Jornadas de Seguimiento de Proyectos, 2007
Programa Nacional de Tecnologías Informáticas

HPSGRID: High Performance Storage for Grid Environments TIN2004-02156

Félix García Carballeira *
Universidad Carlos III de Madrid

Abstract

Data management is one of most important problem in grid environments. Most of the efforts in data management in grids have been focused in data replication. Data replication is a practical and effective method to achieve efficient data access in grids. However, all data replication schemes lack in to provide a grid file system. One important challenge facing grid computing is the design of a grid file system. The Open Grid Forum defines a Grid File System as a human-readable resource namespace for management of heterogeneous distributed data resources, that can span across multiple autonomous administrative domains. This project tries to investigate and to develop technologies that allow to use and integrate the resources and existing solutions to build a high performance, fault tolerance and secure grid file system for grids environment.
Keywords: parallel I/O, grid, replication, grid file system

1 Projet Objectives

The main goal of this project is the development of techniques for improving the performance in data access in grid environments. The specific objectives are the following:

1. Objective 1. To apply the parallelism used in parallel file systems for clusters to grid environments. With this aim, we propose the application of parallel I/O techniques to grid systems..
2. Objective 2. To develop techniques of location, resource management, caching, and load balancing.
3. Objective 3. To develop fault tolerance techniques different from the replication techniques used in grid environments. We will provide new fault tolerance models based on redundancy similar to the employed in RAIDs.
4. Objective 4. To develop a security model for the authentication, confidentiality, and data security.

*Email: `felix.garcia@uc3m.es`

2 Research Results

At the time of writing this report, the project has consumed 24 months out of its whole three year period. At present, a large part of the proposed goals have already been accomplished. Next, we show the level of success level reached in the former objectives:

- Objective 1. We have developed a parallel file system for grids using standard grid technologies. The last year of the project will be dedicated to implement new services in this parallel file system, as for example, a storage service using web services.
- Objective 2. The location is resolved using the grid file system implemented. Last year of the project will be dedicated to study caching and load balancing strategies.
- Objective 3. We have developed a fault tolerant model for parallel file system. This model has been implemented in the cluster version of Expand. The last year of the project will be dedicated to incorporate this model in the parallel file system for grid.
- Objective 4. Security in ensured in the parallel file system implemented because we use the grid security infrastructure (GSI) provided by the Globus Toolkit. This service provides authentitication and security in data transfers.

Next sections summarize the main research achievements reached until now.

2.1 A parallel file system for grids

Many applications in grids, require the access to large amounts of data. For this kind of applications, most of the efforts in data management in grids have been focused in data replication. Data replication is a practical and effective method to achieve efficient data access. File replicas are multiples copies of a file spread across the grid used for improving data access. However all data replication schemes provided in literature lack in to provide a global file system for accessing to files in data grids environments. One important challenge facing grid computing is a true and global file system for grid applications running in different administrative domains.

The Open Grid Forum [1] defines a Grid File System as a human-readable resource namespace for management of heterogeneous distributed data resources that can span across multiple autonomous administrative domains, that can include:

- A logical resource namespace across multiple administrative domains.
- Standard interfaces.
- A virtual namespace of WAN file system.
- Independence of physical data access/transport, authentication mechanisms.

Furthermore, a grid file system must to provide high performance data access. Parallel I/O has been the traditional technique for improving data access in clusters and multiprocessors. Parallelism in file systems is obtained by using several independent servers and striping data among these nodes to allow parallel access to files. There are many parallel file systems for this kind of platforms, however very few parallel file systems has been provided for grid environments.

In this project we have developed a grid file system (objective 1) based on the Expand Parallel File System [2] designed by the research group. Expand is a parallel file system for clusters that uses standard servers and protocols for building parallel and distributed partitions where files are distributed. We have extended this parallel file system for providing a parallel file system for grids. The new version of Expand for grids uses standard grid technologies: GridFTP for data access, and the Resource Namespace Service for naming and data location (objective 2). This system allows the integration of existing heterogeneous servers span across multiple administrative domains in grids for providing parallel I/O services using standard interfaces like POSIX and MPI-IO. The new version of this parallel file system has been implemented using Globus [3], one of the most important middleware for building grid applications. The security is based on the grid security infrastructure (GSI) provided by the Globus Toolkit (objective 4).

Then, we described the main standard technologies used for providing the parallel file system for grids, and described the adaptation of Expand to grid environments.

GridFTP GridFTP [4] [3] is a data transfer protocol defined by the Global Grid Forum Recommendation, that provides secure and high performance data movement in grid systems. The Globus Toolkit provides the most commonly used implementation of that protocol, though others do exist (primarily tied to proprietary internal systems). This protocol extends the standard FTP protocol and includes the following features:

- Grid Security Infrastructure (GSI) support.
- Third-party control and data transfer.
- Parallel data transfer using multiple TCP streams.
- Striped data transfer using multiple servers.
- Partial file transfer and support for reliable and restartable data transfer.

The transfer data in our parallel file system is based on this protocol and the implementation provided by the Globus Toolkit. The access to this protocol in Globus is provided via two libraries: the Globus FTP control library, and the Globus FTP client library. The first library provides low-level services needed to implement FTP client and servers. The API provided by this library is protocol specific. The Globus FTP Client library provides a convenient way of accessing files on remote FTP servers.

Resource Namespace Service The Resource Namespace Service (RNS) is a specification of the Open File System Working Group (GFS-WG) of Open Grid Forum that allows the construction of a uniform, global, hierarchical namespace. It is a web service described by a RNS WSDL [5]

RNS is composed of two main components: *virtual directories* and *junctions*. Directories are virtual because they do not have any corresponding representation outside of the namespace. A junction is an RNS entry that interconnects a reference to an existing resource into the global

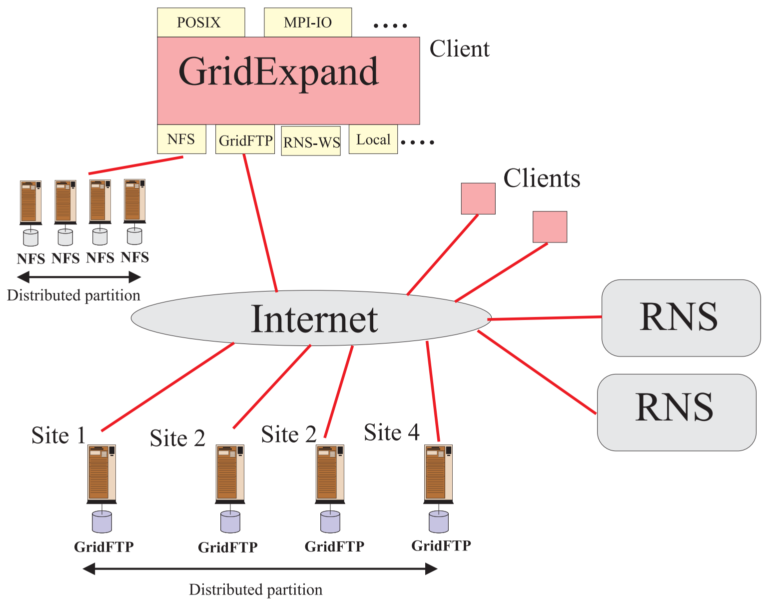

Figure 1: Expand Architecture for grid environments

namespace. There are two types of junction, referral junction and logical reference junction. The first is used to graft RNS namespaces. A logical reference junction contains a unique logical name. For this type of junction, the RNS service returns an endpoint reference (EPR). This endpoint is resolved using the Resource Endpoint Resolution Service or RNS resolver.

RNS is composed of several types of operations: operations for querying namespace entry information; operations for creating, removing, renaming, and updating entries; and operations for managing properties or status of an entry.

The RNS resolver is a service independent of RNS that has the mapping between logical names and endpoint references or address. For each logical name, the RNS resolver may store several endpoint references. The RNS resolver has operations for resolving names to endpoint references, and operations for creating, removing, and updating logical references.

We have implemented a prototype of web service for Globus that incorporates the RNS specification and that has been used in Expand for providing a global namespace in grids (objective 2).

2.1.1 Expand design for grids environments

Figure 1 shows the architecture of Expand for grid environments. This figure shows how Expand can be used for data management in clusters environments and how can be used to access to several sites using the GridFTP protocol. File data are striped by Expand among several servers using different protocols, using blocks of different sizes as stripping unit. Processes in clients use an Expand library to access to an Expand distributed partition. Expand offers an interface based on POSIX system calls. This interface, however, is not appropriate for parallel applications using strided patterns with small access size [6]. For parallel applications, we use ROMIO [7] to support MPI-IO interface [8], implementing the appropriate Expand ADIO.

Expand combines several GridFTP servers (see Figure 1) in order to provide a generic distributed partition. The use of GridFTP allows to use servers allocated in different administrative domains. Each server provide one or more directories that are combined to build a distributed partition through the grid. All files in the system are striped across all GridFTP servers to facilitate parallel access, with each server storing conceptually a subfile of the parallel file.

Expand offers two different interfaces. The first interface is based on POSIX system call. This interface, however, is not appropriate for parallel applications using strided patterns with small access size [6]. Parallel applications can also used Expand with MPI-IO [8]. Expand have been integrated inside ROMIO [7] and can be used with MPICH. Portability in ROMIO is achieved using an *abstract-device interface for I/O* (ADIO).

Parallel access and authentication All file operations in Expand use a *virtual filehandle*. This virtual filehandle is the reference used in Expand to reference all operations. When Expand needs to access to a subfile, it uses the appropriated filehandle. For GridFTP Expand uses the appropriate handle managed by the GridFtp Client Library provided by Globus. To enhance I/O, user requests are split by the Expand library into parallel subrequests sent to the involved servers. When a request involves k GridFTP servers, Expand issues k requests in parallel to the servers, using threads to parallelize the operations. The same criteria is used in all Expand operations. A parallel operation to k servers is divided in k individual operations that are provided by the Globus GridFTP Client Library to access the corresponding subfile.

The access control and authentication is guaranteed in Expand for grid environments, because GridFTP uses the Grid Security Infrastructure provided by Globus. GSI uses public key cryptography as the basis for its functionality and provides secure communication (authenticated and perhaps confidential) between elements of a computational Grid, security across organizational boundaries, thus prohibiting a centrally-managed security system, and support *single sign-on* for users of the Grid, including delegation of credentials for computations that involve multiple resources and/or sites.

2.2 Branch Replication: A New Model for Data Replication in Large Scale Data Grids

Other contribution of this project has been the design of a new replication algorithm called *Branch Replication* [1] . This algorithm tries to bring more scalability through the parallel replication of a replica in several sites, where a replica is divided in subreplicas that reside in several nodes. The main features of this replication method are the following:

- In this algorithm, a storage node support the original file (root replica).
- When we want to create a new replica, n nodes have to be selected, where the subreplicas are going to be supported, but the union of all these subreplicas are the original replica.

[1] Algorithm described in the PhD Thesis *Performance Increase Mechanisms for data access in grids*, José María Pérez Menor

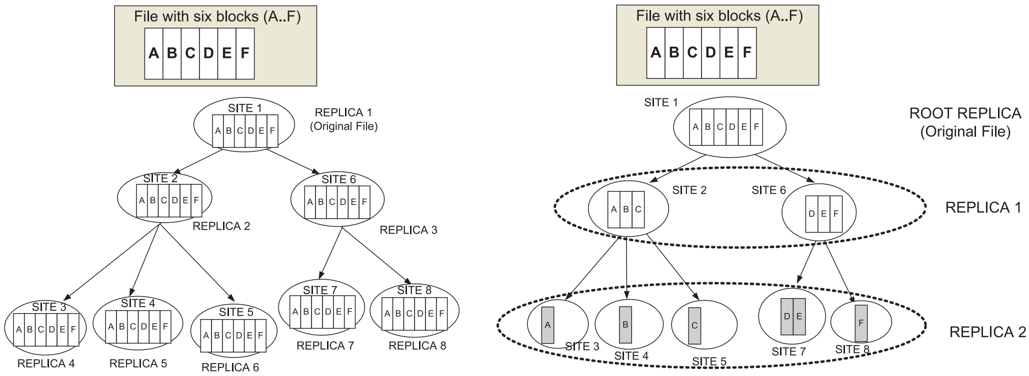

Figure 2: Differences between Hierarchical Replication (left) and Branch Replication (right)

We perform the stripping of the original replica. The subreplicas creation can be performed in parallel, so we can perform the replication in less time that the replica creation in one storage node.

- If we want to make another replica, the subreplicas can be split in several subsubreplicas, and so on.

This replication method allows to use this scheme in grid environments with many data storage resources with different sizes. Moreover, this approach allows parallel access to replicas and it requires less space per node to store the replicas. In traditional approaches, when an entire replica of a file with K GB is made we need K GB per node. Using the *branch replication* with n replicas we require more nodes, 2^{n-1} nodes. This is not a problem in large data grids, however, each node need less space to store data replicas. For an original file with K GB and n replicas, each storage node stores in average:

$$\frac{\sum_{i=1}^{N}\left(\frac{K}{2^{i-1}}\right)}{2^{n-1}} \quad GB$$

Figure 2 shows the differences between a Hierarchical Replication as the proposed in [9] and the branch replication designed in this project.

Replica Location Service Maintaining replicated data in grid environments requires a Replica Location Service (RLS) for providing a mechanism for registering new replicas and discovering them. To support the branch replication described above we must store for each subreplica the following information:

- FR: Replica of upper level.
- CR: Set of child subreplicas. It is stored the location of the files that support the subreplicas and the data replicated in each of them.
- BR: Set of subreplicas at the same level with a common upper level.

In order to provide a standard method for replica location we use the Resource Namespace Service (RNS), a specification of the Grid File System Working Group (GFS-WG) of Global Grid Forum that allows the construction of a uniform, global, hierarchical namespace. It is a web service described by a RNS WSDL [5].

Replica Consistency In order to maintain consistency among updates by clients we use the following mechanism. Clients only can modify the data located in the leaf nodes of the replication tree. So the location of the replica is reduced to the location of the subreplicas of more depth that support the range of data that the application wants to work with. The data update is performed from the child replicas to the parent using parallel and collective I/O techniques.

3 Result Indicators

3.1 PhD Thesis

- Alejandro Calderón Mateos. Fault tolerant techniques for parallel file systems in clusters. Universidad Carlos III de Madrid. 2005.
- José María Pérez Menor. Performance increase mechanisms for data access in grids. Universidad Carlos III de Madrid. 2006.

3.1.1 Thesis in progress

- Borja Bergua Guerra. Parallel file systems for grids and their use in jobs scheduling. Universidad Carlos III de Madrid.
- Margarita Reyes Rodríguez. New techniques for replication and scheduling in grid systems. Universidad Carlos III de Madrid.
- Luis Miguel Sanchez García. Scalability in parallel file systems for large scale clusters.
- Víctor Méndez Muñoz. Applying Artificial Intelligent techniques to replication services in grids. Universidad Carlos III de Madrid.

3.2 Publications

1. F. García, J. Carretero, A. Calderón, J. D. García, L. M. Sánchez. A global and parallel file system for Grids. Future Generation Computer Systems. Vol. 23, 2007. Pags. 116-122
2. M.S. Pérez, J. Carretero, F. García, J. M. Peña, V. Robles. MAPFS: A Flexible Multiagent Parallel File System for Clusters. Future Generation Computer Systems. Vol. 22, 2006. Pags. 620-632
3. David E. Singh, Florin Isaila, Ajelandro Calderón, Félix García, Jesús Carretero Multiple-Phase Collective I/O Technique for improving data access locality Aceptado para publicación en 15th Euromicro International Conference on Parallel, Distributed and Network-based Processing, PDP 2007. Napoles, Italia, 7-9 de febrero de 2007

4. F. Isaila, D. Singh, J. Carretero, F. García, G. Szeder, T. Moschny Integrating Logical and Physical File Models in the MPI-IO Implementation of Clusterfile 6th IEEE International Symposium on Cluster Computing and the Grid. CCGrid2006. Singapur 16-19 mayo de 2006. Pags: 462-469

5. F. Isaila, D. Singh, J. Carretero, F. García. On evaluating decentralized parallel I/O scheduling strategies for parallel file systems. Acepted for publication in Lecture Notes in Computer Science. Revisión posterior al congreso VEPCR'06, 7th International Meeting on High Performance Computing for Computacional Science. 2007.

6. V. Méndez, F. García PSO-Grid Data Replication Service Aceptado para publicación en Lecture Notes in Computer Science. Revisión posterior al congreso VEPCAR'06, 7th International Meeting on High Performance Computing for Computacional Science. A publicar en 2007

7. V. Méndez, F. García PSO vs. ACO, Data Grid Replication Services Performance Evaluation Lecture Notes in Computer Science. Vol 4331. Frontiers of High Performance Computing and Networking – ISPA 2006. International Workshop on Middleware Performance (WOMP). Sorrento, Italy, 4 -7 de diciembre de 2006. Pags. 833-843

8. Mª B. Ibañez, F. García, J. Carretero A Profiling Approach for the Management of Writing in Irregular Applications Lecture Notes in Computer Science. Fourth International Symposium on Parallel and Distributed Processing and Applications (ISPA'2006) Sorrento, Italy ,4-6 de diciembrec. 2006. Pags. 251-259

9. L. M. Sanchez, F. Isaila, F. García, J. Carretero, R. Rabenseifner, P. Adamidis A New I/O Archirecture for Improving the Performance in Large Scale Clusters Lecture Notes in Computer Science. Vol 3984/2006. Computational Science and Its Applications - ICCSA 2006: International Conference, Glasgow, UK, May 8-11, 2006, Proceedings, Part V. Pags. 108-117

10. J. M. Pérez, F. García, J. Carretero, J. D. García, S. Escolar Data Driven Infrastructure and Policy Selection to Enhance Scientific Applications in Grid Lecture Notes in Computer Science. Vol 3458/2005. Scientific Applications of Grid Computing: First International Workshop, SAG 2004, Beijing, China, September 20-24, 2004, Revised Selected and Invited Papers. Pags. 59-74

11. M. B. Ibáñez, J. Carretero, F. Garcia. Symmetrical Data Sieving for Noncontiguous I/O Accesses in Molecular Dynamics Simulations. Lecture notes in Computer Science. Vol 3666/2005. Recent Advances in Parallel Virtual Machine and Message Passing Interface: 12th European PVM/MPI Users' Group Meeting Sorrento, Italy, September 18-21, 2005. Proceedings. Editors: Beniamino Di Martino, Dieter Kranzlmüller, Jack Dongarra. Pags. 441-448

12. M. S. Perez, J. Carretero, F. Garcia, J. M. Peña, V. Robles MAPFS-GRID: A Flexible Architecture for Data-Intensive Grid Applications Lecture Notes in Computer Science series. Vol. 2970/2004. Grid Computing: First European Across Grids Conference, Santiago de Compostela, Spain, February 13-14, 2003. Revised Papers. Pags. 111-118. Publicado en 2004

13. J. M. Perez, F. Garcia, J. Carretero, A. Calderón, J. Fernández A Parallel I/O Middleware to Integrate Heterogeneous Storage Resources on Grids Lecture Notes in Computer Science series. Vol. 2970/2004. Grid Computing: First European Across Grids Conference, Santiago de Compostela, Spain, February 13-14, 2004. Revised Papers. Pags. 124-131

14. D. E. Singh, F. Garcia, F. Carretero. I/O optimization for an air pollution model. Parallel Computing 2005 (Parco2005). 13-16 de septiembre de 2005 Malaga. Spain

15. J.M. Perez, L. M. Sánchez, , F. García, A. Calderón, J. Carretero. High Performance Java Input/Output for Heterogeneous Distributed Computing. 10th IEEE Symposium on Computers and Communications / ISCC 2005. Pp. 969-974. La Manga del Mar Menor, Cartagena, España. Junio 27-30, 2005

16. A. Calderón, F. García., J. Carretero, J. M. Pérez, L. M. Sánchez. A Fault Tolerant MPI-IO implementation using the Expand Parallel File System. 13th Euromicro Conference on Parallel, Distributed and Network-based Processing, PDP 2005. Lugano, Suiza. 9-11 February 2005. ISBN: 0-7695-2280-7. Pags. 274-281

17. A. Calderón, F. García, J. Carretero, L.M. Sánchez, J. M. Pérez. File schemes for fault tolerant in heterogeneous distributed storage systems New Trends in Distributed Data Access. Proceedings of the 2004 International Multiconferece in computer science and computer engineering (PDPTA'04), ISBN: 1-932415-24-6. Págs. 561-567

18. M.S. Pérez, A. Sánchez, J.M. Peña, V. Robles, J. Carretero, F. García. Storage Groups: A New Approach for Providing Dynamic reconfiguration in Data-Based Clusters. Proceedings of the IASTED International Conference on Parallel and Distributed Computing and Networks. February 17-19, 2004, Innsbruck, Austria. Pp: 70-75

19. A. Núñez, J. Carretero, F. García, L.M. Sánchez Sistema de ficheros paralelo basado en servicios web XVII Jornadas de Paralelismo. Albacete. 18-20 de septiembre de 2006. Pags. 191-186

20. D. E. Singh, Mª B. Ibáñez, F. Isaila, F. García-Carballeira, J. Carretero. Optimizaciones de entrada/salida para aplicaciones de dinámica molecular XVI Jornadas de Paralelismo. I Congreso Español de Informática. Granada. 13-16 de septiembre de 2005. Pags. 685-692

21. J. M. Pérez Menor, F. García Carballeira, J. Carretero , J. D. García , A. Calderón Incremento de prestaciones en el acceso a datos en Data Grids XVI Jornadas de Paralelismo. I Congreso Español de Informática. Granada. 13-16 de septiembre de 2005. Pags. 347-354

22. L. M. Sánchez, J. M. Pérez, F. García, A. Calderón, J. Carretero. Arquitectura escalable para E/S de altas prestaciones en sistemas heterogéneos XV Jornadas de Paralelismo. Almería. 15-17 de septiembre de 2004.

3.3 Cooperation with other groups

- Thematic net for the Coordination of Middleware Activities in Grids. This thematic net is coordinated by the Universidad Complutense de Madrid (http://asds.dacya.ucm.es/GridMiddleware/presentacion.php)
- GridiMADRID (http://www.gridimadrid.org), An initiative of grid in the Comunidad de Madrid.

References

[1] Open Grid Forum, Grid File System Working Group, http://www.opf.org

[2] F. Garcia, A. Calderon, J. Carretero, J.M. Perez, J. Fernandez, The design of the Expand parallel file system, International Journal of High Performance Computing Applications, Vol 17, N.1, 2003, pp. 21-37.

[3] The Globus Toolkit, http://www.globus.org

[4] B. Allock, J. Bester, J. Bresnahn, A. Chervenak, I. Foster, C. Keseelman, S. Meder, V. Nefedova, D. Quesnel, S. Tuecke, Secure, efficient data rransport and replica management for high performance data-intensive computing, In Proceedings of the Eighteenth IEEE Symposium on Mass Storage Systems and Technologies - Volume 00, April 17 - 20, 2001, pp. 13-28.

[5] M. Pereira, O. Tatebe, L. Luan, T. Anderson, J. Xu, Resource Namespace Service specification, nov. 2005. http://www.global.http://www.globalgridforum.net.

[6] N. Nieuwejaar, D. Kotz, The Galley parallel file system, In Proceedings of the 10th ACM International Conference on Supercomputing, May 1996.

[7] W. Gropp R. Takhur and E. Lusk, An abstract-device interface for implementing portable parallel-I/O interfaces, In Proceedings of the 6th Symposium on the Frontiers of Massively Parallel Computation, Oct. 1996, pp. 180-187.

[8] A. Calderon, F. Garcia, J. Carretero, J.M. Perez, J. Fernandez, An implementation of MPI-IO on Expand: A parallel file system based on NFS servers, In 9th PVM/MPI European User´s Group, Johannes Kepler University Linz, Austria. sep 29-oct 2, 2002. Pp. 306-313.

[9] H. Lamehamedi, Z. Shentu, B. Szymanski, E. Deelman. Simulation of Dynamic Data Replication Strategies in Data Grids. Proc. 12th Heterogeneous Computing Workshop (HCW2003) Nice, France, April 2003, IEEE Computer Science Press, Los Alamitos, CA, 2003.