WLCG and IPv6 - the HEPiX IPv6 working group

S Campana[1], K Chadwick[2], G Chen[3], J Chudoba[4], P Clarke[5], M Eliáš[4], A Elwell[1], S Fayer[6], T Finnern[7], L Goossens[1], C Grigoras[1], B Hoeft[8], D P Kelsey[9], T Kouba[4], F López Muñoz[10], E Martelli[1], M Mitchell[11], A Nairz[1], K Ohrenberg[7], A Pfeiffer[1], F Prelz[12], F Qi[3], D Rand[6], M Reale[13], S Rozsa[14], A Sciabà[1], R Voicu[14], C J Walker[15] and T Wildish[16]

[1] CERN, CH-1211 Genève 23, Switzerland
[2] Fermi National Accelerator Laboratory, Batavia, Il 60510, U.S.A.
[3] Institute of High Energy Physics, 19B Yuquanlu, Shijingshan District, 100049 Beijing, China
[4] Institute of Physics, Academy of Sciences of the Czech Republic Na Slovance 2 182 21 Prague 8, Czech Republic
[5] The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, 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] 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] Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
[16] Princeton University, Jadwin Hall, Princeton, NJ 08544, U.S.A.

E-mail: `david.kelsey@stfc.ac.uk, 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 (`http://www.ietf.org/rfc/rfc2460.txt`) networking protocols in High Energy Physics (HEP) Computing, in particular in the Worldwide Large Hadron Collider (LHC) Computing Grid (WLCG). RIPE NCC, the European Regional Internet Registry (RIR), ran out of IPv4 addresses in September 2012. The North and South America RIRs are expected to run out soon. 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 1/2 sites are participants in the group's distributed IPv6 testbed and the major LHC experiment collaborations are engaged in the testing. We are constructing a group web/wiki which will contain useful information on the IPv6 readiness of the various software components and a knowledge base (`http://hepix-ipv6.web.cern.ch/knowledge-base`). This paper describes the work done by the working group and its future plans.

1. Introduction

The much-heralded exhaustion of the IPv4 networking address space has finally started, but while the backbone networks have for many years been ready to carry IPv6 traffic there is still a well-known lack of applications using IPv6. To ensure a smooth transition for the applications of interest to the HEP community a full-systems analysis is required. This is a time consuming and complicated endeavour. The HEPiX IPv6 Working Group [**?**] was formed to investigate the many issues feeding into the transition to the use of IPv6 in HEP and in particular by WLCG. The group's activities include the analysis and testing of the readiness for IPv6 and performance of the many different components essential for WLCG 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

The intention of the designers of the IPv6 protocol was to make it full of appealing features, in order to push its adoption widely and quickly. The IPv6 specifications (RFC 1883/RFC 2460) [**?**] were set back in 1995, when the Internet community realized that the classful 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 21st century, the RIRs started warning the Internet community that IPv4 addresses would 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 later, the IPv6 global traffic is gradually increasing but still counts as a very small fraction of the total Internet traffic [**?**].

3. 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 routers to be dual stack. The Network Management System and the Network database had to be made IPv6 aware, all the IPv6 information generated and all the basic network services configured: the Domain Name Service (DNS), the Network Time Protocol (NTP), the Dynamic Host Configuration Protocol (DHCPv6), etc. 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 it is very premature to deploy IPv6-only Virtual Servers. The strategy will be to deploy a 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.

4. The HEPiX IPv6 working group

The HEPiX forum brings together worldwide IT staff, including system administrators, system engineers, and managers from the HEP and Nuclear Physics laboratories and institutes, to foster a learning and sharing experience between sites facing scientific computing and data challenges. At its semi-annual meetings, HEPiX had been considering the issue of migration to IPv6 for a number of years. At the end of 2010 a survey of HEP sites around the world was made asking about their plans for the deployment of IPv6. While it was very clear that there was no requirement for an urgent move to IPv6 a good number of sites were planning such a deployment and a few, particularly CERN, reported a foreseen lack of IPv4 address space in the not too distant future.

It was realised that any decision to deploy IPv6 on the WLCG infrastructure would involve much testing and planning and the decision was therefore taken to start a dedicated working group to investigate the issues. The HEPiX IPv6 working group was formed in 2011 with the following mandate.

Phase 1 of the work was to consider whether and how IPv6 should be deployed in HEP (especially for WLCG). This involved:

- A Readiness and Gap analysis
- The need to include all relevant HEP applications, middleware, security issues, system management and monitoring tools, and end-to-end network monitoring
- Running a distributed HEPiX IPv6 testbed to explore all of the above issues
- An initial report at the end of 2011

Following that initial report it was agreed that the work should continue and that phase 2 of the work should include:

- The proposal of a timetable and an analysis of the resources required for the deployment of IPv6 on WLCG
- The production of an implementation plan including advice to HEP sites on deployment

Since then the group has been working on the mandated tasks and meeting regularly with quarterly face to face meetings at CERN and monthly video/phone meetings to review progress. Full details are available at `http://indico.cern.ch/categoryDisplay.py?categId=3538`.

5. The IPv6 testbed

The impact of IPv6 is not limited to the transport layer but introduces the need for choice and preference in name-to-address resolution, implies multi-homing of all network endpoints (possibly on multiple protocol versions) and requires opaque handling of address information. This broadens the scope of code changes needed to add IPv6 support to existing code and adds to the complexity of testing: continued operation on IPv4 on dual-stack hosts, then preference of IPv6 and options to control it for all network bindings and connections need to be verified with adequate code coverage.

At opposite ends of the spectrum of practical testing options are testing of individual, isolated components and services and the analysis of integrated services on dual-stack nodes. Both approaches are incomplete:

- testing of isolated components misses the interaction with other services at the OS level and usually requires services to be configured differently than for production;
- testing of production-ready, integrated nodes may just be accidentally focusing on normal operation and bring insufficient code and functionality coverage.

This calls for a complementary approach, where individual services are deployed and tested within the scale of available dedicated resources and, once sufficient confidence and knowledge of their level and mode of IPv6 support is built, are watched in the context of a production node with dual-stack network and dual IPv4/IPv6 public address resolution.

Desirable characteristics of a dedicated testbed for single-service testing are:

- Geographical spread covering all ranges of realistic network latencies and as many network providers as possible.
- Uniform authentication/authorization scheme to factor out AA issues.
- Uniform OS installation, to factor out any issue with the custom configuration needed to test isolated services and for easier deployment of new services.

The current list of active testbed nodes can be found at the following URL:
`http://hepix-ipv6.web.cern.ch/testbed-nodes`
While we have at the time of writing a reasonable 9-site/6-National Research and Education Network (NREN) coverage of Europe, the only non-european sites in the testbed are IHEP Beijing in China and Fermilab in the US. More testbed sites are both needed and welcome to join to achieve a better match of our stated testbed goals.

As for OS distributions, testbed nodes are mostly installed with Scientific Linux (CERN) version 5, to replicate production conditions at LCG Tier-X centres. A few testbed nodes have Red Hat Enterprise Linux (RHEL) [**?**] version 6 derivatives installed: this allowed us to discover (and document in our knowledge base, `http://hepix-ipv6.web.cern.ch/knowledge-base`) that, rather unexpectedly, `libc` on RH6 causes unspecified protocol sockets to be bound on IPv4 only, instead of dual-stack as it used to be.

To achieve the simplest possible authentication scheme, a custom Globus Security Infrastructure (GSI) plug-in maps all members of our test Virtual Organisation (VO) (`ipv6.hepix.org`) to one local account, logging any access.

The first service we deployed for standalone testing through the testbed was GridFTP (or, more accurately, `gsiftp`). This was not only because of GridFTP's basic role in WLCG data transfer, but also because the FTP protocol (the GSI extensions don't affect but also suffer from this issue) is a paradigmatic example of how non-trivial IPv6 support can be. The original FTP specifications (RFC 765/RFC 959 [**?**]) used the quad-byte notation for IP addresses *in the syntax of the FTP protocol commands* `PORT` and `PASV`. This required the introduction, with RFC 2428 (September 1998) [**?**] of "extended" versions of the same commands, supporting different address families, and IPv6 in particular. We found on our testbed that support for the 'extended' command forms (and thus *implicitly* for IPv6) is missing from certain FTP client implementations. Retrofitting the clients with these commands is definitely more than a simple change in the transport layer and serves as an example of how ramified the introduction of 'IPv6 support' can be.

Building on this mesh of GridFTP servers, both continuous direct point-to-point file transfer tests and tests of the File Transfer Service (FTS), were successfully carried out. Storage Resource Manager (SRM) endpoints were also added, as described in detail in the next section.

6. IPv6 testing and results

6.1. 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 increasing complexity and should be addressed in this order:

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

6.2. Point-to-point testing

We used the PhEDEx LifeCycle agent [**?**] 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 is running continuously, with more sites joining over time. Finally we have 12 sites transferring 1 GB files among each other.

To date, we have transferred over 2 PB of data between the 12 sites over the 6 months since the testbed started continuous operations. This is 7% of the rate that CMS [**?**] achieve in daily operations, so is quite significant. The average success rate is 87%, which is very high considering that the testbed is operated at-risk. Errors are only detected when someone decides to look for them, and were often left unfixed to aid debugging. So we can conclude that gridftp transfers over IPv6 are very reliable, given adequate hardware to run on.

Figure **??** shows transfer results for the full mesh of sites. The source site is shown along the rows, the destination site is shown in the columns. All plots are scaled to an x-axis of 500 seconds (corresponding 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 Beijing, Chicago) have long tails for transfers out, though transfers to them are more successful.

6.3. Testing with PhEDEx transfers

We also tested with PhEDEx ([**?**]), the CMS data-placement system. We used two WLCG/GridPP sites (Imperial College London and Glasgow) with IPv6-enabled Disk Pool Manager (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, transferring over 120 TB of data. This tells us that PhEDEx can indeed operate to CMS production standards with IPv6-enabled services.

6.4. 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 LDAP. The DNS servers had static IPv6 addresses added. The IPv6 DNS server hostnames were added into `/etc/resolv.conf` on all hosts. AAAA and PTR records were added to the worker nodes and these were made to point to 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`).

7. Software and tools survey

For a successful transition to the use of IPv6 it is necessary to do a full survey of all important applications, middleware and operational tools. We decided to focus on the IPv6 readiness of the important WLCG outward-facing services and all essential applications and management and monitoring tools. We have created an online database of IPv6-readiness where for each software component considered we can store its known state of readiness and details of any testing performed by us or others.

IPv6-readiness is not a simple yes/no question. There are various stages of readiness to be addressed. Does the service break/slow down when used with IPv4 on a dual-stack host with IPv6 enabled? Will the service try using (connecting/binding to) an IPv6 address (AAAA record), when available from DNS? Will the service prefer IPv6 addresses from DNS, when preferred at the host level? Does this need to be configured and how? Can the service be persuaded to fall back on IPv4 if needed? In many ways the most important question is the first one. As long as a service deployed on a dual-stack host behaves properly for IPv4 then we are safe to recommend such a deployment on the WLCG production infrastructure. The current state of our survey may be seen at `http://hepix-ipv6.web.cern.ch/wlcg-applications`.

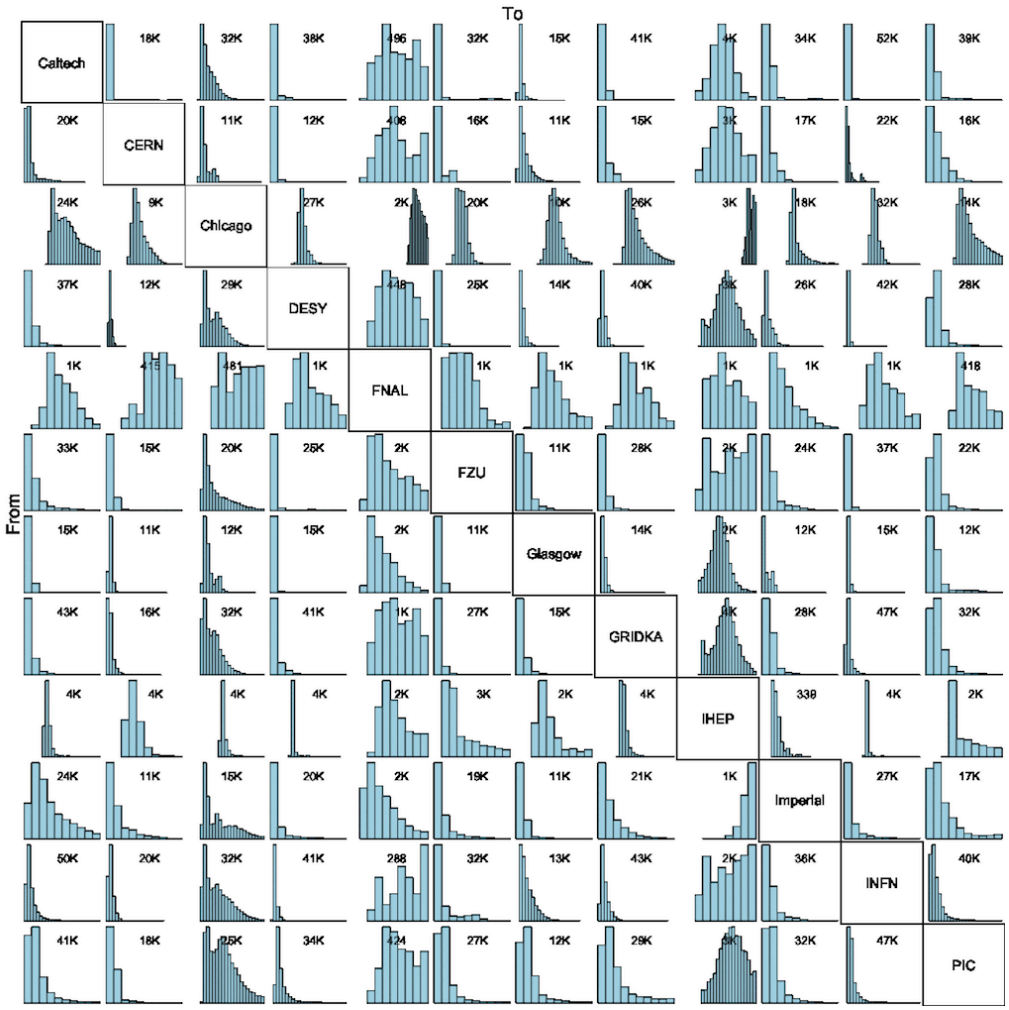

Figure 1. Transfer performance for the IPv6 testbed continuous transfers. A 1 GB file is transferred between each pair of sites, then deleted, then transferred again, continuously. The plots show the distribution of transfer duration times per site pair. The source site is named in the row, the destination 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.

At the time of writing, there are still many packages which need further investigation and testing. Software known not to be ready for IPv6 at this time includes OpenAFS servers and clients, all but the latest release of dCache and many batch systems. Full details of all such problems and investigations will be recorded in our online database.

8. Outlook and future plans

The IPv6 working group has made good progress but has to date only tested a small fraction of the many software components required by WLCG. There are still many more tests and assessments to be made and advice on dual-stack deployment to be formulated before our work can in any way be defined as complete.

Testing will continue in three areas. Firstly we will continue the mesh of data transfer tests

between the testbed sites, expanding to include new types of storage element and allowing for the ongoing assessment of reliability. Some testbed sites are now working on the deployment of larger scale IPv6 testbeds to allow the testing of IPv6-readiness in a more realistic production environment. Finally, we will encourage more Tier 2 sites to repeat the tests at Imperial College (see section 6.4) by enabling dual-stack services on some or all of their production services. Across all of these different scenarios we will gradually work through the list of testing scenarios presented in section 6.1 together with representatives of the experiments. During all of this testing we will continue to update our online database with the details of IPv6 readiness.

At the time of our status report to HEPiX and the WLCG Management Board in 2012 we concluded that the support of IPv6-only clients on WLCG was unlikely to be possible before January 2014. At the time of writing this is still true and indeed it is now clear that it will take much longer. Not only does the working group still have many tests to perform but all of the many Tier 1 and Tier 2 sites need to complete the deployment of an IPv6 infrastructure at their site. This needs to include the revision of local procedures and management tools and the provision of adequate training for network, system operations and security staff. Once we have achieved this we can propose a more general deployment of production-level dual stack services thereby allowing for the eventual support of IPv6-only clients on WLCG.

References

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

[2] 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`

[3] See for instance `http://www.google.com/ipv6/statistics.html`. The 2% global connectivity threshold was crossed in September 2013.

[4] `http://www.redhat.com/products/enterprise-linux/`

[5] C. Aiftimiei, P. Andreetto, S. Bertocco, S. Dalla Fina, A. Dorigo, E. Frizziero, A. Gianelle, M. Marzolla, M. Mazzucato, M. Sgaravatto, S. Traldi, L. Zangrando, Design and Implementation of the gLite CREAM Job Management Service, Future Generation Computer Systems, Volume 26, Issue 4, April 2010, pp. 654-667, doi: 10.1016/j.future.2009.12.006.

[6] M. Cecchi, F. Capannini, A. Dorigo, A. Ghiselli, F. Giacomini, A. Maraschini, M. Marzolla, S. Monforte, F. Pacini, L. Petronzio, F. Prelz, "The gLite Workload Management System, in proceeding of GPC 2009 conference

[7] T Maeno 2008 "PanDA: distributed production and distributed analysis system for ATLAS" J. Phys. Conf. Ser. 119 062036

[8] The gLite File Transfer Service P. Kunszt, P. Badino, R. Rocha, J. Casey, A. Frohner, G. McCance In Workshop on Next Generation Distributed Data Management at HPDC06, Paris, France

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

[10] Status and future perspectives of CernVM-FS J Blomer et al 2012 J. Phys.: Conf. Ser. 396 052013

[11] Field, Laurence, and Markus W. Schulz. "Grid Deployment Experiences: The path to a production quality LDAP based grid information system." Proceedings of the International Conference on Computing in High Energy and Nuclear Physics (CHEP 2004). 2004.

[12] Legrand I, Newman H, Voicu R, Cirstoiu C, Grigoras C, Dobre C, Muraru A, Costan A, Dediu M and Stratan C 2009 MonALISA: An agent based, dynamic service system to monitor, control and optimize distributed systems *Computer Physics Communications, Volume 180, Issue 12, December 2009, Pages 24722498*

[13] T. Wildish, *"Integration and validation testing for PhEDEx, DBS and DAS with the PhEDEx LifeCycle agent"*, also presented at CHEP 2013

[14] The CMS Collaboration 2008 The CMS experiment at the CERN LHC *JINST* **3** *S08004*

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

[16] M. Salichos, O. Keeble, A. Alvarez Ayllon, M. Kamil Simon, *"FTS3 - Robust, simplified and high-performance data movement service for WLCG"*, also presented at CHEP 2013.