Architectures and methodologies for future deployment of multi-site Zettabyte-Exascale data handling platforms

2015 J. Phys.: Conf. Ser. 664 042009

(http://iopscience.iop.org/1742-6596/664/4/042009)

Journal of Physics: Conference Series **664** (2015) 042009 doi:10.1088/1742-6596/664/4/042009

Architectures and methodologies for future deployment of multi-site Zettabyte-Exascale data handling platforms

V Acín[1,11], I Bird[2], T Boccali[6,12], G Cancio[2], I P Collier[8], D Corney[8], B Delaunay[4,13], M Delfino[9,10,11], L dell'Agnello[6,14], J Flix[3,11], P Fuhrmann[5], M Gasthuber[5], V Gülzow[5], A Heiss[7], G Lamanna[4,15], P-E Macchi[4,13], M Maggi[6,16], B Matthews[8], C Neissner[1,11], J-Y Nief[4,13], M C Porto[3,11], A Sansum[8], M Schulz[2] and J Shiers[2]

[1]Institut de Física d'Altes Energies (IFAE), Edifici Cn, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Barcelona), Spain
[2]European Organization for Nuclear Research (CERN), CH-1211 Geneva 23, Switzerland
[3]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, E-28040 Madrid, Spain
[4]Centre National de la Recherche Scientifique (CNRS), 3 rue Michel-Ange F-75794 Paris cedex 16, France
[5]Stiftung Deutsches Elektronen-Synchrotron (DESY), Notkestraße 85, D-22607 Hamburg, Germany
[6]Istituto Nazionale di Fisica Nucleare (INFN), Via Enrico Fermi 40, I-00044 Frascati, Italy
[7]Karlsruher Institut für Technologie (KIT), Steinbuch Centre for Computing (SCC), Kaiserstraße 12, D-76131 Karlsruhe, Germany
[8]Science and Technology Facilities Council (STFC), Rutherford Appleton Laboratory, R89, Harwell, Didcot, Oxfordshire, OX11 0QX, UK
[9]Universitat Autònoma de Barcelona, E-08193 Bellaterra (Barcelona), Spain

E-mail: delfino@pic.es

Abstract. Several scientific fields, including Astrophysics, Astroparticle Physics, Cosmology, Nuclear and Particle Physics, and Research with Photons, are estimating that by the 2020 decade they will require data handling systems with data volumes approaching the Zettabyte distributed amongst as many as 10^{18} individually addressable data objects (Zettabyte-Exascale systems). It may be convenient or necessary to deploy such systems using multiple physical sites. This paper describes the findings of a working group composed of experts from several

[10] Author to whom any correspondence should be addressed.
[11] Also at Port d'Informació Científica (PIC), Campus UAB – Edifici D, E-08193 Bellaterra (Barcelona), Spain
[12] INFN Sezione di Pisa, Edificio C - Polo Fibonacci Largo B. Pontecorvo, 3, I-56127 Pisa, Italy
[13] Centre de Calcul - IN2P3 / CNRS, 21 av Pierre de Coubertin, F-69622 Villeurbanne cedex, France
[14] INFN CNAF, Viale Berti Pichat 6/2, I-40127 Bologna, Italy
[15] Laboratoire d'Annecy-le-Vieux de Physique des Particules, Université de Savoie Mont Blanc, CNRS/IN2P3, F-74941 Annecy-le-Vieux, France
[16] INFN Sezione di Bari, Via E. Orabona 4, I-70126 Bari, Italy

large European scientific data centres on architectures and methodologies that should be studied by building proof-of-concept systems, in order to prepare the way for building reliable and economic Zettabyte-Exascale systems. Key ideas emerging from the study are: the introduction of a global Storage Virtualization Layer which is logically separated from the individual storage sites; the need for maximal simplification and automation in the deployment of the physical sites; the need to present the user with an integrated view of their custom metadata and technical metadata (such as the last time an object was accessed, etc.); the need to apply modern efficient techniques to handle the large metadata volumes (e.g. Petabytes) that will be involved; and the challenges generated by the very large rate of technical metadata updates. It also addresses the challenges associated with the need to preserve scientific data for many decades. The paper is presented in the spirit of sharing the findings with both the user communities and data centre experts, in order to receive feedback and generate interest in starting prototyping work on the Zettabyte-Exascale challenges.

1. Introduction

Several scientific fields, including Astrophysics, Astroparticle Physics, Cosmology, Nuclear and Particle Physics, and Research with Photons, are estimating that by the 2020 decade they will require data handling systems with data volumes approaching the Zettabyte distributed amongst as many as 10^{18} individually addressable data objects. In the paper, we refer to this scale of data handling systems as the Zettabyte-Exascale.

One example of this data growth is illustrated in figure 1, where the number of objects simulated in large-scale cosmological simulations run on supercomputers is shown. Starting under a thousand objects in the seminal simulation by Peebles in 1970 [1], there has been exponential growth in the number of objects simulated, approaching the Terascale (10^{12} objects) by the decade of 2012 [2][3][4].

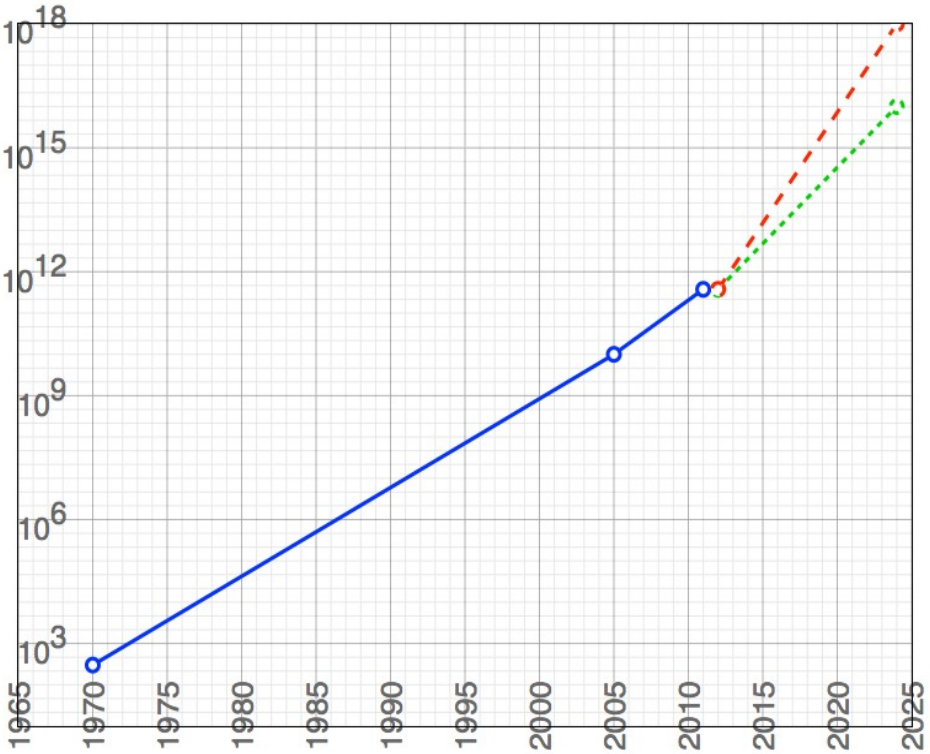

Figure 1: Number of objects in cosmological simulations (vertical axis) versus year (horizontal axis)

The improvements in resolution obtained thanks to this increases are playing a fundamental role in the understanding of existing experiments in observational cosmology, as well as in the design of future ones. The trend is expected to at least continue (green line) or accelerate thanks to the increasing availability of supercomputers with very large memory (red line), so by the year 2025 the expectation is to have 0.01 to 1.0 Exa-objects (10^{16} to 10^{18} objects) per simulation.

The volume of scientific data is also increasing exponentially. The best known example comes from particle physics, where the data volume has grown from hundreds of Gigabytes in the mid-1980s to over a hundred Petabytes stored by the detectors at CERN's Large Hadron Collider. Again, the trend is expected to continue or even accelerate, with expectations of reaching 0.1 to 1.0 Zettabytes (10^{20}-10^{21} bytes) by the year 2025.

2. The ZEPHYR study group

The ZEPHYR study was made in 2014 by a group of people from institutions which run large data centres in Europe [5]. Recognizing the growth in requirements summarized above, the group is analysing the Zettabyte-Exascale challenges in the domains of Astrophysics, Astroparticle Physics, Cosmology, Nuclear and Particle Physics, and Research with Photons. Initially, the group has concerned itself with a number of issues:

- Although computing and data processing technology is known for its fast growth, assembling these technologies into large-scale, reliable and performant systems often takes about a decade. Hence, it is important to start work towards the Zettabyte-Exascale.
- Involvement of the existing and upcoming experiments, projects and user communities which will generate the Zettabyte-Exascale requirements is extremely important in order to guide the development.
- Given the global nature of large-scale science, Zettabyte-Exascale solutions will be needed across data centres in Europe, Asia and the Americas. It is important to start regional initiatives on the subject in order to structure collaboration in developing solutions.
- Experience shows that the eventual development of large-scale production systems critically relies on concrete outputs from simulations and prototypes, which should be shared with the expert community for discussion and improvement.
- Starting with enough lead time gives the opportunity to take a broad look at the architectural choices to build Zettabyte-Exascale systems, taking into account experience from the past without being constrained by it.

The conclusions reached so far by the ZEPHYR study group are presented in this paper, in the spirit of stimulating discussions and further work towards Zettabyte-Exascale systems.

3. Broad requirement features for future Zettabyte-Exascale solutions

3.1. The scientist's environment

The large size of modern scientific collaborations is often emphasized. It is important, however, to also recognize that the granular scientific work is almost always done by teams comprising a relatively small number of scientists which focus on a particular issue. A growing tendency is for the members of such teams to be geographically dispersed. Another important feature is that scientific data analysis is an activity which lasts for years, even decades, and the composition of the teams and the geographical location of their members will vary with time.

The computational and data processing loads generated by these teams are enormous compared to other digital tasks, even those described as "Big Data". A scientific team may routinely access a billion objects with algorithms requiring several minutes of computation per object. The challenge is often compounded by the fact that the exact manner in which the data will need to be analysed is not known *a priori*, hence techniques often used in commercial computing, such as pre-calculation, precise data placement or massive data caching are difficult or impossible to implement.

Nevertheless, the large growth in raw data volume is making it increasingly difficult for individual scientists to routinely access it. In order to circumvent this problem, a great deal of summary data is being generated in a plethora of formats and is being analysed with a large variety of tools and methodologies. In some cases, the value of raw data relative the cost of storing it, the difficulty of accessing it and the possibilities of re-generating it are such that raw data may not be stored at all. This may become the case in genomics, for example, where the advent of inexpensive sequencing machines may turn this past decade's raw data driven supraexponential growth in storage into a non issue, as scientists choose not to store the raw data at all [6]. Abstractly, these summary data looks very similar (though often much larger in volume and number of objects) to the so-called metadata used in "Big Data" applications. In other words, summary data from galaxy redshift observations or nucleotide sequences are not that different from cash-register ticket records or Web access "Big Data".

Another important requirement is that some scientists state the need to preserve a way to "zoom back" from metadata to the raw data. For example, an LHC physicist may search for Higgs bosons using metadata-like summary data; having found instances of interest, she will very likely want to examine the raw data recorded by the particle detector for those particle collisions. Superficially, this may be considered an example of "linked data", but the volumes involved are enormous. Finally, scientific work almost invariably requires a very high level of data integrity, ensuring completeness and reproducibility of the data samples being analysed.

Additional requirements are emerging related to the relation of scientific projects and society. One example is data preservation and provisioning of open access to scientific data. Funding agencies, particularly in the public sector, are increasing requiring that data remain available after the formal end of scientific projects, in order to permit re-analysis in the future by domains scientists as well as ordinary citizens. This is already common in Astronomy, for example, and will become mandatory for all fields in the near future.

Another emerging requirement is data provenance, the systematic recording of the detailed origins of each data object used in scientific analysis, in order to ensure reproducibility and quality control of results. This is already common in drug discovery activities, for example, and again it is expected to become mandatory, or at least considered a best practice, for all fields in the near future.

3.2. The data centre environment

The scientific data centre environment should also be considered in requirements gathering. Including this point of view ensures that the user requirements can be technically fulfilled and enables a link to cost estimation and funding.

Interesting challenges are generated by the convergence of four factors:

- Increased importance of scientific metadata, with an increasing fraction of analysis activities using metadata as its primary input.
- The need to handle technical metadata, such as object creation or last-access time, in an integrated manner with respect the scientific metadata. This should also include support for data provenance and data preservation requirements.
- Linked-data aspects, where support is needed to "zoom" from metadata to other forms of data, including the raw data of experiments.
- The overall increase in scale. The increase in number of objects will affect data and metadata equally, which will need Exascale support. The increase in data volume indicates that when raw data approaches the Zettabyte scale, the metadata handling system will approach the Petabyte scale. This increased volume is implicitly accompanied by an increase in data throughput requirements, which are often difficult for users to fully specify.

Scientific data centres will need to deal with these challenges by the year 2025, or even sooner, and need to work on reliable solutions which contain costs or are able to be tuned for cost optimization, including infrastructure costs such as energy consumption. In addition, the solutions are expected to involve multiple sites, including owned and hired resources (i.e. "private" and "public" cloud systems,

with “private” meaning “owned by public research centres” and “public” meaning “hired from private companies”).

4. Hints at important technical aspects to be studied for the Zettabyte-Exascale

The individual broad requirements described above intersect in various ways the technological space which must be explored to find solutions for the Zettabyte-Exascale. An initial exploration of these intersections yields a number of hints at technical aspects which can start to be investigated now.

4.1. Life-cycle management of users and their roles

The number of users involved in a given scientific project is probably the only parameter which is not growing exponentially. Current solutions for managing users and their data access roles are, however, very far from being satisfactory. Present systems involve a lot of manual intervention by many parties, and therefore are rather error prone and cumbersome. Furthermore, they are poorly (or not at all) integrated with low level data processing services and between different sites.

Present activities, such as *eduGAIN* [7], indicate a path where User Authentication will become a global, federated service which scientific data centres can consume. The challenge will be how to integrate it with the data processing systems, enabling for example support for data provenance.

The situation is far less clear regarding the evolution of “Virtual Organization” membership management and authorization (the management of data access rights), and this is an area which needs further work. Ideally, information should flow in an organized manner, smoothly and automatically, from the units that administrate membership (for example, experiment secretariats or scientific task force coordinators) and the services across data centres which technically enable access rights. Solving this issue will also ease efficient implementations of multi-site accounting, data provenance and preservation.

4.2. Extensible metadata management frameworks, handling Scientific and Technical metadata

Most data has been stored in files since the practical beginning of the computer age. Storage has been based on File Systems, which handle in an integrated manner the storage of the data objects themselves, as well as a limited set of associated metadata (file name, creation and last access date, access control lists, etc.). File systems offer virtually no practical support for associating user-specified metadata to data objects, essentially limited to specifying file names. Although essentially all scientific projects encode some metadata in file names, this is not enough to fulfil their requirements and they maintain in parallel an *ad hoc* metadata management scheme as part of their data management setups. Maintaining these schemes is labour intensive, and would become even more so if support was included for data provenance and preservation.

Recently, a change of paradigm has been put forward where storage systems are based on key-value pairs and metadata management is entrusted to a separate service. This is usually called Object Storage, and has been pioneered by Amazon Web Services through its *S3* service [8] and more recently by Openstack *Swift* [9] and *Ceph* [10]. The main motivation is a simplification of data storage services, allowing for better optimization, more flexibility, higher performance and lower costs. For example, very recently Seagate has launched disk drives which connect directly to Ethernet and whose controllers directly support object storage protocols [11]. It should be noted that object storage is not particularly new to scientific projects using magnetic tape for their storage, since tape has rarely offered file system support and storage on tape has been historically handled by storing objects on tape and maintaining the metadata separately.

The combination of these technological trends with the increasing requirements to manage rich sets of scientific metadata and provide support for linked data indicates the need for extensible metadata management frameworks properly interfaced to object storage systems. It is not clear at this moment whether a few general toolkits will be developed that can fulfil the requirements of all scientific projects. It is important to stimulate work in this area, as services based on such toolkits will have to be integrated into large-scale prototypes in order to ensure the scalability to the Zettabyte-Exascale.

4.3. Site-independent "Data Virtualization" layer: metadata query with redirection to data objects

It is expected that future Zettabyte-Exascale systems for science will be based on multiple sites and multiple providers (see section 3.2). Supporting the data and metadata model described in the previous section across multiple sites will require an additional component, which we call the Data Virtualization layer (DV). The DV is a site-independent service which may itself be implemented in a distributed multi-site manner. It provides a predictable target to receive data access requests from applications, processing metadata queries and returning the results, including handles which redirect input-output to object storage servers.

Currently, there are many data virtualization schemes which are deployed as part of large distributed file systems, such as *dCache* [12], *EOS* [13], *GPFS* [14] and *Lustre* [15]. Deployments are restricted to single sites, with a few exceptions [16], and only manage the technical metadata needed to emulate a traditional file system. In parallel, redirector schemes have been deployed which return handles (in the form of URLs) to file system based storage at multiple sites. Examples are the *xrootd* and *http/WebDAV* redirectors deployed over the last few years by the CERN LHC experiments.

It is clear that much more work is needed in order to attain DV components which fulfil the requirements at Zettabyte-Exascale. Key areas were activity should be stimulated are:

- Information modelling on how to support in an extensible way a coherent view of dynamic scientific and technical metadata.
- End-to-end support to enable global authentication, authorization and accounting.
- Exploration of the use of "Big Data" techniques, such as MapReduce [17] or *NoSQL* databases, to handle queries on complex metadata with volumes ranging from Terabytes to Petabytes.
- High frequency metadata update schemes needed to maintain technical information on a per object basis, such as time of last access, input-output performance, etc.
- Support for bulk operations on groups of objects, such as changing the access control on a complete dataset from being restricted to an experiment to being open-access.

Work in these areas should yield in the next few years a number of suitable candidates for Zettabyte-Exascale DV, so that prototypes can be built and scalability verified before the onset of requirements.

4.4. Smarter use of network capacity and capability

The academic research community has always been at the forefront in the usage of wide area networks. National Research and Academic Networks (NRENs) and international networks such as Géant in Europe and RedClara in Latin America are well established. These networks actively collaborate with industry in order to deploy leading-edge services, such as the LHC Optical Private Network [18] and XSEDENet [19]. There is, however, a widening gap between the possibilities offered by these leading-edge networks and the capabilities that are actually deployed in production applications. Features such as dynamic bandwidth provisioning or traffic prioritization are seldom deployed in production. A fresh look should be taken into how to integrate emerging new features of wide area networks into Zettabyte-Exascale solutions.

Local area network (LAN) technology is also evolving. Widespread use of virtualization places additional demands on the flexibility of LANs. Data centres are utilizing various technologies supported by LAN switches and routers, such a virtual LANs, in order to gain flexibility. A new paradigm of Software Defined Networks is emerging and will be fully supported by 2025. This evolution calls for studies on how to make use of this functionality in Zettabyte-Exascale data processing solutions, making applications more network aware, for example.

Finally, the complete global introduction of version 6 of the Internet Protocol (IPv6) will have a major positive impact. The shortage of Internet addresses under version 4 of IP has been acknowledged since the 1980s. In spite of this, IPv6 deployment has been slow, even in the academic research community. This has mandated the use of various workarounds, such as the reuse of addresses, which have made networks more opaque. Moving fully to IPv6 provides the opportunity to

remove many of these workarounds, and regain the advantage of a transparent Internet where all nodes have unique and fully routable addresses. This should be taken into account when designing Zettabyte-Exascale solutions.

4.5. Gatekeepers for Data Provenance and Data Preservation

There is substantial worldwide activity towards enabling data preservation, sharing and reuse. Examples are the work of the Research Data Alliance [20] and the Preservation and Archiving Special Interest Group [21]. Examples within the domains covered by the ZEPHYR study are the Data Preservation and Long-Term analysis in High Energy Physics group (DPHEP) [22] and a number of Virtual Observatory activities in Astronomy [23].

A full review of proposals in data preservation, sharing and reuse is beyond the scope of this paper, and the focus is on how to integrate such techniques into future Zettabyte-Exascale solutions. Currently, most implementations are based on manual intervention, where data is "deposited" re-creating processes similar to those used for centuries in paper-based libraries. They do not seem well suited to scientific domains with large data environments. Hence, an area where work should be stimulated is in more automated, batch-like, methodologies to tag data with provenance, preservation and other bibliographic information.

These technical processes need to be linked to the analysis techniques and review processes used by scientists. Clearly, all data generated in an experiment or simulation need not be tagged with external bibliographic information (though it may be desirable or even required to tag it with internal information in order to support data provenance). Therefore, the concept of "depositing" data into a given category (internal or external) after the necessary checks have been performed is a useful one. A possible metaphor to be explored is that of *Gatekeepers*, which would use the bulk operations capabilities mentioned in section 4.3 to appropriately tag entire datasets, which in the Zettabyte-Exascale may contain millions of objects amounting to Terabytes of data volume. The specification, design and prototyping of such *Gatekeepers* are areas where work should be stimulated.

5. The role of simulation and prototypes

The ability to build at a reasonable cost future Zettabyte-Exascale crucially depends on industry's ability to lower the cost per capacity of computing equipment as time goes forward. Therefore, implementing even a 1% prototype a decade before the solutions are needed has prohibitive costs.

Simulation of data processing system behaviour is a valuable tool and can be used to evaluate various architectural or technological choices prior to building prototypes. Simulation has been used in the past to guide the early design stages of large systems, for example for the Worldwide LHC Computing Grid [24]. Simulations can also provide guidance for early prototyping work, which will necessarily focus on subsystems or on specific "slices" of the eventual system.

Once a reasonable number of alternatives is identified, the involvement of large scientific data centres on the building of prototypes will be important. Such centres are expected to be able to setup prototypes and tests using resources which are temporarily allocated, thereby reducing costs. This will allow prototyping and testing of enough alternatives to avoid locking into a single solution too early in the development process.

6. Conclusions and outlook

Many scientific projects are forecasting the need to store and analyse data with up to 10^{18} objects and volumes approaching 10^{21} bytes by the year 2025, ushering the Zettabyte-Exascale. The ZEPHYR group has presented a first look at issues raised by this new scale in the domain of Astrophysics, Astroparticle Physics, Cosmology, Nuclear and Particle Physics, and Research with Photons. A number of trends have been identified which impact on the architecture of future scientific data processing systems. Key areas were activity should be stimulated have been identified as follows:

- Life-cycle management of users and their roles
- Extensible metadata management frameworks, handling Scientific and Technical metadata

- Site-independent "Data Virtualization" layer: metadata query with redirection to data objects
- Smarter use of network capacity and capability
- Gatekeepers for Data Provenance and Data Preservation

Simulation is an important alternative for exploring solutions, given the huge cost involved in building substantial size prototypes with present technology. Nevertheless, communities foreseeing Zettabyte-Exascale should build into their planning a program to build over the next decade a set of prototypes of increasing size, functionality and performance.

Acknowledgments
The authors acknowledge the support of their respective institutes and funding agencies, which has enabled the activities of the ZEPHYR working group and the presentation of this paper.

References

[1] Peebles P J E 1970, *Astron. J.*. **75**, 13
[2] Springel V *et.al.* 2005, *Nature* **435**, 629-36
[3] Fosalba P, Crocce M, Gaztanaga E and Castander F J 2015, *Mon.Not.Roy.Astron.Soc.* **448**, 2987-3000
[4] Vogelsberger M, Genel S, Springel V, Torrey P, Sijacki D, Xu D, Snyder G, Bird S, Nelson D and Hernquist L 2014, *Nature* **509**, 177-182
[5] The collaboration between these institutions is being formalized under the name EU-T0. See http://www.eu-t0.eu (accessed on 16/05/2015)
[6] Schatz M C Biological data sciences in genome research 2015 *Genome Res.* **25** DOI: 10.1101/gr.191684.115
[7] See http://services.geant.net/edugain/Pages/Home.aspx (accessed on 16/05/2015)
[8] See http://aws.amazon.com/es/s3/ (accessed on 16/05/2015)
[9] See http://swift.openstack.org (accessed on 16/05/2015)
[10] Weil S A 2007 Ceph: Reliable, Scalable and High-Performance Distributed Storage *Ph.D. thesis, University of California Santa Cruz*. Also see http://ceph.com
[11] See http://www.seagate.com/kinetic (accessed on 16/05/2015)
[12] Millar A P, Behrmann G, Bernardt C, Fuhrmann P, Litvintsev D, Mkrtchyan T, Petersen A, Rossi A and Schwank K 2014 DCache: Big data storage for HEP communities and beyond *Journal of Physics: Conference Series* **513**
[13] Espinal X *et.al.* 2014 Disk storage at CERN: Handling LHC data and beyond *Journal of Physics: Conference Series* **513**
[14] Schmuck F B and Haskin R L 2002 GPFS: A Shared-Disk File System for Large Computing Clusters *FAST* **2** 19
[15] Braam P J and Schwan P 2002 Lustre: The intergalactic file system. *Proc. Ottawa Linux Symp.* 50
[16] As an example, see Behrmann G, Fuhrmann P Grønager M and J Kleist J 2008 A distributed storage system with dCache *Journal of Physics: Conference Series* **119**
[17] Dean J and Ghemawat S 2004 MapReduce: Simplified Data Processing on Large Clusters *OSDI'04: Sixth Symposium on Operating System Design and Implementation*
[18] See https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome (accessed on 16/05/2015)
[19] See https://www.xsede.org/networking (accessed on 16/05/2015)
[20] See https://rd-alliance.org/ (accessed on 16/05/2015)
[21] See http://www.preservationandarchivingsig.org/ (accessed on 16/05/2015)
[22] See http://www.dphep.org/ (accessed on 16/05/2015)
[23] Hanisch R J, Berriman G B, Lazio T J W, Emery Bunn S, Evans J, McGlynn T A and Plante R 2015 The Virtual Astronomical Observatory: Re-engineering Access to Astronomical Data arXiv:1504.02133 [astro-ph.IM] http://arxiv.org/abs/1504.02133
[24] Legrand I C and Newman H B 2000 *Simulation Conf. Proc.* **2** DOI:10.1109/WSC.2000.899171