

## dCache: Big Data storage for HEP communities and beyond

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2014 J. Phys.: Conf. Ser. 513 042033

(<http://iopscience.iop.org/1742-6596/513/4/042033>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

### Download details:

IP Address: 131.225.23.169

This content was downloaded on 26/11/2014 at 20:39

Please note that [terms and conditions apply](#).

# dCache: Big Data storage for HEP communities and beyond

AP Millar<sup>1</sup>, G Behrmann<sup>2</sup>, C Bernardt<sup>1</sup>, P Fuhrmann<sup>1</sup>,  
D Litvintsev<sup>3</sup>, T Mkrtchyan<sup>1</sup>, A Petersen<sup>1</sup>, A Rossi<sup>3</sup>, K Schwank<sup>1</sup>

<sup>1</sup> IT Dept., DESY, Notkestrasse 85, Hamburg, Germany

<sup>2</sup> NORUnet, Copenhagen, Denmark

<sup>3</sup> Fermilab, Chicago, IL, USA

E-mail: paul.millar@desy.de

**Abstract.** With over ten years in production use dCache data storage system has evolved to match ever changing landscape of continually evolving storage technologies with new solutions to both existing problems and new challenges. In this paper, we present three areas of innovation in dCache: providing efficient access to data with NFS v4.1 pNFS, adoption of CDMI and WebDAV as an alternative to SRM for managing data, and integration with alternative authentication mechanisms.

## 1. Introduction

dCache[1] is a Free/Open-Source system that supports storing and retrieving large volumes of data by aggregating the capacity of many servers in a scalable fashion. It provides fast access to this data and scales to many petabytes of storage[2][3][4][5]. dCache offers a number of advanced features like transparent access to tertiary storage systems (including several tape systems), file replication (for resilience and increased performance), hot-spot detection and mitigation and file integrity checking[6].

As dCache is freely available, it is difficult to know exactly how many dCache instances are in use. For WLCG there are around 80 dCache clusters spread over some 60 sites throughout the world[7]. Although dCache is most heavily used in HEP[8][9], sites are increasingly offering dCache storage to non-HEP users[10][11]. This is possible because of dCache's strong emphasis on standards-compliance so that, in addition to HEP-specific protocols, dCache also supports the NFS v4.1/pNFS[12], HTTP/WebDAV[13] and FTP[5] protocols.

This paper describes some of the recent changes in dCache that have been introduced to meet requirements of increasingly diverse dCache user-communities and how these changes are being adopted by HEP users. It will focus on three main areas: managing storage, data analysis and new ways of authentication. The remainder of the paper is structured to follow these topics with a section on CDMI and then by sections on NFS and authentication.

## 2. CDMI and cloud storage

At the current time, WLCG manages about 246 PiB of data spread over some 230 sites running a variety of storage systems[7]. Managing data spread over many distinct storage systems is a formidable task and is only possible if sites present their storage capacity in a standard fashion.



Currently, the standard approach for accessing storage is the SRM v2.2 protocol[14]. Although SRM has been standardised through the Open Grid Forum (OGF)[15], it has had poor uptake outside the HEP community. Moreover, there are several technical and perceived problems with the protocol that are leading the WLCG community to look for alternative approaches, such as WebDAV[16].

WebDAV supports basic management operations, such as renaming and deleting files, but lacks many of the management functions and concepts that are currently used by WLCG. The industry storage standardisation body SNIA has developed the Cloud Data Management Interface (CDMI) protocol[17]. CDMI covers many more of the features needed by WLCG, plus it brings a number of new features that may prove useful[18].

It should be noted that WebDAV allows for simple metadata attached to files and directories. Many of the missing SRM features could be supported by profiling metadata values; for example, properties could describe a file: which pins are active, its access-latency, retention policy and locality. Although such a profile would be outside the specification, a sufficiently standards-complaint clients would be able to interact with storage systems.

### *2.1. Advanced features of CDMI*

In addition to providing features similar to SRM, CDMI includes a number of innovations that may prove useful to the HEP community.

CDMI can expose storage as block-storage, file-storage and object-storage. The file-storage is broadly analogous to SRM and WebDAV. The object-storage provides a mechanism to store data with a single, system-supplied unique ID as a reference. Stored files must be registered in some central catalogue if they are to be used by others. Once registered, the local filename is redundant; indeed, in some cases, the path is auto-generated. Using an object-store removes the need to auto-generate a unique path and simplifies the client and management operations.

CDMI protocol allows the client to store an arbitrary JSON object as metadata against either a file or directory. A CDMI client can then query or update some arbitrary subset of a file's stored metadata. The protocol also supports the possibility of searching for files that match metadata criteria. Information currently stored in the path could be stored in the file's metadata, allowing the storage system to optimise data distribution by using this data to predict likely access patterns. Physics metadata could also be stored in file metadata[19][20]. This would allow jobs to query the storage system for data matching physics predicates[21], thus avoiding the overhead of opening files that the job is not interested in.

### *2.2. dCache support for CDMI*

At the time of writing, the dCache team has a prototype implementation of CDMI providing roughly the same level of support as WebDAV, which we anticipate releasing in early 2014.

After the basic support has been released, work will switch to support for storing and fetching user-defined metadata, initially to allow photon science workflows that require storing both data and the associated metadata[22].

Once this work is completed, more advanced features will be introduced, such as querying for files that satisfy metadata predicates and allowing storing and receiving data via the object store functions.

## **3. NFS in HEP**

The requirements for storing data in HEP has always been challenging[23]. Simply adding more storage to a single server does not scale as certain resources (RAID controller, internal bandwidth, network adaptors) cannot scale indefinitely. To provide storage resources beyond such limitations, multiple servers must be used.

Until recently, the readily available solutions that supported multiple servers provided a single front-end node that mediated access. While functional, this approach prevents storage systems from scaling to the levels demanded by HEP-scale data analysis.

The approach taken by HEP has been to allow direct data transfers between the end client machine and the server node storing the file's data. Various protocols were developed to support this model: DCAP, rfiio, rootd and later xrootd. Such protocols have remained non-standard and proprietary, typically implemented by a single storage system or library. Extensions to FTP that allow direct transfers have been developed and standardised[5] but their use outside HEP remains limited.

The first minor version update of NFS v4[24] introduced an optional feature called pNFS. The pNFS extension allows an NFS client to read and write data directly from servers other than the one mounted. In essence, pNFS should provide the same performance profiles as the HEP-specific protocols, but in a standards-based protocol.

It is worth emphasising the benefits of adopting NFS as a basic protocol. Although dCache supports HEP-specific protocols, such as DCAP and xrootd[25], the clients of such protocols receive support only from a small community of developers. These developers lack the resources to optimise usage with the analysis platform. For example, no HEP-specific protocol has support in the Linux kernel; instead, shims or work-arounds are used, such as pre-load libraries[26] or FUSE plugins[27]. Other potential solutions (such as Lustre[28], GPFS[29] and Ceph[30]) exist but are not widely deployed within WLCG, so not considered by this paper.

### *3.1. Current status*

DESY, as the dCache partner most strongly involved with NFS development, has been building up operational experience with dCache NFS servers over several years[31]. The photon-science community requires the ability to store data through a mounted filesystem, which dCache NFS provides and has been in production for over two years. The analysis software for the Belle II experiment also requires a mounted filesystem and has been using NFS access to dCache for over one year.

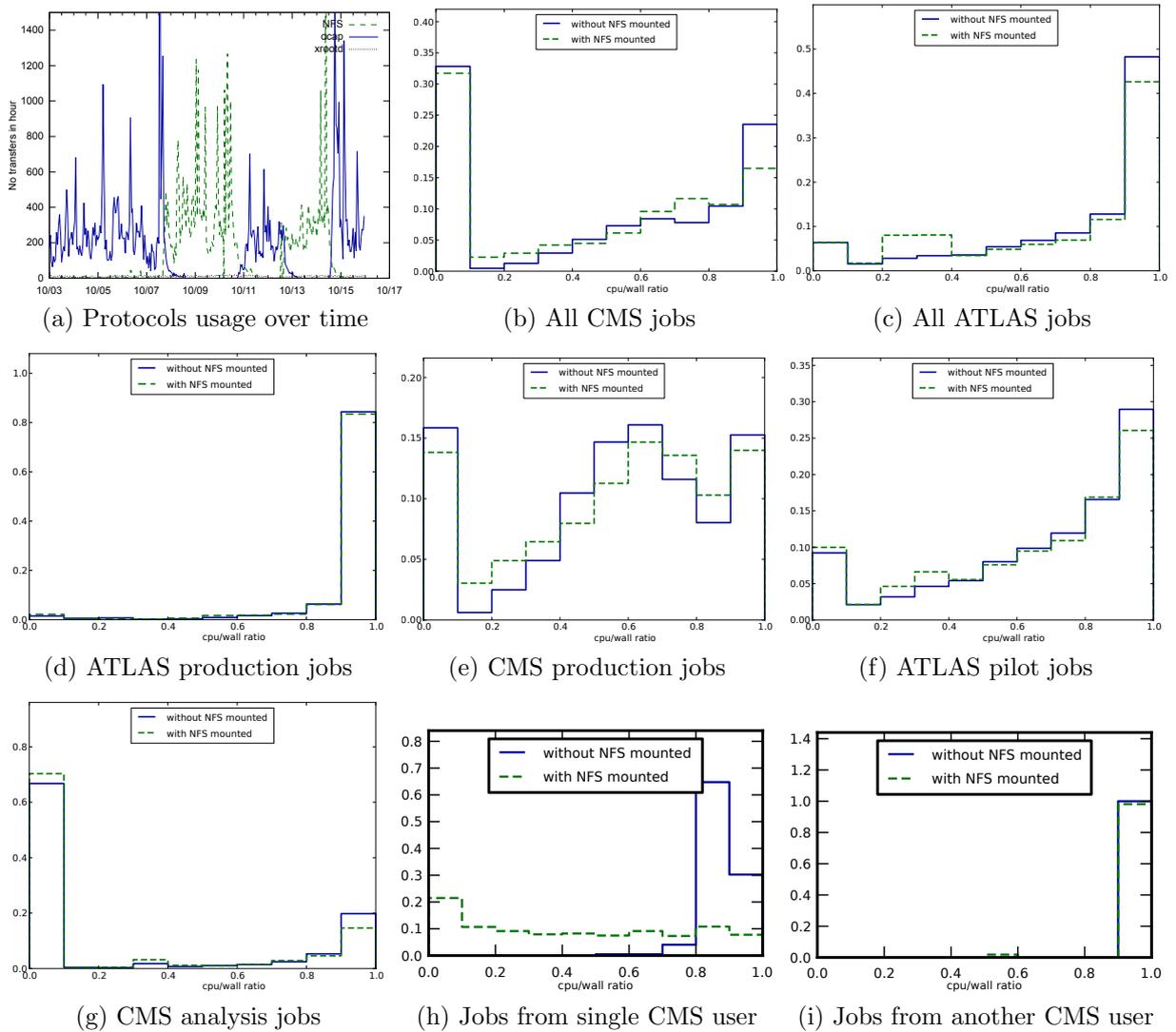
Since the WLCG worker-nodes have been upgraded to Scientific-Linux 6, it has been possible to mount dCache on these machines. DESY grid cluster[11] worker-nodes can run some 7,700 jobs in parallel. This is a far greater load than any dCache NFS server has been exposed to, so we wanted to roll out NFS support slowly to gain experience and watch for problems.

The first stage involved mounting dCache on a single worker-node and updating the CMS Trivial File Catalogue[32] so that any CMS job landing on this node uses NFS for reading. The worker-node continued to receive a mixture of CMS and non-CMS jobs. This allowed the verification that non-CMS jobs were unaffected by the NFS mount and any concurrent CMS jobs that access files through NFS. This deployment was run for 10 days during which CMS and non-CMS jobs continued to function well on this node.

The second stage involved increasing the number of selected worker-nodes to 60 (roughly 1,000 job-slots). In addition to testing dCache at a higher load, the increased number of worker-nodes allowed us to gather better statistics to compare NFS and DCAP performance. For this comparison, a subset of unmodified worker-nodes was selected with hardware and number of job-slots similar to those with NFS mounts. The job efficiency (CPU- to wall-clock ratio) for CMS and ATLAS jobs was collected and compared for various kinds of jobs. This has the advantage of checking real-world behaviour (rather than some synthesised benchmark) but with the disadvantages of being neither reproducible nor under our control.

The preliminary results are shown in Figure 1. The usage of protocols (excluding GridFTP) for the NFS-mounted worker-nodes is shown in Figure 1a with the effect of editing the Trivial File Catalogue clearly visible. The remaining plots are generated from jobs running on three days (2013-10-08 to 2013-10-10, when NFS was heavily used) and compare the efficiency for jobs

Figure 1: Preliminary NFS performance results. (a) shows the effect of changing the Trivial File Catalog; (b) to (i) show the job efficiency for different groups of jobs with values normalised so the area is 1.



for the two sets of worker-nodes. The graphs are normalised so the area under each curve is 1.

Although the results are preliminary and require further analysis to detect possible systematic errors from selection effects, early results seem promising. There are two points worth noting: some ATLAS jobs appear to show some deviation (see Figures 1c and 1f) and some specific user-submitted jobs appear to show a strong decrease in performance when using NFS (see Figure 1h). Further data and analysis is needed to clarify whether these effects are real, or whether they are instead some statistical fluctuation or artefact of our analysis.

During testing a race condition in dCache (the pool service) was discovered which resulted in an incorrect response from the pool when under heavy load. The client (the worker-node) then attempted to recover from this, but did so in a way that dCache did not expect, namely, by making alternate OPEN requests to the NFS server and READ requests to the pool node in a tight loop, which was processed at  $\sim 120$  Hz. As each open file requires the NFS server to use some memory, after  $\sim 40$  hours and with some  $\sim 17,000,000$  concurrent open files, the NFS server

ran out of memory, died and was automatically restarted. It is worth noting that the cluster survived this restart and recovered completely automatically. Although some jobs (~100) failed, other jobs and all subsequent jobs were unaffected by the restart.

The experience gained from this problem resulted in fixing a bug in dCache and modifying dCache to be more robust against such client activity. The CMS dCache cluster at DESY will be updated with these changes at the next scheduled downtime to allow further testing.

Previous analysis[12] suggests that NFS should out-perform DCAP for HEP analysis. While the initial results are encouraging, they do not reflect this anticipated enhanced performance. We will investigate this discrepancy and invest effort into improving the NFS performance, based on our real-world experience.

However, assuming there are no “show stoppers”, DESY plans to update all CMS jobs to use only NFS for reading by the end of 2013. Once this is achieved, DESY will look at migrating other VOs running at DESY to use NFS and support other dCache sites that want to transition to using NFS.

#### 4. Authentication

Traditionally, authentication has been closely associated with the service that uses the information gathered through authentication: typically users will present a user-name and password when they log into some service. There are several alternative methods of authentication (for example, two-factor authentication, one-time passwords, asymmetric encryption), but these do not alter the requirement to identify oneself with the service.

With the arrival of single-sign-on services, such as Kerberos[33], responsibility for authentication becomes somewhat blurred: the service with which the user authenticates, the Key Distribution Centre (KDC), issues a token that is then used, transparently, when interacting with a service. A strong trust relationship exists between the server providing a reliable and secure service and the KDC as the server now also depends on the KDC functioning correctly.

Within WLCG, a solution is needed to allow users to authenticate with services in many institutes. This requires the same trust relationship to exist between the service and the institute that issued the user’s credential. The IGTF[34] is successful in establishing this trust, providing a world-wide set of trusted Certificate Authorities (CAs).

Users new to grid computing often experience problems in acquiring an X.509 certificate, as the CA must vet the users identity. For new communities, it is harder as they often have no access to an IGTF-accredited CA and establishing a new CA is prohibitively expensive.

However, often this overhead is unnecessary as the user’s home institute has already vetted the user with at least the same level of scrutiny as IGTF-membership requires. Because of this, several projects have appeared that allow the user to obtain their certificate by authenticating against their home institute[35]. While this is an improvement, it still requires the use of X.509 certificates. This is problematic as most software does not make handling X.509 easy for the user.

LSDMA[36] (Large Scale Data Management and Analysis) is a project funded by the German government to link research of the Helmholtz Association of research centres in Germany with community specific Data Life Cycle Laboratories (DLCL). The DLCLs work in close cooperation with scientists and they process, manage and analyse data during its whole life cycle. The joint research and development activities in the DLCLs lead to community-specific tools and mechanisms. The DLCLs are complemented with a Data Services Integration Team (DSIT). This provides generic technologies and infrastructures for multi-community use, based on research and development in the areas of data management, data access and security, storage technologies and data preservation.

dCache, as major partner within the LSDMA DSIT team, is working on solving these authentication problems through authenticating via SAML. This is non-trivial as almost all

SAML usage has been for web-driven activity while most data transfers are not initiated by a web-browser.

Within DSIT, the plan is to establish a federation of Identity Providers (IdPs) for Germany: LSDMA-AAI. This federation is similar to the existing DFN federation (DFN-AAI)[37] and we anticipate LSDMA-AAI becoming part of DFN-AAI in the future. The initial work will be on providing web-portals that allow a user to authenticate via SAML. An X.509 credential is built from that information, which is used when authenticating with data servers. This follows existing work in this direction[38].

Group-membership is a common concern across many projects as often it is desired to authorise some operation based on that user's group membership. A common group-membership service will be run within Germany to allow services to discover group membership of a user.

Later, support in dCache for certain transfer protocols (e.g., FTP and NFS) will be updated to allow direct SAML-based authentication. This will allow clients to access their data directly without a web-portal using their home institute credentials in a secure fashion.

## 5. Conclusion

We have shown that dCache continues to focus on standards-compliance, which will allow new communities to more easily adopt dCache services. Existing dCache sites have started moving existing users over to these standards to better support them, reduce the site's support load and to improve performance by benefiting from other's work.

By expanding into new scientific user base (for example, through dCache's participation in LSDMA) and solving the storage requirements of new user-groups, dCache continues to evolve. This keeps dCache at the forefront of managed storage for scientific users.

The new abilities introduced into dCache, backed by standards, allow both existing and new HEP communities to push for the most efficient usage of storage by removing barriers and enabling scientists to best extract meaning from their collected data.

## Acknowledgments

Work described in this paper was funded by LSDMA, DESY, Fermilab and NDGF.

## References

- [1] The dCache project website <http://www.dcache.org/> accessed: 2013-10-30
- [2] Oleynik G, Alcorn B, Baisley W, Bakken J, Berg D, Berman E, Huang C H, Jones T, Kennedy R, Kulyavtsev A *et al.* 2005 *22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies* pp 73–80 doi:10.1109/MSST.2005.16
- [3] Deatrach D, Liu S, Payne C, Tafirout R, Walker R, Wong A and Vetterli M 2008 *22nd International Symposium on High Performance Computing Systems and Applications* pp 167–171 doi:10.1109/HPCS.2008.27
- [4] Haupt A, Leffhalm K, Wegner P and Wiesand S 2011 *Journal of Physics: Conference Series* Doi:10.1088/1742-6596/331/1/012007
- [5] Behrmann G, Fuhrmann P, Grønager M and Kleist J 2008 *Journal of Physics: Conference Series* vol 119 (IOP Publishing) p 062014 doi:10.1088/1742-6596/119/6/062014
- [6] Fuhrmann P and Güllow V 2006 *Euro-Par 2006 Parallel Processing* (Springer) pp 1106–1113 doi:10.1007/11823285\_116
- [7] REBUS: C5 report <http://wlcg-rebus.cern.ch/apps/gt/c5report> accessed: 2013-10-30
- [8] Stewart G A, Cameron D, Cowan G A and McCance G 2007 *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68* (Australian Computer Society, Inc.) pp 69–77
- [9] Donno F and Litmaath M 2008 Data management in WLCG and EGEE Tech. Rep. IT-Note-2008-002 CERN
- [10] Stoffers H, van de Sanden M, Raber J, Langborg T, Tsouloupas G, Rouchon O and Marceteau F 2012 Storage and long term preservation strategies in PRACE tier-1 datacentres Tech. rep. PRACE-RI
- [11] Gellrich A 2010 *Journal of Physics: Conference Series* vol 219 (IOP Publishing) p 062048
- [12] Elmsheuser J, Fuhrmann P, Kemp Y, Mkrtchyan T, Ozerov D and Stadie H 2011 *Journal of Physics: Conference Series* vol 331 (IOP Publishing) p 052010

- [13] Bernabeu G, Martinez F, Acción E, Bria A, Caubet M, Delfino M and Espinal X 2011 *Journal of Physics: Conference Series* vol 331 (IOP Publishing) p 072003
- [14] Shoshani A, Sim A and Gu J 2004 *Grid resource management* (Springer) pp 321–340
- [15] Donno F, Abadie L, Badino P, Baud J, Corso E, Witt S, Fuhrmann P, Gu J, Koblitz B, Lemaitre S *et al.* 2008 *Journal of Physics: Conference Series* vol 119 (IOP Publishing) p 062028
- [16] Serfon C, Vigne V, Beermann T, Goossens L, Garonne V, Nairz A, Lassnig M, Stewart G and Barisits M 2013 Atlas DQ2 to Rucio renaming infrastructure Tech. rep. ATL-COM-SOFT-2013-053
- [17] Cloud data management interface (CDMI) [http://www.snia.org/tech\\_activities/standards/curr\\_standards/cdmi](http://www.snia.org/tech_activities/standards/curr_standards/cdmi) accessed: 2013-10-30
- [18] Jensen J Heads in the cloud? GSM-WG at OGF-31 <http://www.ogf.org/OGF30/materials/2199/gsm-wg.pptx> accessed: 2013-10-30
- [19] Cranshaw J, Doyle A, Kenyon M, Malon D, McGlone H and Nicholson C 2008 *Journal of Physics: Conference Series* vol 119 (IOP Publishing) p 042008
- [20] Doherty T, Cranshaw J, Hrivnac J, Slater M, Nowak M, Quilty D and Zhang Q 2012 *Journal of Physics: Conference Series* vol 396 (IOP Publishing) p 052028
- [21] Mambelli M, Cranshaw J, Gardner R, Maeno T, Malon D and Novak M 2010 *Journal of Physics: Conference Series* vol 219 (IOP Publishing) p 072042
- [22] Flannery D, Matthews B, Griffin T, Bicarregui J, Gleaves M, Lerusse L, Downing R, Ashton A, Sufi S, Drinkwater G *et al.* 2009 *e-Science, 2009. e-Science'09. Fifth IEEE International Conference on* (IEEE) pp 201–207
- [23] Hoschek W, Jaen-Martinez J, Samar A, Stockinger H and Stockinger K 2000 *Workshop on Grid Computing (GRID 2000), Bangalore, India*
- [24] Shepler S, Eisler M and Noveck D 2002 Network File System (NFS) version 4 minor version 1 protocol <http://tools.ietf.org/html/rfc5661>
- [25] Behrmann G, Ozerov D and Zangerl T 2011 *Journal of Physics: Conference Series* vol 331 (IOP Publishing) p 052021
- [26] Agarwal A, Enge R, Fransham K, Kolb E, Leavett-Brown C, Leske D, Lewall K, Reitsma H, Rempel E and Sobie R 2010 *Journal of Physics: Conference Series* vol 219 (IOP Publishing) p 072024
- [27] Peters A 2008 *Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research* vol 1 p 41
- [28] Schwan P 2003 *Linux Symposium* p 380
- [29] Schmuck F B and Haskin R L 2002 *FAST* vol 2 p 19
- [30] Weil S A, Brandt S A, Miller E L, Long D D and Maltzahn C 2006 *Proceedings of the 7th symposium on Operating systems design and implementation* (USENIX Association) pp 307–320
- [31] Haupt A, Gellrich A, Kemp Y, Leffhalm K, Ozerov D and Wegner P 2012 *Journal of Physics: Conference Series* vol 396 (IOP Publishing) p 042026
- [32] Fanfani A, Afaq A, Sanches J A, Andreeva J, Bagliesi G, Bauerdick L, Belforte S, Sampaio P B, Bloom K, Blumenfeld B *et al.* 2010 *Journal of Grid Computing* **8** 159–179
- [33] Neuman B C and Ts'o T 1994 *Communications Magazine, IEEE* **32** 33–38
- [34] Simmel D, Rea S and Stolk A 2012 *Proceedings of 2012 Latin American Conference on High Performance Computing (CLCAR'12)*
- [35] Basney J, Fleury T and Welch V 2010 *Proceedings of the 9th Symposium on Identity and Trust on the Internet* (ACM) pp 1–11
- [36] van Wezel J, Streit A, Jung C, Stotzka R, Halstenberg S, Rigoll F, Garcia A, Heiss A, Schwarz K, Gasthuber M *et al.* 2012 *arXiv preprint arXiv:1212.5596*
- [37] DFN-AAI website <https://www.aai.dfn.de/> accessed: 2014-1-20
- [38] Groeper R, Grimm C, Piger S and Wiebelitz J 2007 *Software Engineering and Advanced Applications, 2007. 33rd EUROMICRO Conference on* (IEEE) pp 367–374