Journal of Grid Computing manuscript No.
(will be inserted by the editor)

The Back-end of a 2-Layer Model for a Federated National Datastore for Academic Research VOs that Integrates EGEE Data Management

Brian Coghlan · John Walsh · Stephen Childs · Geoff Quigley · David O'Callaghan · Gabriele Pierantoni · John Ryan · Neil Simon · Keith Rochford

May 11, 2010

Abstract This paper proposes an architecture for the back-end of a federated national datastore for use by academic research communities, developed by the e-INIS (Irish National e-InfraStructure) project, and describes in detail one member of the federation, the regional datastore at Trinity College Dublin. It builds upon existing infrastructure and services, including Grid-Ireland, the National Grid Initiative and EGEE, Europe's leading grid infrastructure. It assumes users are in distinct research communities and that their data access patterns can be described via two properties, denoted as *mutability* and *frequency-of-access*. The architecture is for a back-end – individual academic communities are best qualified to define their own front-end services and user interfaces. The proposal is designed to facilitate front-end development by placing minimal restrictions on how the front-end is implemented and on the internal community security policies. The proposal also seeks to ensure that the communities are insulated from the back-end and from each other in order to ensure quality of service and to decouple their front-end implementation from site-specific back-end implementations.

Keywords Digital Repositories · Data Storage · EGEE · Data Management · Grid

1 Introduction

There is a trend towards shared academic infrastructures at national and international levels. These are well established for computation; are the subject of specific policy efforts, e.g. the US Cyberinfrastructure Committee, the European Strategy Forum on Research Infrastructures (ESFRI) and the e-Infrastructure Reflection Group (e-IRG); and have strong and well established flagship infrastructure projects, e.g. US Teragrid, US Open Science Grid (OSG), Distributed European Infrastructure for Supercomputing Applications (DEISA) and its successor, the Partnership for Advanced Computing in Europe (PRACE), EU Enabling Grids for E-sciencE (EGEE) and its successor the European Grid Initiative (EGI). But until recently there has been very little by way of data equivalents. The existing users need to continue to be supported, as they are doing large-scale science, like the CERN LHC experiments, that is dependent on the existing infrastructures. Hence the existing infrastructures need to be extended or integrated rather replaced. The support needs to be extended to encompass many other disciplines, including the arts and humanities, and also large-scale data management to support large national and international data-intensive science. To do this, existing sites within these infrastructures will need to be extended. Here we focus on large-scale data management and propose a suitable integrating architecture.

In recent years, various international bodies have issued recommendations and position statements on e-infrastructures and data repositories, such as those from the Organisation for Economic Co-operation and Development (OECD) [7], European Strategy Forum on Research Infrastructures (ESFRI) [5]) and e-Infrastructures Reflection Group (e-IRG) [14]). The U.S. National Science Foundation have put forward a Cyberinfrastructure Vision for 21st Century Discovery [6] and in Australia the Australian National Data Service (ANDS) has been established to influence national policy in the area of data management in the Australian research community, inform best practice for data-curation and guide the transformation of disparate collections of research data into a cohesive collection of research resources. Long term cura-

B. Coghlan · J. Walsh · S. Childs · G. Quigley · D. O'Callaghan · G. Pierantoni · J. Ryan · N. Simon
School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland. E-mail: geoff.quigley@cs.tcd.ie

K. Rochford
Dublin Institute of Advanced Studies, Dublin 2, Ireland. E-mail: rochfordk@cp.dias.ie

tion of data is also being addressed at national and international levels by efforts such as the Digital Curation Centre in the UK and the European CASPAR project (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval).

The OECD identified a series of principles for access to research data from publicly funded sources, briefly:

Openness: openness means access on equal terms for the international research community at minimal cost.

Flexibility: dealing with the current diversity of requirements and rapid/unpredictable changes in IT and research.

Transparency: what data exists, its source, documentation and conditions of use should be available transparently.

Legal conformity: respect legal rights, requirements and legitimate interests of all stakeholders, protect privacy and data with commercial, legal or national security issues.

Protection of intellectual property: consider the applicability (or not) of copyright and IP laws.

Formal responsibility: promote formal, explicit practices regarding the roles and responsibilities of the parties.

Professionalism: professional standards and values should be the basis for management of data.

Interoperability: interoperability at both technological and semantic levels are key considerations.

Quality: value and utility of data are largely dependent on the quality of the data itself; record provenance of data.

Security: specific attention should be paid to ensuring the integrity and security of research data.

Efficiency: increase efficiency of publicly funded research and prevent costly and unnecessary duplication.

Accountability: performance of data access should be subject to periodic review.

Sustainability: consideration should be given to the sustainability of access to publicly funded research data.

A useful example of the direction in which policy is being guided, and one that has influenced decisions in the e-INIS project, is the position paper [5] from the ESFRI working group about digital repositories. They state that digital repositories must provide availability, permanency, quality, right of use and interoperability. ESFRI projects must have a policy on each of these five areas. At the time of writing this involves 44 projects with funding of ∼€20 billion for construction and ∼€2 billion per-year operation. The recommendations on each of the 5 areas are as follows.

Availability: Keep repositories close to the data source. Maintain availability of data.

Permanency: Data must remain available from the creation to the time any user may need them.

Quality: Data must be quality-proved, curated.

Right of use: Public funded research produces data for public access and use. Restriction may apply temporarily to prepare publications or as per prior contract.

Interoperability: Digital repositories must be interoperable, i.e. using open standards and cross-references. Application of universal naming/referencing should be considered.

The eIRG data management taskforce [14] have surveyed existing data management approaches and further examined the quality and interoperability issues in data management. Acquisition of high-quality raw data is a sine qua non for useful derivative science, but it must also be attributed with high-quality metadata (meaning *data about data* or *data describing resources*, depending on usage) that describes resources/services and thereby semantically classifies data, as well as capturing provenance, etc. At present data and related metadata are on many kinds of systems using various data models. The interpretation of metadata is very discipline-specific, and appears to be the principal barrier to interoperable access to data from even cognate sub-disciplines. The analysis does not argue that all data must be reachable and usable by all, as this may not scale in effort and cost, but data should be widely accessible by design.

In this context it has also become clear that the really significant problem is the definition of semantic metadata and its wide-scale provision. This process has yet to begin for most scientific disciplines.

Most suppliers, when approached, ask for details of the expected access patterns and then refer to commercial datastore examples and the access patterns that they are designed for. This would often be the simplistic split of either a database-optimised server for structured data or a filestore for unstructured data. It would appear in this case that the large number of sub-disciplines and their very specific data and metadata access patterns imply there is no predominant access pattern or, at least, that none can be assumed. Nonetheless, access to the metadata generally involves querying a relational database (e.g. when using software such as AMGA or iRODS) and hence may involve many random accesses, just as is the case in most commercial settings. However the use of underlying storage is rather different than in most commercial environments. The bulk storage of large amounts of raw scientific data tends to result in once-off sequential writes followed by numerous reads of contiguous storage (e.g. ATLAS and CMIP5, as described in Section 8). Subsequent evolution of the datasets is often focussed on building and extending the metadata, which need not be co-located with the raw data. Here we focus on adding value to the bulk data storage.

As a result of the various considerations, the architecture described here addresses access to the back-end bulk data storage by a range of methods and leaves the front-end – the interface to the end-user – as a task for individual user communities. The scope of the project is also restricted to a pilot project and not, therefore, long-term (e.g. 50-year) curation/preservation, another significant problem. The focus

is also not on replacing local storage that could be cheaply provided with off the shelf consumer solutions but on providing storage for distributed communities of users. The e-INIS storage will add value by providing Internet-accessible storage and support for virtual organisations, by providing metadata and catalogue services and by providing management of the back-end and expertise.

The context in which this architecture is being proposed is the support of academic research communities in Ireland. The quality of shared Information Technology resources available to Irish researchers has been significantly advanced in recent years. Initiatives such as the National Capability Computing Service [4] and the e-INIS project have led to considerable investment in areas such as High Performance Computing and advanced computer networking. This national research infrastructure is being further developed and integrated under the e-INIS Federated National Data Store pilot activity. This data storage and management resource is an essential component of the overall e-Infrastructure being developed under e-INIS. It is critical to facilitating maximum re-use of shared data resources and extending the national capacity for data-driven research.

e-INIS is a federation of core electronic infrastructure providers dedicated to the provision of a sustainable national e-infrastructure supporting advanced research activities in Ireland. The project is funded under the Irish Higher Education Authority's Programme for Research in Third-Level Institutions (PRTLI), a component of the National Development Plan. By coordinating and consolidating the activities of national research ICT providers, e-INIS aims to provide a cohesive e-infrastructure of a scale that enables internationally competitive research. The e-INIS federation offers:

- Access to advanced capacity and capability computing facilities
- Specialist expert user support and training
- Secure network and grid services
- Pilot national data management services

Given the e-INIS project objective of extending national research capacity, the provision of large-scale data storage and associated management services is seen as a critical activity that will underpin the e-INIS national research infrastructure by facilitating maximum re-use of shared data resources and extending the national capacity for data-driven research. The e-INIS funding provides for a pilot national datastore, distributed across a number of regional datastores. This paper proposes an architecture and describes its initial implementation at the regional datastore in the OpsCentre at TCD.

The architecture proposed here attempts to solve the problem of providing diverse communities of academic researchers secure access to data storage and management facilities in a scalable manner that lends itself to distribution and incremental upgrades and yet supports the very different access patterns of those communities as well as the traditional groups such as high performance computing and grid users. This architecture builds on existing infrastructure and aims to extend the facilities available both to the current users and to new user communities. The federated architecture was selected in response to the recommendations mentioned above with respect to availability and keeping data close to the source where possible. The architecture is designed to be as flexible as possible in response to the recommendations on interoperability and right of use. Security and specifically the insulation of communities from each other's activities is seen as critical not only to ensuring end-users only access the data to which they are entitled but also to ensuring high availability of the data that they are entitled to access.

The e-INIS regional datastore at Trinity College Dublin is run by the Grid-Ireland operations centre (OpsCentre). Grid-Ireland is the Irish NGI (National Grid Initiative). It is actively involved in EGEE (Enabling Grids for E-SciencE), Europe's leading grid computing project, and its planned successor EGI (European Grid Initiative). EGEE provides infrastructure for over 10,000 researchers world-wide, from such diverse fields as high energy physics, earth and life sciences. Grid-Ireland uses EGEE's gLite middleware [21], which includes highly scalable *grid-enabled* data management components. Grid-Ireland's OpsCentre runs a set of national services, such as catalogues and resource brokers [11], as well as a grid site that hosts a 768-core cluster and more than 300TB of grid-enabled storage, the latter being the result of a recent upgrade and forming the initial basis for the regional datastore. The current facility supports bulk transfers of scientific data from grid-aware communities well, although there is anecdotal evidence to suggest that the current grid access methodologies (particularly the strong security) are not acceptable to all user communities.

The paper is organized as follows: it starts at a high-level based on distinct user communities and access patterns, describes the pre-existing data storage and then moves on to propose an architecture to cost-effectively support the two major access patterns and have the flexibility to support other access patterns in the future. Both the software infrastructure and hardware are described. The paper concludes by describing some projects that are early-adopters of the datastore.

2 Access to the e-INIS Datastore

e-INIS has decided that each user community of one or more researchers will be considered as a virtual organization (VO). A single unified global e-INIS data address space will be supported, with the second level delineated by VO: /einis/<VO>/..., e.g. /einis/gene/... Each community will have to appoint a VO Manager, who will manage

VO membership. e-INIS will introduce a peer review process for allocation of space, based on the Irish Centre for High-End Computing (ICHEC) model for allocating computing resources, described in Section 7.4. Grid-Ireland already runs a VO membership service for Irish VOs and this standing infrastructure is being used to support the new, non-grid, user communities also.

The datastore has a two-layer hardware architecture, where the OpsCentre hosts a common back-end that can be accessed by user-community-specific front-end servers. In this proposed back-end architecture, to secure access via protocols unsupported by the pre-existing infrastructure, each user community will interface to the common back-end via a bridge server, see Figure 1, which controls which parts of the datastore they have access to as well as providing their preferred access protocols. It is envisaged that each community would have one or more bridge servers but common services could be shared amongst multiple communities so long as they remain sufficiently insulated from each-other's activities. The bridge servers are key to providing a solution that is both generic and secure. Although the conceptual model of the national datastore has two layers, with a multiplicity of front-ends and a common back-end, this proposed back-end architecture has two-layers, yielding in effect a three-layer model. Direct access to the back-end via the existing grid protocols will be maintained.

2.1 Access patterns

Since some data will never be modified, only deleted, remote read caching and/or replication will provide very high performance without consistency issues, allowing use of bottom-of-the-range 'Just a Bunch of Disks' (JBOD) technology. This can be far cheaper than the storage technology required for frequently changing data with its attendant consistency issues. The terms *immutable* and *mutable* are therefore used here to indicate this difference, i.e. data that can be edited in place is described as *mutable*. Also, some data needs to be available for frequent access, some is accessed less often and some rarely. Therefore the distinction is made between *online*, *nearline* and *offline* storage; *online* being used to denote storage ready for immediate access, *offline* for data in an archive (e.g. a tape in a safe) and *nearline* data that lies between the two such as a tape in a library or a spun-down disk. This yields six cases: *online-mutable* and *online-immutable*, *nearline-mutable* and *nearline-immutable*, and *offline-mutable* and *offline-immutable*, where in the *nearline* and *offline* cases the access is infrequent even if the data is *mutable*.

The distinction between access patterns will allow assignment of space reservations to the most cost-effective back-end technologies. The users will still see a single unified global e-INIS data address space, delineated at the third level by access pattern:

/einis/<VO>/...
/einis/<VO>/<immutable>/...

where /<immutable>/ is defined by the user community as the top level of the user community's *online-immutable* storage space. A similar approach may be taken for *nearline* space.

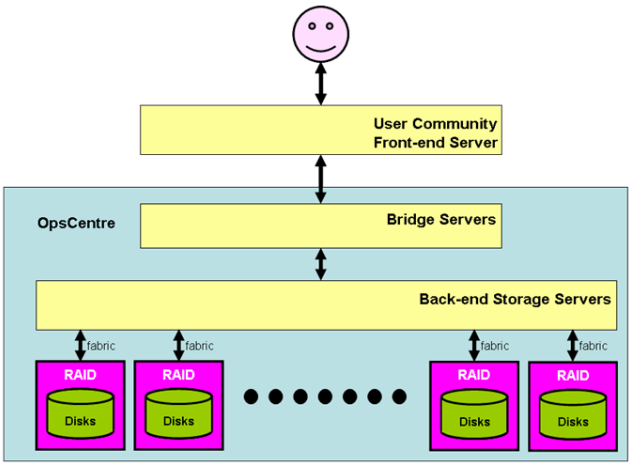

Fig. 1 Datastore high-level architecture

2.2 Data Lifetimes

The SRM [27] protocol used in EGEE's gLite data management defines three storage lifetime categories:

- *Volatile:* delete after expiry
- *Durable:* raise error after expiry
- *Permanent:* no expiry

Thus *volatile* data (e.g. temporary files) persist only during the expiry lifetime, *durable* data persists beyond the expiry lifetime, while *permanent* data persists. These are useful categories that the datastore may employ.

2.3 Authentication, Authorization and Accounting (AAA)

A federated access-management pilot, called EduGate, is currently being assembled by the Irish national research and education network provider, HEAnet. This uses SAML 2.0 [24] (e.g. using freely available software such as Shibboleth or simpleSAMLphp) assertions about user identity. Once federated access-management is widely deployed in Ireland, the OpsCentre will implement a Short Lived Credential Service (SLCS), probably based on the experiences

with the SWITCH SLCS [30] or similar. Accesses to the datastore will result in a Service Provider (SP) requesting a Where-Are-You-From (WAYF) server to query the user's institution for assertions about the user's identity, which the SLCS will convert to short-term certificates to secure accesses to the datastore. e-INIS has decided to adopt the Grid-Ireland public-key infrastructure (PKI)[1] until federated access-management is deployed as a production infrastructure. The project partners have agreed a policy that all write access to the datastore back-end will be secured to this level (PKI or federated identity).

The VO Manager of a user community will have to obtain a Grid-Ireland user certificate, a Grid-Ireland host certificate for their front-end server and, potentially, a Grid-Ireland robot certificate for the VO. The other users in that community will only require certificates if the community so decides and then the robot certificate can be used on behalf of the whole community when connecting to the back-end. In this case, the community (and VO manager) will implement their own access controls. One example would be for the community as a whole to have read-only access via the bridge and for the VO manager to handle all write access through their Grid-Ireland user certificate, thus allowing the community to manage themselves but ensure that the e-INIS security policy is adhered to and the community is identified as the source of the data written to the back-end.

The connection between the bridge host and the back-end storage will have to be secured by the related host certificates, but that is internal to the site.

e-INIS has decided that VO membership services will be provided by Grid-Ireland's existing Virtual Organization Membership Service (VOMS). This also allows VOMS-attributed certificates to convey VO information about the users to the various datastore instances.

e-INIS has also decided that VOMS information will be mirrored by a Lightweight Directory Access Protocol (LDAP) service and so the OpsCentre is running an experimental installation of VOMS-to-LDAP synchronisation software, developed by LAL. There is extensive web server-side support for LDAP, so this allows standard browsers and community portals to import some VO information. It will also be necessary to have an LDAP server to manage accounts on NFS storage and clients, in order to keep account information consistent, and an investigation is underway to determine whether the two tasks should use the same LDAP server.

The EGEE compute and data accounting software is already deployed by the OpsCentre. Compute accounting is quite mature, but the data accounting is as yet very undeveloped and only supports LFC.

[1] Grid-Ireland's uses the globus Grid Security Infrastructure(GSI)[15]

3 Existing Grid Infrastructure

Various necessary functions are already provided by Grid-Ireland. In addition to the VOMS services mentioned above, Grid-Ireland has its own CA (Certification Authority) and both central and distributed services for data management and computation using the gLite middleware suite, developed by EGEE. Two of the Grid-Ireland sites (TCD and DIAS) actively participate in EGEE (and will do in its successor EGI) as well as supporting Irish national research.

This EGEE-based infrastructure represents a platform that supports Irish participation in large-scale international research projects, e.g. the CERN Large Hadron Collider ATLAS and LHCb high-energy physics experiments and the HELIO solar physics collaboration, each of which includes EU and non-EU participants. Ideally its data management should be an integral part of the new datastore architecture, preferably without needing customisations. The proposed datastore architecture will integrate and leverage this global infrastructure and continue to support the existing user communities.

In its current configuration, the Grid-Ireland data management only supports *online-immutable* data (the EGEE gLite middleware only supports *immutable* data), but some EGEE sites outside Ireland have *nearline* and *offline* storage.

The OpsCentre in TCD has the largest storage resources currently connected to Grid-Ireland (hundreds of terabytes) and five other sites have $>$1TB storage available, including the e-INIS datastore partner sites in University College Cork, Dublin Institute of Advanced Studies and National University of Ireland Galway. All the remaining Grid-Ireland sites also have a storage element present, although there are presently only tens of gigabytes storage attached for test purposes.

4 Providing Online-Mutable Storage

The existing Grid-Ireland infrastructure currently only offers *online-immutable* storage which, until now, has only been accessible using the standard grid tools and a small number of web portals. To cover the full range of possible applications it is necessary to provide storage for data that may need to be edited in place while maintaining high availability, i.e. *online-mutable* storage.

Online-mutable storage could be accessed in a number of ways. Here we consider:

- catalogue access
- Fedora Commons access
- filesystem access, both by web-based export from a bridge (WebDAV, Davis, OPeNDAP) and via POSIX-like I/O (no code modification, standard POSIX semantics)

– streaming access
– block-oriented access

This list is neither exhaustive nor exclusive. For example, what is denoted here as catalogue access logically combines with the other methods in the list.

For each user community there will be a separate bridge Virtual Machine (VM) hosted on a bridge host (although some communities will warrant a complete physical server rather than a VM). The bridge VM will in effect mount a subset of the back-end storage, and will export that space to the user community front-end server. The connection to the front-end server will be secured by a virtual private network (VPN), see Figure 2. In specialised cases of small known user communities that do not need a front-end, their user clients and the front-end may be collapsed to direct connections from known IP addresses over a VPN to their bridge server VM.

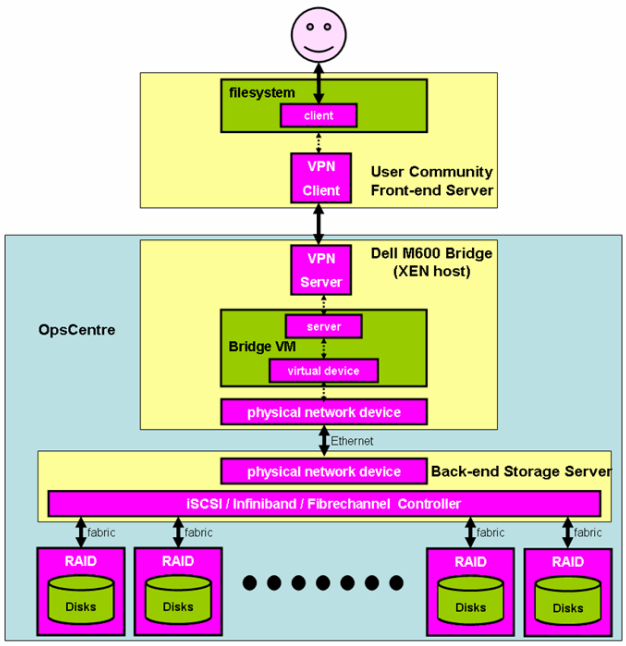

Fig. 2 Datastore *online-mutable* access architecture

As stated above, to use the bridge the VO Manager will have to obtain three Grid-Ireland certificates; a user certificate, a host certificate for their front-end server, and a robot certificate for the VO. The VO manager's user certificate will be used to secure the VOMS administration. The host certificate will be used to secure the VPN. The robot certificate will be used to secure accesses to the storage space on behalf of the community. The connection between the bridge host and the back-end storage will have to be secured by the related host certificates.

The intention is to present each user community with a view only of the storage it has access to; i.e. to insulate the various user communities from each other and from the back-end storage. To each community, its storage space will appear on a private network. The user community can use that storage space as it sees fit. Crucially, they have no execute capabilities on the back-end storage servers. This is reminiscent of a network "de-militarized zone" (DMZ) where those hosts are treated as being outside the trusted network but not as untrusted as a host under third-party control. The hosts will be a combination of blades and virtual machines connected by 10Gbit ethernet to the back-end storage using the top-level network switch.

4.1 iRODS Catalogue Access

The Integrated Rule-Oriented Data System (iRODS) system from University of North Carolina at Chapel Hill (the developers of the SRB, the Storage Resource Broker), is designed to support digital libraries, persistent archives, and real-time data systems. Importantly, iRODS is not only used by scientists and engineers - it has also seen extensive use in the curation and preservation of data from the arts and humanities, for example in the UK[16, 9].

Through its catalogue, the iCAT, iRODS provides a global namespace for digital repositories (which it calls *collections* in a manner analogous to *directories* in a filesystem), and allows communities to define policies for their collections in the form of rules (sets of assertions implemented as *microservices*) and state information. The rule engine interprets these to decide how the system is to respond to various requests and conditions. For example, the SRM storage lifetime categories could be implemented via appropriate rules and microservices.

The physical storage (*resources*) can be on multiple non-iCAT-enabled iRODS servers and data can be automatically distributed across resources, specifically placed on a particular resource or replicated across multiple resources. The namespace is separate to the resources and so one collection can exist across multiple resources. Rules can also be used to target particular collections to specific resources or resource groups. iRODS currently supports three classes of resource

– Cache - generally, storage resources with short access latency - which can be mapped to *online*
– Archival - storage resources with longer access latency - which can be mapped to *nearline*
– Compound - storage resources where data access I/O calls such as open, read, write, lseek, close, etc, are not readily available. Instead, *put* and *get* calls are used to transfer entire files. In the current implementation, a cache class storage resource must exist in the same

resource group as the compound resource. Data in the compound resource cannot be accessed directly but via the cache resource using staging/synchronisation.

The iRODS catalogue service (iCAT) will be hosted by an iCAT-enabled iRODS server in the OpsCentre in Dublin, although this will not host any data resources. The data collections will be hosted by a number of non-iCAT-enabled iRODS servers hosted on the back-end datastore servers. Each e-INIS datastore site will host its own back-end storage resources and associated data collections. The iRODS catalogue service will federate these collections. For access to the catalogue the bridge host will be enabled to access iRODS collections, and it will enable access to a subset of the collections from the bridge VM, so that a community can access its collection. See Figure 3.

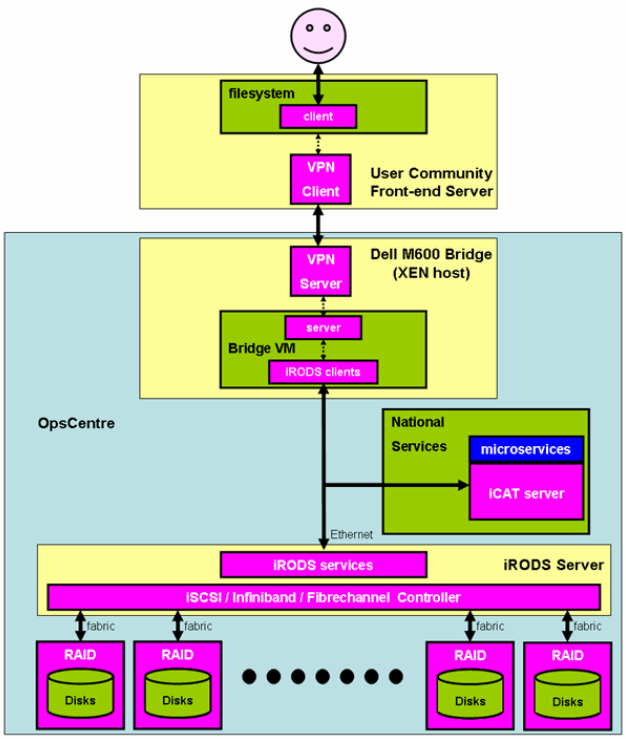

Fig. 3 Datastore iRODS catalogue access architecture

The data address space will be of the form: /einis/<VO>/... where /einis/<VO>/ is the top level of the user community's storage space.

Advantages: global namespace, rule-oriented, user-definable microservices.

Disadvantages: User unfamiliarity – iRODS is not widely used in Ireland.

4.2 Fedora Commons Access

Fedora Commons is a general-purpose, open-source digital object repository system. There are many ways in which iRODS could be integrated with Fedora Commons, as described in the iRODS webpage on Fedora Commons and in [8]. One example is to use iRODS to store Fedora Commons objects; see Figure 4. In this example iRODS provides a distributed storage environment for Fedora Commons, which can store both data objects and data streams as iRODS files that can be geographically distributed and are managed by iRODS. Upon request from Fedora Commons, iRODS will send data streams to Fedora Commons. This uses iRODS merely for storage, ignoring some of its core features such as rule processing. This is very worthwhile exploring.

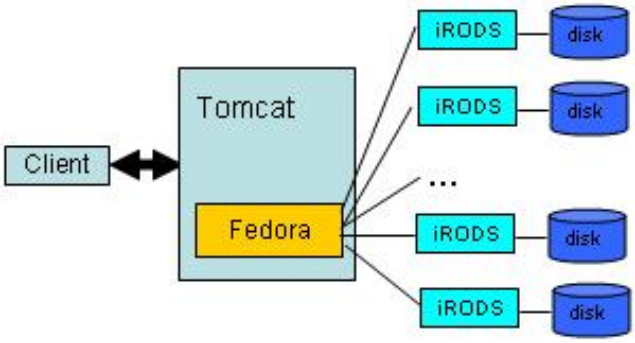

Fig. 4 Example architecture for datastore iRODS access from Fedora Commons (from the iRODS website on Fedora Commons)

Advantages: Federated storage for Fedora Commons.

Disadvantages: Does not maximally exploit iRODS' own metadata capabilities.

4.3 Web-based Filesystem Access

WebDav (Web-based Distributed Authoring and Versioning [29]) is being investigated as the initial implementation of filesystem access to the datastore. WebDAV is widely supported and is an extension of HTTP. It is proposed that a filesystem on the datastore will be exported to the bridge VM (e.g. using NFS), which will in turn export that over WebDAV to the user community's front-end server.

Work is also underway evaluating use of the Davis system [31] from University of Adelaide to enable WebDav access to iRODS. In this case the access is rather simpler - the bridge VM will be running Davis and this will connect directly to an iRODS server. In this case it is expected that GSI will be used to secure the connection.

Further work is exploring HTTP export via OPeNDAP (Open-source Project for a Network Data Access Protocol, [13].

Advantages: User friendly, web/firewall transparency.

Disadvantages: Users need create/delete permissions.

4.4 POSIX-like Filesystem Access

Front-ends may employ Filesystem in USEr space (FUSE) and/or Parrot [28, 18]. Both allow the datastore to leverage iRODS and/or filesystems. FUSE uses a user-space daemon, invoked by the front-end administrator, to mount a remote filesystem. There are many FUSE daemons available, including one for iRODS. Once the filesystem is mounted, all accesses to paths under the mount-point are passed by the front-end's Linux VFS to the daemon which then handles communications, etc. With Parrot, commands that will access storage are simply prefixed with the *parrot* command and then filesystem calls are captured (using ptrace) and handled. In both cases, no user-code modification is necessary – programs use standard I/O that meets most POSIX standards but run in a specific environment. One downside to these solutions is that they are tied to Linux. Parrot is only available on Linux and efforts to port FUSE to MacOSX and Microsoft Windows lack the maturity of the main project.

NFS access may be provided by a bridge VM re-exporting NFS-mounted datastore filesystem data. This allows front-ends to employ the industry standard networked filesystem with which their administrators are most likely to be familiar, using the same tools and protocols that are prevalent in their own organisations. This provides POSIX semantics, well understood behaviour (which meets most POSIX standards) and allows the use of most applications without any modification. Currently, research is ongoing into NFS re-exporting and associated performance issues. Similarly, an investigation is also ongoing into provision of CIFS (Commons Internet File System[17]) support via SAMBA (as an alternative to NFS re-export), in order to better support communities using Microsoft Windows.

Distributed/clustered filesystems such as AFS, GPFS and Lustre can also be used by individual sites to provide the underlying storage, potentially improving performance by spreading load across many servers while giving a single filesytem.

Advantages: Developer friendly, application I/O.

Disadvantages: For FUSE and Parrot, the drivers must be loaded/unloaded before/after usage (or at login/logout); this may not satisfy some user communities. For NFS and CIFS, re-export is non-standard. For clustered filesytems, the greater complexity and possible restrictions on choice of operating system are seen as a possible disadvantage.

4.5 Streaming Access

High-performance ($>$500MB/s) streaming access to iRODS has been demonstrated by the CineGrid project [19] at the University of Amsterdam by using alternate stream-oriented network drivers. Again this allows the datastore to leverage iRODS.

Advantages: High-bandwidth streaming I/O.

Disadvantages: Non-standard drivers.

4.6 Block-oriented Access

For block-oriented access, a partition of the datastore will be exported to the bridge VM as an iSCSI [25] target (other protocols may be considered in future) and an iSCSI initiator on the bridge VM will mount that partition and will in turn export that space as a network block device (e.g. iSCSI or gNBD) to the user community's front-end server, as per Figure 2 but with the iSCSI initiator as the virtual device. Alternatively, the bridge VM can export a block device backed by a file stored on the datastore. The front-end's iSCSI or gNBD initiator will therefore see an iSCSI or gNBD target and access it as it would normally do so, providing a block device for appropriate use by the front-end.

This allows for certain types of workload that cannot otherwise be catered for with traditional file type access, or other proposed solutions. Additionally, it may prove the most appropriate solution for some user communities, depending on their existing methods of operation.

The latency of access will be much greater than for local (including iSCSI) storage. User communities will be encouraged to use other methods of access for their front-ends in preference to using this type of access.

Advantages: The bridge VM can aggregate blocks and in theory there are no execute permissions whatever, only read/write.

Disadvantages: Limits potential to add value. Longer access latencies. Security concerns have been expressed regarding iSCSI command execution.

4.7 Supporting Open Access Policies

For communities that require minimal or no security, a level of indirection via the bridge host may be introduced to facilitate the communities in providing such an open access policy. For example, instead of the bridge VM directly mounting an NFS filesystem, the bridge host can mount the filesystem and export all or part of that filesystem to the bridge VM.

Advantages: Support for very open access policies.

Disadvantages: Severe performance penalties.

4.8 Other types of access

It is possible that a user community may wish to employ another form of access which has not been evaluated or, hitherto, has been regarded as low importance. In such an eventuality, the OpsCentre will attempt to evaluate their request with an eye to cost of maintenance and implementation, and also the potential benefit. The model proposed here, using virtual machines to provide a bridge from the back-end to the community's front-end server, should be sufficiently open to extensions that the support of these unanticipated requests will be possible.

4.9 Use of iRODS to provide federated datastore

iRODS has been selected by the e-INIS project partners as the means to provide a federated datastore. This choice was guided by various factors:

- iRODS is freely available for a range of platforms, actively supported and widely used globally
- iRODS has integrated support for metadata
- There are many actively developed clients and interfaces to iRODS so integration with community services can, in many cases, be achieved without developing new software
- It can easily be customised and tasks automated using rules and microservices
- Individual sites retain a high degree of choice in how they provide their storage resources — so long as it is compatible with export via iRODS.
- iRODS has some built-in support for GSI authentication

Some access patterns may be supported locally, at the individual partner sites, by direct export of storage e.g. by NFS, iSCSI in which case the allocation process still needs to be followed and the security model adhered to but the storage would not form part of the federation.

5 Provision of Online-Immutable Storage

5.1 Integrating EGEE Data Management

iRODS is well suited to the *online-mutable* patterns that are expected to be used in a significant proportion of data-sets. Nonetheless, it is also expected that a substantial proportion of data-sets will obey an *immutable* pattern, for example raw datasets from instruments, or backups/archives. iRODS can support this pattern directly and in entirety, but in that case could not avail of or leverage the existing global grid infrastructure or associated technologies. Hence an attempt will be made to enable *online-immutable* access from iRODS to EGEE data management by writing microservices that call the EGEE *gfal* grid file access library, see Figure 5. These microservices and associated rules are not customisations of iRODS but a normal instance-specific item of configuration. This would also allow the catalogue to include external *immutable* catalogues, for example the CERN LHC experiment catalogues. The microservices would provide the address translation: /einis/<VO>/<grid>/... → /grid/<VO>/... Whilst an attractive proposition, this faces the major problem that iRODS lacks support for delegated GSI credentials[2].

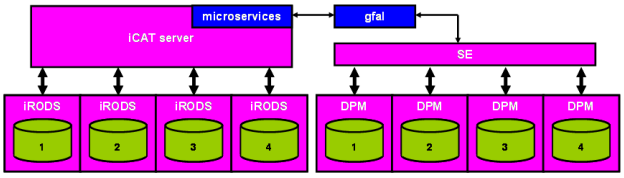

Fig. 5 Datastore *online-immutable* access architecture

However, if successfully implemented, the result will be that all *online-immutable* accesses will ultimately access a grid *storage element* (SE). In the Grid-Ireland infrastructure an SE resides at each site *gateway*, see [11]. Grid users may access any of the SEs using a number of grid tools, services and libraries. For example, data accesses can be made using the EGEE *lcg-utils* command-line tools, the EGEE *FTS* file transfer service, or from applications using the EGEE *gfal* library. If users have the requisite authorizations they can access SEs elsewhere in the world using the same tools.

The EGEE LHC File Catalogue (LFC) service provides a global namespace for *immutable* file repositories. Given that files are only written once, it specifically supports file replication to multiple SEs with a related namespace of instances. The associated EGEE AMGA metadata service [20] allows grid users to store various metadata in a database but in a file-oriented fashion, and is often used in conjunction with the LFC catalogue, where AMGA stores metadata about the files in LFC. It may be useful to attempt to enable iRODS to access AMGA metadata by writing microservices that invoke AMGA webservices.

Both LFC and AMGA assume the SE adheres to the SRM protocol [27]. Several SE implementations support this protocol, e.g. DPM, d-Cache and StoRM (details are readily available on the world-wide-web). Obviously the SRM storage lifetime categories are built-in.

The Grid-Ireland OpsCentre has deployed LFC since 2001, has deployed DPM to all its gateways since 2006, deployed AMGA in 2008, and expects to deploy StoRM in 2010.

[2] Grid components such as the LFC and DPM can be configured to trust particular hosts and services but there is a question over whether this can be made to fit within the security policy

If the user invokes LFC directly (not through iRODS), then as a byproduct of being grid-secured, bridging is not needed. Accesses are set up by SRM calls from the user's client to the SE, but the actual transfers are usually conducted using gridFTP [22] directly from the back-end SRM storage server (e.g. a DPM disk server) to the user's client, or vice versa, see Figure 6. These are known as *third-party transfers*, and the datastore hardware architecture specifically supports these with 10Ge network paths from the external network to the back-end storage servers. gridFTP can exploit this to achieve high transfer bandwidths by concurrently transferring multiple files over multiple channels *(streams)* between a client and a server and dynamically adjusting TCP properties. The transfers may also be done using HTTPS but not all SRM storage servers implement third-party transfers with anything but gridFTP (e.g. DPM doesn't), so a severe performance loss is incurred.

Advantages: Catalogues, storage pools, space reservation, storage categories, 3rd party transfers.

Disadvantages: Some individuals are averse to using grid services.

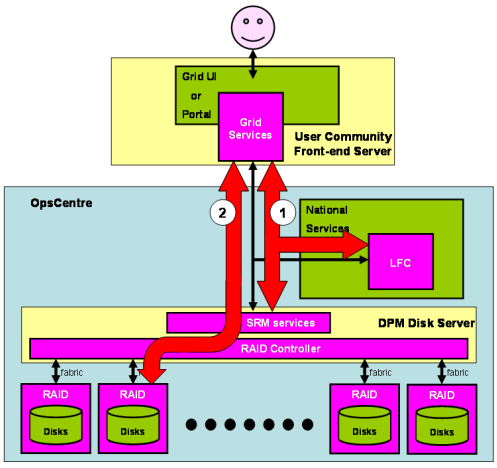

Fig. 6 Datastore leveraging of 3rd party transfers to/from online-immutable storage

5.2 Use of EGEE Data Management for ALL online-immutable Accesses

While iRODS could support all non-grid *online-immutable* accesses directly and the microservices could map /einis/<VO>/<immutable>/... to bottom-of-the-range JBOD technology, raw data (e.g. raw datasets from instruments) stored this way will not be accessible from the global grid infrastructures. Hence initially *ALL* such accesses will be mapped to EGEE data management, i.e. the microservices will provide the address translation: /einis/<VO>/<immutable>/... → /grid/<VO>/... If successful this will further leverage the existing global grid infrastructure and thereby guarantee wider accessibility of raw data. It will also allow iRODS to maximally exploit the much less expensive *online-immutable* LFC technology developed largely by the CERN LHC community. If unsuccessful then other avenues will be explored to provide access from the global grid infrastructures to iRODS data.

6 Nearline and Offline Storage

For the first services offered by the pilot the emphasis has been on providing storage that is available for immediate access as much of the time as can be achieved. There are two other important cases that are being considered at this stage (large dark-archives and curation being outside the remit of the project). These are backups of important data and the efficient storage of infrequently accessed data.

6.1 Backups

There are two main challenges for providing backups in this architecture. The first is the logistical problem of providing sufficient suitable storage for the backups. Few resources are available in the pilot specifically for this purpose so tape libraries are provided that only have the capacity for a subset of the data store. Both the grid middleware and iRODS support replication and this is seen as sufficient for the bulk of the data published in the national datastore. The second challenge is one of ensuring that the backup is a consistent snapshot if the filesystem is changing during the period that the backup is being generated.

In the case of WebDAV or POSIX-like filesystem access where filesystems are being (perhaps multiply) re-exported, attempting to backup the back-end filesystems could potentially result in corrupt images. Some responsibility therefore needs to be placed with the community to ensure their storage is in a safe state for a backup. Once a filesystem is flushed and a backup is invoked by the front-end administrator, the formatted filesystem can be backed up at the OpsCentre. Clearly there is potential for customisations and automations to streamline this process.

For block-oriented access, file-level backups will not be possible as the OpsCentre just sees blocks, no metadata, and hence will not have understanding of the data stored on the blocks. In mitigation, the OpsCentre could provide *immutable* catalogue space for user-generated backup files.

At a later date other filesystems will be investigated which support journalling and snapshots (*live archiving*).

6.2 Infrequently Accessed Data

Rarely or infrequently accessed data storage can be provided with greater packaging density and energy efficiency. If the disks comply with the de-facto MAID (Massive Array of Idle Disks) specification [12], they can be instructed to change state to reflect the access frequency. If the expected frequency of accesses is specified by users when requesting space then this will allow assignment of space reservations such as to maximise energy efficiency.

The initial specification, MAID 1.0, only considers disks can adopt two states: full operation (i.e. full power dissipation, full access speed), and idle (all but essential electronics off, for, e.g. 60% less power, $<$60 sec recovery). With MAID 2.0 [26], disks that comply can adopt multiple states, for example, withdraw heads in state 1 (e.g. 15% less power, $<$1 sec recovery), reduce revolutions in state 2 (e.g. 35% less power, $<$15 sec recovery), shut down to idle state in state 3, all to reduce energy demands and hence allow increased packaging density. Typically, drive spin-up is sequenced to reduce power surges, and drives are restored for periodic surface scans to ensure data integrity. As a result, 48 disks can be packed into a 4U or 5U rackmounted chassis, compared to the usual 9U or 10U required.

This technology costs more than storage for *online-immutable*, but is available in implementations that cost less than the highest-performance *online-mutable* storage. Generally the configuration is highly customisable by the user, and so it can support both *online-mutable* and *online-immutable* partitions, each configured as necessary. This leads to the typical categorization shown in Table 1.

At present MAID is marketed as an optimization/enhancement, and while this is true it will cost more than the most cost-effective *online-immutable* technologies. In particular this is likely to remain the case for implementations that automatically categorise accesses on the fly. However, it is possible that a public domain solution will arise, particularly for the simpler MAID 1.0 specification.

Advantages: Packaging density and energy efficiency.

Disadvantages: Cost, and not often provided in conjunction with storage virtualisation (e.g. thin provisioning).

7 Deployment

The first full deployment of this architecture is the Regional Datastore at TCD and is presently a work in progress. This first deployment is described here to provide a concrete example and, although some details are site-specific to TCD, the same model can be applied elsewhere.

The proposed characteristics of the TCD Regional Datastore are shown in Table 2. For the two main supported access paths (iRODS-related and LFC/SRM/DPM) that can be used to access the datastore, the support is asymmetric: iRODS-related protocols can provide both *online-mutable* and *online-immutable* access to storage space, but LFC/SRM/DPM protocols can only provide *online-immutable* access. It may be feasible to provide LFC access to iRODS-managed space via BeStMan (Berkley Storage Manager) or similar SRM software. Access to tapes will be limited to OpsCentre staff. Backup/restore to/from tapes will be agreed on a case-by-case basis.

7.1 Initial OpsCentre Deployments

The support for *online-immutable* accesses (using AMGA, LFC, SRM, DPM, gridFTP, gfal) is already deployed. The iRODS support for *online-mutable* accesses is deployed in pilot form. Support for infrequent/nearline accesses (MAID) will not be implemented until mid 2010, so initially these accesses will be mapped to use iRODS or LFC/SRM/DPM as appropriate. For AAA, the Grid-Ireland PKI and GSI authentication, the VOMS authorization, and the EGEE compute and data accounting are already deployed.

WebDav access to iRODS is already deployed in pilot form using Davis. Fedora Commons access to iRODS is also being explored. It is intended that provision of mount-points will be explored via Davis mounting on all SEs, P-GRADE and the Migrating Desktop. Alternative authorization will be explored via a VOMS-to-LDAP mirror, which is already deployed in pilot form. The use of microservices to store and use filesystem metadata (attributes) will be actively explored.

7.2 Client Tools

Client tools are outside the OpsCentre's orbit. The user communities will decide what client tools they will deploy. However, the use of Fedora Commons, iRODS' Explorer for Windows, and FUSE and Parrot for POSIX access could all be promoted within an adaptive e-Learning course [10] on the use of the e-INIS datastore.

7.3 Space Reservations

When requesting space the users must specify whether the data will be *mutable* or *immutable* and the expected read and write access frequencies. Initially LFC space will be allocated to VOs as pools, i.e. a pool will be created for each VO that is granted space. Pool space is round-robin distributed across disks. Within the pool, the VO can request space reservation tokens for use by specific VO roles, e.g. to guarantee one or more VO members with a specific role has sufficient reserved space to store raw datasets that all

immutable	idle state	no.writes	max.access time	class	deployment
NO	state 0	>1	<0.1 sec	Online-Mutable	initial
NO	state 1-3	>1	<60 sec	Nearline-Mutable	future (MAID)
NO	N/A	>1	>60 sec	Offline-Mutable	initial
YES	state 0	1	<0.1 sec	Online-Immutable	initial
YES	state 1-3	1	<60 sec	Nearline-Immutable	future (MAID)
YES	N/A	1	>60 sec	Offline-Immutable	initial

Table 1 Storage classes

access type	online-mutable	online-immutable	nearline	offline
organization	filesystems	pools	mixed	tape set
abstraction	TBD	space reservations	mixed	volumes
archiving	backup	backup	backup	on request
sharing	per VO	per VO	per VO	per VO mgr
allocation	on request	on request	on request	on request
max.size	as allocated	as allocated	as allocated	800GB tapes
priority	national	national	national	OpsCentre
software	iRODS	LFC/SRM/DPM	mixed	backup/restore
storage servers	ExDS	DPM	MAID(future)	PowerVault
typical usage	online processed data	online raw data	infrequently-accessed data	offline critical data

Table 2 TCD datastore characteristics

the VO members can then access. A token is limited to its pool. To simplify management, the allocation quantum will be one RAID6 filesystem of fifteen 1TB disks. One or more of these filesystems can be added to each pool and, should a further allocation be granted, pools can easily be extended by adding more filesystems.

An analogous mechanism has yet to be fully deduced for iRODS space but it is possible to use resources and resource groups in a manner analogous to the filesystems and pools in LFC. The rule engine also gives the possibility that finer grained quotas could be assigned by the correct configuration. This is under investigation and the initial implementation is planned for the first half of 2010.

7.4 Applications for Space

For all the e-INIS National Datastore storage, the allocation process is to be modelled on the ICHEC allocation process for compute resources. Applications for space fall into three classes: A,B and C. 10% of the total space is reserved for class C projects which will entirely be managed by ICHEC. This space is for the many computational science projects that need to store model data on a national basis. Up to a further 20% of the storage capacity may be used for exploratory and small-scale (class B) projects. These projects will be approved and reviewed by the e-INIS executive on a case-by-case basis using light-weight procedures. The remaining bulk of the storage will be used to support major (class A) research projects on the basis of applications which will be peer reviewed and evaluated against criteria such as whether the project has a national or international dimension, whether there is a comprehensive strategy for data management and access control and whether there is an education/outreach component.

Initial space allocations will be as shown[3] in Table 3 ($1\text{TB} = 2^{40}$ bytes).

VO	Online-Mutable	Online-Immutable	Nearline
	(iRODS)	(LFC/SRM/DPM)	(mixed)
ATLAS	-	47.3TB	-
LHCb	-	35.5TB	-
HELIO	-	11.8TB	*
DHO	12TB	-	*
Gene	-	11.8TB	-
Cosmo	-	11.8TB	-
CMIP5	-	11.8TB	*
Other	10TB	8TB + 7TB	*
Reserved	remainder	remainder	-
Total TB	232TB	375TB	-

Table 3 Initial space allocations

7.5 Hardware Architecture

The bridge hosts are 8-core blade servers within two blade chassis. Up to 32 of these could be provisioned within the existing configuration, see Figure 7.

The datastore hardware architecture necessary to support *online-mutable* accesses is initially based on a HP ExDS9100 that presents a highly available (polyserve)

[3] "Other" is a pair of separate pools, one of which is used as a default pool and the other which was created just for Irish VOs with no other allocation giving a total of 15TB storage that they can access. Nearline storage is not available to applications yet but applications that may have data that may be able to utilise this space are marked with *

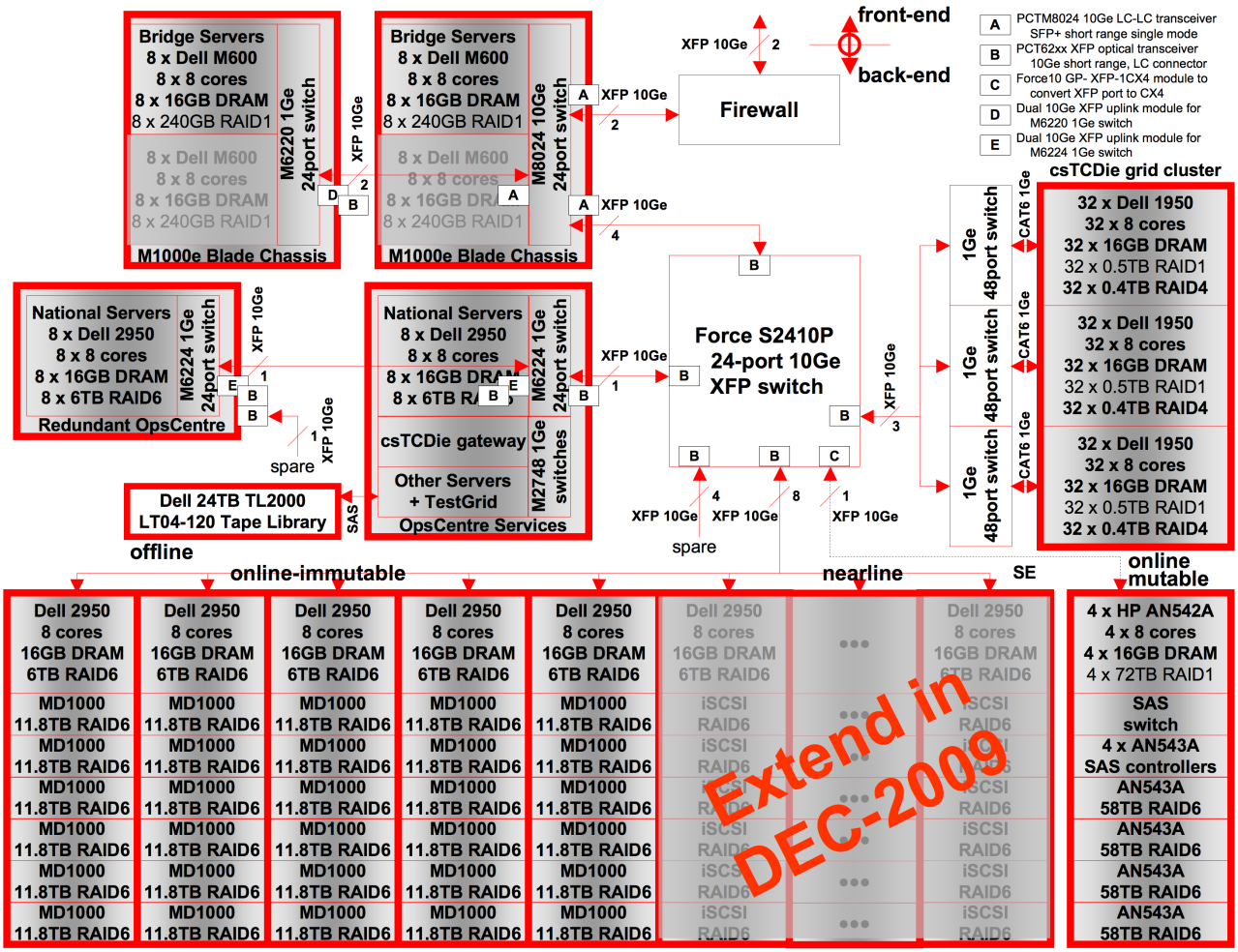

Fig. 7 Proposed TCD datastore hardware architecture

filesystem across 328TB of raw storage, yielding 232TB usable storage, again see Figure 7.

The detailed hardware architecture for *online-immutable* accesses is well defined and has been deployed for a long time. DPM is used to provide very inexpensive *online-immutable* storage direct-attached SAS RAID chassis, each of which contains up to fifteen 1TB disks, see Figure 7. Within each chassis, the disks are configured in one RAID6 group, yielding 11.8TB usable storage. Six such chassis are controlled from each disk server. This represents a 70.8TB storage block that can be exported. Five blocks are provided (354TB). The disk servers contain an extra 4TB RAID6 (six 1TB disks) that can be exported. Performance can be kept high using large write-through caches at all levels, since the proportion of writes will be very low, and no updates will ever occur, only deletes. In early to mid 2010 the 1TB disks will be upgraded to 2TB, plus more storage blocks will be added.

MAID is relatively new technology, and is marketed as an enhancement to more expensive *online-mutable* technologies, so will not be provided within the TCD datastore until the second quarter 2010, when the costs will probably be lower. There is rack space reserved for sixteen 4U or twelve 5U MAID rack-mounted storage units, with future expansion to another thirty 4U or twenty four 5U units.

A 24-tape tape library is provided. These incorporate an IBM Ultrium LTO-4 drive. LTO-4 is a high-speed tape technology that can backup 432GB of data per hour (LTO technology is based on a tape industry standard that specifies backward read-and-write compatibility with LTO-4 & LTO-3 generation media (*n*,*n-1*) and read compatibility with LTO-2(*n-2*) media). LTO-4 tapes store 800GB (uncompressed).

7.6 Network Upgrade

It is to be expected that a greatly increased storage capacity will lead to greater demands being placed on the network infrastructure in order to move data around. Lifetimes of data vary greatly but consider a simplified example where,

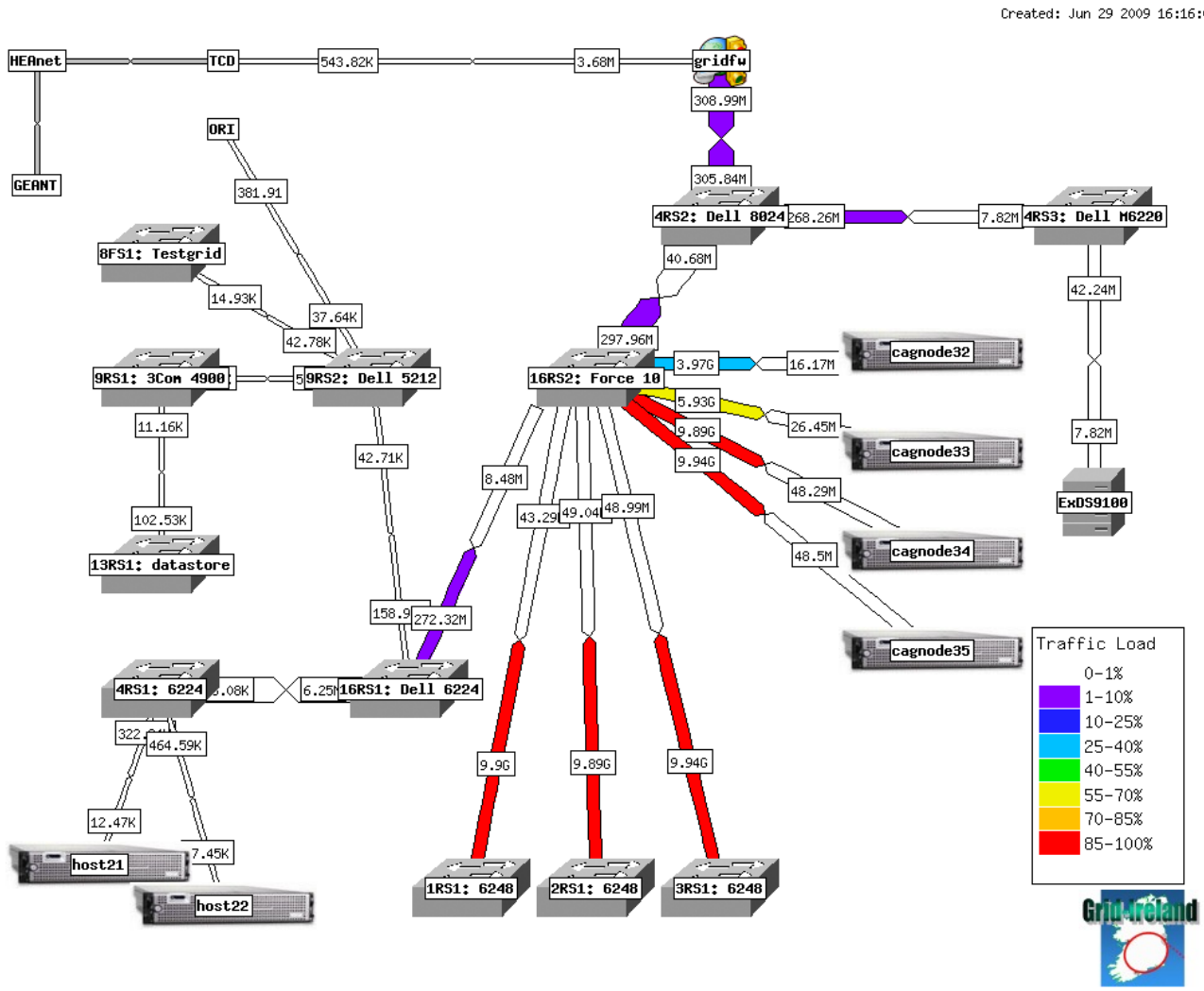

Fig. 8 Example network weathermap during testing of high loads from the cluster using netperf

over a period of 6 months, 300TB data (approximate sum of estimated use of pilot projects prior to inception of the pilot) is written once and read once (2 transfers). Ignoring overheads, 300 TB data being stored in this manner gives approximately 0.6 PB traffic every six months or roughly 0.3 Gb/s. This is an average transfer rate and in practice one would expect transfers to occur in a less uniform manner, for the network protocols to add overhead and for files to be read multiple times. Therefore the expected bandwidth required, to avoid bottlenecks, is expected to be several times greater: >1 Gb/s. The two main options for achieving >1Gb/s networking were to either aggregate multiple 1Gb/s ethernet (1Ge) links or to upgrade to 10Gb/s ethernet (10Ge). It was decided to upgrade the core of the network to 10Ge using 10GBASE-SR components - 850nm wavelength through OM3 multimode fibre.

The uplink from the firewall to the internet is currently 1Ge, throttled to 500Mb/s to ensure fair-sharing. The connection from the firewall in to the top level switch, which resides in one of the bridge blade chassis, is 10Ge. This top level switch is a layer-3 switch and serves to isolate the bridge zone (ultimately perhaps a DMZ) from the bulk data storage, where layer-2 10Ge switching is employed. The ExDS 9100 (high availability storage) is temporarily linked via a switch in the second blade chassis until CX4 ports can be provisioned in the same network zone as the grid storage — a necessary separation for when the service goes into production.

A PHP network weathermap is used, in combination with Cacti (RRDTool-based graphing solution), to monitor network traffic via SNMP connections to the network switches. An example weathermap plot is show in Figure 8. The nodes labelled cagnode3X are the disk servers for grid storage.

8 Applications of the Datastore

Below we consider four initial datastore applications, the first two from physics at small and large scales, one from climate prediction, and the last a planned future application from humanities. These projects are all at different stages in their life-cycles and so represent current use as well as future use.

8.1 CERN LHC ATLAS Experiment

The subset of the datastore architecture that is accessible using LFC/SRM/DPM has been thoroughly tested by the data challenges conducted by ATLAS, which simulate the real loads that ATLAS data transfers and jobs will apply. The LHC STEP09 data challenge aimed to exercise all aspects of their computing model, to test the infrastructure and to identify and understand any bottlenecks. It involved:

- Distribution of data to sites
- Production of simulated data
- Execution of “user” analysis jobs

The Grid-Ireland OpsCentre at TCD was involved as a Tier-2 site in the Dutch cloud, associated with the Netherlands Tier-1 at SARA. During the challenge, TCD received a defined proportion of the datasets, ran analysis jobs on those datasets and ran production jobs. This was a chance for the OpsCentre to see bulk data transfer in conjunction with cluster access to stored data and was a stringent test of data paths into and within the site infrastructure, testing the new infrastructure against higher sustained loads than ever previously experienced. Data transfers into TCD from SARA in the Netherlands were throttled to 500 Mbit/s but nevertheless peaked at 440 Mbit/s. Data transfers from storage to the cluster were observed to exceed 7 Gbit/s, the first time since testing began that an application has put a sustained load of this magnitude on the new 10Gb/s optical network. To support these loads it was necessary to update the disk servers’ Linux kernel to one with MSI-X support for the network cards as the default kernel was sending all the RX interrupts to a single CPU core, which was blocking. A kernel from the Scientific Linux 5 distribution was back-ported to Scientific Linux 4, solving this problem. Also, at the start of STEP09, the network switches for the cluster only had a single 1Ge uplink that proved insufficient. This was upgraded to a LAG and the switches have since been upgraded to models that have 10Ge uplinks.

The network weathermap in Figure 9 is a snapshot of the network utilisation during the STEP09 challenge. The *cagnodeXX* hosts are the DPM disk servers and the Dell 2748 switches, near the bottom of the figure, were the switches for the cluster nodes at that time. At the time of this snapshot, a 3-channel (3Gb/s) LAG was being used to uplink those switches to the Dell 6224 switch that also hosted the links from the national servers and other local servers. This Dell 6224 has a 10Ge uplink and, in this snapshot, was transferring data to the cluster at 7.31Gb/s. This load was mostly from jobs reading data using RFIO.

It can be seen that the amount of communication with cagnode32 was far greater than that for the other storage servers. This is because, at the start of STEP09, all the filesystems in the ATLAS pool were hosted on cagnode32. As the exercise progressed, it became clear that this was far from optimal and two out of the three filesystems were made read-only and additional space on the other two hosts added to the pool. While this has distributed subsequent data across the three hosts, much of the analysis used data that had already been written to cagnode32 - hence the continued higher loading of that host.

Figure 10 shows the varying proportions of compute capacity used by different types of job as STEP09 progressed. The start of the graph is blank as that period predates this monitoring being enabled. The graph shows how, early in STEP09, the ATLAS analysis jobs ramped up and then were gradually superseded by production jobs. LHCb jobs can also be seen to become more significant towards the end of the challenge. The variation in the number and type of jobs had direct consequences on the way the storage was stressed.

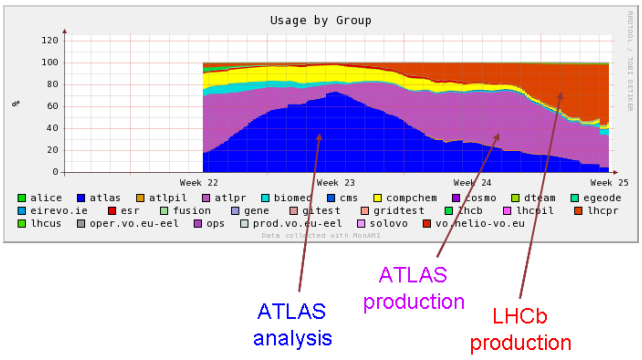

Fig. 10 TCD cluster utilisation during WLCG STEP09 challenge

From the perspective of the TCD site the main outcomes of this test were:

- Monitoring is crucial to understanding what’s going on:
  - Weathermap for quick visual check
  - Cacti for detailed information on network traffic
  - LEMON/Ganglia for host load, cluster usage, etc.
- At the start of the challenge there were a large number of analysis jobs running on cluster nodes
- These accessed large datasets directly from storage
- This caused heavy load on the network and disk servers
- This caused problems for other jobs accessing storage

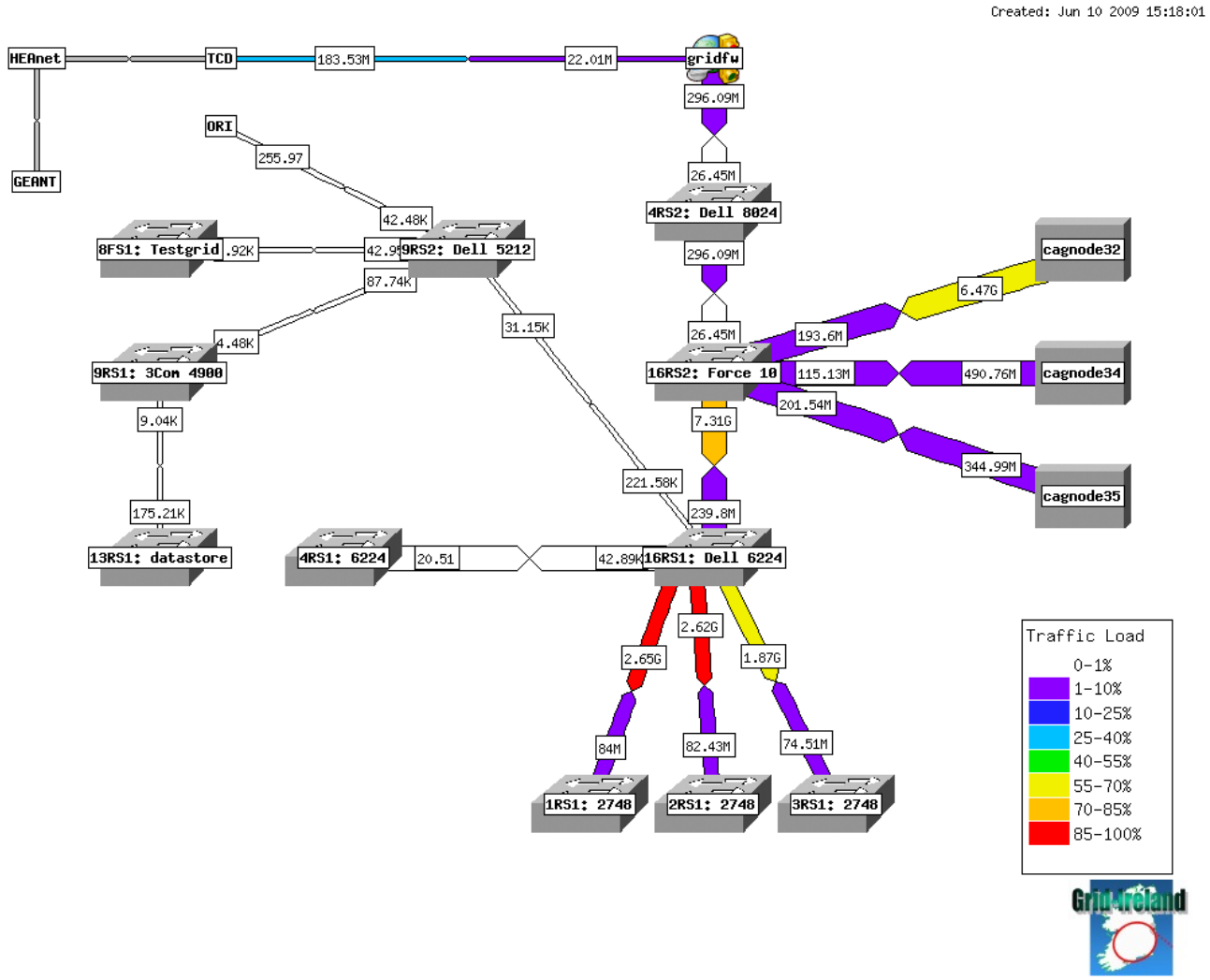

Fig. 9 Example network weathermap during STEP'09, not the final configuration

Due to the usefulness of this stress testing, the Ops-Centre plans to not only continue to take part in these service challenges but also to use the 'Hammercloud' tests between challenges as a way of testing the robustness of the infrastructure after upgrades.

8.2 HELIO

HELIO [3] is an EU funded project for the study of the heliosphere and its influences on the Earth. HELIO will provide integrated access to metadata from the various domains that constitute heliophysics in order to identify interesting phenomena and track them as they propagate through inter planetary space and affect the planetary environments. It will provide services to locate and retrieve observations and return them to the user in the format they require. To do this, HELIO will deploy the TAVERNA workflow engine [23] as a service, and mirror and store replicas of heliospheric data. TAVERNA will be used to define and execute workflows, but as it only offers basic authentication and authorization, its access must be controlled. One proposed solution assumes that users run predefined workflows via a portal that invokes the workflow over a VPN connected to a TAVERNA service running on a bridge server VM, that itself accesses data mirrors on the datastore. A pilot *Archiver* is already deployed that mirrors various data sources to the e-INIS datastore; it uses fetches the data over HTTP to a cache on a bridge server VM, and then copies and registers the data to the datastore through the gLite *lcg-cr* command to invoke LFC/SRM/DPM. Privileged users define the list of sources and the frequency of the data transfers through a Wiki page hosted on a bridge server VM. Some statistics from the first proof-of-concept run are given in table 4. The 'registration' referred to in the table is the lcg-cr command which registers a file in the LFC and copies it into storage. For larger

files, the time spent transferring the files is more significant than the time to register them in the catalogue, so the highest registration speed gives an indication of the transfer performance of the local network. In contrast, for small files, the transfer time can be less significant than the time to register in the catalogue and so the lowest registration speed is more indicative of the performance of the catalogue than the performance of the network.

Total amount of data moved	361.12 Gigabytes
Total number of files moved	476
Total time	12 hours and 7 minutes
Average download speed	10.55 megabytes/second
Average registration speed	45.22 megabytes/second
Average File size	775.24 megabytes
Largest File	6.09 gigabytes
Smallest File	6.1 kilobytes
Highest download speed	10.97 megabytes/second
Lowest download speed	0.28 megabytes/second
Highest registration speed	170.75 megabytes/second
Lowest registration speed	0.001 megabytes/second

Table 4 HELIO proof-of-concept transfer statistics

8.3 Climate Prediction (CMIP5)

Between June 2009 and October 2010 the climate modelling community will measure how the main prediction models compare [1]. This activity will involve a number of computing centres, including the Irish Centre for High-End Computing (ICHEC), which will generate substantial quantities of metadata and result data. The core data generated by each site will be stored at primary sites and that and non-core data will reside at secondary sites. The data that is generated at ICHEC will be stored on the e-INIS datastore, and the metadata will be stored using a GeoNetwork OpenSource spatial information metadata catalogue on a community front-end server at ICHEC. Of the order of 200TB of the *online-immutable* result data will be stored using LFC/SRM/DPM on the e-INIS datastore, with replication for fault tolerance, but this data has an expected lifetime of only 18 months. It is proposed that public web-based access will be provided by read-only export via OPeNDAP running on a bridge server.

8.4 Digital Humanities Observatory (DHO)

The Royal Irish Academy is developing a digital repository for e-humanities scholarship in the island of Ireland [2], based on the use of Fedora Commons. The DHO was established to manage and co-ordinate the increasingly complex e-resources created in the arts and humanities. It will enable research and researchers in Ireland to keep abreast of international developments in the creation, use, and preservation of digital resources. The repository is currently in a pilot stage and while it is expected that the demands placed on the storage infrastructure in terms of capacity will initially be considerably less than many of the supported scientific disciplines (but still of order terabytes), the requirements are expected to grow rapidly as additional multi-media collections are added. The desire to replicate digital object collections across multiple geographically distributed sites in the interest of date protection is one of the motivating factors in the investigation of layering the repository upon the federated data infrastructure.

The repository is being implemented in parallel to the data storage infrastructure and although currently independent at the technical level, close collaboration between the DHO and the e-INIS project has resulted in an integration plan to allow the storage of digital objects on the data storage infrastructure while all cataloguing, metadata, and disseminations tasks will remain the within the repository. The federations aims to take advantage of the existing work in the area of iRODS/Fedora integrations such as the storage module developed at the San Diego Supercomputer Center. It is expected that in the next phase this will exploit the datastore via one of the iRODS supported mechanisms, and so will be a good test of that functionality and of a remote front-end service accessing the common back-end.

One example of the digital collections that have been so far been ingested into the DHO repository is an archive of Irish dialect recordings made between 1928 and 1931 by Dr Wilhelm Doegen. Known as the Doegen Records, the collection was commissioned by the Irish government in 1926 and includes early Irish language recordings of folk-tales and songs among other material. Through catalogues using the Fedora Commons repository, the collection is made available to the public via a custom front-end portal developed by the DHO.

9 Future Work

At the time of writing, *online-immutable* storage is in place and in daily use, iRODS is installed and under evaluation and a tool-kit is being assembled for the various bridge machines.

Various tasks (and challenges) remain. So far the iRODS configuration is very basic and only one site in the federation (TCD) has a running service. Further resources will be installed at other sites (initially just University College Cork and Dublin Institute of Advanced Studies) and will use the same iCAT catalogue. Subject to successful tender, further resources will be acquired supporting MAID technologies and these will also be integrated.

Rules need to be developed to manage the different resource classes and data lifetimes to ensure optimum data

placement. Experimental iRODS microservices will be developed to attempt to map part of the namespace to EGEE services (possibly as iRODS compound resources). This will require significant expenditure of development effort as iRODS currently has little support for delegated GSI credentials - a requirement of the current security policy.

Significant testing remains to be done. While the majority of the individual software components currently deployed have been tested, not every combination has been tried yet and problems will inevitably arise. Transfers between sites need to be tested. The whole system needs to be stressed in a similar way to the STEP09 tests on the EGEE infrastructure.

Examples are needed of applications making use of the full iRODS functionality and also of applications that take advantage of combined iRODS/LFC functionality. Early adopters will be needed in order to develop these case studies.

10 Conclusions

We have proposed a federated national datastore architecture suitable for large-scale data management that provides Internet-accessible storage for diverse communities of academic researchers across science and the humanties. It supports these communities as virtual organisations, and provides digital repository services and integrates grid-enabled data management.

This 2-layer architecture recognises that the communities themselves are best able to define metadata to describe their data and are best able to develop a front-end interface for their users whilst a common back-end can take the best advantage of economies of scale. An architecture has been proposed for the back-end based on the use of iRODS for the digital repository services and EGEE LFC/SRM/DPM for the grid-enabled data management. The proposal is designed to facilitate front-end development by placing minimal restrictions on how the front-end is implemented and on the internal community security policies. Bridge servers are introduced to ensure that the communities are insulated from the back-end and from each other in order to ensure quality of service and to decouple their front-end implementation from site-specific back-end implementations. The use of back-end technology such as iRODS and DPM allows extra storage servers to be added as needed and ensures that the e-INIS financial drawdown schedules and budget limitations can be met by allowing inexpensive technologies to be used in appropriate places - such as where data is immutable.

The first e-INIS datastore site to be installed following the proposals here is that at Trinity College Dublin. The hardware architecture of this site has been described in detail and some initial results are given from applications using the datastore.

11 Acknowledgements

The TCD Regional Datastore and the pilot National Datastore are activities of the e-INIS project, aimed at building a sustainable national e-Infrastructure for the Irish research community. e-INIS is funded under the Irish Higher Education Authority's Programme for Research in Third-Level Institutions (PRTLI), a component of the National Development Plan. This work is co-funded by the European Commission through the EGEE-III project (www.eu-egee.org), contract number INFSO-RI-222667,and the results produced made use of the EGEE grid infrastructure.

References

1. CMIP5 - Coupled Model Intercomparison Project Phase 5. URL `http://cmip-pcmdi.llnl.gov/cmip5/index.html`
2. Digital Humanities Observatory. `http://dho.ie`
3. Heliophysics Integrated Observatory (HELIO). `http://www.helio-vo.eu`
4. Irish Centre for High End Computing (ICHEC). `http://www.ichec.ie/services`
5. ESFRI working group about digital repositories, ESFRI Position Paper. ESFRI (2007). Also available as `ftp://ftp.cordis.europa.eu/pub/esfri/docs/digital_repositories_working_group.pdf`
6. NSF07-28 Cyberinfrastructure Vision for 21st Century Discovery. "National Science Foundation" (2007). Also available as `http://www.nsf.gov/pubs/2007/nsf0728/index.jsp`
7. OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD (2007). Also available as `http://www.oecd.org/dataoecd/9/61/38500813.pdf`
8. Enabling Inter-Repository Access Management between iRODS and Fedora. In: 4th International Conference on Open Repositories, OR09. Conference Presentations. Georgia Institute of Technology (2009). URL `http://hdl.handle.net/1853/28494`
9. Blanke, T., Hedges, M., Dunn, S.: Arts and humanities e-science—current practices and future challenges. Future Generation Computer Systems **25**(4), 474–480 (2009). DOI 10.1016/j.future.2008.10.004. URL `http://dx.doi.org/10.1016/j.future.2008.10.004`
10. Cassidy, K., McCandless, J., Childs, S., Walsh, J., Coghlan, B., Dagger, D.: Combining a virtual grid testbed and grid elearning courseware. In: Proc. Cracow Grid Workshop 2006 (CGW06). Academic Computer Centre CYFRONET AGH, Cracow, Poland (2006)
11. Childs, S., Coghlan, B., O'Callaghan, D., Quigley, G., , Walsh, J.: Centralised fabric management for a national grid infrastructure. In: Cracow Grid Workshop (CGW'05). Cracow, Poland (2005)
12. Colarelli, D., Grunwald, D., Neufeld, M.: The case for massive arrays of idle disks (maid). In: In The 2002 Conference on File and Storage Technologies, p. 2002. On (2002)
13. Cornillon, P., Gallagher, J., Sgouros, T.: Opendap: Accessing data in a distributed, heterogeneous environment. Data Science Journal **2**, 164–174 (2003). DOI 10.2481/dsj.2.164. URL `http://dx.doi.org/10.2481/dsj.2.164`
14. Data Management Task Force: e-IRG Report on Data Management. e-Infrastructure Reflection Group (2009)
15. Foster, I., Kesselman, C., Tsudik, G., Tuecke, S.: A security architecture for computational grids. In: Proc. 5th ACM Conference on Computer and Communications Security Conference, pp. 83–92 (1998)

16. Hedges, M., Blanke, T., Hasan, A.: Rule-based curation and preservation of data: A data grid approach using iRODS. FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF GRID COMPUTING-THEORY METHODS AND APPLICATIONS **25**(4), 446–452 (2009). DOI 10.1016/j.future.2008.10.003. 3rd IEEE International Conference on e-Science and Grid Computing, Bangalore, INDIA, DEC 10-13, 2007
17. Hertel, C.R.: Implementing CIFS: The Common Internet File System. Prentice Hall PTR (2003). URL `http://www.ubiqx.org/cifs/index.html`
18. Klous, S., Frey, J., Son, S.C., Thain, D., Roy, A., Livny, M., van den Brand, J.: Transparent access to grid resources for user software. Concurrency and Computation: Practice and Experience **18**(7), 787–801. DOI 10.1002/cpe.961. URL `http://dx.doi.org/10.1002/cpe.961`
19. Knopper, S., Koning, R., Roodhart, J., Grosso, P., de Laat, C.: Amsterdam cinegrid exchange – a distributed high-quality digital media solution. Available at `http://www.science.uva.nl/research/sne/reports/` (2009). SNE technical report SNE-UVA-2009-01
20. Koblitz, B., Santos, N., Pose, V.: The amga metadata service. Journal of Grid Computing **6**(1), 61–76 (2008). DOI 10.1007/s10723-007-9084-6
21. Laure, E., Gr, C., Fisher, S., Frohner, A., Kunszt, P., et al.: Programming the grid with glite. In: Computational Methods in Science and Technology, vol. 12, pp. 33–45 (2006). URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.93.1312`
22. Mandrichenko, I., Allcock, W., Perelmutov, T.: GridFTP v2 Protocol Description. Also available as `http://www.ogf.org/documents/GFD.47.pdf` (2005). GGF Document Series GFD.47
23. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., Li, P.: Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics **20**(17), 3045–3054 (2004). DOI 10.1093/bioinformatics/bth361. URL `http://dx.doi.org/10.1093/bioinformatics/bth361`
24. Saklikar, S., Saha, S.: Next Steps for Security Assertion Markup Language (SAML). In: SWS'07: PROCEEDINGS OF THE 2007 ACM WORKSHOP ON SECURE WEB SERVICES, pp. 52–65. ASSOC COMPUTING MACHINERY, 1515 BROADWAY, NEW YORK, NY 10036-9998 USA (2007). ACM Workshop on Secure Web Services, Fairfax, VA, NOV 02, 2007
25. Satran, J., Meth, K., Sapuntzakis, C., Chadalapaka, M., Zeidner, E.: Internet Small Computer Systems Interface (iSCSI). RFC 3720 (Proposed Standard) (2004). URL `http://www.ietf.org/rfc/rfc3720.txt`. Updated by RFCs 3980, 4850, 5048
26. Schulz, G.: MAID 2.0:Energy Savings without Performance Compromises. `http://www.storageio.com/Reports/StorageIO_WP_Jan02_2008.pdf`
27. Sim, A., Shoshani, A., Badino, P., Barring, O., Baud, J., Corso, E., Witt, S.D., Donno, F., Gu, J., Haddox-Schatz, M., Hess, B., Jensen, J., Kowalski, A., Litmaath, M., Magnoni, L., Perelmutov, T., Petravick, D., Watson, C.: The storage resource manager interface specification version 2.2. Also available as `http://www.ogf.org/documents/GFD.129.pdf` (2008). GGF Document Series GFD.129
28. Thain, D., Livny, M.: Parrot: Transparent user-level middleware for data-intensive computing. In: In Workshop on Adaptive Grid Middleware (2003). URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.8435`
29. Whitehead, J.E., Wiggins, M.: Webdav: Ietf standard for collaborative authoring on the web. IEEE Internet Computing **2**(5), 34–40 (1998). DOI 10.1109/4236.722228
30. Witzig, C.: Shibboleth Interoperability Through a Short Lived Credential Service. EGEE-II (2006). Report EGEE-II-MJRA1.4-770102-v0.96.doc
31. Zhang, S., Coddington, P., Wendelborn, A.: Davis: A Generic Interface for SRB and iRODS (2009). DHPC Technical Report DHPC-197