IPv6 in production: its deployment and usage in WLCG

Marian Babik[1], *Martin* Bly[2], *Tim* Chown[3], *Jiři* Chudoba[4], *Catalin* Condurache[2], *Alastair* Dewhurst[2], *Xavier* Espinal Curull[1], *Thomas* Finnern[5], *Terry* Froy[6], *Costin* Grigoras[1], *Kashif* Hafeez[2], *Bruno* Hoeft[7], *Hironori* Ito[8], *David P.* Kelsey[2*], *Fernando* López Muñoz[9,10,], *Edoardo* Martelli[1], *Raja* Nandakumar[2], *Kars* Ohrenberg[5], *Francesco* Prelz[11], *Duncan* Rand[12], *Andrea* Sciabà[1], *Ulf* Tigerstedt[13], and *Dennis* Van Dok[14]

[1]European Organization for Nuclear Research (CERN), CH-1211 Geneva 23, Switzerland
[2]STFC Rutherford Appleton Laboratory, Harwell Campus, Didcot, Oxfordshire OX11 0QX, United Kingdom
[3]JISC, Lumen House, Library Avenue, Harwell Campus, Didcot, Oxfordshire OX11 0SG, United Kingdom
[4]Institute of Physics, Academy of Sciences of the Czech Republic, Na Slovance 2 182 21 Prague 8, Czech Republic
[5]Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, D-22607 Hamburg, Germany
[6]Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
[7]Karlsruher Institut für Technologie, Hermann-von-Helmholtz-Platz 1, D-76344 Eggenstein-Leopoldshafen, Germany
[8]Brookhaven National Laboratory, Upton, NY 11973-5000, U.S.A.
[9]Port d'Informació Científica, Campus UAB, Edifici D, E-08193 Bellaterra, Spain
[10]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
[11]INFN, Sezione di Milano, via G. Celoria 16, I-20133 Milano, Italy
[12]Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
[13]Helsinki Institute of Physics, Gustaf Hällstriömin katu 2, FI-00014 Helsinki, Finland
[14]Nikhef, Science Park 105, NL-1098 XG Amsterdam, The Netherlands.

Abstract. The fraction of general internet traffic carried over IPv6 continues to grow rapidly. The transition of WLCG central and storage services to dual-stack IPv4/IPv6 is progressing well, thus enabling the use of IPv6-only CPU resources as agreed by the WLCG Management Board and presented by us at CHEP2016. By April 2018, all WLCG Tier-1 data centres should have provided access to their services over IPv6. The LHC experiments have requested all WLCG Tier-2 centres to provide dual-stack access to their storage by the end of LHC Run 2. This paper reviews the status of IPv6 deployment in WLCG.

1 Introduction

The HEPiX IPv6 Working Group [1] has been investigating the many issues involved in the deployment and use of IPv6 in HEP in general and more specifically in WLCG. The group's paper at CHEP2016 [2] presented the status of the work to allow sites to deploy IPv6-only CPU resources. Driven by the requirements of the LHC experiments, the WLCG Management Board, in September 2016, had approved the requirement that all WLCG Tier-2

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

storage services should aim to support dual-stack IPv6/IPv4 by the end of 2018. Since then the group has worked with others to encourage, support and monitor that transition and to identify and help solve any technical issues as they arise.

2 Status of the transition

2.1 Status of the Tier-0 and Tier-1's

The approved timeline for IPv6 readiness of the Tier-1 sites was defined to be April 2018, but not all succeeded. Some Tier-1 sites like PIC in Spain and the nordic NDGF Tier-1 were IPv6-ready long before the deadline; other sites reached the goal on time while others are still not fully IPv6-ready.

Some Tier-1 sites encountered technical problems with their transition. For example at DE-KIT the IPv6 deployment was designed as a dual-homed architecture and began with the IPv6 deployment for CMS only. The motivation for dual-homing resulted from their current IPv4 deployment, of one internal private address, and hosts/servers with a second public address applied to a different (virtual) interface. Since all hosts should receive only one IPv6 address, the first approach was to assign the IPv6 address to the so-called "internal" interface with the private address. In this scenario, a large number of transfers were failing. Analysis showed that the dCache-Doors require a dual-stack deployment instead of dual-homed. After discussion with the dCache developers, and also given problems in gridftp, KIT decided to implement a dual-stack only environment.

Early in 2018, EOS [3] at CERN was upgraded to a version that supported IPv6. As soon as the EOS instances for LHCb and ALICE were configured with AAAA DNS records for their public nodes, a large increase of IPv6 traffic was noticed on the CERN central firewall, moving from few hundreds of Mbps to almost 10Gbps. Figure 1 shows the traffic increase when the EOS dual-stack version was introduced.

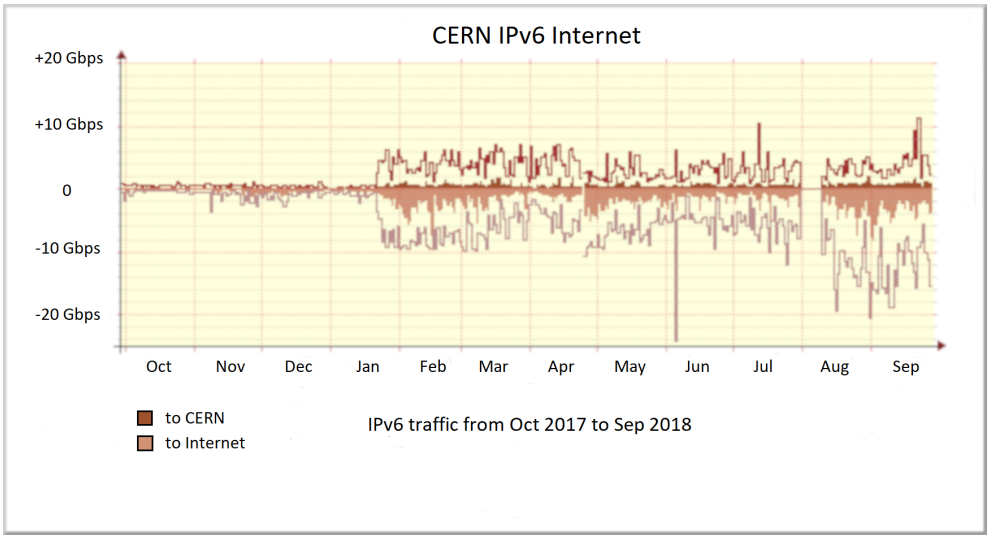

Figure 1. IPv6 traffic through the CERN central firewall, period Oct 2017 to Sep 2018

This increased traffic began to saturate CERN's IPv6 firewall, so in June 2018 a firewall bypass was deployed for IPv6, thereby allowing the activation of dual stack access also for the ATLAS and CMS EOS instances.

The status of the deployment at the Tier-0 and Tier-1's is shown in figure 2. The only sites not yet configured with dual-stack storage are the two Tier-1's in the USA, namely US-T1-BNL and US-T1-FNAL, and one of the Russian Sites, RRC-KI-T1. BNL is already dual-stack enabled, but they currently prefer IPv4 over IPv6; FNAL will enable their IPv6 dual-stack storage early in 2019; the Russian site is also preparing its IPv6 environment.

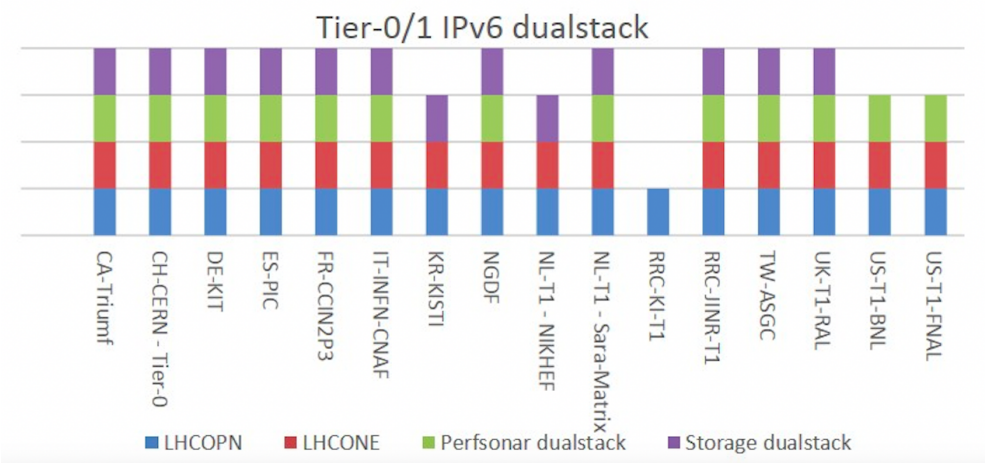

Figure 2. LHC Tier-0 and Tier-1 sites - Dual-stack readiness of LHCOPN, LHCONE, PerfSONAR and Storage

2.2 Status of the Tier-2's

According to the WLCG IPv6 plan, Tier-2 sites shall have their storage systems and perfSONAR installation working in dual stack by the end of 2018. To this purpose, a ticketing campaign via GGUS was launched in November 2017, targeting all WLCG Tier-2 sites managed by EGI, while for OSG sites the coordination for the IPv6 deployment was delegated to OSG operations.

The text of the tickets explained the goals and the motivation for the IPv6 deployment, and asked the sites to provide an estimated time scale for the deployment and some details about the required steps. The last step would be a check by perfSONAR and experiment experts that the services are working as expected via IPv6.

The status of the tickets was periodically reviewed and questions from the sites addressed by the experts in the HEPiX IPv6 working group. The tickets are classified as *no reply*, *on hold*, *in progress* and *done*. Following the decommissioning of the OSG operations team, the information about OSG sites is now sourced from the LHC experiments themselves. The status of the deployment at the start of October 2018, both globally and by region, is depicted in figure 3.

The time evolution of the site status shows a steady increase of the number of sites that have deployed IPv6. A detailed analysis of the tickets shows that, in many cases, sites need to wait for the IPv6 deployment on site, which often depends on people different from the WLCG site staff, while typically the deployment on the services happens quickly and

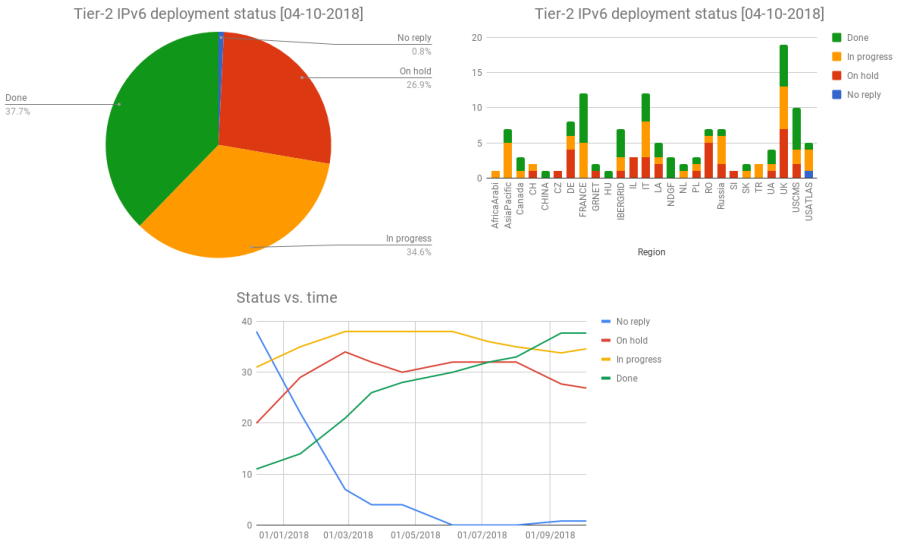

Figure 3. (left) Tier-2 deployment status by site globally, (right) by region, and (bottom) time evolution

painlessly. It is interesting also to see the fraction of the Tier-2 storage that is accessible via IPv6, which can be simply calculated from the amount of storage space available to each of the supported experiments (table 1). Given that only very few sites declared that they will

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

ALICE	ATLAS	CMS	LHCb	Global
41%	33%	55%	33%	42%

not be able to meet the deadline, and that it is natural to expect more activity closer to the deadline, it is expected that the Tier-2 deployment goals will be successfully met.

2.3 Experiment services

2.3.1 ALICE

The legacy Grid services of ALICE (AliEn [4]) are built around Perl 5.10 and Xrootd[5] 3.0 that don't support IPv6 natively. The new set of services (jAliEn [6]) are written in Java and using the latest Xrootd for data transfers, both fully supporting IPv6. Central services run both software stacks with functionality being migrated in steps from the legacy to the new version. One important step was the migration of data transfer tools to the new version, in particular all RAW data and calibration files are exported from Tier-0 to Tier-1's with the new tools. In progress is the migration of the experiment software to ROOT6 and a recent Xrootd library for data access, allowing job deployment on IPv6 resources. Because of the fully federated storage under the experiment computing model a full IPv6 storage deployment is required before IPv6-only worker nodes can run jobs. At the moment of writing this article a 54% of the storage volume is dual stacked with less than 5% of the volume running old server software versions that are not IPv6-ready.

2.3.2 CMS

The submission infrastructure in CMS is based on glideinWMS [7] and HTCondor [8], which are fully IPv6-compliant. The deployment of IPv6 on the relevant service nodes has been completed at CERN but not at FNAL, where it is scheduled by the end of the year. It is expected that, after a period of testing in the Integration Testbed (ITB), the production infrastructure will see IPv6 enabled across the board.

2.3.3 LHCb

The LHCb infrastructure is built around DIRAC [9] which is fully IPv6-compliant and the servers are based at CERN. All the servers running DIRAC services are dual-stacked and happy to listen to connections on either IPv4 or IPv6 as they come. The LHCb experiment is now waiting for the sites to deploy the pledged storages in dual-stack mode. With the storages already available (33% of LHCb Tier-2D, Tier-2 providing local storage, by volume and 64% Tier-0/1 by volume) and the automatic failover mechanisms in place, LHCb is already running on IPv4 or IPv6 or dual-stacked worker nodes without any manual intervention.

3 Monitoring

3.1 SAM

WLCG Service Availability Monitoring (SAM) [10] has been one of the primary tools to track availability and reliability of the WLCG sites. The monthly reports are based on profiles, which define how exactly a specific set of metrics are aggregated and which site services are taken into account.

The underlying measurement tool that provides the metrics to SAM is the WLCG Experiments Test Framework (ETF) [11]. ETF provides a low level monitoring framework used to test core site services at regular intervals, performing basic "ping-like" tests on remote compute, worker nodes and storage. A separate dual-stack ETF infrastructure has been setup during 2017 and runs alongside the production IPv4-only infrastructure.

In February 2018, the ETF dual-stack infrastructure became production ready and started publishing its results to SAM. The final step in the migration process was to understand how to define profiles that would mix both IPv4 and IPv6 results and this was achieved by introducing separate IPv4 and IPv6 service types that can be used in the aggregation algorithm to combine the results coming from ETF. Once evaluated and deployed in production, it will make it possible for SAM to generate the WLCG monthly reports for both IPv4 and/or IPv6 depending on the profile setup by the experiments.

3.2 perfSONAR

The WLCG has adopted the perfSONAR toolkit [12] for the monitoring of its network infrastructure and its deployment and configuration is being coordinated by the WLCG Network Throughput working group [13]. perfSONAR offers a very good way of checking that the migration to IPv6 hasn't caused any network/routing issues as it clearly separates IPv4 and IPv6 network measurements while supporting full range of network testing including throughput, latency, packet loss, packet re-ordering and duplicates, network path and packet retransmits.

Testing within WLCG is organised by a central configuration system (PWA), which operates around groups of sites also referred to as meshes. Currently, there are meshes for each experiment such as ATLAS, CMS and LHCb as well as network groupings such as LHCONE

and LHCOPN and since 2017 a dedicated dual-stack mesh was introduced, which groups together all dual-stack hosts on the network. Following the working group campaign to deploy dual-stack services at all Tier-1s and Tier-2s, the number of dual-stack hosts has grown to cover almost 50% of the infrastructure, currently there are 132 dual-stack perfSONARs out of total 268 available.

The successful adoption rate of dual-stack perfSONAR has led to a proposal to integrate IPv6 testing directly into the existing production meshes, which would significantly improve the test coverage and eventually bring it on par with IPv4. This migration is currently in progress and was already implemented for LHCOPN, LHCONE, USATLAS, USCMS and ATLAS.

In addition to the configuration changes, there has also been good progress in the area of visualisation. Recently, a new Grafana-based dashboard has been deployed to complement the existing Maddash [14] and a dedicated IPv6 view was setup [15] to help sites compare side-by-side their IPv4 and IPv6 perfSONAR measurements.

3.3 FTS

In WLCG bulk data transport is carried out predominantly by the File Transfer Service (FTS3) [16]. The monitoring data reported by each FTS3 server for such data transfers indicate whether IPv6 was used during the transfer. Figure 4 shows FTS data transfers for all VOs (not just the WLCG experiments) during the month of October 2018. It can be seen that whilst IPv4 still dominates IPv6 is used approximately a quarter of the time. Interestingly, transfers over IPv6 appear to be more reliable, hovering around the 90% mark whilst transfers over IPv4 average approximately 75% success rate. The reasons for this are not yet understood. In order for FTS transfers to go over IPv6 both of the storage elements and the FTS server itself need to be dual-stack. Currently, of the eight FTS servers in regular use by the WLCG, five are dual-stack. It is planned that at least one of the remaining three will enable IPv6 at the end of LHC Run 2.

3.4 XrootD

The other major method by which data is accessed, often remotely, is XrootD[5] most notably by the ALICE and CMS experiments. Currently, the Monit WLCG Transfers Monitoring Service at CERN does not display whether this data was transferred over IPv4 or IPv6. However, the XrootD monitoring data has now been updated to report whether IPv4 or IPv6 was used and work is in progress to include this information in the Monit monitoring.

4 Future plans and conclusions

4.1 Obstacles and potential show-stoppers

Our experience with the transition of WLCG Tier-1 and Tier-2 centers so far has identified various cases where the IPv4→IPv6 transition has consequences that exceed the simple replacement of the IP *transport* layer. These broadly fall in the following categories:

1. Software components and protocol with fixed-size storage for network addresses, for example in (Grid-)FTP and AFS. This can be overcome by appropriate protocol extensions (e.g. the introduction of 'extended' FTP commands `EPRT`, `EPSV`, etc.), with a large development effort required. In certain cases (AFS) this effort was determined to be too large and not worthwhile, while the GrifFTP 'v2' extensions were issued with no IPv6 support.

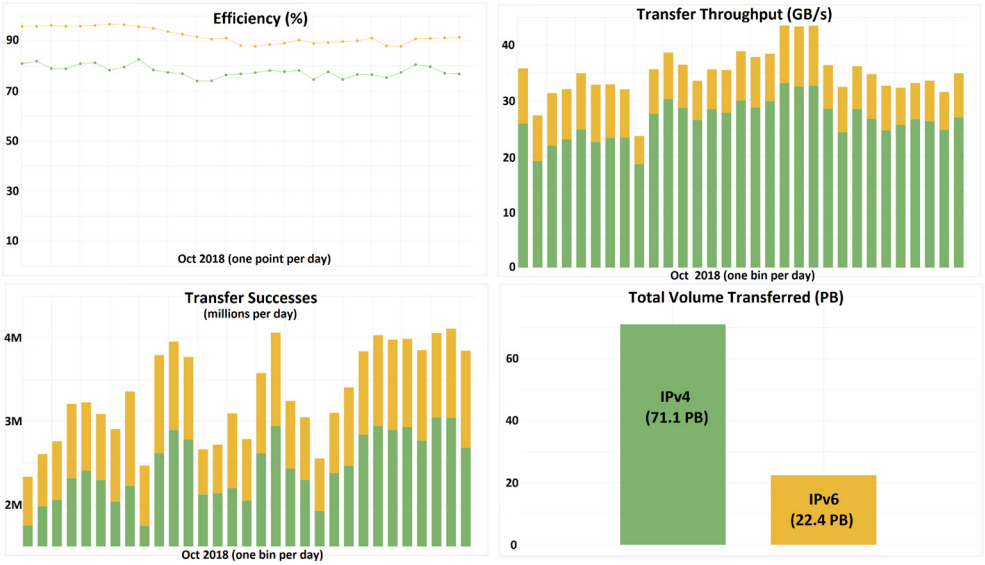

Figure 4. FTS transfer throughput for Oct 2018 and success rate according to whether IPv6 was used. IPv4 is green; IPv6 is amber

2. Software components and protocols that assume single addresses or a single IP protocol for network endpoints (to various extents, all the components in the WLCG software matrix). While all Operating Systems do provide hybrid network stacks and prioritized/configurable source and destination address selection[1], applications should always iterate over multiple possible results (belonging to multiple IP protocol versions) and provide configurable overrides and preferences.

3. Software components and network infrastructures providing asymmetric handling, or separate code stacks, for IPv4 and IPv6 traffic: these also have the ability to load the network transport choice with measurable performance consequences.

Detecting and explaining these IPv4/IPv6 asymmetries and fostering improved symmetry across the WLCG software matrix has been a long-standing activity for our group. As there is inherent risk in changing, testing and rolling out (sometimes sizable) software changes when no functional issue requires immediate attention, these changes are often not shortlisted for deployment by software development teams. We need to continue tracking these issues down to the level where no measurable performance asymmetry between the two protocols is seen: this can be seen as the backdrop of any future activity.

4.2 Further steps

Once the data transfer monitoring infrastructure described in Section 3 is completed and covers all services, two cases need to be consistently monitored and, where needed, investigated and resolved:

1. cases where the fraction of data transferred over IPv6 is lower than expected: the preference for IPv6 over IPv4 has to established throughout the system;

[1] Address selection is regulated in most cases by RFC6724.

2. cases where the transfer performance is either significantly worse or significantly better on IPv6 over IPv4: asymmetries in the routing and transport should be identified, especially in the LHCONE and LHCOPN networks.

Over a stable and understood data transfer network IPv6-only worker nodes should encounter no more operational anomalies than other types of worker nodes. The job failure rate on IPv6-only, dual-stack and IPv4-only worker nodes should be monitored statistically and deviations properly troubleshot.

Once the use-case of IPv6-only worker nodes operates smoothly, the next step in the transition roadmap is moving services and network segments to IPv6-only operation.

4.3 Conclusion

The Tier-1 deployment of IPv6 did not complete by the original planned date of April 2018, but by December the majority of large Tier-1's should be IPv6-capable, including more than two-thirds of the total storage capacity, except for the two USA TIer-1's who were unable to deploy during LHC running. The Tier-2 migration has gone well but will not reach completion in 2018. It is expected that more than half will be completed during 2018. The main reason for delay is the local networking team not being ready to support IPv6. Based on input from the delayed sites, we expect that the deployment of dual-stack storage will happen at most of the Tier-2 sites during 2019.

Data from a complete and pervasive monitoring infrastructure are crucial in building confidence in the new transport layer, which is expected to provide the *same* level of performance and reliability as the one it replaces: these come from various sources (see Section 3), with a few remaining blind spots being covered.

Feedback from the actual operation of IPv6-only computing farms will then determine the time-scale and viability of driving the transition process to its natural ending, when the burden of operating a duplicated transport infrastructure will be eventually lifted.

References

[1] S. Campana et al, J. Phys. Conf. Ser. **513**, 062026 (2014)
[2] M. Babik et al, J. Phys. Conf. Ser. **898**, 082033 (2017)
[3] A. J. Peters et al, J. Phys. Conf. Ser. **664(4)**, 042042 (2015)
[4] S. Bagnasco et al, J. Phys. Conf. Ser. **119(6)**, 062012 (2008)
[5] L. Bauerdick et al, J. Phys. Conf. Ser. **396** 042009 (2012)
[6] A. Grigora et al, J. Phys. Conf. Ser. **523**, 012010 (2014)
[7] I. Sfiligoi, J. Phys. Conf. Ser. **119(6)**, 062044 (2008)
[8] D. Thain et al, Concurrency - Practice and Experience, **7(2-4)**, 323 (2005)
[9] A. Tsaregorodtsev et al, J. Phys. Conf. Ser. **513**, 032096 (2014)
[10] A. Aimar et al, J. Phys. Conf. Ser. **898(9)**, 092033 (2017)
[11] Marian Babik, CERN, http://etf.cern.ch/docs/latest/
[12] A. Hanemann et al, In Boualem Benatallah, Fabio Casati, and Paolo Traverso, editors, *Service-Oriented Computing - ICSOC 2005*, pages 241–254, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
[13] S. McKee et al, J. Phys. Conf. Ser. **664(5)** 052003 (2015)
[14] perfSONAR Consortium, http://psmad.grid.iu.edu/maddash-webui/
[15] CERN monitoring team, https://monit-grafana.cern.ch/
[16] A. A. Ayllon et al, J. Phys. Conf. Ser. **513(3)** 032081 (2014)