Journal of Physics: Conference Series **513** (2014) 032058 doi:10.1088/1742-6596/513/3/032058

Implementation of Grid Tier 2 and Tier 3 facilities on a Distributed OpenStack Cloud

Antonio Limosani[1] (speaker), Lucien Boland[2], Paul Coddington[3], Sean Crosby[2], Joanna Huang[2], Martin Sevior[2], Ross Wilson, Shunde Zhang[3]

1. ARC Centre of Excellence in Particle Physics at the Tera-scale, School of Physics, University of Sydney, Camperdown NSW 2006, Australia
2. ARC Centre of Excellence in Particle Physics at the Tera-scale, School of Physics, University of Melbourne, Parkville 3010, Australia
3. eResearch SA, Hedge House 14 Little Queen Street, University of Adelaide, Thebarton SA 5031, Austrlia

E-mail: antonio.limosani@sydney.edu.au

Abstract. The Australian Government is making a $AUD 100 million investment in Compute and Storage for the academic community. The Compute facilities are provided in the form of 30,000 CPU cores located at 8 nodes around Australia in a distributed virtualized Infrastructure as a Service facility based on OpenStack. The storage will eventually consist of over 100 petabytes located at 6 nodes. All will be linked via a 100 Gb/s network. This proceeding describes the development of a fully connected WLCG Tier-2 grid site as well as a general purpose Tier-3 computing cluster based on this architecture. The facility employs an extension to Torque to enable dynamic allocations of virtual machine instances. A base Scientific Linux virtual machine (VM) image is deployed in the OpenStack cloud and automatically configured as required using Puppet. Custom scripts are used to launch multiple VMs, integrate them into the dynamic Torque cluster and to mount remote file systems. We report on our experience in developing this nation-wide ATLAS and Belle II Tier 2 and Tier 3 computing infrastructure using the national Research Cloud and storage facilities.

1. Introduction

The Large Hadron Collider (LHC), located at the CERN laboratory, is the worlds most powerful particle accelerator [1]. Data collected in 2011 and 2012 by both the ATLAS [2] and CMS [3] experiments situated on the LHC ring provided the discovery of the long sought after Higgs Boson particle. Data thence and into the future will be used to elucidate Higgs Boson properties and attempt to make new fundamental discoveries of other hypothetical particles or phenomena.

The Australian high energy physics (HEP) groups based at the Universities of Adelaide, Melbourne and Sydney are institutional members of the ATLAS experiment and play significant roles in analysis of LHC data.The Australian ATLAS community consists of 40 researchers including academics and postgraduate students. Given the growth in numbers over the last decade, the community is expected to consist of at least 80 researchers by 2020. Australian researchers collaborate under the umbrella of the ARC Centre of Excellence in Particle Physics at the TeraScale (CoEPP) [4]. CoEPP is comprised of 80 physicists as it also includes theoretical particle physicists not involved in the analysis of ATLAS data.

CoEPP hosts a research omputing team responsible for providing the computing resources that contribute to ATLAS. Primarily this involves maintaining Australia's ATLAS Tier 2 grid site that is part of the WLCG, but also includes running local computing clusters for researchers based at the CoEPP nodes in Adelaide, Melbourne and Sydney. In ATLAS computing parlance local computing clusters are referred to as Tier 3 sites.

To help meet the needs of the growing pool of Australian researchers CoEPP has secured funding to implement a solution for either in-part or all of its future computing needs to be met in the Australian Research Cloud currently being commissioned with Compute and Storage nodes.

NeCTAR (National eResearch Collaboration Tools and Resources) [5] is a $47M Australian government funded project aiming to build new infrastructure specifically for the needs of Australian researchers. One of NeCTAR's four programs is the creation of a 30,000 CPU cores Infrastructure-as-a-service cloud spanning 8 locations around Australia, which is deployed on OpenStack-Grizzly.

Research Data Storage Infrastructure (RDSI) Project [6] is a $50M Australian government funded project to enhance data centre development and support retention and integration of nationally significant data assets into the national collaboration and data fabric. RDSI will deploy 100 petabytes of storage over a distributed network.

Nodes hosting Compute and Storage will connected via 100 gigabit per second links provided by Australia's Academic and Research Network (AARNet) [7]. AARNET is the not-for-profit company that operates Australia's National Research and Education Network (NREN).

CoEPP hosts a ATLAS Tier-2 site with about 1000 CPU cores and a petabyte of storage in Melbourne, and three independent Tier-3 clusters at Adelaide, Melbourne and Sydney, with collectively comprise of about 120 CPU cores and 110 terabytes of storage. CoEPP was awarded funding through NeCTAR's eResearch Tools program for a project titled "High Throughput Computing for Globally Connected Science" to provide research software for the Australian particle physics community and address specific research needs. It has a strong focus on enhancing existing tools and applications to be more collaborative, accessible and support research workflows. The eResearch tools will be deployed on NeCTAR'S Research Cloud. Our project has running for since March 2012 and is due to terminate in March 2014. Specifically the two main goals of the project are to:

- Augment Australian ATLAS Tier 2 capacity
- Build a federated Tier 3 for high throughput data analysis and processing.

The CoEPP Research Computing Centre will fully take over operations and maintenance after the end of March 2014. Practically these outcomes will result from the integration and interoperability between national Australian infrastructures of NeCTAR, RDSI and AARNET and existing grid resources through the deployment of a cloud Tier 2 and Tier 3. Currently the NeCTAR research cloud consists of about 4200 operational CPU cores hosted at the University of Melbourne, Monash University and the Queensland Cyber Infrastructure Foundation (QCIF). Currently CoEPP has been allocated 400 of those CPU cores. RDSI Storage storage has been provided to CoEPP at QCIF in Brisbane (150 TB - production system), Intersect in Sydney (1 TB - pilot system), and eResearch South Australia (eRSA) in Adelaide (4 TB - production system, 20 TB - pilot system). It's expected that by the end of 2014 CoEPP would have secured allocations of 1000 CPU cores and 1 petabyte of storage from NeCTAR and RDSI, or in other words, would have doubled the capacity provided by in-house resources.

2. Australian ATLAS Tier 2

To deploy cloud resources into the Australian ATLAS Tier 2, and specifically to run ATLAS MC production and analysis jobs, the following has been done:

- Integration with the ATLAS PANDA [8] job framework.
- Inter-operability with the OpenStack cloud [9].
- Dynamic scalability using the Condor [10] batch system and Cloud Scheduler [11] software, where the latter has been done in collaboration with University of Victoria (Canada).
- Deployed the CERN VM File System [12] (CVMFS) to distribute software libraries.
- Automated service configuration via Puppet [13] e.g. VM contextualisation.

The system framework and indicative workflow is depicted in Fig. 1.

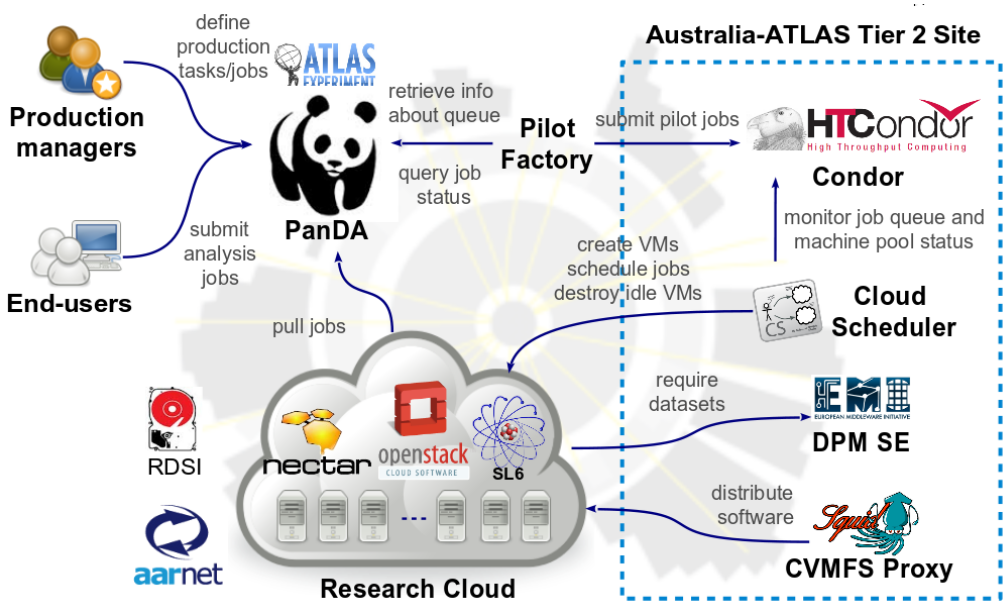

Figure 1. Tier 2 Cloud System Framework.

A key achievement of the project has been the development of software to efficiently launch VMs with all needed services, so-called VM contextualisation, such that the VM is a worker node fit to handle jobs. ImageFactory [14] is an open-source software which builds guest images for various operating systems and cloud providers, including Scientific Linux 6 (SL6) and OpenStack. We prepare the template definition file for ImageFactory to build them into the guest image. Once the image is built, we then use ImageFactory to upload it to the OpenStack Glance image service via its OpenStack plugin. Custom Puppet modules and manifests are used to further contextualise the VM. This VM is a so-called Golden VM, of which a snapshot is taken. This snapshot image is used to instantiate new worker nodes in the cloud. Workflow for the contextualisation of the VM is depicted in Fig. 2.

To date 160 CPU cores have been deployed into the Australian ATLAS Tier 2. These are distributed as 20 8 CPU core SL6 machines on the NeCTAR cloud. These resources are labeled as "Australia-Nectar" queue in Fig. 3, which shows running jobs over the period of a day in the Canada queue pool.

Efforts are ongoing to ensure system robustness and reliability through careful study and subsequent tuning of filesystem, network, and disk I/O configurations. Future work plans include supporting the Belle II collaboration as another Virtual Organisation (VO) in the Tier 2 and expansion of compute and storage resources into other geographic locations *i.e.* fragmenting or deploying separate Tier 2s.

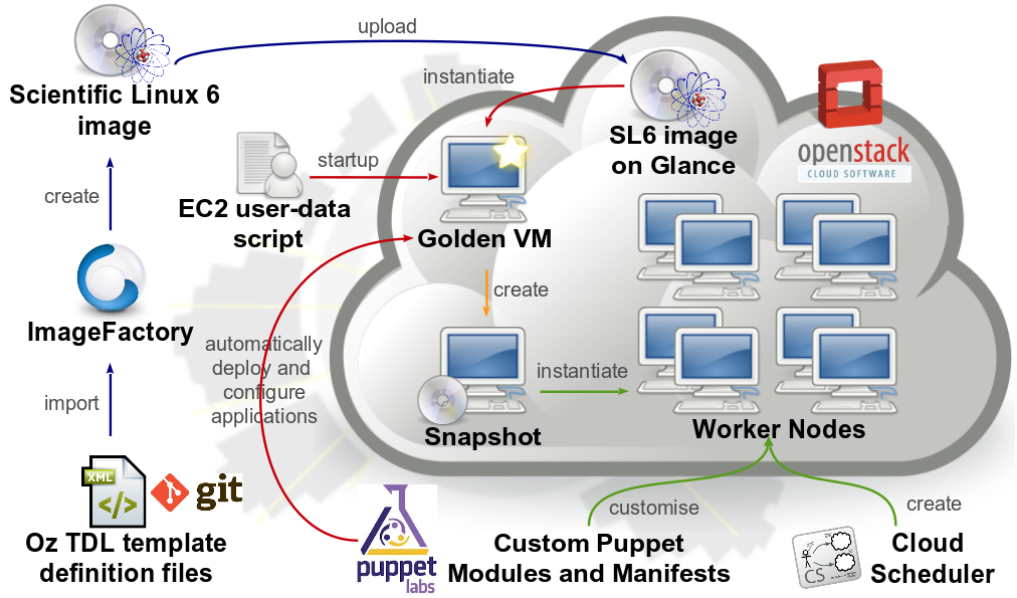

Figure 2. Virtual machine contextualisation

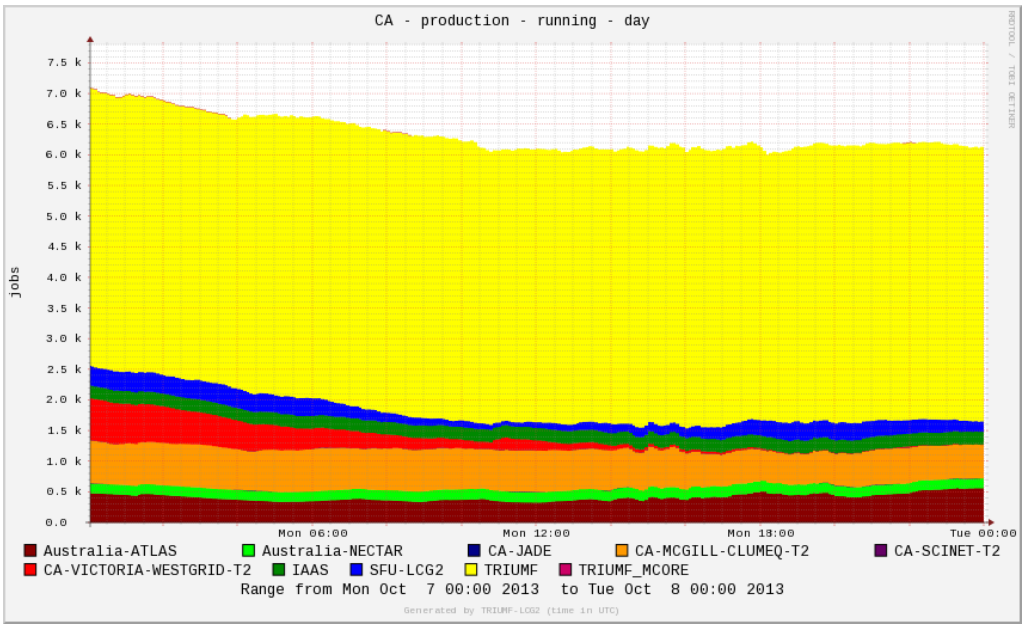

Figure 3. ATLAS jobs over the period of a day in the Canada queue pool. Australia-Nectar shows the number of jobs running on the NeCTAR OpenStack cloud and thereby adding to the number run by the pre-existing Australia-ATLAS queue.

3. Tier 3

Currently a Tier 3 (computing and storage cluster) relying on cloud resources for interactive use and batch processing is currently in production and accessible to all CoEPP users. Currently 20 terabytes of storage is exported via NFS to the worker nodes in the cloud in a partition named /data. Two log in nodes for interactive use and for job submission are provisioned, to

ensure high performance, with 16 CPU cores, 8 GB of memory, and 480 GB local storage each. Batch processing utilises the Torque resource manager [15] and the Maui cluster scheduler [16] services, coupled with an in-house extension to Torque that allows for dynamic resource and provisioning [17]. A CVMFS server is used to host and distribute software to meet local users' requirements. Currently the /home area is located only at the Melbourne node. We are investigating filesystems such as PVFS [18] and CEPH [19], to allow users to access local resources to reach their home areas. Xrootd [20] is currently being tested and tuned in preparation for deployment across RDSI nodes and will be used to provide a common namespace for a distributed large data area using XrootdFS. The latter will store data sets commonly required or outputs produced by data analysis, simulation and processing work. LDAP and Kerberos are the authentication and authorisation architecture. Xrootd is attractive due to ease of implementation and flexibility which allows clients to easily identify, transfer and process data directly or via the batch system, and can use the File residency manager (FRM) service to allow for caching of data sets, and so can improve throughput between geographically separated sites.

Using central control services we can build a new Tier 3 worker node on the cloud within minutes. Currently there are 94 static and 40 dynamic single CPU core worker nodes in the cloud. If the job queue reaches the point where the 94 static CPU cores are utilised and more jobs are queued dynamic torque will provision additional VMs to handle the overflow. Therefore the maximum capacity is 134 CPU cores. The use of the Tier 3 is demonstrated in Fig. 4 and Fig. 5, where the aggregate wall time for user jobs on a weekly basis is plotted along with the number of CPU cores in use by the batch system on a given day. The latter demonstrates overflow demand being met by additionally provisioned VMs.

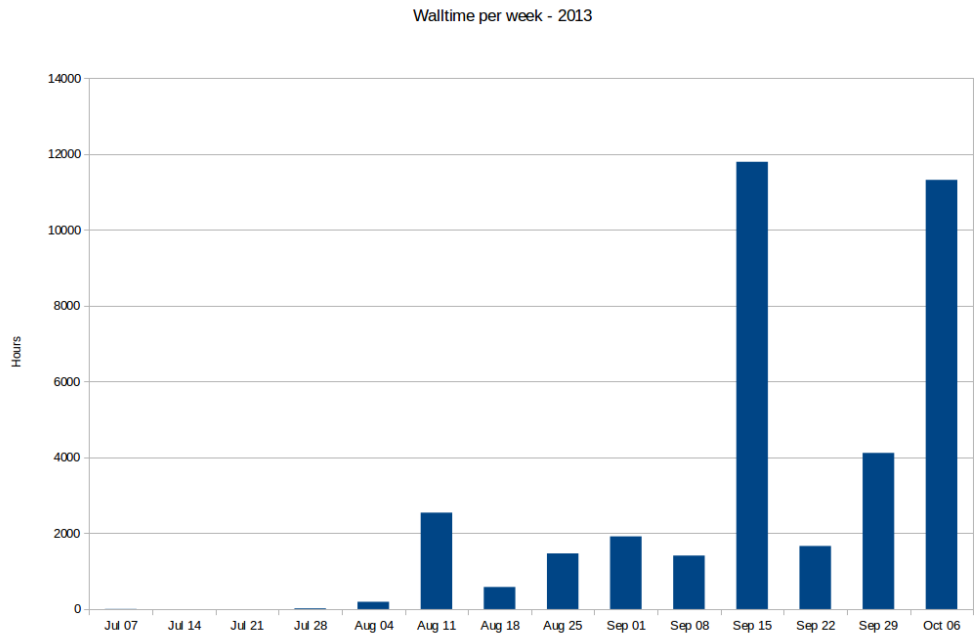

Figure 4. Australian cloud Tier 3 CPU wall time hours per week since July 2013.

The Tier 3 system framework goal is depicted in Fig. 1.

The following list summarises the federation wide central control services used to deliver our production Tier 3 cloud system

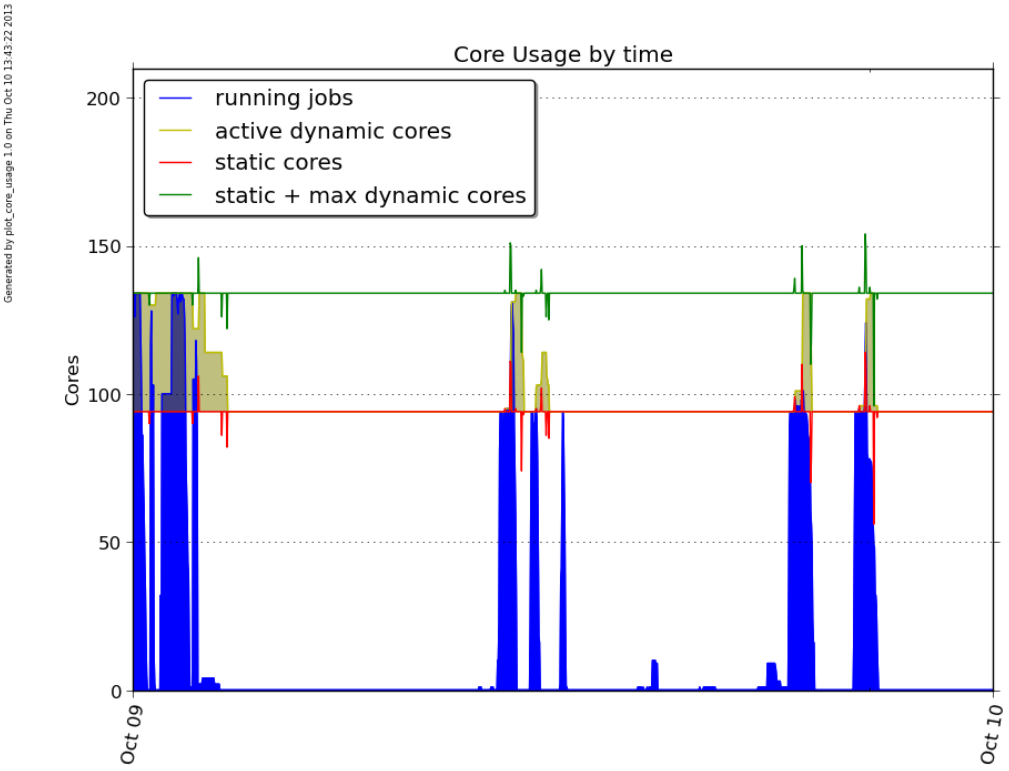

Figure 5. Australian cloud Tier 3 CPU core usage by time on October 9th, 2013.

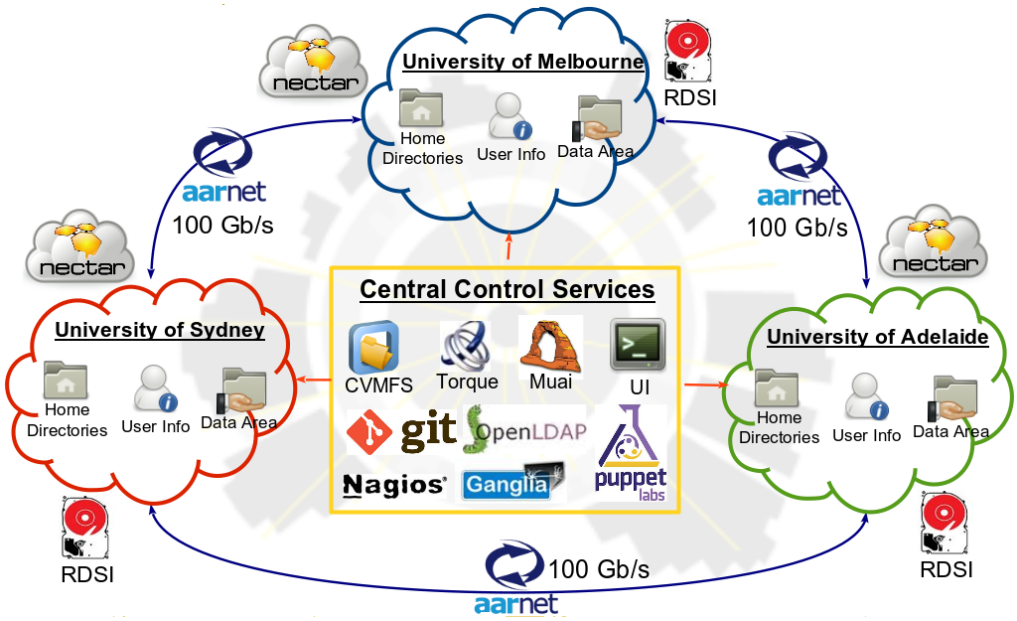

Figure 6. Tier 3 Cloud System Framework.

Puppet Configuration management and automated deployment.

CVMFS Experiment software repositories.

LDAP+Kerberos Authentication and authorisation architecture.

Torque+Maui+extension Dynamic computer cluster manager.

Distributed home area Currently using the Netowrk File System (NFS).

Nagios+Ganglia Service availability and performance monitoring.

User Interface Batch job submission node.

4. Summary

The Australian Tier 2 cloud is now fully functioning with operations to run large-scale production MC simulation for ATLAS. The CoEPP federated Tier 3 cloud is now in production providing extra computing resources on cloud to all collaborators of CoEPP. We are continuing to improve and manage our cloud infrastructure in a reliable and automated fashion and are simplifying operational processes to reduce maintenance costs. Current infrastructure is cloud-ready and allows for future expected growth in the number of CPU cores from NeCTAR and storage capacity from RDSI.

References

[1] http://home.web.cern.ch/about/accelerators/large-hadron-collider
[2] http://atlas.web.cern.ch
[3] http://cms.web.cern.ch
[4] http://www.coepp.org.au
[5] http://nectar.org.au
[6] http://rdsi.uq.edu.au
[7] http://www.aarnet.edu.au
[8] T. Maeno *et al.* "PanDA: distributed production and distributed analysis system for ATLAS" J. Phys.: Conf. Ser. 119 062036 (2008)
[9] http://www.openstack.org
[10] http://research.cs.wisc.edu/htcondor
[11] http://cloudscheduler.org
[12] http://cernvm.cern.ch/portal/filesystem
[13] http://puppetlabs.com
[14] http://www.aeolusproject.org/imagefactory.html
[15] http://www.adaptivecomputing.com/products/open-source/torque
[16] http://www.adaptivecomputing.com/products/open-source/maui
[17] See proceeding in this volume "Dynamic VM provisioning for Torque in a cloud environment" S. Zhang *et al.*
[18] http://www.pvfs.org
[19] http://ceph.com
[20] http://xrootd.slac.stanford.edu