13: 978-1-58705-594-2

[2] *"Integration and validation testing for PhEDEx, DBS and DAS with the PhEDEx LifeCycle agent"*, CHEP 2013

[3] *"Egeland R, Wildish T and Metson S 2008 Data transfer infrastructure for CMS data taking"*, XII Advanced Computing and Analysis Techniques in Physics Research (Erice, Italy: Proceedings of Science)