



CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Integration of Homogenous Sites and Improvements in
Stability of MC-GPFS for Production

Deliverable ID: DEISA-SA2-3A
Due date: October, 31st, 2005
Actual delivery date: November 25, 2005
Lead contractor for this deliverable: RZG, Germany

Project start date: May 1st, 2004
Duration: 4 years

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	X
RE	Restricted to a group specified by the consortium (including the Commission	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of Contents

Table of Contents.....	1
List of Figures.....	1
List of Tables.....	1
1. Introduction.....	2
1.1 Executive Summary.....	2
1.2 References and Applicable Documents	2
1.3 Document Amendment Procedure	2
1.4 List of Acronyms and Abbreviations	2
2. Overview on Multi-Cluster GPFS for DEISA.....	4
3. Preparing the Integration of the Homogeneous Production Systems.....	4
3.1 General Scalability and Reliability Tests	4
3.2 Test Configuration at the DEISA EU Review Meeting.....	6
4. Integrating the Production Systems of the Core Sites	6
4.1 Tests and Integration of the CINECA Production System	6
4.2 Tests and Integration of the FZJ Production System	7
4.3 Results and Experience with the Core Sites	7
4.4 Configuration of the Core Sites	7
5. Integration of CSC	10
6. Performance, Stability and Availability.....	10
7. Work for Future Configurations and Roadmap	11
7.1 Preparation for Future Integration of Heterogeneous Sites.....	11
7.2 Future Configuration of the MC-GPFS throughout DEISA	12
7.3 Preparation of a Showcase for the Supercomputing SC05 in Seattle	12

List of Figures

Figure 1: Communication paths for GPFS	5
Figure 2: Configuration for Testing as Used during Review Meeting in Paris.....	6
Figure 3: CINECA Production Configuration.....	8
Figure 4: FZJ Production Configuration	9
Figure 5: IDRIS Production Configuration.....	9
Figure 6: RZG Production Configuration.....	10
Figure 7: GPFS provision throughout DEISA.....	12

List of Tables

Table 1: Current hardware configuration of the core sites providing MC-GPFS.....	8
--	---

1. Introduction

1.1 *Executive Summary*

One of the main objectives of DEISA SA2-TB1 is to provide a Global File System, namely the new Multi-Cluster version of GPFS (General Parallel File System), on all the AIX-computers participating in DEISA. This document, "Integration of Homogenous Sites and Improvements in stability and Performance of Multiple-Cluster-GPFS", is the third SA2 deliverable, describing the availability and the performance of the MC-GPFS among the four "core" sites (IDRIS, RZG, CINECA and FZJ) and the integration of CSC as another IBM site.

After a short overview this document documents the full integration of the two core sites CINECA and FZJ into the DEISA wide MC-GPFS. The complete configuration of the four core sites IDRIS, CINECA, FZJ and RZG is given, too. A short section describes the integration of the homogeneous (IBM-AIX) system of CSC. Availability and stability, as well as a short indication of the performance are the content of the next section, while more details about performance issues can be found in the accompanying deliverable D-SA2-3B [3]. In the last section work in progress and some future aspects are highlighted.

1.2 *References and Applicable Documents*

- [1] DEISA home-page: <http://www.deisa.org/>
- [2] Deliverable D-SA2-2A
- [3] Deliverable D-SA2-3B
- [4] Acronyms and Abbreviations:
<http://cgi.snafu.de/ohei/user-cgi-bin/veramain-e.cgi>
- [5] DEISA User Guide <http://www.deisa.org/userscorner/>

1.3 *Document Amendment Procedure*

The initial document amendment procedure is via communication between members of DEISA SA2 team. The document is then submitted for review to the DEISA Executive and an Executive appointed DEISA reviewer. The document is then amended according to comments received from the Executive and the DEISA appointed reviewer. It is subsequently re-submitted to the DEISA Executive for submission to the EU.

1.4 *List of Acronyms and Abbreviations*

AIX	Advanced Interactive eXecutive (IBM's derivative of UNIX OS)
ATA	Advanced Technology Adapter (Hard Drive Technology)
CPU	Computing Processor Unit
CRPP	Centre de Recherches en Physique des Plasmas

DEC	DEISA Executive Committee
FC	Fibre Channel (disk-connection protocol)
GA	General Availability
GID	Group IDentification (UNIX Group)
GPFS	General Parallel File System
HPC	High Performance Computing
HPS	High-Performance Switch (Fast Interconnect for IBM-Computers) IBM Official name for the Federation Switch
HW	Hardware
IBM	International Business Machines (Computer Manufacturer)
IA32	Intel 32Bit processor architecture (also known as x86)
IA64	Intel 64Bit processor architecture (also known as Itanium)
I/O	Input/Output
IP	Internet Protocol
IPP	Max-Planck Institut für Plasma-Physik (hosting RZG).
LAN	Local Area Network
Linux	Free UNIX-like Operating System
LPAR	Logical Partition (subset of a larger system)
MC-GPFS	Multi-Cluster GPFS
ML	Maintenance Level
NSD	Network Shared Disk, a component of GPFS
ORB	Simulation Code for Global Turbulence
OS	Operating System
P655, P690	High performance computing nodes built by IBM
RAID	Redundant Array of Independent (Inexpensive) Disks
RTT	Round Trip Time

SAN	Storage Area Network
SATA	Serial Attached ATA
SW	Software
TCP	Transmission Control Protocol
UID	User IDentity (UNIX User)
UNIX	An Operating System
VSD	Virtual Shared Disk, a component of GPFS
WAN	Wide Area Network

2. Overview on Multi-Cluster GPFS for DEISA

The four DEISA core sites operate IBM HPC systems and share a Multi-Cluster General Parallel File System (MC-GPFS) to provide a Grid File System. In DEISA this Grid file system provides transparent access to data just like with a local file system. The availability and first performance test of the MC-GPFS between two core sites (IDRIS and RZG) were described in the last deliverable D-SA2-2A [2] along with the details of the principal configuration and interaction with the underlying network infrastructure. This previous deliverable also discussed performance issues.

In this latest deliverable the integration of the other two core sites (CINECA and FZJ) is documented along with the final configuration details across all four core sites. The switching of the system into production use across all four core sites is also documented. The system was proven using a real application and simple copy tests during the EU review meeting in Paris in June 2005.

DEISA now therefore offers a distributed European wide file system spanning the four core sites and CSC that are currently connected by the 1Gbit/s DEISA network. It offers a total of almost 12TB of disk space for DEISA and 3720 Power4/5 processors with a total of nearly 12 TB of memory for running grand challenge applications. The shared memory units of either 8 or 32 CPUs are connected with the High Performance "Federation" Switch at each site.

3. Preparing the Integration of the Homogeneous Production Systems

3.1 General Scalability and Reliability Tests

With the integration of the production systems from the two core sites, CINECA and FZJ, almost 300 nodes, consisting of almost 4000 CPUs, are now connected together with the MC-GPFS. The concept of MC-GPFS requires the ability for each pair of nodes to be able to directly communicate. The performance of GPFS is mainly achieved by the fact that much of the communication is delegated to the "clients". The servers are primarily

responsible for the disk-I/O, while the clients themselves are responsible for e.g. locking files. Thus it is important to have an extremely stable underlying network infrastructure, which is used by all nodes for their communication.

With an ever increasing number of nodes and sites participating in the GPFS it may happen, that there are problems in the client-client communication while the client-server communication is still functioning. To overcome these situations, there should be some routing features integrated into the servers. An example of the principal communication is shown below in Figure 1.

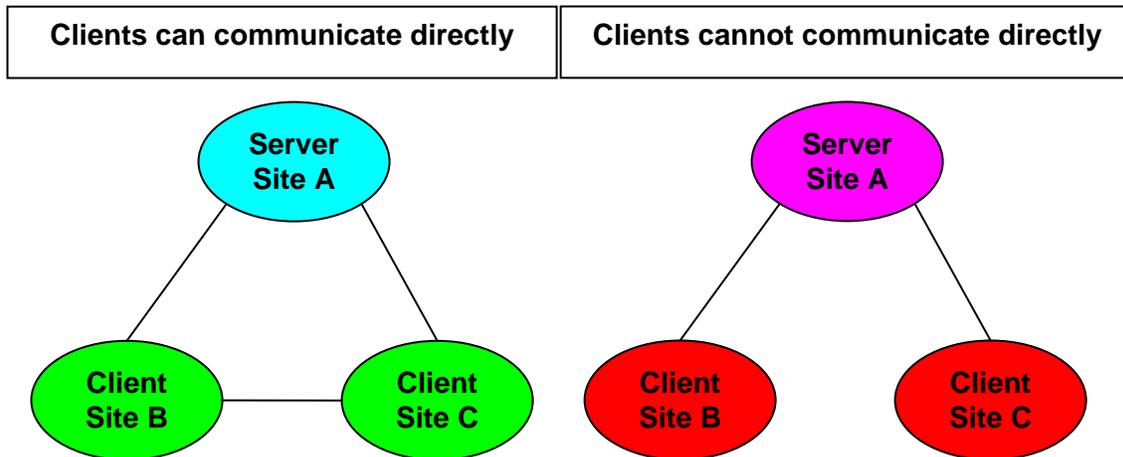


Figure 1: Communication paths for GPFS

Currently, after some initial meta-data communication between a client and the server, the clients accessing overlapping parts of the file system communicate directly for e.g. locking information during access of a directory. In the case of an interruption of the direct connection between clients, both clients can still send requests to the server but clearly if the server asks the clients to coordinate themselves then this is not possible.

In a first version of the GPFS software this led to a flip-flop like mounting/un-mounting of the GPFS on the clients. In a later version one of the clients lost the file system until the direct communication was possible again. However it was unclear, which of the sites would lose the file system.

Since this behaviour is on a per client basis, it is possible that some clients on a site can mount the file system, while others on the same site cannot. In such a case of interrupted client communication, it would be a major improvement, if the server would route the communication between the clients, so that both sites can see the file system. Although this routed communication is less efficient than direct communication, it would offer a significant gain in availability. Due to this behaviour a request has already been lodged with IBM for a future version of GPFS to include such features. In addition, the DEISA partner BSC, who has a system with thousands of clients, requires similar functionality even for the local GPFS.

Despite the lack of these features in the current release, the standard scenarios one can imagine are all properly handled by the MC-GPFS. Below is a list of the tests, along with their outcomes, undertaken to investigate these scenarios.

- Increasing the number of nodes showed no effect on stability or performance
- Reboots of clients have no effect on the total GPFS
- Network drops are recovered after the network is back again
- Fixes and updated versions for GPFS can be installed on a per site basis
- Reboots of I/O-servers are handled without interruption, if the servers are configured redundantly
- Outages of single sites do not affect the other parts of GPFS

3.2 Test Configuration at the DEISA EU Review Meeting

During the DEISA EU review meeting in Paris in June 2005, the following scenario illustrated in Figure 2 was showcased.

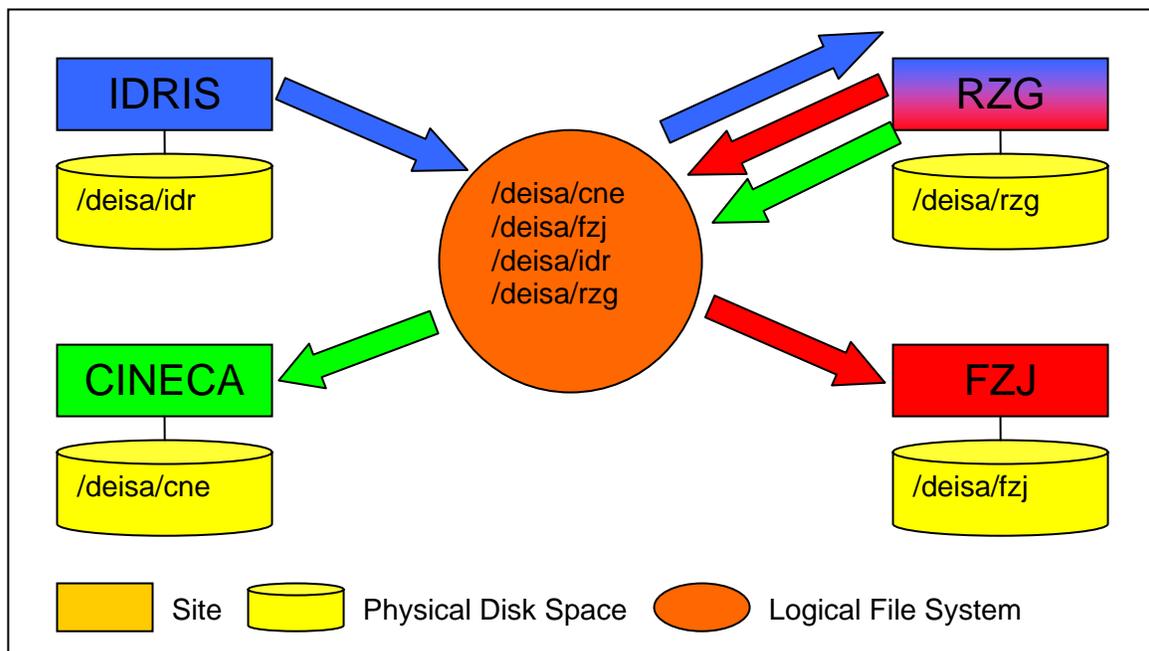


Figure 2: Test configuration used during Review Meeting in Paris in June 2005

- (1) Gigabytes of data were read from /deisa/idr into the RZG computing system
- (2) Some computation was performed at RZG with the code T-ORB
- (3) Gigabytes of restart-data were written from RZG to /deisa/fzj
- (4) Result data was written to /deisa/rzg
- (5) The newly calculated data from /data/rzg was read and displayed with a program running at CINECA

4. Integrating the Production Systems of the Core Sites

4.1 Tests and Integration of the CINECA Production System

After the successful set-up of the MC-GPFS between the production systems of IDRIS and RZG the next production system to integrate was CINECA. As stated in the last

deliverable the CINECA system underwent a complete hardware upgrade before it reached its final production state. This meant a completely new hardware configuration, now based on Power5 processors, had to be integrated.

First GPFS was set-up locally in CINECA after the upgrade by IBM. It was seen that this version of GPFS was fully compatible with the new hardware. Since the installed version of this software already contained the MC-GPFS required for DEISA, CINECA could be integrated straight forward. With help from IDRIS and RZG but without help from IBM, due to the contract, the production system of CINECA was integrated by exchanging the site public keys for MC-GPFS. The integration required no stop or reboot, nor any unmount of any already existing MC-GPFS. The CINECA file system could be mounted immediately after the keys had been exchanged and implemented at the other sites. Also CINECA could mount the MC-GPFS file systems of the production systems of IDRIS and RZG.

4.2 Tests and Integration of the FZJ Production System

The MC-GPFS was available at FZJ almost from the very beginning of DEISA, but only on a test system. This file system, however, was already used at the other sites in the production environment. For example, it was used as part of the showcase during the DEISA EU Review Meeting in Paris in June 2005.

However as stated in the last deliverable, the FZJ production system could not be fully integrated into the DEISA file system structure, since an important piece of software that FZJ relies on, was not available for the OS-level required for the MC-GPFS. Once this additional software became available, FZJ could install the MC-GPFS software on its production systems, as well as provide a MC-GPFS to the others and mount the MC-GPFS file systems from the others after the creation, exchange and implementation of the keys required for secure communication.

4.3 Results and Experience with the Core Sites

By end of September 2005 all four core sites offer a MC-GPFS, which is mounted on the production systems of all these sites. This now provides a transparent global file system spanning over the four core sites. The configurations of each these sites can be seen below in section 4.4 of this deliverable.

The four core sites IDRIS, CINECA, FZJ and RZG now offer a European wide file system, where the data can be accessed from the production systems of each site transparently, thus offering an easy way to run an application on each of the core sites. The availability of this MC-GPFS file system is part of the test regularly run by SA3 when testing the status of the environments of all DEISA core sites.

4.4 Configuration of the Core Sites

Each of the four core sites has a locally configured MC-GPFS, which is exported to the other core sites. Thus, strictly speaking DEISA does not use a single MC-GPFS over all core sites. Instead a number of MC-GPFS's are interwoven in such a way that they appear to be a single, shared file system. They are mounted as `"/deisa/<site>`". Below

this the MC-GPFS provides locally a “home” and a “data” either in one file system or two separate ones. For redundancy all sites distribute the file system on two servers. During job start up the user can access his data easily using environment variables pointing to the unique location in the common MC-GPFS (details in the User Guide [5]).

The current situation is summarized in Table 1. Only the resources available to the end users, including login nodes, are shown, while file and backup servers are not considered.

Site	Fileservers	Storage	Compute-CPU's	Memory
CINECA	2	1.1 TB	480 Power5 (1.9 GHz)	1152 GB
FZJ	2	2.2 TB	1288 Power4 (1.7 GHz)	5152 GB
IDRIS	2	1.4 TB	1024 Power4 (1.3 GHz)	3136 GB
RZG	2	7.0 TB	928 Power4 (1.3 GHz)	2368 GB

Table 1: Current hardware configuration of the core sites providing MC-GPFS

The configurations of the production systems of the core sites, CINECA, FZJ, IDRIS and RZG, are displayed in Figure 3, Figure 4, Figure 5, and Figure 6 respectively. All systems including I/O and backup systems are shown. As one can see, all the configurations are different with respect to the size of a single system and the maximum memory available per processor. Thus it is very likely, that an application will find the best fitting configuration for optimal computational performance in DEISA. This is an enormous benefit compared to the possibilities offered by a site acting alone.

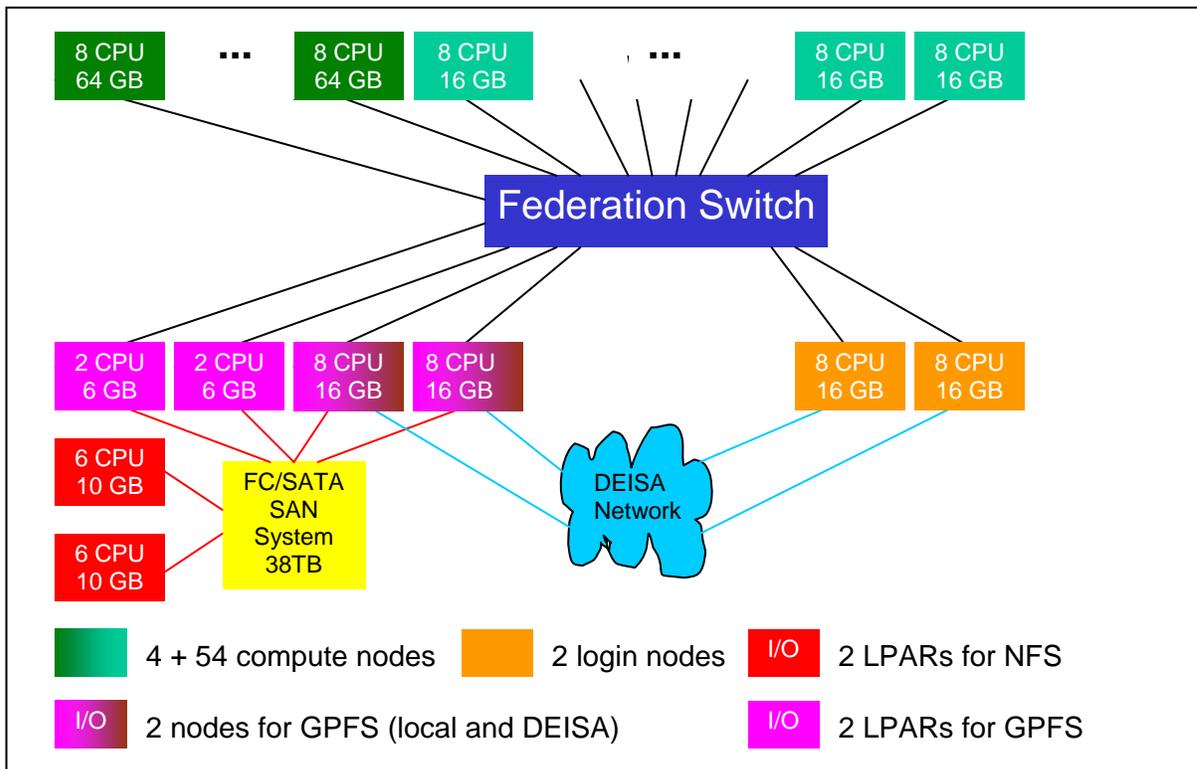


Figure 3: CINECA Production Configuration

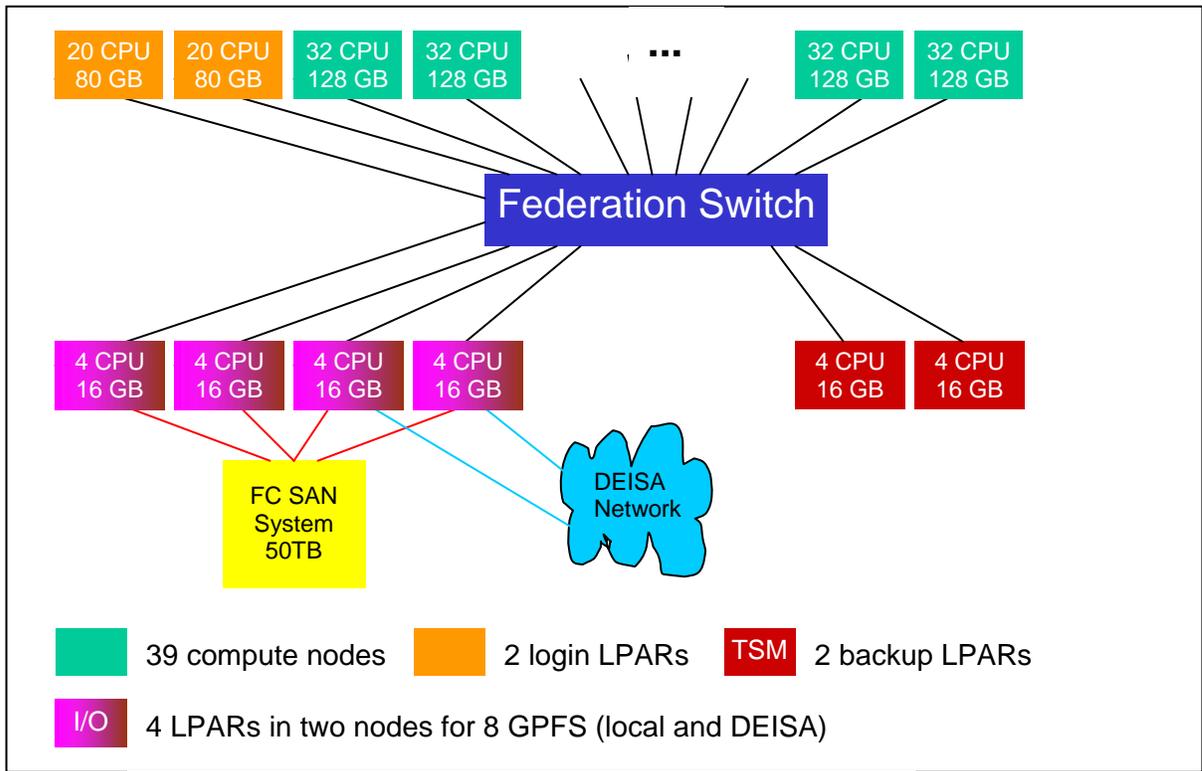


Figure 4: FZJ Production Configuration

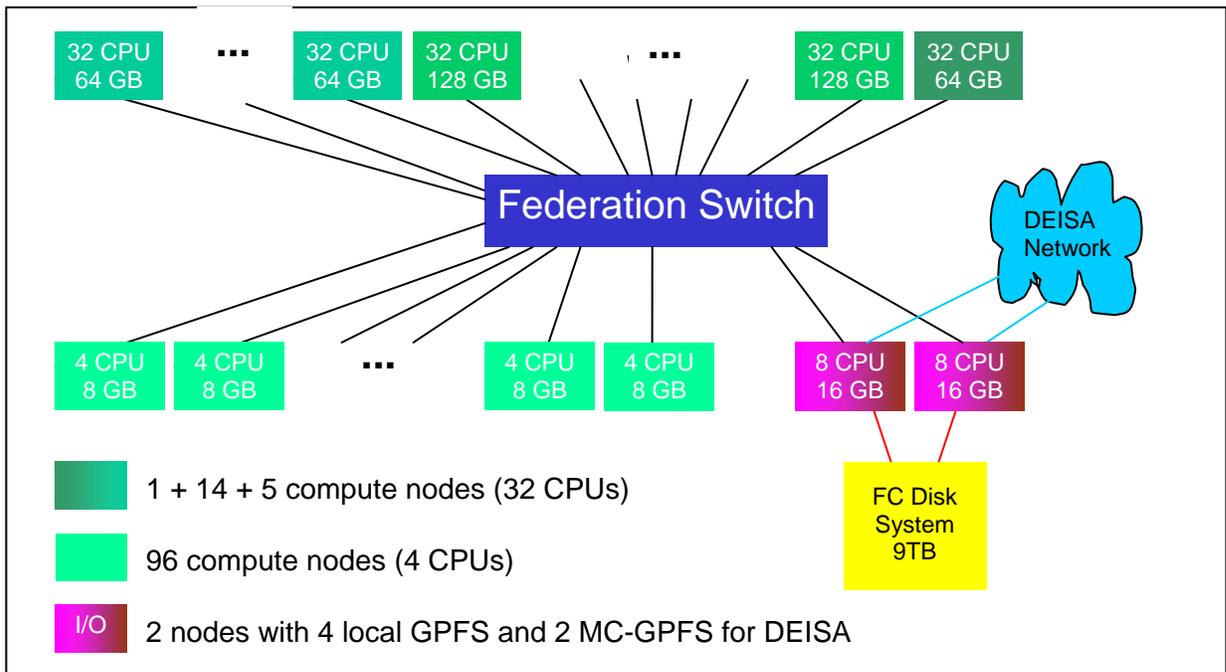


Figure 5: IDRIS Production Configuration

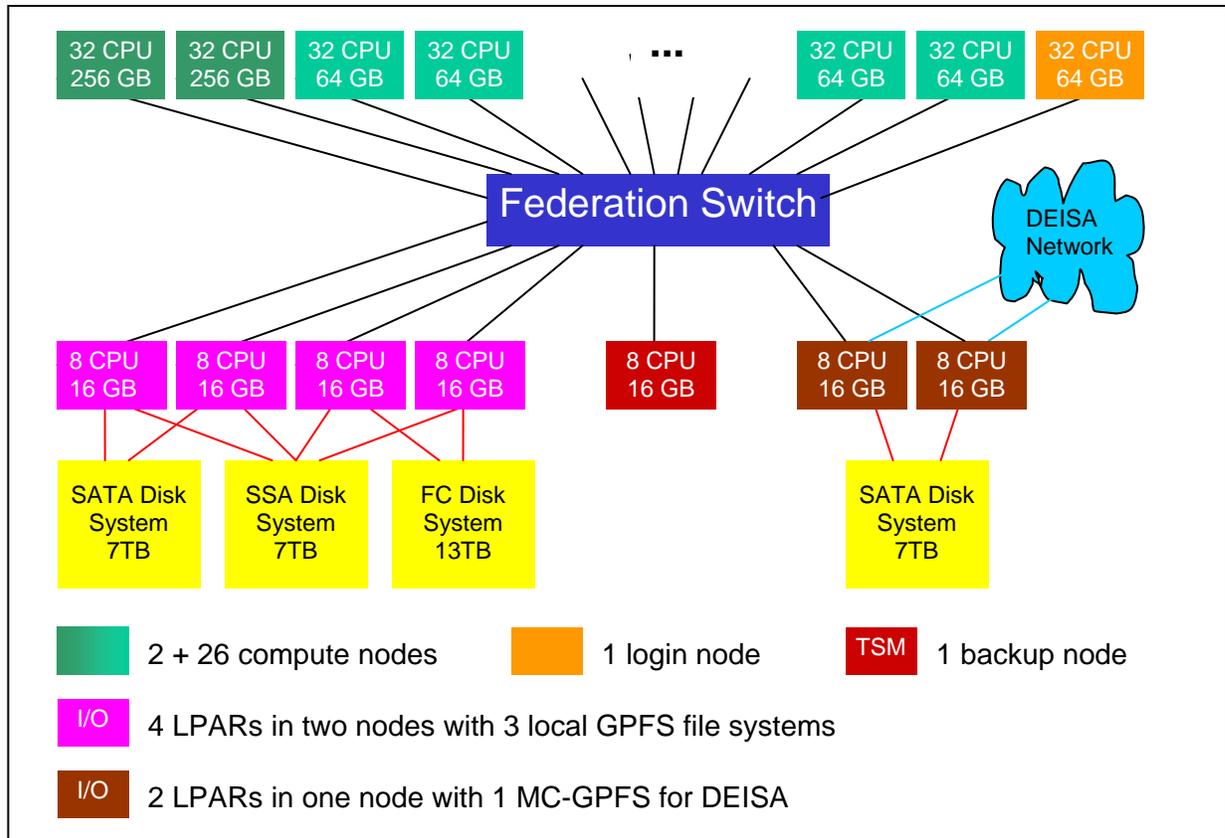


Figure 6: RZG Production Configuration

5. Integration of CSC

The DEISA partner, CSC, runs the same large IBM-AIX systems and hardware as the four core sites. The integration of CSC's system is thus straight forward. However, it still requires some major reconfiguration, since the CSC system was not originally set-up with a later integration into the DEISA grid in mind. The CSC system is small compared to the others and so CSC will not provide their own disk server but rather will use the disk space of RZG. This means that CSC will mainly act as a client, mounting all the other MC-GPFS file systems.

6. Performance, Stability and Availability

During the DEISA EU review meeting in Paris in June 2005 the showcase (see above in section 3.) proved using a simple copy test and a real application program, that MC-GPFS could saturate almost fully the 1 Gbit/s bandwidth of the network using the production systems of IDRIS and RZG and the test systems of FZJ and CINECA. A copy from the local part of a GPFS to a remotely located GPFS, e.g. from RZG to IDRIS, filled the bandwidth by about 80%. The real application, T-ORB, a plasma-physics code,

which was running on 256 processors in RZG, reading data from IDRIS and writing the restart file to FZJ, filled the bandwidth by about 60%.

Since the MC-GPFS is now an integrated part of the production systems the availability is as good as if not better than that of the production systems. This is because the disk servers are often independent from the compute nodes and thus can provide the MC-GPFS to the remote sites even if the local compute nodes are down due to e.g. maintenance.

Despite interruptions to network connections had some interruptions, the restarting of machines at individual sites and even complete maintenance at some sites, the GPFS in the production environments of IDRIS and RZG has remained stable and had no side effects or any other negative influences. This proves the stability and availability of the MC-GPFS as an outstanding distributed file system.

The status of the MC-GPFS is continuously monitored by a suite of software tools, which show high availability and good performance (see deliverable DEISA-SA2-3B [2]). There have been no outages due to software problems. Only maintenance of the single systems resulted in the non-availability of parts of the MC-GPFS. This confirms the appropriateness of the global file system concept for the coupling of HPC sites.

7. Work for Future Configurations and Roadmap

7.1 Preparation for Future Integration of Heterogeneous Sites

A similar approach as that taken with CSC (c.f. section 5.)5. , where CSC does not provide its own disk space to DEISA, will be used for the integration of heterogeneous sites. This is appropriate since these sites will not be able to provide disk servers on their production systems. Thus RZG or another of the four core sites will offer the disk space for individual non-IBM sites.

In order to proceed with the integration of these heterogeneous sites, in particular the SGI-Altix systems, contract negotiations with IBM have been established. The main difficulties faced and discussed here, have been the following:

- Old RedHat Linux release with 2.4 Kernel due to necessity of CXFS
- Switch to SLES Linux with 2.6 Kernel when SGI requests the update
- Required size of the Altix system for the porting
- Setup of a test environment, able to test the port of the GPFS

The contract covers the following main objectives

- Port of GPFS version 2.3 for SGI Altix with Kernel 2.4 under RedHat Linux with "client" functionality by IBM
- Testing of this functionality and the stability of this port by IBM
- Support and fixes for this ported version of GPFS by IBM
- Set-up of a suitable sized Altix system with required software setup by SARA
- Set-up of file serving and testing hardware environment by DEISA/SARA

The contract was signed in October 2005. According to the contract, the integration of the Altix systems will happen in the time frame according to the deliverable D-SA2-5A.

7.2 Future Configuration of the MC-GPFS throughout DEISA

Figure 7 shows the general view on the logical and physical provision of disk space and file systems for the whole DEISA cooperation, including the heterogeneous extension.

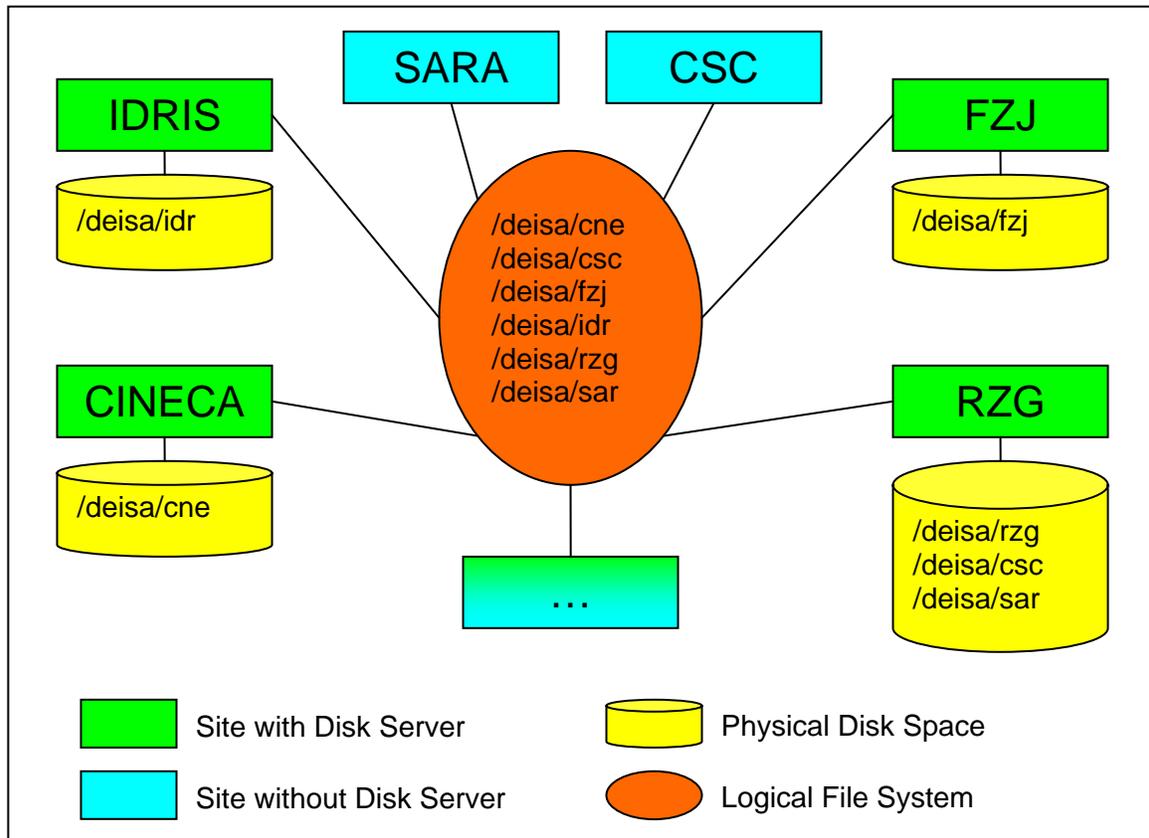


Figure 7: GPFS provision throughout DEISA

7.3 Preparation of a Showcase for the Supercomputing SC05 in Seattle

In order to prove the advantages of the global file system concept in supercomputing, a collaboration between DEISA and TeraGrid, the US counterpart to DEISA, will give a live demonstration during the Supercomputing Conference in Seattle in November 2005. This will show applications running on different systems in TeraGrid and DEISA with reading and writing of data to and from the file systems provided by TeraGrid and DEISA.

This will extend the showcase from the DEISA EU Review Meeting in Paris in June 2005. There will be an integration of the DEISA MC-GPFS with the GPFS from TeraGrid, located at the SDSC in San Diego, USA. Thus the GPFS will be a world-wide transparent file system, covering the four core sites of DEISA for Europe and the SDSC for the United States.

There will be four use cases during the Supercomputing SC05 in Seattle. These utilise different parts of the GPFS and the computing resources. These are:

- ROSETTA (Protein Structure Prediction)
computing at TeraGrid and DEISA, data at TeraGrid
- GADGET (Cosmological Simulation)
computing at DEISA, data at TeraGrid
- ENZO (Cosmological Simulation)
computing at TeraGrid, data at DEISA
- TORB (Gyrokinetic Turbulence Simulation)
computing at DEISA, data at DEISA and TeraGrid

All these codes will make use of the transatlantic connection of the DEISA and the TeraGrid systems.