Jornada de Seguimiento de Proyectos, 2010
Programa Nacional de Tecnologías Informáticas

New scalable storage techniques for high-performance computing (TEALES)[1]

TIN2007-63092

Jesús Carretero[*]
Universidad Carlos III de Madrid

Abstract

Nowadays there is a great interest on developing new technologies for high-performance computing using grid and cluster environment integrated together. Storage systems are currently one of the major problems in this kind of installations for high-performance computing applications, involving aspects related to performance and usability. Access to data is still an unresolved problem, especially when the whole storage hierarchy is considered because of problems of heterogeneity and the always growing unbalance between the storage system bandwidth and the computing power of the compute nodes. The last problem is becoming especially important with the arrival of new multicore architectures.
In this project we propose to develop a new data access technology for grids and clusters relying on the integration of the already existing input/output mechanisms for the whole storage hierarchy. A triple goal is pursued: to develop new techniques to provide scalable storage in grid environments, to propose new storage techniques and architectures for clusters, and to develop, from the applications point of view, techniques for distributing and organizing data to enhance the performance of input/output operations.

Keywords: Grid computing, cluster computing, manycore, storage systems, I/O optimizations, fault tolerance

1 Project goals

In this project we propose to develop a new data access technology for grids and clusters relying on the integration of the already existing input/output mechanisms for the whole storage hierarchy. A triple goal is pursued: to develop new techniques to provide scalable storage in grid environments, to propose new storage techniques and architectures for clusters, and to develop, from the applications point of view, techniques for distributing and organizing data to enhance the performance of input/output operations.
For the topics of the Project, today systems use several technologies that have some clear problems concerning input/output systems. On the one side, Clusters have bottlenecks in data access paths.

[1] **Nuevas técnicas de almacenamiento escalable en computación de altas prestaciones**
[*] Email: jesus.carretero@uc3m.es
[*] Web: http://arcos.inf.uc3m.es/~teales/doku.php

On the other side, Grid data storage technologies are still in a very early stage of development, provided that basically they have a file transfer protocol, like GridFTP, and data replication techniques to provide reliability. Moreover, in both cases, data replication and hardware support are used as the basic ways of increasing storage reliability.

The project has the following specific goals:

1.- To develop new storage techniques for grid environments to enhance performance in data accesses and to achieve a good usability for those environments from the data management, resource scheduling and job execution perspectives. To achieve that, we´ll propose new grid scheduling, data localization, resource management, global data access, and load distribution techniques. Moreover, we´ll propose techniques to provide fault-tolerance and an enhanced reliability in grid environments.

2.- To propose new input/output models and architectures for clusters to solve some of the existing bottlenecks, to provide high scalability and performance, at the same that they increase the reliability through new methods of fault-tolerance and data replication adapted to the architectures proposed.

3.- To study the influence of multicore processors on the input/output mechanisms of the cluster architectures from two points of view: impact on the bandwidth increasing from each node over the efficiency of the input/output system and application of multicore processors to the input/output nodes.

4.- To propose new methods to optimize the input/output in massively parallel applications. We´ll propose optimization techniques base on data distribution over compute nodes and optimization techniques based on the data locality into the processor.

5.- To integrate the developed solutions into a prototype of the Expand parallel file system, developed by the research group over the last years. Expand is a parallel file system based on standard servers. Currently it uses NFS as base protocol for clusters and GridFTP to be used with Globus in Grid environments. On one side we plan to enhance the existing prototype, and on the other we plan to validate with a real prototype the new techniques and solutions developed into the project.

We intend to apply parallel and distributed I/O techniques to define geographically distributed metafiles. This kind of solutions a huge storage capacity may be available as resources from different domains are aggregated and parallel access to data of files composing the metafile can be performed. Besides, the usage of these metafiles will allow implementing by software and in an economical way fault tolerant mechanisms on the global distribution whose model will be specified by the user, allowing increasing system reliability. To solve security problems, used techniques will be integrated into the Globus GSI security model, an industry standard to ensure authentication, integrity and access control.

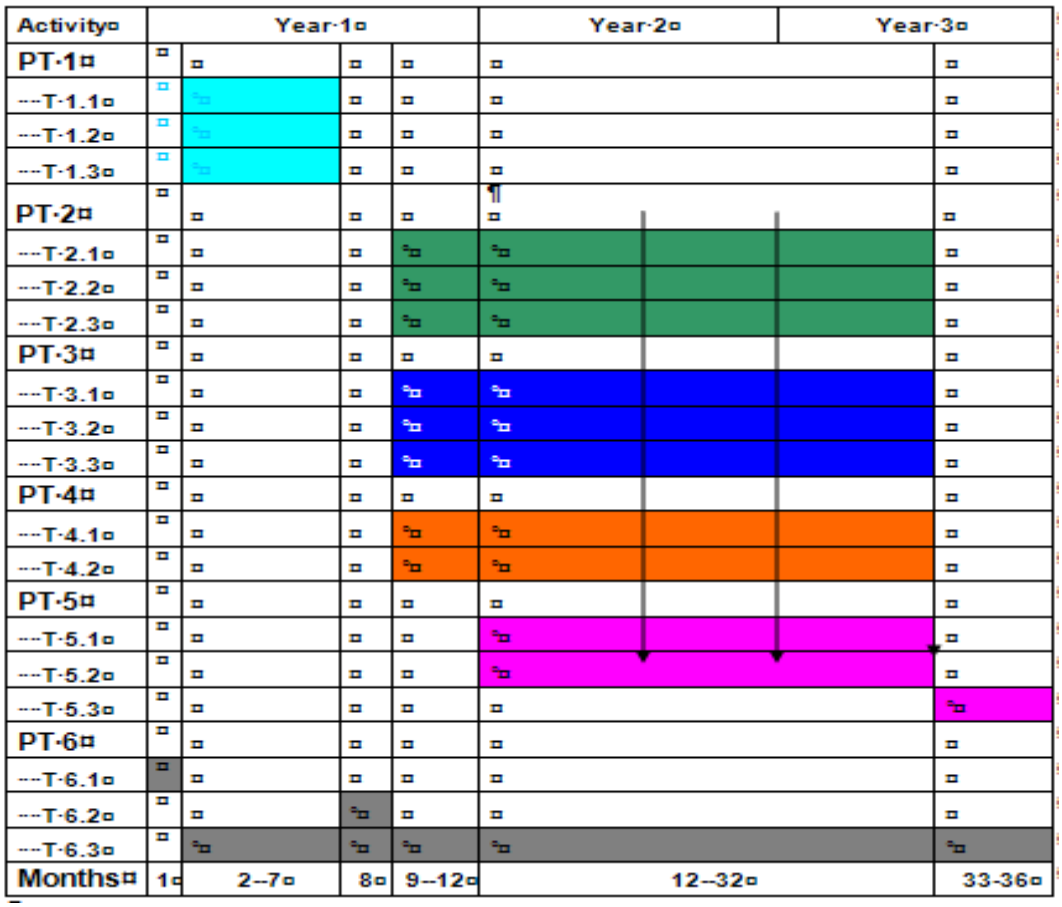
The Expand parallel file system was an initial platform to integrate the new techniques and solutions to be developed into the Project, as it is already demonstrated the feasibility of including

existing resources, as NFS for Clusters and GridFTp for Grid environments, to build a high-performance storage system in Clusters and Grids.

The **work plan** proposed has a length of 36 months and is organized in six work packages:

1. Bibliographic survey and alternatives study.
2. New scalable storage techniques for Grid environments.
3. New scalable storage techniques for Cluster environments
4. Data distribution and organization techniques for performance improvement in Input/Output operations.
5. Systems integration, final prototype and evaluation.
6. Coordination and planning activities.

The temporal planning of the activities of the project is show below.



2 Success level of the project

This section shows an analysis of the research work done in the TEALES project, the goals already achieved, and the most relevant scientific and technological results of the project.

Below, we describe the former aspects for every work package of the work plan included in the project proposal.

2.1 Bibliographic survey and alternatives study

The major goals of this task were to perform an exhaustive bibliographic survey of existing solutions on Grid and clusters I/O and storage systems and techniques in Input/Output optimization for high performance applications. It was successfully accomplished. Actually, the studies done have been very useful to complete the state of the art of the PhD thesis going on. Moreover, they clarified very well, the initial starting point of the next work packages.

2.2 New scalable storage techniques for Grid environments

I/O has been enhanced in **Grid environments** by using transparently parallel file systems, branch-tree replication schemas, user- defined fault-tolerant distribution for files, etc. I/O scheduling techniques have been applied to parallel GridFTP and parallel HTTP data retrieval. A prototype, named GridExpand, has been implemented that includes those features.

At a higher level, we have also developed the HIDDRA tool to enhance massive global data access with data load distribution, multiprotocol channels, and super-peers nodes. It has been also integrated on top of GridExpand. HIDDRA includes new data location and management techniques in grid and Wed systems based on peer to peer features. It´s multiprotocol channels can be open in parallel to retrieve data simultaneously from many servers in a very efficient way, thus including efficient load sharing algorithms. Retrieving data from several servers also allows to provide reliability and fault-tolerance for massive data retrieval, which is far ahead the existing solutions.

Both tools have been tested in ESA ESOC center and systems of Red Española de E-Ciencia.

Relevant results of this work package:

- Publications:
    - Daniel Higuero, Juan M. Tirado, Jesús Carretero, Fernando Félix and Antonio de la Fuente, *HIDDRA: a highly independent data distribution and retrieval architecture for space observation missions*, Journal of Astrophysics and Space Science. Volume 321, 3-4, 2009.
    - José M. Pérez, Félix García-Carballeira, Jesús Carretero, Alejandro Calderón, Javier Fernández, *Branch replication scheme: A new model for data replication in large scale data grids*, 26, 1, January, 2010, Future Generation Computer Systems, Elsevier.
    - Alejandro Calderon, Felix Garcia-Carballeira, Luis Miguel Sanchez, Jose Daniel Garcia, Javier Fernandez Muñoz. *Fault tolerant file models for parallel file systems:*

introducing distribution patterns for every file. Journal of Supercomputing, Volume 47, Number 1, March 2009**.** Springer, ISSN: 0920-8542.

- Technological results:
  - HIDDRA and GridExpand.
  - Technology transfer to European Space Agency and INSA company.

2.3 New scalable storage techniques for Cluster environments

For large scale **clusters and supercomputers**, we have provided new input/output models relying on adaptive caching and prefetching solutions that provides high scalability and performance removing some bottlenecks on massively parallel applications. We have also developed new reliability and fault-tolerance techniques based on adaptive dynamic replication of virtual I/O servers and applying them to large scale systems by creating clusters of virtual servers.

Those solutions have been included into the AHPIOS system, a parallel I/O system that can be deployed on-demand on several cluster configurations, adapting dynamically the number of I/O aggregators and servers required to enhance I/O bandwidth. We heavily rely on collective I/O techniques to do that. By creating replicated aggregators, reliability can be increased. The AHPIOS system has been developed and tested in Cooperation with IBM and PVFS group in Argonne Labs, lead by Prof Rob Ross, and Northwestern university (leaded by Prof. Choudhary), both in Chicago, USA.

In the **multicore/manycore** area, we have provided optimizations in two different research directions: processor and memory intensive applications. Both converge on the Project goals of studying the influence of multicore processors on the input/output mechanisms to study the impact on the bandwidth increasing from each node over the efficiency of the input/output system. We have succeed to enhance the data locality on each processor, which alleviates the communication and memory load due to I/o data distribution, and to design efficient data prefetching for irregular codes. Those techniques has been implemented and tested on PRiSM systems (Versailles-St-Quentin University) in collaboration with Prof. William Jalby.

Relevant results of this work package:

- Publications:
  - Florin Isaila, Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary, *A Scalable Message Passing Interface Implementation of an Ad-Hoc Parallel I/O System*, International Journal of High Performance Computing Applications, October, 2009.
  - Javier García Blas, Florin Isaila, Jesús Carretero, Robert Latham and Robert Ross, *Multiple-level MPI file write-back and prefetching for Blue Gene systems*, 16th EuroPVM/MPI, Finland, September, 2009
  - Florin Isaila, Francisco Javier Garcia Blas, Jesus Carretero, Rob Latham, Sam Lang, Rob Ross, *Latency hiding file I/O for Blue Gene systems*, Nineth IEEE International Symposium on Cluster Computing and the Grid (CCGRID), Shanghai, May, 2009

- Technological results:

- AHPIOS, SIMCAN.
- Technology transfer to ANL.

2.4 Data distribution and organization techniques for performance improvement in Input/Output operations.

We are cooperating with Argonne labs to propose new methods to **optimize** the input/output in massively **parallel applications**, at application or MPI library level. As a result, we have developed new collective I/O techniques based on file views that provide better performance that 2PIO solution.

We have developed optimization techniques to provide data locality aware of the compute node distributions (LARD) that reduce communication overhead due to data exchange, specially for irregular applications, and on-the-fly data compression techniques that reduces message size thus providing more scalability with the same data channels.

All those techniques have been included into MPICH2 distribution, thus they can be applied transparently to every application.

- Publications:
  - Florin Isaila, Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary, *A Scalable Message Passing Interface Implementation of an Ad-Hoc Parallel I/O System*, International Journal of High Performance Computing Applications, October, 2009.
  - Florin Isaila, Walter Tichy. Mapping *Functions and Data Redistribution for Parallel Files*. Journal of Supercomputing, Volume 46, Number 3. December 2008. Springer, ISSN: 0920-8542.
  - Rosa Filgueira, David E. Singh, Alejandro Calderón, and Jesús Carretero, *CoMPI: Enhancing MPI based applications performance and scalability using run-time compression*, Espoo, Finland, January, 2009, Euro PVM/MPI
- Technological results:
  - Enhanced MPICH2 version.
  - Technology transfer to ANL.

2.5 Systems integration, final prototype and evaluation.

The project goals were to integrate the research results in prototypes to test them and to allow technology transfer to companies and research centers. As shown in the next section, we have implemented advanced prototypes for Grid and Cluster I/O systems. The integration and evaluation stages started before the date scheduled in the workplan due to the success of the research work packages.

The integration goal is currently going on. We are making now a second round of optimizations into the prototypes developed, taking into account the result obtained until now. The promising results obtained are leading us to increase the functionality level of the prototypes developed

2.6 Coordination and planning activities.

The global management of the project is focused on the research strategy, the achievements of goals, and the efficient usage of resources. The management and team coordination effort made has been fundamental for the successful scientific, technological, and team cohesion results achieved. Budget for personnel and equipments is being executed without any kind of deviations from the project proposal.

The program of regularly scheduled meetings for the project members to share scientific and technological ideas has been fundamental to achieve a good scientific and technological success. The role of work package coordinators is also being very important for the success of the project.

As a result of the research work done, and the international cooperation, we have more than 40 publications, several grants, etc. Especially important for the project diffusion success have been the strong planning of target journals an international conferences, which has allow to double the number of papers published in three years.

3 Results achieved

Below, we describe several aspects related to the results achieved in the project by the research group ARCOS. More detailed information can be found in TEALES Web page.

Level of achievement of the project goals

In this section, we summarize the results achieved in the project, categorized by the goals defined in the project proposal.

Scientific goals:

- The project had a triple goal: to develop new scalable storage techniques for Grid environments, to propose new storage techniques for cluster environments, and to develop, from the application optimization point of view, techniques to distribute and to organize data to enhance the performance on input/output operations.
- We believe that the objectives set for the project are being accomplished successfully. Actually, goals 1 to 4 described in Section 1 are mostly covered. Goal 5, integration, is now entering into a second exploitation and optimization round, as prototype development was made along the project. The work plan is being covered well in advance.

Scientific and technological contributions:

- Publications. 2 book chapters, 10 journal papers, 28 papers in international conferences, and 3 papers in national conferences. Two more papers are already accepted, but still not published: one in IEEE TPDS and one in IJHPC.
- Events organization. We have organized two international workshops in 2008, one accepted in 2010. Chairs

- Prototypes and tools. We have developed 5 applications directly related to the project research. Three of them are now being testing on preproduction platforms at ESA and ANL.
- Web page. As proposed, the web page of TEALES is the major diffusion center for the project.

Utility of the results for the socio-economic environment:

- Relation with companies: INSA, IBM and ESA.
- Technology transfer: GridExpand (ESA), HIDDRA (ESA), AHPIOS (ANL).

Educational results:

- PhD graduated: 1 in 2009
- PhD students: 9. 4 graduation scheduled in 2010.
- International grants: 1 IBM PhD Fellowship grant, 3 HPC-Europe.
- National grants: 1 FPU grant form MEC, 1 PIF grant from UC3M.

International collaborations:

- International research stays of ARCOS members: 4 USA, 3 EU (all with grants)
- Visiting professors: 5
- Research cooperation: ANL, NWU, IBM Research, PRISM
- Standardization committees: ISO/IEC JTC1/SC22/; AEN/CTN71/
- Technological platforms: NESSI, eMobility

The results obtained until now into the project are described in more detail in the next sections.

Scientific and technological production

We have **published** 2 book chapters, 10 journal papers, 28 papers in international conferences, and 3 papers in national conferences. Two more papers are already accepted, but still not published: one in IEEE TPDS and one in IJHPC. See References section for details.

All the journals are referenced in the computation area of the JCR index, and almost 40% are in the first quarter of their raking. For conferences, some of our results have been presented in top conferences, with acceptance lower than 20%, like EURO PVM/MPI, CCGRID, or IPDPS. Those results show the quality and novelty of them.

Special relevance has the papers:

- Daniel Higuero, Juan M. Tirado, Jesús Carretero, Fernando Félix and Antonio de la Fuente, *HIDDRA: a highly independent data distribution and retrieval architecture for space observation missions*, Journal of Astrophysics and Space Science. Volume 321, 3-4, 2009.
- José M. Pérez, Félix García-Carballeira, Jesús Carretero, Alejandro Calderón, Javier Fernández, *Branch replication scheme: A new model for data replication in large scale data grids*, 26, 1, January, 2010, Future Generation Computer Systems, Elsevier.

- Florin Isaila, Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary, *A Scalable Message Passing Interface Implementation of an Ad-Hoc Parallel I/O System*, International Journal of High Performance Computing Applications, October, 2009-

As a result of the project, ARCOS researchers have got several **grants** for making international research stays. We have got 3 HPC_Europa2 grants, 1 Jose Castillejo program grant, and 4 UC3M Mobility program grants. They are detailed below:

- David Exposito Singh. HPC-Europa2 grant. PRiSM "Research laboratory in computers sciences", Versailles-St-Quentin University. September-decembe r 2009.
- Beca HPCE2. Programa Europeo Supercomputación. Javier Fernández Muñoz. EPCC. Edimburgh. UK.
- Beca HPCE2. Programa Europeo Supercomputación. Alberto Núñez Covarrubias. HRLS. Sttutgart. Germany.
- Beca José Castillejo. Ayuda movilidad MICIN. Flotin Isaila. Argonne National Laboratory - Mathematics and Computer Science Division
- Ayuda programa movilidad UC3M. Laura Prada Camacho Computer Engineering Group. Texas A&M University
- Ayuda programa movilidad UC3M. María Soledad Escolar Díaz Sensor Networks and Pervasive Computing. Universidad de Bonn (Alemania).
- Ayuda programa movilidad UC3M. Francisco Javier García Blas Argonne National Laboratory - Mathematics and Computer Science Division.

One of the engineers hired in the project as technical support, Juan Manuel Tirado Martín, made the Master in Ciencia y Tecnología Informática of Universidad Carlos III de Madrid, and got a 4 years FPU grant (reference AP2007-03530) starting in January 2009.

Special relevance has the IBM Ph.D. Fellowship award received by Alberto Nuñez Covarrubias in 2009. This is an intensely competitive worldwide program, which honors exceptional Ph.D. students.

Events organization. We have organized and contributed to several international research events. Specifically, some ARCOS members have organized the following international workshops related to the project activities:

- SDMAS 2010. SDMAS'10 is an international workshop that will be held within the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications. Las Vegas. USA. Accepted.
- International Workshop on Wireless Sensor Networks Architectures, Simulation and Programming (WASP'09) within MOBILWARE 2009. Berlín, Germany.
- SDMAS 2008. SDMAS'08 is an international workshop held within the 2008 International Conference on Parallel and Distributed Processing Techniques and Applications. Las Vegas. USA.
- HiperIO 2008. Second International Workshop on High Performance I/O Systems and Data Intensive Computing held within IEEE Cluster 2008. Tsukuba. Japan.

Moreover, we have participated as chairs in the following conferences:

- Jesús Carretero. Keynote and Plenary sessions chair of 2010 IEEE symposium on Computers and Communications (ISCC'10). Riccione. Italy. June 2010.
- Jesús Carretero. Industrial chair of the 12th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM), and to beautiful Tenerife, Canary Islands, Spain. October 2009.
- Jose Daniel García. General Chair del 10th International Conference on Algorithms and Architectures for Parallel Processing. Busan, Korea. Del 21 al 23 de mayo de 2010.
- Jose Daniel García. Workshop Chair del 12th IEEE International Conference on Computational Science and Engineering (CSE 2009). Vancouver, Canada. Del 29 al 31 de agosto de 2009.

At the **technological level** we have done several software prototypes, and the first three of them are being publicly shown or transferred to industry:

- **GrideXpand**: Parallel File System for Clusters and Grids. Parallel file system for heterogeneous general purpose distributed environments. To satisfy this goal, we use NFS in clusters, GridFTP on Grids, etc. to have a PFS portable to every system without modifying data servers or protocols. It has been applied at ESOC ESA center in Villafranca del Castillo to access satellite data. We have also adapted it to the FITS standard format to provide high-performance data access to that kind of files.
- **HIDDRA**: Highly Independent Data Distribution and Retrieval Architecture. HIDDRA is composed by two main components: a publish/subscriber and an efficient parallel multiprotocol engine. The product notification system automatically notifies end users about new available products matching their interests. The efficient parallel multiprotocol engine, improves bandwidth consumption reducing the download time by using several protocols at the same time (HTTP, HTTPS, GridFTP). Using HIDDRA the problem of data distribution becomes transparent to end users, who are only interested in getting access to new products as fast as possible. The architecture also offers a finer control of the distribution system, allowing the administrator to optimize its behaviour based on the characteristics of the deployment scenario. HIDDRA is a generic solution that can be easily adapted to a large variety of data distribution scenarios. It has been applied at ESOC and ESRIN ESA centers (Spain and Italy) for parallel massive distribution of satellite data.
- Ad-hoc parallel I/O system (**AHPIOS**) is a light-weight dynamically configurable parallel file system. AHPIOS can be started on-demand on a computing system with distributed storage resources. The approach is useful in several scenarios. First, it can be employed as a middle layer between the applications and an existing parallel file system. The configuration can be done on demand, at application start-up, allowing for the tailoring of the system parameters to the application needs. Second, a system with dynamical availability of resources may employ AHPFS for storage virtualization on-the-fly. Third, different parallel file systems can be virtualized in order to increase the parallelism degree. Fourth, it allows the virtualization and the parallel use of different storage resources in a distributed system, in which no parallel file system is running. It has been included into

ROMIO and tested on Blue-Gene supercomputers. It is the base of our collaboration with Argonne Labs (EE.UU.)

- **SIMCAN**: A Simulator Framework for Computer Architectures and Storage Networks. The main goal of SIMCAN is to simulate large complex storage networks. Moreover, with this simulator, high performance applications can be modeled in large distributed environments. Thus, SIMCAN can be used for evaluating and predicting the impact of high performance applications on overall system performance. By using parallel simulations we are not limited to the resources that a single computer can supply using sequential models. This system has been applied to data systems simulation in the EPCC art Edimburgh University (UK).
- **OSAL**: Operating System Abstraction Layer for WSN. This platform is aimed to challenge the conceptual barrier posed for large WSN applications, by means of abstracting the hardware services in order to facilitate the writing of applications, portability and extend the WSN usage to a community of nonexpert programmers through the usage of OSAL middleware. We have validated our ideas by developing and evaluating an OSAL prototype for several WSN platforms.

Utility of the results for the socio-economic environment

Our research is targeted to optimize applications needing massive data transfer in both parallel and distributed environments. Data I/O optimizations can be applied to enhance many fields of scientific (biomedicine, astronomy, linear algebra, …) and industrial (large scale simulations, finite structures, content delivery networks, …) problems. As a result, we have got direct support from three companies and research labs:

- Ingeniería de Sistemas Aeroespaciales (INSA) is providing support to test and to enhance the HIDDRA platform prototype in the ESOC center of the European Space Agency (ESA). We are currently testing it also with ESRIN ESA center for massive data distribution. We have also testing a version of the EXPAND parallel file system to storage and retrieve massive astronomical data in FITS format.
- IBM Research is providing support for our parallel I/O activities and the application of the AHPIOS system to the Blue-Gene supercomputers storage system. As a result, Alberto Covarrubias has got an IBM Ph.D. Fellowship award. 2009, which is competitive worldwide.
- Argonne National Laboratory - Mathematics and Computer Science Division is also supporting us to include some of the techniques developed into the project to the new PVFS version for supercomputers. We have access to the new Blue-Gene Q system.

We are distributing some of the prototype versions of our systems as **open source software**. Optimizations provided for MPI has been included in MPICH2 distribution, and they will become part of the distribution.

For technology transfer of the results of this project, we have got the following projects with companies:

- Control Automatizado de Procesos Agrícolas (COPA). NETHALIS SOLUTIONS S.L., 2008-2010. Aplicación de tecnologías de captación de datos usando redes de sensores.
- Métodos Avanzados de Distribución de Conjuntos de Datos "Calientes" de Misiones de Observación de la Tierra. INSA SA. 2009-2011. Aplicación de HIDDRA para distribución masiva de datos de satélite.
- Sistema de Diseño y Cálculo de Infraestructuras Ferroviarias. ADIF. 2007-2010. Aplicación de cálculo de estructuras intensiva en cálculo y en datos.

Educational results

Related to the project, we have graduated a PhD thesis in 2009:

- Técnicas de optimización de E/S en sistemas de computación masivamente paralelos. Luis Miguel Sánchez Garcia. Universidad Carlos III de Madrid. Tutores: Jesús Carretero y Félix García.

Before july 2010, we plan to graduate 4 PhD Thesis, all of them result of the cooperation with international research centers, and all of them granted with "European Mention":

- Collective I/O Techniques for Chip Multiprocessor Clusters Rosa Filgueira Vicente.
- A scalable view-based collective I/O optimization for large-scale parallel applications Francisco Javier García Blas.
- A generic software architecture for portable applications in heterogeneous Wireless Sensor Networks María Soledad Escolar Díaz.
- New strategies for characterizing and improving high performance I/O architectures Alberto Núñez Covarrubias.

We have three more PhD students participating into the project on their second PhD year.

For the hired engineers granted to the project, one has got a FPU grant from the Ministry of Education, and other has got a PhD grant from Universidad Carlos III de Madrid. Both of them are starting the PhD studies into the research group. Another one has got a Research Personnel Formation (PIF) grant at Universidad Carlos III de Madrid.

We are also starting cooperation with Universidad del Zulia in Maracaibo (Venezuela) to develop 3 PhD thesis in topics related to the project.

The former results clearly show the educational capacity of the group to form new researchers.

European and international research collaborations

Members of the project have made the following **research stays** in European and USA centers:

- Alberto Nuñez Covarrubias. Internship at IBM Almaden Research Center, Storage Department, from June 22 to September 17 2009. San Jose, California, United States.

- David Expósito Singh. Internship at PRiSM "Research laboratory in computers sciences", Versailles-St-Quentin University, more specifically with the Architecture et Parallélisme (ARPA) group. September to December 2009.
- Florin Isaila. Argonne National Laboratory - Mathematics and Computer Science Division. Chicago. USA. June to December 2008.
- Laura Prada Camacho Computer Engineering Group Texas A&M University. USA.
- María Soledad Escolar Díaz Sensor Networks and Pervasive Computing. Bonn University. Bonn. Germany. March to July 2008.
- Francisco Javier García Blas Argonne National Laboratory - Mathematics and Computer Science Division. Chicago. USA.
- Alberto Núñez Covarrubias. HRLS. Stuttgart. Alemania.

We have also received some **visiting researchers** at Universidad Carlos III along 2008 and 2009.

- Adriana Iamnitchi - University of South Florida. USA.
- Raffaele Montela - Università degli Studi di Napoli "Parthenope". Italy.
- Vlad Olaru. Cluj-Napiuca university. Rumania.
- Prof. Hamid Arabnia. Georgia University. USA.
- María Cristina Marinescu. IBM TJ Watson research Center. USA.

As a result, we have established or reinforced **research collaborations** with several European and US research groups and centers. Among them, we can mention:

- IBM Almaden and IBM TJ Watson Research Center, Storage Department. Prof. Cristina Marinescu, from Watson, is visiting at Carlos III for 2 years.
- Bonn University - Sensor Networks and Pervasive Computing Group. Prof. pedro Marron is coming in February 2010 to impart a research seminar.
- William Jalby. PRiSM "Research laboratory in computers sciences", Versailles-St-Quentin University, more specifically with the Architecture et Parallélisme (ARPA) group. Prof. Jalby is coming in May 2010 to impart a research seminar and to define an exchange program for researchers.
- Argonne National Laboratory - Mathematics and Computer Science Division. Chicago. USA.
- Center for Distributed Systems Research. Northwestern University. Chicago. USA. Prof. Choudhary is coming in November 2010 to impart a research seminar.

ARCOS research group members are also committed with **standardization activities** to the international and national level. Those activities are closely related to the Project activities, as they are mostly developed on the parallelism area.

- International standardization activities.
    - ISO/IEC JTC1/SC22/WG21: The C + + Standards Committee. Since July 2008, Jose Daniel Garcia, represents Spain (through AENOR) as Head of Delegation. This working group is currently performing a thorough review of the C++ programming language with important implications for the development of parallel software.

  - ISO/IEC JTC1/SC22: Programming languages, their environments and system software interfaces. Since October 2009, Jose Daniel Garcia, represents Spain (through AENOR), as Head of Delegation. This subcommittee integrates all working groups within ISO/IEC dealing with programming languages (C, C++, ADA, Fortran, ...), as well as other aspects such as portability vulnerabilities of programming languages or operating systems portable interfaces (POSIX).
- National standardization activities.
  - AEN/CTN71/GT21 - C++ Language. Since September 2009, Jose Daniel Garcia, chairs this national working group in which both industry (IBM, Telefonica I + D, INDRA, BBVA, Goal 4, ...) and academia (University Carlos III, Politecnica de Catalunya, Oviedo and Cadiz) are represented. Jose Daniel Garcia has led the creation of this working group.
  - AEN/CTN71 - Information Technology. Since September 2009, Jose Daniel Garcia, is a member of National Technical Committee of Standardization in Information Technology, which coordinates all subcommittees and working groups in this domain.

ARCOS research group is an active member of two **European Technological platforms**:

- NESSI, the European Technology Platform dedicated to Software and Services. Its name stands for the Networked European Software and Services Initiative.
- eMobility. the European Technology Platform dedicated to mobile and wireless communications and services.

We also cooperate with the Network of Excellence HIPEAC, the High-performance and Emedded Architecture and Compilation, which faces the application of high-performance information systems to European chand challenges.

This project has been very important for ARCOS group, as two major **national level** cooperation links are directly related to the project activities:

- We are strongly cooperating with group ACCA, lead by Prof. Jose Duato from UPV. This research group, including researchers from UPV, UV, UCLM, and UM, is very complementary with ARCOS. Thus, we are making effort to drive both groups towards a stronger integration. Regular meetings are held between both groups.
- Moreover, ARCOS group participates in the "Red Española de eCiencia", a major forum to coordinate and to enhance scientific computation and data repositories in Spain. This scope is giving more social visibility to the results of our project.

Project management

The management of the project is focused on the research strategy, the achievements of goals, and the efficient usage of resources. The management and team coordination effort made has been fundamental for the successful scientific, technological, and team cohesion results achieved.
The program of regularly scheduled meetings for the project members to share scientific and technological ideas has been fundamental to achieve a good scientific and technological success.

Budget for personnel and equipments is being executed without any kind of deviations from the project proposal.

4 References

Book chapters

[1] Javier García Blas, Florin Isaila, Jesús Carretero, *A general parallel I/O architecture for massively parallel supercomputers*, ACACES 2009, Terrasa, Spain, July, 2009, Academia Press, 978 90 382 14, 277-280.

[2] Javier García Blas, Florin Isaila, Jesús Carretero, A view-based approach for collective I/O operations, Transnational Access Meeting 2008, Bologna, Italy, June, 2008.

Journals

[3] Soledad Escolar, Florin Isaila, Alejandro Calderón, Luis Miguel Sánchez, David E. Singh,*SENFIS: a Sensor Node File System for increasing the scalability and reliability of Wireless Sensor Networks applications*, The Journal of SuperComputing, DOI: 10.1007/s11227-009-0275-8

[4] Alberto Núñez, Javier Fernández, Jose D. Garcia, Félix Garcia and Jesús Carretero, *New techniques for simulating high performance MPI applications on large storage networks*, The Journal of SuperComputing, DOI: 10.1007/s11227-009-0279-4

[5] José M. Pérez, Félix García-Carballeira, Jesús Carretero, Alejandro Calderón, Javier Fernández, *Branch replication scheme: A new model for data replication in large scale data grids*, 26, 1, January, 2010, Future Generation Computer Systems, Elsevier, 0167-739X, 12-20,

[6] Daniel Higuero, Juan M. Tirado, Jesús Carretero, Fernando Félix and Antonio de la Fuente, *HIDDRA: a highly independent data distribution and retrieval architecture for space observation missions*, Journal of Astrophysics and Space Science, Volume 321, 3-4, 2009, Springer Netherlands, 0004-640X, 169-175,

[7] Florin Isaila, Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary, *A Scalable Message Passing Interface Implementation of an Ad-Hoc Parallel I/O System*, International Journal of High Performance Computing Applications, October, 2009

[8] Javier Garcia Blas, Florin Isaila, Jesus Carretero, David E. Singh, Felix Garcia, *Write-back and prefetching in an MPI-IO implementation for GPFS*. International Journal of High Performance Computing Applications, January, 2009,

[9] Alejandro Calderon, Felix Garcia-Carballeira, Luis Miguel Sanchez, Jose Daniel Garcia, Javier Fernandez Muñoz. *Fault tolerant file models for parallel file systems: introducing distribution patterns for every file*. Journal of Supercomputing, Volume 47, Number 1, March 2009. Springer, ISSN: 0920-8542.

[10] David E. Singh, Florin Isaila, Juan Carlos Pichel and Jesús Carretero. *A collective I/O implementation based on Inspector-Executor paradigm*. Journal of Supercomputing, Volume 47, Number 1, March 2009. Springer, ISSN: 0920-8542.

[11] Florin Isaila, Walter Tichy. Mapping *Functions and Data Redistribution for Parallel Files*. Journal of Supercomputing, Volume 46, Number 3. December 2008. Springer, ISSN: 0920-8542.

[12] David E. Singh, Alejandro Miguel, Félix García and Jesús Carretero. *Mobile Agent Systems Integration into Parallel Environments*. Scalable Computing: Practice and Experience (SCPE), 2008. Nova Science Publishers, ISSN: 1895-1767.

Conferences.

[13] Juan A. Lorenzo, Juan C. Pichel, David LaFrance-Linden, Francisco F. Rivera and David E. Singh*, Lessons Learnt Porting Parallelisation Techniques for Irregular Codes to NUMA Systems,* Pisa, Italy, January, 2010, The 18th Euromicro International Conference on Parallel, Distributed and Network-Based Computing.

[14] Borja Bergua-Guerra, Félix García-Carballeira, Luis Miguel Sánchez, Alejandro Calderón, Alejandra Rodríguez, Jesús Carretero, *Architecture for improving data transfers in Grid using the Expand parallel file system,* 3rd Iberian Grid Infraestructure Conference (IBERGRID'2009), May 20-22, Valencia, Spain, May, 2009

[15] Javier García Blas, Florin Isaila, Jesús Carretero, Robert Latham and Robert Ross, *Multiple-level MPI file write-back and prefetching for Blue Gene systems*, 16th EuroPVM/MPI, Finland, September, 2009

[16] Javier García Blas, Florin Isaila and Jesús Carretero, *A General Parallel I/O Architecture for Clusters and Supercomputers*, IEEE International Parallel & Distributed Processing Symposium (IPDPS), TCPP PhD Forum, Rome, Italy, May, 2009

[17] Florin Isaila, Francisco Javier Garcia Blas, Jesus Carretero, Rob Latham, Sam Lang, Rob Ross, *Latency hiding file I/O for Blue Gene systems*, Nineth IEEE International Symposium on Cluster Computing and the Grid (CCGRID), Shanghai, May, 2009

[18] Javier García Blas, Florin Isaila y Jesús Carretero, *Arquitectura de E/S paralela de alta prestaciones para sistemas Blue Gene*, XX Jornadas de Paralelismo, A Coruña, Spain, September, 2009

[19] Laura Prada, Jose Daniel Garcia, Jesus Carretero, and Felix Garcia, *Saving power in flash and disk hybrid storage system*, MASCOTS 2009, London, England, September, 2009, Proceedings of the 17th Annual Meeting of the IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems, 978-1-4244-49, 632-634

[20] Laura Prada, Jose Daniel Garcia, Jesus Carretero, and Felix Garcia, *Ahorro energético en un sistema de almacenamiento híbrido compuesto por un disco duro y varias memorias flash*, La Coruña, Spain, September, 2009, Actas de las XX Jornadas de Paralelismo, 84-9749-346-8, 259-264

[21] Alejandra Rodriguez, Jesus Carretero, Borja Bergua, Felix Garcia Carballeira, *Resource Selection for Fast Large-Scale Virtual Appliances Propagation*, 14th IEEE Symposium on Computers and Communications Program (ISCC'09), July 5-8, Sousse, Tunisia, July, 2009

[22] Alejandra Rodriguez, Jesus Carretero, Alberto Nunez, Borja Bergua, Felix Garcia, Jose-Daniel Garcia, *An Efficient Deployment Strategy for Large Sets of Virtual Appliances*, International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'09), July 13-16, Las Vegas, Nevada, USA, July, 2009

[23] Borja Bergua, Félix García-Carballeira, Alejandro Calderón, Luis Miguel Sánchez, Jesús Carretero, *Improving the performance of the BOINC volunteer computing platform using the Expand parallel file system*, 5th IEEE International Conference on e-Science (e-Science 2009), December 9-11, Oxford, United Kingdom, December, 2009.

[24] Rosa Filgueira, David E. Singh, Alejandro Calderón, and Jesús Carretero, *CoMPI: Enhancing MPI based applications performance and scalability using run-time compression*, Espoo, Finland, January, 2009, Euro PVM/MPI.

[25] B. Bergua, F. Garcia, A. Calderón, L. M. Sánchez, J. Carretero, *Comparing grid data transfer technologies in the Expand parallel file system,* 16th Euromicro International Conference on Parallel, Distributed and network-based Processing, PDP-2008. 13-15 de febrero, Toulouse, Francia. , 2008.

[26] Alberto Núñez, Javier Fernández, Jesús Carretero, J. D. García and Laura Prada, *New Techniques for Modelling File Data Distribution on Storage Nodes*, 41th Annual Simulation Symposium, Ottawa, Canada. Abril, 2008. ANSS 2008Págs. 8 pages.

[27] Alberto Núñez, Javier Fernández, Jesús Carretero, Jose D. García, and Laura Prada, *SIMCAN: A Simulator Framework for Computer Architectures and Storage Networks*, First International Conference on Simulation Tools and Techniques for Communications, Networks and Systems, Marseille, France. March, 2008. SIMUTools 2008 ISBN: 978-963-9799-. Págs. 8 pages.

[28] Alberto Núñez, Javier Fernández, Jose D. García, Laura Prada, Jesús Carretero, *M-PLAT: Multi-Programming Language Adaptive Tutor*, The 8th IEEE International Conference on Advanced Learning Technologies, Santander, Spain. Julio, 2008. ICALT 08Págs. 3 pages.

[29] Jose Daniel Garcia, Laura Prada, Javier Fernandez, Jesus Carretero, Alberto Nunez, *Using black-box modeling techniques for modern disk drives service time simulation.* The 41th Annual Simulation Symposium (ANSS'08), Abril, 2008. Proccedings of the 41th Annual Simulation Symposium

[30] Soledad Escolar, Jesús Carretero, Florin Isaila and Giacomo Tartari, *A MDA-based development framework for sensor networks applications*, 4th IEEE International Conference on Distributed Computing on Sensor Systems, Santorini Island, Greece. Junio, 2008. ISSN: 0302-9743.

[31] Soledad Escolar, Jesús Carretero, Florin Isaila, Stefano Lama, *A lightweight storage system for sensor nodes*, Int. Conference on Parallel and Distributed Processing Tecniques and Application PDPTA'08, Vol. Volume II, Las Vegas, Nevada, USA. Julio, 2008. ISBN: 1-60132-082-5. Págs. 638-644.

[32] Juan C. Pichel, David E. Singh and Jesús Carretero, *Reordering Algorithms for Increasing Locality on Multicore Processors*, 10th IEEE Int. Conference on High Performance Computing and Communications (HPCC), Dalian, China. Septiembre, 2008.

[33] Rosa Filgueira, David E. Singh, Juan C. Pichel, Jesús Carretero. *Exploiting Data Compression in Collective I/O Techniques*. IEEE International Conference on Cluster Computing, HiperIO Workshop (CLUSTER), Tsukuba, Japan. September, 2008.

[34] Alejandra Rodríguez, Javier Fernández, Jesús Carretero, *Model for on-demand virtual computing architectures – OVCA*, IEEE Symposium on Computers and Communications, 2008. ISCC, Marrakech, Moroco. Septiembre, 2008. ISBN: 978-1-4244-27. ISSN: 1530-1346. Págs. 447 - 454.

[35] Alberto Núñez, Javier Fernández, Jose D. García and Jesús Carretero, *Analyzing Scalable High-Performance I/O Architectures*, PDPTA'08, The 2008 International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, Nevada (USA). Julio, 2008. Proceedings of The 2008 International Conference on Parallel and Distributed Processing Techniques and Applications ISBN: 1-60132-082-5. Págs. 631-637.

[36] Alberto Núñez, Javier Fernández, Jose D. García and Jesús Carretero, *New techniques for simulating high performance MPI applications on large storage networks*, HiperIO'08. The Second International Workshop on High Performance I/O Systems and Data Intensive Computing held within IEEE Cluster 2008, Tsukuba, Japan. October, 2008. Proceedings of the 2008 IEEE International Conference on Cluster Computing.Págs. 1-9.

[37] Rosa Filgueira, David E. Singh, Juan C. Pichel, Florin Isaila and Jesús Carretero, *Data locality aware strategy for Two-Phase Collective I/O*, International Meeting High Performance Computing for Computational Science (VECPAR), Toulouse, France. , 2008. Lecture Notes in Computer Science.Springer-Verlag.

[38] Javier García Blas, Florin Isaila, David E. Singh and Jesús Carretero, *View-based collective I/O for MPI-IO*, IEEE International Symposium on Cluster Computing and the Grid (CCGRID), Lyon, France, 2008.

[39] Javier García Blas, Florin Isaila, Jesús Carretero and Thomas Grossemann, *Implementation and evaluation of an MPI-IO interface for GPFS in ROMIO*, The 15th Euro PVM/MPI 2008 conference, Dublin, Ireland. Septiembre, 2008.

[40] Javier García Blas, Florin Isaila, Jesús Carretero, *WiP-FAST08 View-based collective I/O for MPI-IO*, 6th USENIX Conference on File and Storage Technologies (FAST '08), San Jose, California, EEUU. , 2008.

[41] Javier García Blas, Florin Isaila, Jesús Carretero, *Implementación y evaluación de una interfaz para GPFS en ROMIO*, Jornadas de paralelismo de Castellon, Castellón, España. Septiembre, 2008.

[42] Florin Isaila, Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, Alok Choudhary, *AHPIOS: An MPI-based ad-hoc parallel I/O system*, 14th Intl Conference on Parallel and Distributed Systems, Melbourne, AUSTRALIA. , 2008.

[43] David E. Singh, Alejandro Miguel, Félix García and Jesús Carretero, *MASIPE: A tool based on mobile agents for monitoring parallel environments*, International Conference on Parallel Processing and Applied Mathematics (PPAM'07), Poland, 0, 2007, Lecture Notes in Computer Science., Springer--Verlag Editorial., conference

[44] Alejandro Calderón, Félix García, Florin Isaila, Rainer Keller, Alexander Schulz, *Fault tolerant file models for MPI-IO parallel file systems*, EuroPVM/MPI 2007, 4757/2007, París, Francia, September, 2007, 978-3-540-754, 0302-9743, 153-160, Lecture Notes in Computer Science.

Web

[45] TEALES. http://arcos.inf.uc3m.es/~teales/doku.php

Performance analysis of high performance application on large storage networks is a very complex and time-consuming task. However, modelling and studying the behaviour of any application on complex network architectures is crucial to obtain good performance. The goal of this work is to predict both scalability degree and performance of any high computing appli-cations on any network architecture. A very interesting feature of this work is that our approach does not require to modify the application in order to simulate its behaviour.Â This work has been partially funded by project TIN2007-63092 of Spanish Ministry of Education and project CCG07-UC3M/TIC-3277 of Madrid State Governement. REFERENCES. New techniques for simulating high performance MPI applications on large storage networks, Alberto NÃºÃ±ez, Javier FernÃ¡ndez, Jose D. GarcÃa and JesÃºs Carretero, HiperIOâ€™08. The Second International Workshop on High Performance I/O Systems and Data Intensive Computing held within IEEE Cluster 2008, Tsukuba, Japan, October, 2008, Proceedings of the 2008 IEEE International Conference on Cluster Computing., 1-9, conference. Data locality aware strategy for Two-Phase Collective I/O, Rosa Filgueira, David E. Singh, Juan C. Pichel, Florin Isaila and JesÃºs Carretero, International Meeting High Performa... Developing High-Performance, Scalable, cost effective storage solutions with IntelÂ®Cloud Edition Lustre* and Amazon Web Services. Authors: Gabriele Paciucci Intel High Performance Data Division Solution Architect. Steve Paper Intel Technical Account Manager. Ian Meyers Amazon Principal Solution Architect.Â Designed specifically for high performance computing, the open source Lustre parallel file system is one of the most popular, powerful and scalable data storage systems currently available, and is in widespread use today in super-computing scenarios where high performance and enormous storage capacity is required. 60% of the Top 100 clusters in the world1 are currently running Lustre. Executive Summary. High Performance Computing (HPC) as a technology is no longer just a researchersâ€™ tool; now more and more companies are discovering the compet-itive advantages of HPC for their own business models. They generate huge volumes of data and use high-perfor-mance data processing applications to analyze and derive value from their data flows.Â RAIDIX consists of software services designed to create high-performance storage systems using widely used Intel processor based hardware platforms.Â High-Availability & Scalable Cluster-In-The-Box HPC Storage Solution: Using RAIDIXÂ® Storage Software Integrated with IntelÂ® Enterprise Edition for Lustre*. 4. Figure 1 - Recommended Deployment Scheme for Typical HPC Application.