

Deploying a distributed data storage system on the UK National Grid Service using federated SRB

Manandhar A.S., Kleese K., Berrisford P., Brown G.D.

CCLRC e-Science Center

Abstract

As Grid enabled applications enter into production, there is a growing need for a distributed data access systems that are able to provide an efficient and large storage capabilities to Grid applications which have the capacity to migrate between multiple machines. The SDSC Storage Resource Broker (SRB) provides a software infrastructure to supports access to data distributed across multiple storage repositories. It virtualizes the data space and aids users in accessing data on multiple possibly geographically dispersed storage space thus acts as a Data Grid platform for data collaboration between organizations.

This paper explains the details behind the data federation strategy between the NGS sites for improving efficiency and robustness. It discusses about how the system has been designed for interoperability with grid applications and describes the ways in which the system can be utilized. Finally, the future directions for the NGS SRB are described.

1. Introduction

The SDSC Storage Resource Broker (SRB) provides a software infrastructure to supports access to data distributed across multiple storage repositories. It virtualizes the data space, provides multi user support and aids researchers in accessing data on multiple geographically dispersed storage space using multiple authentication protocols thus acts as an effective data grid platform for data collaboration between organizations.

Through the year 2003, SRB was initially piloted and deployed for E-minerals project, E-materials project and the CERN CMS project at UK. The initial responses from the user communities were very positive. At Cambridge, the user groups utilized SRB with Condor-G interface to PBS for their applications and used SRB space as their project file system [1]. The stability and performance provided by the SRB implementation were also very impressive where an average authenticated connection consistently remained within 0.2 seconds [Intel P4 3.0 GHz using Encrypt1 Authentication mechanism, recent results using GSI on similar hardware show 0.3 seconds], handled more than 161,000 connection over an typical 8 hour period [Dual server clustered Oracle on IBM X440].

As Grid enabled applications enter into production, SRB provides a reliable multi user,

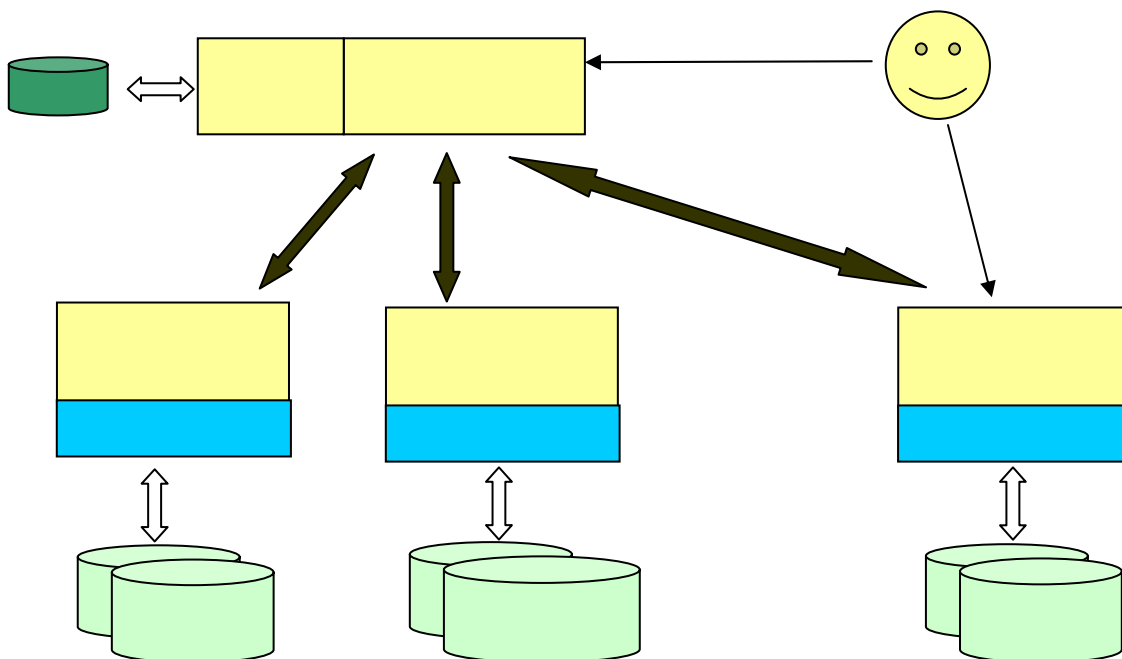
distributed data access infrastructure that provides an efficient and large storage capabilities to Grid applications which have the capability to migrate between multiple machines.

The next section describes a general structure of a SRB network. Section 3 describes the new federation concepts recently introduced into SRB for improving reliability, scalability and performance. Section 4 describes the requirements for National Grid Service (NGS) SRB that were deemed necessary and the subsequent sections describe its deployment structure and use by the projects hosted at NGS.

2. Background

An SRB Network [2] is made of multiple SRB servers whereby the servers intercommunicate to provide a coherent data service to the user. Each participating SRB server in the network provide one or more storage resource and one of the SRB servers provides the Meta data catalogue (MCAT) service which interacts with a relational database to manage the persistence of the system for user management, resource management and logical namespace management for stored files. A general deployment of SRB can be viewed as in figure 1.

A user can access the SRB system through any one of the participating servers. The user may use a Unix shell like interface (Scommands),



Windows explorer style interface (inQ) or a web browser (mySRB) to interact with the system. The system would look much like a traditional file system to the user as it arranges the files in an hierarchical structure. The user may also utilize C, Java and Webservice API for a more advanced interaction with SRB and their application.

The primary features provide by SRB are:

- Logical abstraction to multiple heterogeneous storage resources.
- Ability to inter-operate with Grid applications with GSI authentication mechanisms
- Provides user/group management features
- Fine grained access control mechanism for file access
- Provides metadata support for managing files
- Provides access to SRB via Unix, Windows and Macintosh systems

Behind the scenes SRB provides users with

- Device driver interface to disk arrays, and to mass storage systems such as Atlas and HPSS.
- Container management for efficiently managing files on Mass Storage systems

- Parallel file transfers for performance
- Bulk file transfer for performance enhancements while transferring multiple small files
- Backup and replication tools for File management

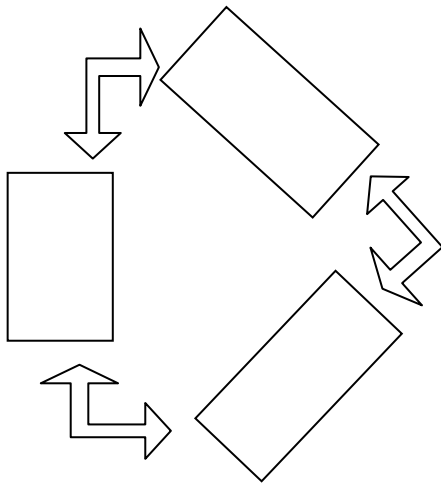
3. Federation Strategy

With the growth of SRB usage, there was a need for more manageability, reliability and performance. A federation concept [3] was brought in to SRB whereby a SRB network has the capability to contain more than one MCAT while still providing one single logical view to the data space. This provided better resilience against any single point of failure which could occur at SRB, database, operating system or network level and supported the notion of virtual organizations whereby each collaborating groups could have control over their SRB federation while still being able to create trust relations and collaborate seamlessly in a Grid like fashion.

The terminology used for a single MCAT setup is a 'Zone'. A zone may contain one or more SRB servers providing resource with a single MCAT for persistence management. A

federated SRB has two or more zone making the complete hierarchical tree.

In a federated environment, a geographic location would have its own MCAT which maintains its own resources, users and file metadata catalog. Each of the participating



zones could independently manage its network and collaborate to provide a single unified view to the data space as in a non federated environment.

In the event of any failure on one of the zone only that specific zone would be temporarily affected with the rest of the network continuing to provide the service.

For example on the National Grid Service an MCAT can be setup for each region. There could be an MCAT at RAL, Oxford, Leeds and Manchester. If for instance if the RAL's network remained down, the users still would be able to access information at Oxford, Leeds and Manchester and when RAL is back up it would be able to continue to provide the service. Also management of resources, users and projects can be simplified as each zone can manage its own network as it has the best knowledge of its area. Trust relations are then created between the zones for the level of cooperation.

Physically each participating MCAT can perform database replication for reliability of its segment of the tree at a layer below SRB (database layer) for additional reliability from the database layer.

4. Requirements

The aim of the NGS SRB is to provide an easy to use file storage management service for e-Science projects. For providing this distributed file management service, the requirements deemed necessary are:

- Location transparent wide area network access

As projects and resources span across different regions, it would be necessary to access the files irrespective of the user's location. The system should be able to provide a multi user secure access to files in a wide area environment.

- Single Logical Namespace

The Logical Namespace would provide a logical view to the files irrespective of its stored location across different servers. Providing a single logical view to the files would remove the ambiguity of the location of the files. This would make it easy for applications to retrieve the files from any computational machine as they all would have the same view to the file location.

- Easy interoperability with Grid Applications

As this service is for Grid applications, easy interoperability with Grid applications would be necessary. Much of the e-Science Grid applications utilize GSI, hence, support for GSI authentication mechanism would be necessary for easy interoperability.

- Reliability

The system would need to be reliable with the ability to ensure integrity of the overall system and provide an acceptable level for fault tolerance

- Manageability

The system would also be able to manage multiple users and resources and be able to provide service to multiple projects.

- Expandability

The system would need to be able to cope with addition of new resources and projects. It would also need to be able to cope with addition of new sites.

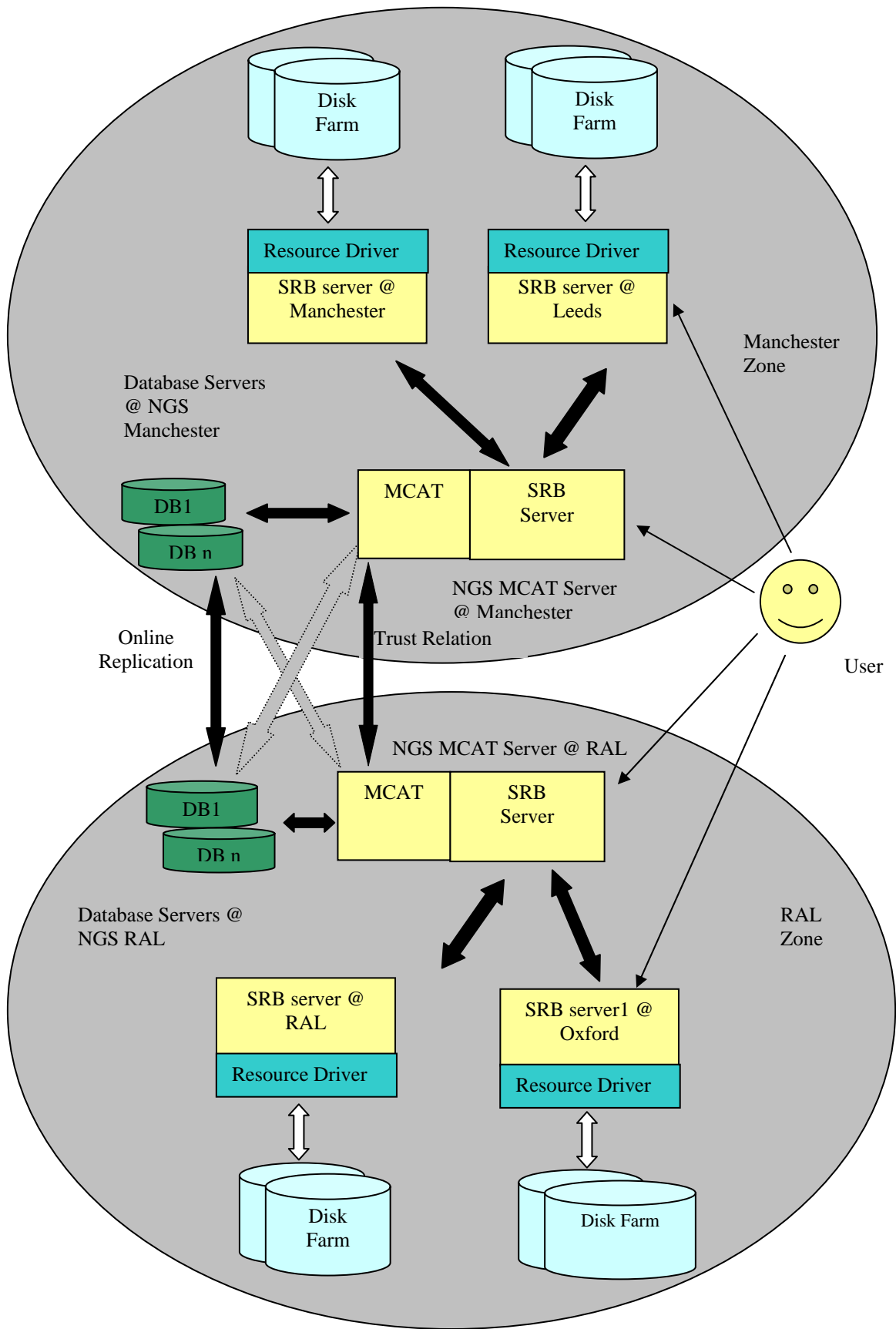


Figure 3: NGS SRB deployment structure using federation

5. Deployment Structure

While deploying the NGS SRB, federation concept is being used as its architectural benefits provides better reliability, manageability and room for easy future expansion.

Currently the NGS SRB is being deployed as a two zone federation centralized around the two data nodes of NGS at RAL and Manchester. Each of the zones hosts its own MCAT and manages its own part of the network and resources. A trust relation is established for the information between the two MCAT so that users may seamlessly switch between zones.

Oxford and Leeds clusters utilize the MCAT service provided by RAL and Manchester respectively as depicted in the diagram. Ideally it may have been suitable for each of these clusters to also have its own MCAT so that it would be able to continue to provide a service even if the MCAT at RAL or Manchester were to go down. However due to physical and human resources and manageability issues it seemed more suitable for these computational clusters to utilize the MCAT service provided by the data clusters at RAL and Manchester.

Logically each of the federating members is a child node of the root node and so a user continues to see it as one hierarchical tree. In this context logically 'ral' and 'man' are two child nodes of the root node 'ngs'. Hence as the user changes directory from say 'ral' to 'man', the user seamlessly switches between zones. In this way, even though there are multiple zones participating in the federation, the user would continue to see it as one logical structure.

At each site, each MCAT utilizes a clustered Oracle database service whereby if one of the nodes of the Oracle cluster fails the other nodes continue to provide the service. For additional reliability online database replication is being performed. The MCAT database at RAL is being replicated to Manchester and similarly the MCAT database at Manchester is being replicated to RAL. In the even of one the database service going down the MCAT server continues to get its service from the backup database.

6. Benefits by the use of federation

By using the federation feature provided by SRB, the NGS SRB benefits in the following issues:

Reliability

The MCAT being a central point the system, can also be a point of concern. With federation the persistence information is shared between multiple MCATs present in the federation. In the event of one of the MCAT temporarily going down it does not bring the complete system down as would have been the case if there would be only one MCAT. Only segment of SRB that is controlled by the offline MCAT would be temporarily affected. Given the level of importance of the data, by use of replication techniques within the federation it would be possible to further minimize the impact of an MCAT going down.

Manageability

The MCAT is also the central point of administration. Hence addition of new users and resources to the system is centrally managed by the MCAT administrator. As new storage spaces are needed, request would be made for modification on to the MCAT to support the new resources and its access privileges. Likewise for user management and creation of new users are performed centrally.

In a federated SRB each Zone is a virtual organization. As SRB network grows, there is an increase of users, groups/projects and resources which makes a central point of administration more difficult. Projects may want to add more of its physical resources or users to SRB. It would be more manageable if each zone manages its own users. In a federated network each Zone has the responsibility to maintain its segment of the network and it is also in the best position to manage its resources and user communities as maintenance of an SRB network involves hardware, OS, network, database and user interaction.

Scalability

As projects grow they may need more autonomy while still being able to collaborate with other projects and interact with NGS services. In a federated SRB it is possible for projects to start their own SRB service and

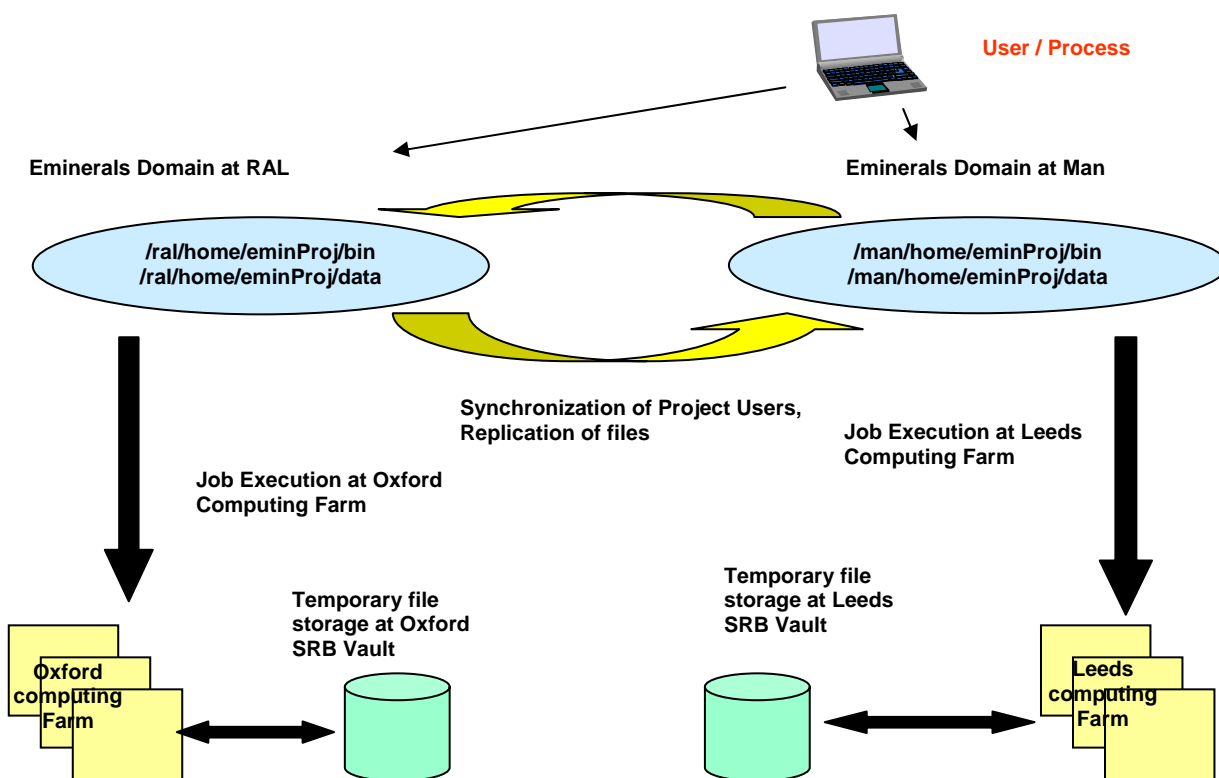


Figure 4: NGS SRB deployment structure using federation

create trust relations to join the NGS SRB federation.

Likewise Universities may want to join in to SRB federation. Universities can start their own SRB network and join in to NGS federation with different groups at NGS being able to collaborate with the groups at Universities. Hence in this way Data Grid deployment is made very scalable

Performance

Over the past few years, SRB user communities have grown both in numbers and physical distance. Wide area network performance has become one of the issues. Network bandwidth has been continuously increasing faster than the pace other components of a computer system. However the initial access time is still limited. In order for SRB to satisfy a user request, it makes multiple interactions with the MCAT system. These requests have small payloads so increase in bandwidth may not necessarily have much effect but the physical distance between the physical resources and the MCAT can affect the performance of the system as there are multiple queries for each request. Having a

MCAT close to the resource improves the overall performance.

7. SRB Usage by Projects

Every project will have its own requirements and possibly its own creative way of utilizing the service based upon their requirements. Figure 4 describes the SRB usage scenario by the e-minerals project on the NGS. The minerals utilizes the SRB space as its online project file system that can be access from any of its sites. The setup has two primary directories namely '/eminProj/bin' for program executables and '/eminProj/data' for result data within its domain 'eminProj'. For code execution the user logs into a computational farm (say Oxford node) and retrieves the program executables and input data from the SRB space. During execution it may utilized the SRB storage resource at Oxford itself for temporary data storage and on completion of the job it then returns the output data back into the SRB space.

For additional reliability an identical directory structure '/eminProj/bin' and '/eminProj/data' is created under the 'man' subdirectory and is regularly synchronized. In this way a complete

set of information is also located at Manchester zone and in the case of RAL zone being temporarily offline the users can continue to work from the other zone.

The e-minerals project also hosts its own SRB storage space at Cambridge, UCL and Reading servers external to the NGS servers. For longer term storage of files or if the execution of jobs is more frequent at its computational clusters it may physically move the necessary files from the NGS storage space to the storage space at the other sites. This is also another benefit of the logical namespace provided by SRB as even if the files are physically moved between places to the applications it would still seem to be on the same logical location.

9. Future Additions

The current SRB setup on the NGS is still relatively new and is continually shaping. The ongoing modifications will continue to be made based on new upcoming requirements as more projects start to use the service.

In the near future we are looking to build new tools for easier inter operability between computational Grids and data Grids. We are also looking to provide transfer management tools that maybe used by the many projects, introduce tools being built by the SRB community and modify as SRB evolves.

10. References

- [1] *Dove M., Calleja, M., et al*, .Environment from the molecular level: an science test bed project, AHM 2003, Nottingham UK
- [2] *Rajasekar A., Wan M., et al*, Storage Resource Broker – Managing distributed data in a Grid, Computer Society of India Journal, Vol. 33, No.4 Oct 2003
- [3] *Rajasekar A, Wan M, Moore R., Schroeder W.*, Data Grid Federation, PDPTA, Las Vegas , June 2004.