Grid Infrastructures for Secure Access to and Use of Bioinformatics Data: Experiences from the BRIDGES Project

Prof R. Sinnott[1], Dr M. Bayer[1], A. Stell[1], Dr J. Koetsier[2]
National e-Science Centre,
[1]University of Glasgow,
[2]University of Edinburgh
ros@dcs.gla.ac.uk

Abstract

The BRIDGES project was funded by the UK Department of Trade and Industry (DTI) to address the needs of cardiovascular research scientists investigating the genetic causes of hypertension as part of the Wellcome Trust funded (£4.34M) Cardiovascular Functional Genomics (CFG) project. Security was at the heart of the BRIDGES project and an advanced data and compute Grid infrastructure incorporating latest Grid authorisation technologies was developed and delivered to the scientists. We outline these Grid infrastructures and describe the perceived security requirements at the project start including data classifications and how these evolved throughout the lifetime of the project. The uptake and adoption of the project results are also presented along with the challenges that must be overcome to support the secure exchange of life science data sets. We also present how we will use the BRIDGES experiences in future projects at the National e-Science Centre.

1. Introduction

With the completion of the sequencing of the human and several other eukaryotic genomes, as well as more than a hundred microbial genomes, and the development of modern post-genomic high-throughput technologies allowing comprehensive studies of mRNA, protein and metabolite complements of biological samples, the life and biological sciences are experiencing an era of exponential data growth [1]. These enormous and highly heterogeneous, distributed data sets require well-designed data standards for their discovery, linkage and further analysis. Grid technology offers a paradigm which can support such requirements allowing access to and usage of the large scale computational resources needed for comparison and analysis of such data sets [2].

The life science domain offers new challenges with regard to security that are not immediately associated with other domains such as high energy physics. Data or more accurately information and knowledge associated with these data sets, can be intellectually and commercially sensitive, offering considerable exploitation possibilities. Whilst large scale public genomic databases are sources for the wider life sciences community to access and use, numerous other data classifications exist in the life science domain which are not so freely available. The Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES) project[1] [3] was funded by the UK Department of Trade and Industry to develop a Grid based computational infrastructure to support the needs of the Wellcome Trust funded (£4.34M) Cardiovascular Functional Genomics (CFG) project [4]. The CFG consortium themselves were investigating possible genetic causes of hypertension, one of the main causes of cardiovascular mortality. The consortium involves five UK and one Dutch site (depicted in Figure 1) and is pursuing a strategy combining studies on rodent models of disease (mouse and rat) contemporaneously with studies of patients and population DNA collections.

A characterisation and classification of the security requirements associated with the CFG project data sets was made at the start of BRIDGES. This classification evolved throughout the course of the project. In this paper we review the infrastructure that was developed and how it evolved to meet the needs of the scientists from a security perspective. The challenges in developing secure infrastructures within the BRIDGES project and encouraging their uptake are typical across the research community.

[1] BRIDGES successfully completed in December 2005 and involved the National e-Science Centre at the Universities of Glasgow and Edinburgh, and IBM.

Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06)
0-7695-2567-9/06 $20.00 © 2006 IEEE

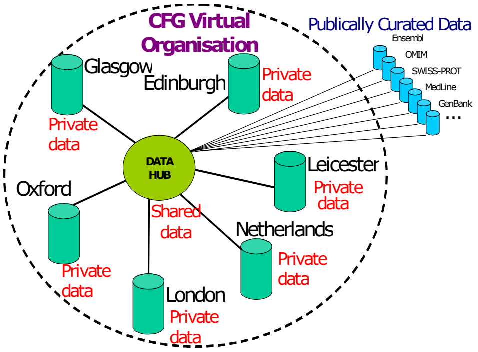


Figure 1: Data Distribution and Security of CFG Partners

2. BRIDGES Data Classification

At the project outset, it was identified that various significant data with different security aspects would be accessed and integrated to support the CFG research activities. This included:

- Public data: data from public sources, such as SwissProt and EMBL. We recognised that these could be accessed directly or be held as local copies for performance reasons.
- Processed public data: public data that has additional annotation or indexing to support the analyses needed by CFG. This kind of data must be held within the consortium, but one copy could serve the entire consortium. They may be of interest to and made available to other consortia.
- Sensitive data: the data about individuals in the cohorts of patients and the data derived from animal experiments. This kind of data would require careful enforcement of privacy and may be restricted to one site, or even part of a site.
- Special experimental data: these data sets fall into a particular category, e.g. quantitative trait loci (QTL) or microarray data, which has special arrangements for its storage and access already agreed. Given the cost associated with conducting microarray experiments and generating the data sets, these kinds of data would have security restrictions at the discretion of the consortia partners generating the data;
- Personal research data: data specific to a researcher as a result of experiments or analyses that that researcher is performing. This kind of data may not even be shared among the local team. The data may however later become team research data, e.g. once results have been published.
- Team research data: data that is shared by the team members at a site or within a group at a site. It may later become consortium research data, e.g. when the researchers are confident of its value or have written about its creation and implications.
- Consortium research data: data produced by one site or a combination of sites that is now available for the whole consortium.
- Personalisation data: metadata collected and used by the bioinformatics tools pertinent to individual users. This data is normally only needed to support the specific user to which it pertains, but it may need to move between sites, e.g. when bioinformaticians visit sites or work together.

The rich variety of data requirements is typical of the needs of a large biomedical research project. As a result of these classifications it was recognised that there would be obvious sensitivities about where data could be stored and how it may be accessed. For example, privacy mechanisms acceptable to the clinical researchers would need to involve access controls and it was considered that messages and certain data sets would need to be encrypted when in transit. It was considered that there would also be issues as to what data might be used by the BRIDGES software/Grid developers. In some cases, the data may need to be systematically randomised, anonymised or encrypted, while preserving distribution properties, before it could be made available for testing.

Based on this, different classes of integration were identified and were expected to be supported by the Grid infrastructure. This included integration of the many sources of data: public data, such as SwissProt, the mouse, rat and human gene sequences, etc.), project data, such as mouse, rat, congenic rat and human population phenotypical data (clinical data, RNA, microarray data, proteomic gel data, etc.) and derivative data, e.g. QTL.

The Data Hub identified in Figure 1 was to form the fulcrum of this data access and integration activity, where public data would be combined with shared and private data according to security policy. Two versions of the Data Hub itself were to be created and compared, both of them based upon IBM DB2 technology [5]. The first based upon a commercial data integration technology solution, IBM DiscoveryLink and later remarketed as IBM Information Integrator[2] [6]; the second based on the Grid communities open source Open Grid Service Architecture Data Access and Integration (OGSA-

[2] To be renamed as IBM Masala.

DAI) software [7]. An evaluation and comparison of these technologies including their performance and overall usability in the functional genomics domain was made and is documented in [8].

The Data Hub provided a single repository through which access to numerous other federated genomic repositories and scientific research data sets could be made. The public genomic data sets of particular interest to the scientists included Ensembl (rat, mouse, human databases) [9]; Mouse Genome Informatics (MGI) [10]; Online Mammalian Inheritance in Man (OMIM) [11], Human Genome Organisation (HUGO) [12], Rat Genome Database (RGD) [13] and the Gene Ontology (GO) data base [14]. We note that of these, it was identified in the course of establishing the Data Hub that the programmatic access needed for the Grid data integration technologies was only available for the Ensembl and MGI databases. For the other data sets alternative solutions were required including downloading the data (often with no schema being provided), parsing the flat files and developing solutions to trigger the population of these files into the DB2 database. These issues are described in [8] along with the challenges and solutions that were adopted to handle changes in the remote database schemas.

The Data Hub supported various client side tools through which queries could be issued and used to access these remote databases. Given that the scientists based much of their research upon results from microarray experiments, these queries were typically based upon returning all information associated with a given gene (or sets of genes). This information was dependent upon the schemas and data sets associated with the remote databases themselves.

3. BRIDGES Data Access Client Tools

The initial data access application developed for the scientists was MagnaVista [15]. This application provided a completely configurable environment through which the scientists could navigate to and access a broad array of life science data sets of relevance to their research. Specifically through MagnaVista the user could input the genes that they were most interested in. Based upon this input, MagnaVista would invoke a stored procedure on DB2 which would build up the query[3], federate the query to the remote databases and subsequently join the results back together for display in the MagnaVista client application.

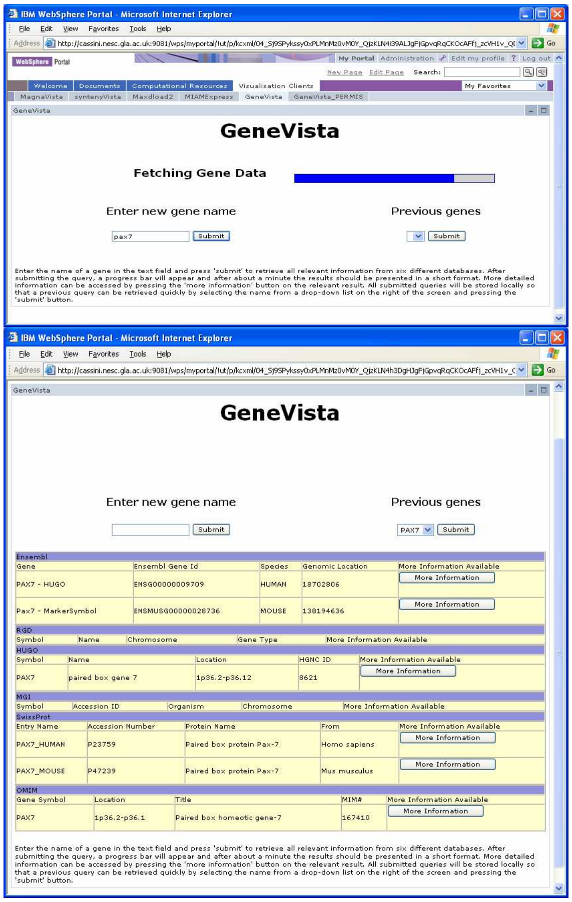
Thus rather than the user manually hopping to each of these remote resources, a single query issued by MagnaVista was used to deliver collections of data associated with the genes of interest. To support the specific targeted data needs of the scientists, the MagnaVista application could be personalised in various ways. For example, users could select specific (remote) databases that should be interrogated; select various data sets (fields) that should be returned from those databases; store specific genes of interest, and personalise the look and feel of the application itself.

The actual MagnaVista application itself was Java based and delivered to the users using Sun Web Start technology. Through launch buttons on the portal web page, a single mouse click could be used to automatically deliver the application and associated libraries, including the Web Start environment if it is not already present. However due to anomalies in Web Start with non-Internet Explorer versions of browsers used by the scientific community and issues of local firewalls blocking Web Start traffic, it was decided that a simpler version of this application was needed. It was also the case that the scientists were uncomfortable with the personalisation possibilities and having multiple panels and windows. In short, the application was not immediately intuitive and simple to use. The GeneVista was produced to address these issues.

GeneVista is a portlet based application. Portlets are Java-based Web components, managed by a portlet container, that process requests and generate dynamic content. Portals use portlets as pluggable user interface components that provide a presentation layer to information systems which enable modular and user-centric Web application access. Through a portlet based approach, the issues in firewalls and problems with Web Start with non-Internet Explorer browsers were overcome.

In essence the functionality of GeneVista is very similar to MagnaVista. However, it does not support the richness of personalisation. We note that this was at the request of the scientific end users. They simply wanted to be able to select a collection of gene names and retrieve all available information. Few of them bothered with personalisation possibilities. A Google-like front end to GeneVista was designed to reflect this (top part of Figure 2). The GeneVista portlet simply requires that the scientist input the gene names that they are interested in and selects submit. Following this, HTML based data sets are returned and presented within the browser window as shown in bottom part of Figure 2.

[3] The query was built using the GO database to address potential circular referencing that exist when querying multiple related databases, e.g. Ensembl data may include references to MGI data which references Ensembl etc.


IBM WebSphere Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://cassini.nesc.gla.ac.uk:9081/wps/myportal/!ut/p/kcxml/04_Sj9SPykssy0xPLMnMz0vM0Y_QjzKLN4i39ALJgFjGpvqRqCKOcAFfj_zcVH1v_QI
Go
WebSphere Portal
My Portal Administration Edit my profile ? Log out
New Page Edit Page Search:
Welcome Documents Computational Resources Visualisation Clients My Favorites
MagnaVista syntenyVista Maxdload2 MIAMExpress GeneVista GeneVista_PERMIS
GeneVista

GeneVista

Fetching Gene Data

Enter new gene name: pax7 [Submit]

Previous genes: [Submit]

Enter the name of a gene in the text field and press 'submit' to retrieve all relevant information from six different databases. After submitting the query, a progress bar will appear and after about a minute the results should be presented in a short format. More detailed information can be accessed by pressing the 'more information' button on the relevant result. All submitted queries will be stored locally so that a previous query can be retrieved quickly by selecting the name from a drop-down list on the right of the screen and pressing the 'submit' button.

IBM WebSphere Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://cassini.nesc.gla.ac.uk:9081/wps/myportal/!ut/p/kcxml/04_Sj9SPykssy0xPLMnMz0vM0Y_QjzKLN4h3DgHJgFjGpvqRqCKOcAFfj_zcVH1v_C
Go
GeneVista

GeneVista

Enter new gene name: [Submit]

Previous genes: PAX7 [Submit]

Ensembl

Gene	Ensembl Gene Id	Species	Genomic Location	More Information Available
PAX7 - HUGO	ENSG00000009709	HUMAN	18702806	More Information
Pax7 - MarkerSymbol	ENSMUSG00000028736	MOUSE	138194636	More Information

RGD

Symbol	Name	Chromosome	Gene Type	More Information Available

HUGO

Symbol	Name	Location	HGNC ID	More Information Available
PAX7	paired box gene 7	1p36.2-p36.12	8621	More Information

MGI

Symbol	Accession ID	Organism	Chromosome	More Information Available

SwissProt

Entry Name	Accession Number	Protein Name	From	More Information Available
PAX7_HUMAN	P23759	Paired box protein Pax-7	Homo sapiens	More Information
PAX7_MOUSE	P47239	Paired box protein Pax-7	Mus musculus	More Information

OMIM

Gene Symbol	Location	Title	MIM#	More Information Available
PAX7	1p36.2-p36.1	Paired box homeotic gene-7	167410	More Information

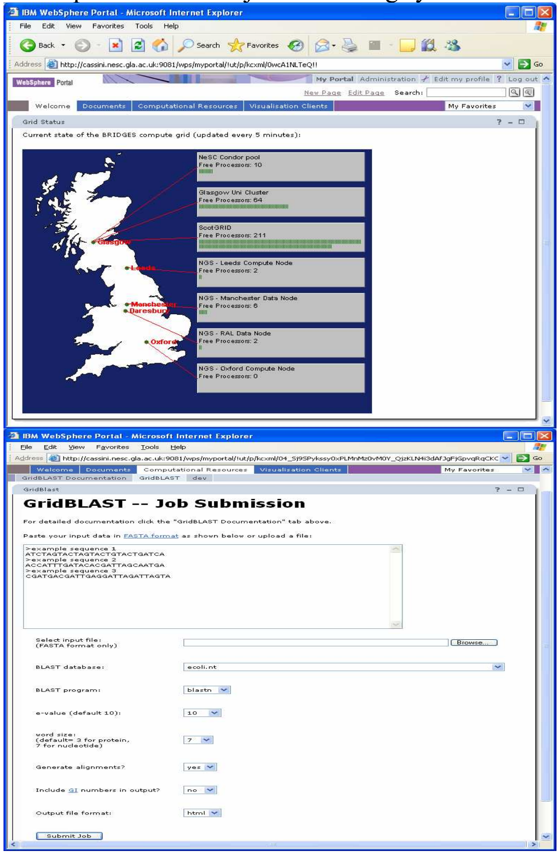
Enter the name of a gene in the text field and press 'submit' to retrieve all relevant information from six different databases. After submitting the query, a progress bar will appear and after about a minute the results should be presented in a short format. More detailed information can be accessed by pressing the 'more information' button on the relevant result. All submitted queries will be stored locally so that a previous query can be retrieved quickly by selecting the name from a drop-down list on the right of the screen and pressing the 'submit' button.

Figure 2: GeneVista Basic Usage for Gene Query

4. BRIDGES Client Compute Grid Tools

In their pursuit of novel genes and understanding their associated function life scientists often require access to large scale compute facilities to analyse their data sets, e.g. in performing large scale sequence comparisons or cross-correlations between large biological data sources. The Basic Local Alignment Search Tool (BLAST) [16] has been developed to perform this function. Numerous versions of BLAST currently exist which are targeted towards different sequence data sets and offer various levels of performance and accuracy metrics. BLAST involves sequence similarity searches, often on a very large scale, with query sequences being compared to several million target sequences to compute alignments of nucleic acid or protein sequences with the goal of finding the *n* closest matches in a target data set. BLAST takes a heuristic (rule-of-thumb) approach to a computationally highly intensive problem and is one of the fastest sequence comparison algorithms available.

It was recognised in BRIDGES that users should not have to learn the often complex options associated with job submission to job schedulers such as Condor [17] or OpenPBS [18]. In addition, one of the primary benefits of Grid technology is the ability to dynamically select and use a variety of heterogeneous resources is essential. This in turn requires that meta-schedulers are available that can dynamically schedule jobs across a variety of heterogeneous resources utilising a variety of local job schedulers. The BRIDGES Grid BLAST service which provides such a simplified BLAST based job submission system, enabling access to and usage of an extensible collection of HPC facilities is shown at the top of Figure 3 and is described in detail in [19]. This service was based upon the Globus technology (version 3) [20] with wrappers developed for external job scheduling systems.


IBM WebSphere Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://cassini.nesc.gla.ac.uk:9081/wps/myportal/!ut/p/kcxml/0wcA1NLTeQ!!
Go
WebSphere Portal
My Portal Administration Edit my profile ? Log out
New Page Edit Page Search:
Welcome Documents Computational Resources Visualisation Clients My Favorites
Grid Status
Glasgow
Leeds
Manchester
Daresbury
Oxford

Current state of the BRIDGES compute grid (updated every 5 minutes):

NeSC Condor pool
Free Processors: 10

Glasgow Uni Cluster
Free Processors: 64

ScotGRID
Free Processors: 211

NGS - Leeds Compute Node
Free Processors: 2

NGS - Manchester Data Node
Free Processors: 6

NGS - RAL Data Node
Free Processors: 2

NGS - Oxford Compute Node
Free Processors: 0

IBM WebSphere Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://cassini.nesc.gla.ac.uk:9081/wps/myportal/!ut/p/kcxml/04_Sj9SPykssy0xPLMnMz0vM0Y_QjzKLN4i3dAfJgFjGpvqRqCKC
Go
Welcome Documents Computational Resources Visualisation Clients My Favorites
GridBLAST Documentation GridBLAST dev
GridBlast

GridBLAST -- Job Submission

For detailed documentation click the "GridBLAST Documentation" tab above.

Paste your input data in FASTA format as shown below or upload a file:

```
>example sequence 1
ATCTAGTACTAGTACTGTACTGATCA
>example sequence 2
ACCATTTGATACACGATTAGCAATGA
>example sequence 3
CGATGACGATTGAGGATTAGATTAGTA
```

Select input file: (FASTA format only) [Browse...]

BLAST database: ecoli.nt

BLAST program: blastn

e-value (default 10): 10

word size: (default= 3 for protein, 7 for nucleotide) 7

Generate alignments? yes

Include GI numbers in output? no

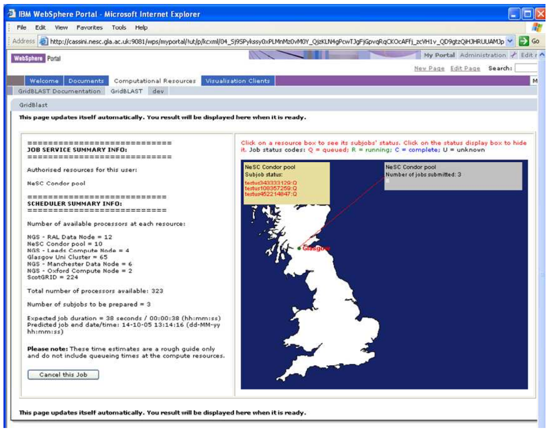
Output file format: html

[Submit Job]

Figure 3: Grid enabled BLAST service

The BRIDGES GridBLAST service makes use of the ScotGrid cluster [21], other HPC clusters at the University of Glasgow, Condor pools at the National e-Science Centre and all nodes of the National Grid Service (NGS) [22]. The status of these resources is shown at the top of Figure 3 and the user interface itself presented below. Intelligent default settings are automatically selected for the users.

When used, the service checks what resources are available, where the jobs are best run and subsequently provides a prediction of how long the complete BLAST job will take to complete. In addition, monitoring of the status of the various sub-jobs is undertaken (Figure 4) and staging of the various input and output files onto the compute resources is provided. Users can see where their jobs have been submitted and their status at any given time.


IBM WebSphere Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://cassini.nesc.gla.ac.uk:9081/wps/myportal/!ut/p/kcxml/04_Sj9SPykssy0xPLMnMz0vM0Y_QjzKLN4gPcwTJgFjGpvqRqCKOcAFfj_zcVH1v_QD9gtzQiHJHRUUAMJp
Go
WebSphere Portal
My Portal Administration Edit
New Page Edit Page Search:
Welcome Documents Computational Resources Visualisation Clients
GridBLAST Documentation GridBLAST dev
GridBlast
Glasgow

This page updates itself automatically. You result will be displayed here when it is ready.

```
=============================
JOB SERVICE SUMMARY INFO:
=============================

Authorised resources for this user:

NeSC Condor pool

============================
SCHEDULER SUMMARY INFO:
============================

Number of available processors at each resource:

NGS - RAL Data Node = 12
NeSC Condor pool = 10
NGS - Leeds Compute Node = 4
Glasgow Uni Cluster = 65
NGS - Manchester Data Node = 6
NGS - Oxford Compute Node = 2
ScotGRID = 224

Total number of processors available: 323

Number of subjobs to be prepared = 3

Expected job duration = 38 seconds / 00:00:38 (hh:mm:ss)
Predicted job end date/time: 14-10-05 13:14:16 (dd-MM-yy
hh:mm:ss)
```

Please note: These time estimates are a rough guide only and do not include queueing times at the compute resources.

Cancel this Job

Click on a resource box to see its subjobs' status. Click on the status display box to hide it. Job status codes: Q = queued; R = running; C = complete; U = unknown

NeSC Condor pool
Subjob status:
testus343333129:Q
testus108357259:Q
testus452214847:Q

NeSC Condor pool
Number of jobs submitted: 3

This page updates itself automatically. You result will be displayed here when it is ready.

Figure 4: Monitoring the Status of the GridBLAST service

5. Grid Security and CFG Needs

Fine grained security was essential to encourage the uptake of the Grid infrastructure by the CFG scientists. Most Grid solutions today are based upon X.509 certificates to support public key infrastructures (PKIs) [23].

The central component of a PKI is a Certificate Authority (CA). A CA is a root of trust which holders of public and private keys agree upon. CAs have numerous responsibilities including issuing of certificates, often requiring delegation to a local Registration Authority (RA) used to prove the identity of users requesting certificates. CAs are also required amongst other things to revoke older or compromised certificates through issuing Certificate Revocation Lists (CRL). A CA must have well documented processes and practices which must be followed to ensure identity management.

The UK e-Science efforts are based around a centralised CA at Rutherford Appleton Laboratory [24]. However the process of applying for certificates is off-putting for many of the wider less-IT focused research community (like the CFG scientists) since it required them to convert the certificate to appropriate formats understandable by Grid (Globus) middleware, e.g. through running commands such as: *$> openssl pkcs12 -in cert.p12 -clcerts -nokeys -out usercert.pem* which is often not available on Windows desktops as typically used by the scientists.

We note that the UK CA now suggests for researchers with Windows based PCs that they can use a Windows openSSL based solution [25] but this in turn requires them to install and configure additional software etc. In some circumstances this is not possible, for example if they do not have sufficient privileges on their PC (root access etc) – a not uncommon practice in certain departments and faculties at Glasgow University for example. In this case the researchers will instead have to refer to a local system administrator to help with the installation and configuration.

Assuming researchers have managed to obtain a certificate which they have converted into the appropriate format, they are then expected to remember strong 16-character passwords for their private keys with the recommendation to use upper and lower case alphanumeric characters. The temptation to write down such passwords is apparent and an immediate and obvious potential security weakness. The weakest link adage of security is exacerbated in a Grid environment.

This process as a whole does not lend itself to the wider research community which the e-Science and Grid community needs to reach out to and engage with. It is a well known adage that the customer is always right. Usability and addressing researcher requirements is crucial to the uptake and success of Grid technology. End user scientists require software which simplifies their daily research and not make this more complex. Given the fact that the initial user experience of the Grid currently begins with application for UK e-Science certificates, this needs to be made as simple as possible, or potentially removed completely. Alternative solutions which do not require any user certificates are thus sought.

There are other issues with PKIs and Grid certificates as currently applied in the UK and wider Grid community. Thus for example, security is typified via access control list approaches. In the Globus solution for example, *grid-mapfiles* are manually updated and managed based upon individual user requests. The dynamicity of this manual approach is also not conducive to the Grid-idea for establishing new short term VOs. Instead users have to statically have their DNs registered at collaborating sites which have previously made available/allocated local accounts.

The fundamental issue with PKIs however, is trust. Sites trust their users, CAs and other sites. If the trust between any of these is broken, then the impact can be severe, especially since users are currently free to compile and run arbitrary code. With the now global PKI and associated recognition of international CAs through efforts such as the International Global Trust Federation [26], this basic trust model is naïve.

Advanced authorisation infrastructures which support definition and enforcement of what users are allowed to do on resources are thus needed. One of the leading authorisation infrastructures today that is closely aligned with Grid development is the Privilege and Role Management Infrastructure Services Validation (PERMIS) software [27].

The PERMIS software realises a Role Based Access Control (RBAC) authorisation infrastructure. It offers a standards-based Java API that allows developers of resource gateways (gatekeepers) to enquire if a particular access to a resource should be allowed. The PERMIS RBAC system uses XML based policies defining rules, specifying which access control decisions are to be made for given virtual organisation resources. These rules include definitions of: subjects that can be assigned roles; source of authorities (SOA), e.g. local managers trusted to assign roles to subjects; roles and their hierarchical relationships; what roles can be assigned to which subjects by which SOAs; target resources, and the actions that can be applied to them; which roles are allowed to perform which actions on which targets, and the conditions under which access can be granted to roles.

Roles are assigned to subjects by issuing them with X.509 Attribute Certificate(s). A graphical tool called the Privilege Allocator (PA) has been developed to support this process. Once roles are assigned, and policies developed, they are digitally signed by a manager and stored in one or more LDAP [28] repositories. When requests are made to access and use a given service, e.g. GridBLAST, checks on the authorisation of the user invoking the service are made by cross-checking with the signed policy in the LDAP service (in X.812 parlance the policy enforcement point interacts with the policy decision point). Depending upon the result from the policy, the decision to allow or reject is made. It should also be noted that if a given action is not explicitly allowed, i.e. included in the policy, then it is rejected.

5.1 Advanced Security with PERMIS in BRIDGES

Both the GridBLAST and the GeneVista services were based upon a fine grained Grid security model utilising the PERMIS technology. The architecture of the security infrastructure is shown in Figure 5.

IBM WebSphere was used as the portal technology.

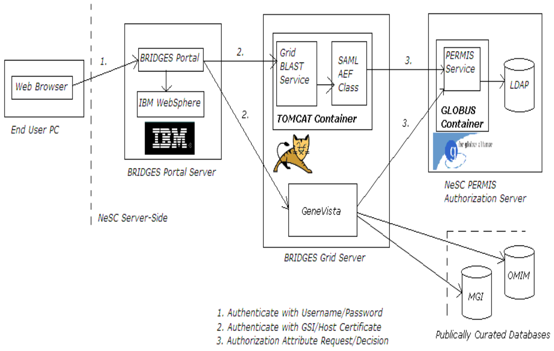


Figure 5: BRIDGES Security Infrastructure

5.1.1 GridBLAST with Advanced Security

It is the case in the Grid community right now that in order to access large scale HPC resources such as those made available through the NGS end users are expected to have a valid UK e-Science X.509 certificate. In the experiences of the BRIDGES project, this was not something that the biological end users were comfortable with (and they did not do!). To address this, we provided a solution which did not mandate that the users have their own X.509 certificates instead we exploited X.509 certificates for the *server* on which the GridBLAST service was hosted. User authentication via username/password to the portal was supported. Once authenticated (logged in), usernames were, through the PERMIS infrastructure, used to retrieve the policies that applied to that user. This information was then fed to the meta-scheduler and job submission system. The BRIDGES project supported three policies:

- If they are unknown users the job will only be submitted to the local Condor pool (we allow anyone access to the portal, however we restrict what they are allowed to do once there).
- If we recognise the users but they do not have a local ScotGrid account the job will be submitted to the Condor pool and NGS.
- If we recognise the users and they have an account on ScotGrid then the job will be submitted potentially to the Condor pool, the NGS and to ScotGrid (based on job numbers).

Given that we do not mandate that end users have a UK e-Science certificate, but provide services which allow access to resources such as NGS through server certificates requires that detailed logging of user actions is made. We also note that since users interact with the Grid resources via graphical user interfaces for the services they are not able to compile and run arbitrary code. This greatly simplifies the authorisation infrastructure.

5.1.2 GeneVista with Advanced Security

With regard to data security, PERMIS policies were defined and implemented restricting access to certain databases offered via the Data Hub to certain users. This was achieved through extensions to the GeneVista software to support queries of the PERMIS based LDAP policies. These policies distinguish CFG users from other users of the BRIDGES software. Specifically, the policies allow CFG scientists access to all of the data sets that are accessible from the federated repository. Other non-CFG users are allowed to create accounts on the portal, however they are only entitled access to the remote (federated) data sets accessible via the portal. It is important to note that both GeneVista and the GridBLAST security authorisation are completely transparent to the end users. They issue queries and receive results without any knowledge that a security policy has been enforced and that they are potentially only seeing a subset of the complete data sets depending on their role.

Through the course of the BRIDGES project, the richness of the classification of the data sets identified previously and how the infrastructure might allow for their secure sharing never fully materialised. It was and is especially difficult to convince scientists to exchange and share data that they regard as having value. Colleagues are also competitors and a philosophy of keeping data until research results have been published in journals remains. Whilst certain journals now require publication of MIAME compliant data sets for example, the data repositories are more likely to include older data sets. However it is the case that funding bodies in the UK are moving to a model whereby funding is given for life science research with the proviso that data sharing and longer term data curation considerations are incorporated [1]. It is only through changing funding models that scientists can be made to share their data since social, economic and political aspects of data sharing do exist (as demonstrated through the course of the BRIDGES project).

6. Conclusions

Security is fundamental to the success of bioinformatics and life science research. This includes both computational and data security. Experience has shown in the BRIDGES project however that usability is also crucial to the uptake and success of Grid technology. End user scientists require software which simplifies their daily research and not make this more complex. The idea of getting training on use of Grid software and resources or learning how to acquire and manage certificates and subsequently use them within a PKI is quite simply not something many scientists have the time or inclination for. Grid application and software developers need to address this fact.

The BRIDGES project has developed real data Grid and real compute Grids which have taken into account real biological user demands and explicitly targeted ease of use with fine grained security. The BRIDGES services are helping to shape the wider UK Grid activities – for example helping to define the biological data sets being deployed across the NGS.

It is a fact that the customer is always right. Whilst BRIDGES has developed much richer services in terms of functionality such as MagnaVista, end user scientists did not feel comfortable with these services hence simpler services have been engineered. Simpler and more intuitive user interfaces are crucial for the success of Grid applications. Similarly, solutions which help to overcome existing requirements on Grid infrastructures, e.g. possession of X.509 certificates, are required. Why should a biologist need an X.509 certificate when they only want to run BLAST jobs on available HPC resources? Such ideas are being taken forward in many other projects at the National e-Science Centre at the University of Glasgow where fine grained security is required, but client side software has to be trivial (and not include any complex Grid middleware.

The experiences within the BRIDGES project have shown that scientists need to be encouraged to share their research data sets. Waiting until papers are published in journals before access to MIAME compliant microarray data sets are made for example is not conducive to timely research. The BBSRC funded Grid Enabled Microarray Expression Profile Search (GEMEPS) project [29] involves a collaboration with Cornell University, US [30] and the Riken Institute in Japan [31] and is addressing these areas directly. Establishing security infrastructures across these sites so that scientists can securely share their microarray data sets so that for example they can find who has run experiments and generated similar results. When such similarities are found, cross-site research into the relevant genes can be explored. Thus the scientists have to be seen to benefit from sharing of their data sets. Thus rather than feel they are "giving away" their data sets, they should feel they are gaining access to other data sets instead. This change in paradigm is crucial for the success of any security infrastructure.

The functional genomics domain has degrees of security, however the NeSC in Glasgow are involved in other more security focused domains. For example the Virtual Organisations for Trials and Epidemiological Studies (VOTES) project [32] is exploring Grid technologies for the recruitment, data collection and study management activities of clinical trials. This includes access to patient data sets. The Genetics Healthcare Initiative [33] at NeSC is also involved in linking genetic information from people across Scotland with their medical records via a Grid infrastructure. Once again very fine grained security infrastructures are needed to enforce access control decisions. This is pushing the boundaries of advanced authorisation infrastructures in the Grid domain.

6.1 Acknowledgements

This work was supported by a grant from the Department of Trade and Industry. The authors would also like to thank members of the BRIDGES and CFG team including Prof. David Gilbert, Prof Malcolm Atkinson, Dr Dave Berry, Dr Ela Hunt and Dr Neil Hanlon. Magnus Ferrier is acknowledged for his contribution to the MagnaVista software. Derek Houghton is acknowledged for his contribution in establishing the federated data repository. Acknowledgements are also given to the IBM collaborators on BRIDGES including Dr Andy Knox, Dr Colin Henderson and Dr David White. The CFG project is supported by a grant from the Wellcome Trust foundation.

7. References

[1] P. Lord, A. MacDonald, R. Sinnott, Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models, The Joint Data Standards Study, http://www.mrc.ac.uk/pdf-jdss_final_report.pdf

[2] I. Foster, C. Kesselman, and S. Tuecke, The anatomy of the grid: Enabling scalable virtual organizations, International Journal of High Performance Computing Applications, vol. 15, pp. 200-222, Sage Publishers, London, UK, 2001.

[3] BioMedical Research Informatics Delivered by Grid Enabled Services project (BRIDGES), www.nesc.ac.uk/hub/projects/bridges

[4] Cardiovasular Functional Genomics project, www.brc.dcs.gla.ac.uk/projects/cfg

[5] IBM DB2 software family, www.ibm.com/db2

[6] IBM Information Integrator, http://www-306.ibm.com/software/data/

[7] Open Grid Service Architecture Data Access and Integration (OGSA-DAI) project, www.ogsadai.org.uk

[8] R.O. Sinnott, D. Houghton, Comparison of Data Access and Integration Technologies in the Life Science Domain, Proceedings of UK e-Science All Hands Meeting, September 2005, Nottingham, England.

[9] EMBL-EBI European Bioinformatics Institute, http://www.ebi.ac.uk/ensembl/

[10] Mouse Genome Informatics (MGI), www.informatics.jax.org/

[11] NCBI Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/OMIM/

[12] Human Genome Organisation (HUGO), www.gene.ucl.ac.uk/hugo

[13] Rat Genome Database (RGD), http://rgd.mcw.edu/

[14] Gene Ontology (GO), http://www.ebi.ac.uk/GO/

[15] R.O. Sinnott, M. Bayer, D. Houghton, D.Berry, M. Ferrier, Development of a Grid Infrastructure for Functional Genomics, Proceedings of Life Science Grid Conference (LSGrid 2004), June 2004, Kanazawa, Japan.

[16] NCBI Tools for Bioinformatics, Basic Local Alignment Search Tool, ncbi.nih.gov/Tools

[17] Condor, www.cs.wisc.edu

[18] OpenPBS, www.openpbs.org

[19] R.O. Sinnott, M. Bayer, Distributed BLAST in a Grid Computing Context, Proceedings of First International Workshop on Distributed Data Mining in Life Science, Konstanz, Germany, Sept. 2005.

[20] Globus toolkit, www.globus.org

[21] ScotGrid, www.scotgrid.ac.uk

[22] National Grid Service, www.ngs.ac.uk

[23] R. Housley, T. Polk, Planning for PKI: Best Practices Guide for Deploying Public Key Infrastructures, Wiley Computer Publishing, 2001.

[24] UK Certification Authority, www.grid-support.ac.uk/ca

[25] Windows openSSL solutions, http://www.openssl.org/related/binaries.html

[26] International Global Trust Federation www.gridpma.com

[27] D.W.Chadwick, A. Otenko, The PERMIS X.509 Role Based Privilege Management Infrastructure, Future Generation Computer Systems, 936 (2002) 1–13, December 2002. Elsevier Science BV.

[28] Lightweight Directory Access Protocol (LDAP), www.openldap.org

[29] Grid Enabled Microarray Expression Profile Search, www.nesc.ac.uk/hub/projects/gemeps

[30] Computational Biology Service Unit, Cornell University, Ithaca, New York, http://www.tc.cornell.edu/Research/CBSU/

[31] Riken Genomic Sciences Centre Bioinformatics Group, Yokohama Institute, Yokohama, Japan http://big.gsc.riken.jp/

[32] Scottish Bioinformatics Research Network (SBRN), www.nesc.ac.uk/hub/projects/sbrn

[33] Virtual Organisations for Trials and Epidemiological Studies (VOTES), www.nesc.ac.uk/hub/projects/votes

[34] Genomics and Healthcare Initiative (GHI), www.nesc.ac.uk/hub/projects/ghi