EPJ Web of Conferences **295**, 07036 (2024) https://doi.org/10.1051/epjconf/202429507036
CHEP 2023

Overcoming obstacles to IPv6 on WLCG

Marian Babik[1], *Martin* Bly[2], *Nick* Buraglio[3], *Tim* Chown[4], *Dimitrios* Christidis[1], *Jiri* Chudoba[5], *Phil* DeMar[6], *José* Flix Molina[7], *Costin* Grigoras[1], *Bruno* Hoeft[8], *Hiro* Ito[9], *David* Kelsey[2*], *Edoardo* Martelli[1], *Shawn* McKee[10], *Carmen* Misa Moreira[1], *Raja* Nandakumar[2], *Kars* Ohrenberg[11], *Francesco* Prelz[12], *Duncan* Rand[13], *Andrea* Sciabà[1], and *Tim* Skirvin[6]

[1]European Organization for Nuclear Research (CERN), CH-1211 Geneva 23, Switzerland

[2]UKRI STFC Rutherford Appleton Laboratory (RAL), Harwell Campus, Didcot OX11 0QX, UK

[3]Energy Sciences Network (ESnet), Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA

[4]Jisc, Portwall Lane, Bristol BS1 6NB, UK

[5]Institute of Physics, Academy of Sciences of the Czech Republic, Prague 8, Czech Republic

[6]Fermi National Accelerator Laboratory (FNAL), P.O. Box 500, Batavia IL 60510, USA

[7]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain

[8]Karlsruhe Inst. of Technology, Hermann-v-Helmholtz-Pl. 1, D-76344 Egg.-Leopoldshafen, Germany

[9]Brookhaven National Laboratory (BNL), 98 Rochester St., Upton NY 11973, USA

[10]University of Michigan Physics, 450 Church St, Ann Arbor, MI 48109, USA

[11]Deutsches Elektronen-Synchrotron (DESY), Notkestraße 85, D-22607 Hamburg, Germany

[12]INFN, Sezione di Milano, via G. Celoria 16, I-20133 Milan, Italy

[13]Imperial College London, South Kensington Campus, London SW7 2AZ, UK

Abstract.

The transition of the Worldwide Large Hadron Collider Computing Grid (WLCG) storage services to dual-stack IPv6/IPv4 is almost complete; all Tier-1 and 94% of Tier-2 storage are IPv6 enabled. While most data transfers now use IPv6, a significant number of IPv4 transfers still occur even when both endpoints support IPv6. This paper presents the ongoing efforts of the HEPiX IPv6 working group to steer WLCG toward IPv6-only services by investigating and fixing the obstacles to the use of IPv6 and identifying cases where IPv4 is used when IPv6 is available. Removing IPv4 use is essential for the long-term agreed goal of IPv6-only access to resources within WLCG, thus eliminating the complexity and security concerns associated with dual-stack services. We present our achievements and ongoing challenges as we navigate the final stages of the transition from IPv4 to IPv6 within WLCG.

*e-mail: david.kelsey@stfc.ac.uk

1 Introduction

The Worldwide Large Hadron Collider Computing Grid (WLCG) infrastructure spans over 170 computing centers in over 40 countries. Over the past few years, the WLCG community has been working towards increased use of the Internet Protocol version 6, known as "IPv6", alongside the old Internet Protocol version 4, called "IPv4", in a mode known as "dual-stack", with the ultimate goal of, at some point in the future, running "IPv6-only".

The main drivers for the use of IPv6 in WLCG continue to be the lack of routable IPv4 addresses and the requirement for WLCG to support the use of compute resources that only communicate using IPv6. Another driver for the use of IPv6 in data transfers is the requirement to analyze traffic flows and understand the use of the LHC experiment network. The Scitags [1] initiative has developed an IPv6-specific solution to identify the owner (community) and associated purpose (activity, such as rebalancing) of network traffic, as described in another CHEP2023 paper.

To help this initiative, the HEPiX IPv6 Working Group [2] has investigated many issues related to moving WLCG services to dual-stack IPv6/IPv4 networking and to enable the use of IT resources that only communicate over IPv6 as agreed by the WLCG Management Board and presented by us at CHEP2018 [3].

Dual-stack deployment is a more complex networking environment than just using IPv6. The agreed endpoint of the WLCG transition is to have IPv6-only sites and services, to remove the complexity and security concerns of operating dual stacks. This was presented by us at CHEP2019 [4].

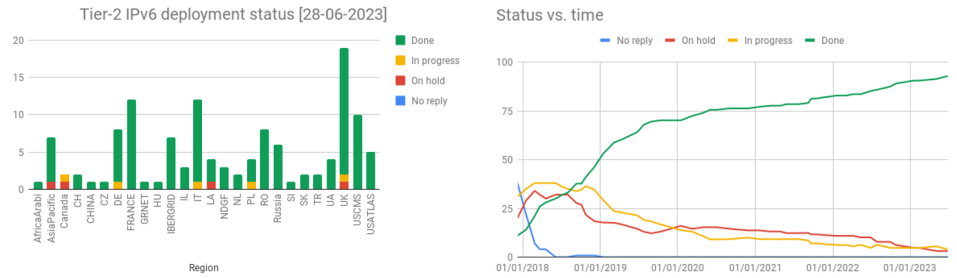
The deployment of dual-stack on WLCG is now nearing completion. During the last year, the group has continued to investigate the remaining obstacles to the use of IPv6. One very interesting finding is that a significant number of transfers still use IPv4 even when both endpoints have been made dual stack. Thus, an important activity has been fixing the reasons for such data transfers between dual-stack endpoints that still take place over IPv4.

This paper reviews the status of dual-stack storage deployment in WLCG as it nears 100% completion, discussing the obstacles found and how they were addressed, while also presenting those that have not yet been resolved and noting those that are outside the control of the WLCG.

2 Status of the IPv6/IPv4 dual-stack transition at WLCG sites

IPv6 deployment in the Tier-1 storage systems has been completed and has also made substantial progress at the Tier-2 sites, even if we are now years past the original implementation deadline of the end of 2018, which was never officially extended. Many WLCG sites have experienced difficulties meeting the objective; while it has been possible to deploy IPv6 within WLCG facilities at sites, it has often proven more challenging for the campuses through which they connect to do so, given the additional complexities there and the refresh cycles of campus networking hardware. However, both the WLCG and the campus IT teams have been largely successful and now the fraction of sites that have completed IPv6 deployment in their storage systems is now 93% and is still increasing, as can be seen in Figure 1.

The fraction of Tier-2 storage capacity, not the number of Tier-2 sites, accessible via IPv6 is shown in Table 1 for each experiment and for WLCG as a whole; two experiments have completed the deployment (CMS and LHCb) and the other two are in very good shape, at more than 90%, with the overall figure for WLCG at 94% at the time of writing.



(a) By region. **(b)** Time evolution.

Figure 1: Tier-2 deployment status.

Table 1: Fraction of Tier-2 storage available over IPv6.

	ALICE	ATLAS	CMS	LHCb	WLCG
Tier-2 storage	91%	90%	100%	100%	94%

Another important indicator of progress is monitoring, to quantify the level of adoption of IPv6, for example, in data transfers. There are several systems providing information: perfSONAR, which tests network links between sites separately for IPv4 and IPv6; ETF, which submits functional tests to site services, optionally via IPv6; the FTS monitoring, which is supposed to show which IP version was used for each data transfer, and custom network utilization plots, e.g., specifically for LHCOPN/LHCONE traffic.

3 Obstacles to IPv6 found

Several obstacles have been delaying the deployment of IPv6 and identifying them has been an important task of the HEPiX IPv6 Working Group. These are the main obstacles that we have been addressing:

- Certain applications and middleware that do not yet support IPv6
- WLCG sites (or their campus networks) that did not yet deploy IPv6 networking
- The campus site networks have IPv6 but the Tier-2 has no dual-stack storage
- IPv6 monitoring is not available or broken
- A service is dual-stack but IPv4 is still being used.

Support for IPv6 in worldwide LHC Computing Grid (WLCG) applications was one of the first obstacles tackled by the Working Group. Applications were first tested in a dual-stack environment, and problems were reported to developers. We maintain a list of applications with an indication of their IPv6 compliance. It took a long time, but most of the incompatibilities have been removed, and today the most important applications work properly with IPv6.

The lack of IPv6 deployment at the sites has been another major obstacle. Deploying IPv6 is relatively easy in a small site, but deploying it at production level in a large site with all the IPv4 functionalities (DNS, firewalling, address management, etc.) is a large and expensive

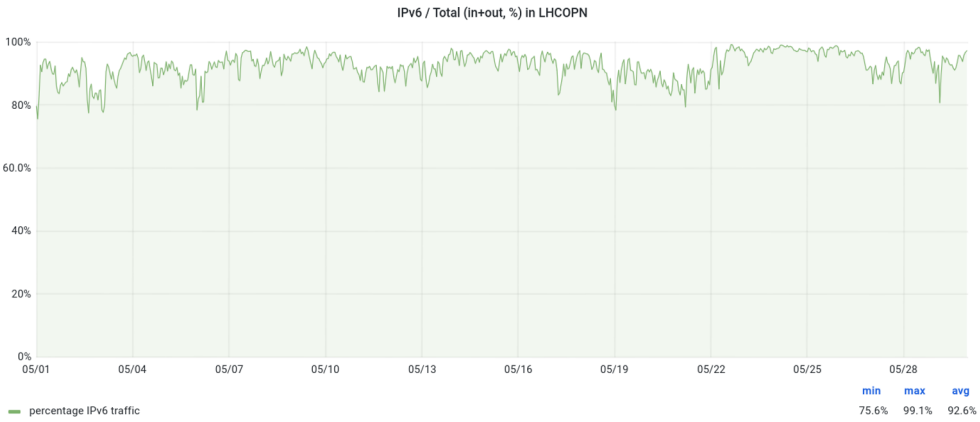


Figure 2: Percentage of IPv6 traffic on LHCOPN seen at CERN

task. Fortunately, the management of some large sites understood the risk of delaying this deployment and made dual-stack networks a reality, often by focusing the deployment on the WLCG elements of their sites (rather than a full campus-wide deployment).

But even when sites had IPv6-capable networks, IPv6 traffic was lacking. We soon realized that the most important service to provide IPv6 capabilities was storage, so the Working Group proposed to WLCG to mandate the deployment of dual-stack storage. WLCG accepted and we started a Global Grid User Support (GGUS) ticket campaign asking every site to report on their progress. The campaign is still ongoing, but a critical mass was reached quickly and today more than 90% of the Large Hadron Collider Optical Private Network (LHCOPN) traffic is IPv6.

We realized that it was not always possible to count data transfers using IPv6. The File Transfer Service (FTS) developers made changes so that the monitoring could distinguish the IP protocol used for the transfers.

More was done to improve network monitoring, to distinguish traffic by protocol on the main network links. In LHCOPN we first used the sflow data generated by the CERN routers, but once they were replaced to support 400 Gbps links, reliable sflow data was no longer available. So more work was done to separate IPv4 and IPv6 in different VLANs on all LHCOPN routers. Figure 2 shows the new Grafana visualization panel showing the percentage of IPv6 traffic in total. The percentage was encouragingly high compared to the data collected several months earlier, when we were able to use sflow data.

Another subtle obstacle encountered was that although all the pieces were in place (dual-stack network, IPv6-capable software, service DNS names with correct AAAA records) clients were still preferring IPv4. In most cases, it was some hidden or forgotten setting within the application that forced the use of IPv4. As an example, Figure 3 shows data transfers to US / ATLAS Great Lakes Tier 2 (AGLT2) where it is possible to see the moment the variable *java.net.preferIPv6Addresses* in dCache was set to *True*.

4 Obstacles to IPv6 - found and fixed

In this section we describe how data available in the monitoring platform at CERN can be harvested to identify machines configured with IPv4-only or IPv4 preference. This allows corrections to be made to the IPv6 configuration in readiness for IPv6-only operation.

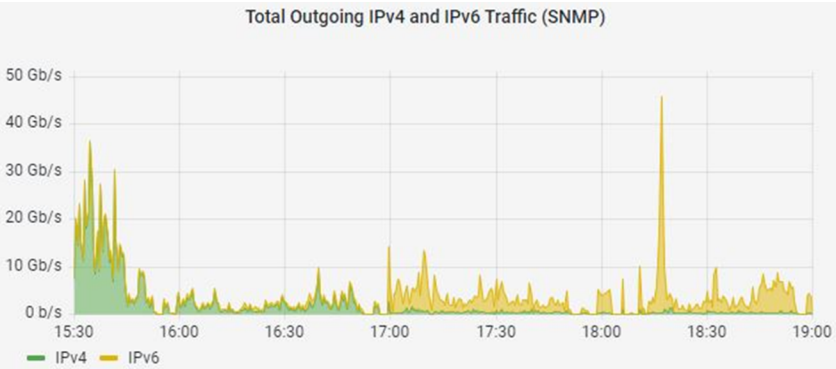


Figure 3: Data transfers into US / ATLAS Great Lakes Tier 2 (AGLT2)

4.1 Analysis of data transfer applications and data management systems

In this way, by taking the data transfers and available resources within CERN Monit as a starting point, we conducted an extensive analysis of the HTCondor, XRootD and FTS data exchange events. The main objective of this analysis is to identify endpoints that lack dual-stack configuration. In fact, this is a crucial step in removing obstacles and misconfigurations in the network, since any machine located at either end of the data transfer that is not IPv6-ready will not be executed via IPv6. Therefore, it is necessary to modify the configuration of these machines to enable the utilization of IPv6 and also establish a preferential mode for IPv6. At this point, it can be mentioned that in this workload, when we declare that a machine is lacking dual-stack configuration, it means one of these scenarios:

1. The DNS server is not configured to respond with an IPv6 address despite the machine having an assigned address.

2. The machine itself does not have an IPv6 address configured as part of its network configuration.

Firstly, regarding the condor job analysis, we examined the *monit_prod_condor_raw_metric** queries, which contain a record of more than 500 lines of information for each submitted job. In particular, we focus on the *schedd_name* and *startd_principal* pairs, which represent the source and destination endpoints involved in data transfers.

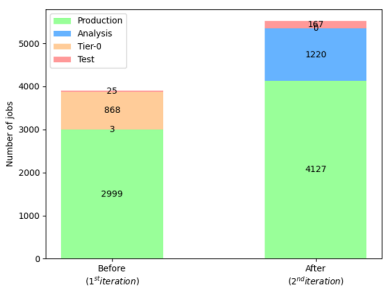


Figure 4: HTCondor data transfers where the machines at both ends were dual-stacked.

In this analysis, we inspected a total of 17813 data exchange events involving various source and destination endpoints, aiming to verify their status regarding the dual-stack configuration. We discovered that 3895 machines were already dual-stack configured at both ends of the transfer where, 2999 belong to production jobs, 3 to analysis jobs, 25 to test, and 868 to Tier-0 jobs, as specified in the job submission classification query. Consequently, based on the results, we generated a comprehensive list of machines exhibiting non-dual-stack configurations that were provided to the corresponding network administrator to initiate the necessary configuration modification to be IPv6-ready in those machines.

Subsequently, after a few weeks, we re-executed the analysis where, in this second iteration, we analyzed a total of 16072 data exchange events, revealing a significant improvement in dual-stack implementation as shown in Figure 4. Remarkably, 6382 of those events now featured dual-stack endpoints where 4127 belong to production jobs, 1220 to analysis jobs, 167 to test and 0 to Tier-0 jobs, which represents a notable increase of nearly 2500 endpoints and events moving from IPv4-only to dual-stack configuration.

Similarly, we analyze the XRootD data transfers *monit_prod_xrootd_enr_transfer** queries, with more than 100 lines of information for each job submitted. In this case, we considered the parameters *client_host* and *server_host* to check whether the machines at both ends of the transfer were dual-stack configured or not. In this analysis, we inspected a total of 6660 data exchange events, involving diverse source and destination endpoints, including read and write operations. Among these events, we found that in 6625 data exchange events, both end machines were already dual-stacked belonging to ATLAS, CMS, and LHCb experiments. However, contrary to the HTCondor analysis, here extensive changes in the network configuration were not required, as the vast majority of machines were already correctly configured as dual stack. In fact, several IPv6-only events were detected where the machines in both are dual stacked, which is our desired scenario.

Lastly, we conducted an analysis of FTS transfers using *monit_prod_fts_enr_complete** queries, which provide more than 150 lines of information for each job submitted. In particular, we focus on the *src_hostname* and *dst_hostname* fields to verify whether both endpoints involved in the data transfer were configured as dual stacked or not. We analyze a total of 1094 data exchange points, involving diverse origin and destination points. Among these events, we found that 814 of them were already configured with dual-stack belonging to ATLAS, CMS, and LHCb experiments. Once again and contrary to the HTCondor analysis, given that the overwhelming majority of machines were already correctly configured, no major changes were required.

4.2 Analysis of LHCONE/LHCOPN toptalkers

After addressing and reducing obstacles related to misconfigured machines in data transfer applications and data management systems, we proceeded with an analysis of IPv4 toptalkers within the LHCONE/LHCOPN networks. Similarly, our goal is to identify all machines that were not configured with dual stack and subsequently notify the service manager to prepare the machines for IPv6-ready and also establish IPv6 as the preferred mode.

In this analysis, we considered the IPv4 toptalkers from the last year, a total of 1200 machines with 100 toptalkers recorded for each month. We identified 12 machines with misconfigurations and contacted the service managers, who made the appropriate adjustments, leading to a substantial increase in transfers moved from IPv4 to IPv6. To attain greater granularity, we further focused on the 100 IPv4 toptalkers each week.

5 IPv4 to IPv6 Worker Node farm migration at KIT

As an example of the way sites have been migrating, in this section we present a detailed description of the migration activities at the German WLCG Tier-1 GridKa at KIT, where the local team started to build an IPv6-only Worker Node (WN) testbed. Unfortunately, issues were encountered with several components of the testbed including DNS, WN installation, squids, CernVM File System (CVMFS) as well as monitoring. Therefore, a new approach was adopted where instead the production WN farm was migrated to IPv6. Detailed monitoring was needed to capture all incoming and outgoing communication in each WN. The

package packetbeat was installed on all WNs and the data are forwarded to OpenSearch, the former Elasticsearch, for storage. The stored data are then analyzed and visualized with Kibana. All details of the WN traffic are available including IPv4 and IPv6 utilization, and protocols and ports. The monitoring system grew from a small WN subset to the complete WN farm. Keeping the size of the storage database below 0.5 TB allows for only the storage of about 6 days.

The amount of IPv6 traffic has grown from 25% of the total in May 2022 to 80% one year later, as shown in Table 2.

Table 2: IPv6 ratio change over time

Date :	May 2022	July 2022	Dec. 2022	May 2023
IPv6 percentage :	25%	53%	67%	80%

The KIT team found that even with WNs and the domain name server being dual-stack, the name resolving happened over IPv4 98.5% of the time and only a small amount over IPv6. There is a file *resolve.conf* in the configuration directory */etc/*, of each WN. If the first lines point to the IPv4 address, communication will hardly run over IPv6. After first changing the order to IPv6 addresses, the IPv6 addresses of the DNS server ratio will gradually change over time. Today, almost 85% of the DNS services are already running on the IPv6 protocol.

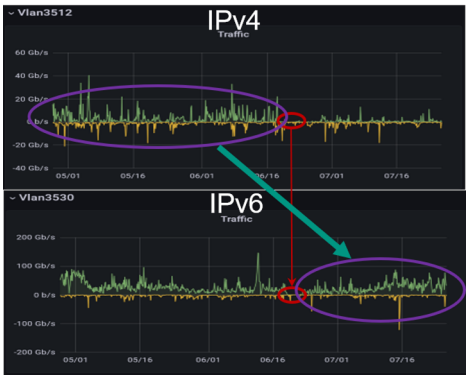


Figure 5: dCache upgrade 6.2.34 to 7.2.15

One major step during the migration towards IPv6 was an upgrade of the storage system, dCache. In June 2022 dCache was upgraded from version 6.2.34 to 7.2.15. The new release brought several significant IPv6 improvements. Even with gridsite delegation, HTTP-"third part copy" now prefers IPv6 addresses. The handling of storage resource reporting requests can be handled over IPv6, and the storage resource manager logging over IPv6 is fixed. These changes, moving storage to storage communication, were a big step toward IPv6, and even communication between WNs and storage was positively impacted as shown in Figure 5.

Although all squid web caches in the CVMFS were dual stack, the communication was still using IPv4. After configuring the flag *"cvmfs_ipfamily_prefer=6"* at all WNs, almost all communication between WNs and squids now uses IPv6. Only the CVMFS frontier ignores that flag, and communication is still running via IPv4. Disabling IPv4 will smoothly move traffic to IPv6.

Initially, the local resource management, LRMS/batch, was not dual-stack enabled. After dual-stack deployment and setting preferences towards IPv6, the earlier two-third IPv4 transfer volume was shrinking to one tenth. The team found that most IPv4 communication is between WNs and outside resources, and only 20% of IPv4 traffic is local to their site. WN hibernation state control via the HTCondor rooster daemon was also moved from IPv4-only to IPv6 at the end of 2022 when the server was deployed as dual-stack.

As described above, several issues have been solved. However, there are more tasks ahead. The Network File System (NFS) migration is still pending, since the NFS connecting driver for the underlying General Parallel File System (GPFS) must be thoroughly tested for a smooth transition with the IPv6 protocol.

The GridKa monitoring system can show very detailed WN communication behavior. Kibana allows for a differentiated search regarding all required subjects; for example, a list of all host communication to special destination ports which differentiates between IPv4 and IPv6. This will allow the team to identify and fix more areas where the IPv6 migration is not yet complete.

6 Further obstacles - still to be fixed

We have identified additional obstacles to the deployment of IPv6. These are either outside of our control and therefore require encouragement or persuasion, or they are obstacles that we have not yet fully addressed. These are as follows:

Non-storage services, for example "compute", are not yet dual-stack. Approximately 60% of all WLCG services are currently dual-stack enabled. Sites that use dual-stack services are not having any problems. The working group will seek WLCG Management Board support for a new campaign of operational tickets to request that all sites deploy IPv6 on all their services.

WLCG client CPU (worker nodes, virtual machines, or containers) that are still IPv4-only. Sites running dual-stack CPU clients do not experience problems. We plan to include the request to deploy dual-stack CPU in the ticket campaign described above.

Services/clients outside of WLCG Tier-1/Tier-2 have not yet been studied. The working group has concentrated on Tier-1 and Tier-2 sites. LHC computing has additional resources under investigation, including Tier-3 resources, Public/Commercial Clouds, and new Analysis Facilities [5]. The working group must investigate the status of IPv6 deployment in these services and encourage IPv6 deployment, especially for access to other WLCG services.

Use of new or evolving technologies not yet tested or tracked. The WLCG infrastructure is constantly evolving. New CPU architectures, e.g. GPUs, non-X86 CPU, etc. have not yet been tested for IPv6 compliance. Problems are not expected, but the systems should be tested before we encourage the use of IPv6.

Conflicting priorities can be the obstacle. It is common for service and client managers to have multiple priorities and sub-projects to work on concurrently. In some cases the staff at a site prioritise other items, which delays IPv6 deployment. We will continue to persuade and encourage sites.

Use of old software to analyze old data. Some experiments, e.g., ALICE, need to use old versions of their analysis software to re-analyze old data, e.g., LHC Run2. Old software may not be IPv6-capable. This is a tough problem to fix and this requirement must be part of the decision on the date for services to move from dual-stack to IPv6-only.

7 Conclusions

In this paper we have shown that the deployment of dual-stack IPv6/IPv4 storage throughout WLCG is now very close to completion. The working group has successfully identified many obstacles to the implementation of IPv6 and has been systematically investigating and fixing

problems. This has allowed WLCG data transfers to take place over IPv6 for more than 90% of the time on the LHCOPN network.

The HEPiX IPv6 working group will continue to study and fix more obstacles by, for example, running a new ticket campaign to move all services and all CPU to dual stack. We also plan to perform IPv6-only testing of WLCG clients/services. The LHCOPN network is a good candidate to be the first to turn off IPv4 peering. WLCG storage services were the first to move to dual-stack mode and will no doubt be the last services to remove IPv4.

References

[1] S. McKee et al., Identifying and Understanding Scientific Network Flows, submitted to this conference (CHEP2023)

[2] S. Campana et al., J. Phys. Conf. Ser. **513**, 062026 (2014)

[3] M. Babik et al, J. Phys. Conf. Ser. **214**, 08010 (2019)

[4] M. Babik et al., J. Phys. Conf. Ser. **245**, 07045 (2020)

[5] D. Ciangottini, A. Forti, L. Heinrich, N. Skidmore et al, HSF Analysis Facilities Forum White Paper. To be published https://hepsoftwarefoundation.org/activities/analysisfacilitiesforum.html