



CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Proof of Functionality of Multiple-Cluster-GPFS (MC-GPFS)

Deliverable ID: DEISA-SA2-1A
Due date: November, 30, 2004
Actual delivery date: November 18, 2004
Lead contractor for this deliverable: RZG, Germany

Project start date : May 1st, 2004
Duration: 5 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 Content

Table of Content	1
List of Figures	2
1. Introduction	3
1.1 Executive Summary	3
1.2 References and Applicable Documents	3
1.3 Document Amendment Procedure	3
1.4 List of Acronyms and Abbreviations	3
2. Multi-Cluster GPFS for DEISA	6
2.1 Introduction	6
2.2 GPFS (General Parallel File System)	6
2.3 GPFS Solution in a Local Environment	8
2.4 MC-GPFS: a Solution for a European Wide Environment	10
2.5 Setup of Dedicated Test Systems	11
2.6 Installation	13
2.7 Testing Results and Conclusions	13
2.8 Test MC -GPFS Scenario	15
2.9 Application Tests	16
2.10 Future Investigations	16

List of Figures

- Figure 1.....Overview of the components of GPFS
- Figure 2GPFS Cluster configuration in a LAN
- Figure 3Single GPFS Super-Cluster configuration
- Figure 4MC-GPFS configuration with remote mounts
- Figure 5Testbed Configuration of the four Core Sites
- Figure 6Common File System Structure at the Four Core Sites

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, "Proof of Functionality of Multiple-Cluster GPFS", is the first SA2 deliverable, showing that the described functionality has been achieved among the four "core"-sites (IDRIS, RZG, CINECA and FZJ).

1.2 *References and Applicable Documents*

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

1.3 *Document Amendment Procedure*

1.4 *List of Acronyms and Abbreviations*

AIX	Advanced Interactive eXecutive (IBM's derivative of UNIX OS)
CPU	Computing Processor Unit
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)
I/O	Input/Output

IP	Internet Protocol
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
SAN	Storage Area Network
SW	Software
UID	User IDentity (UNIX User)
UNIX	An Operating System
VSD	Virtual Shared Disk, a component of GPFS
WAN	Wide Area Network

2. Multi-Cluster GPFS for DEISA

2.1 Introduction

The Multi-Cluster General Purpose File System (MC-GPFS), an extension of GPFS which is described below, is intended to be the solution for the Global File System ("Grid File System") between IBM's HPC systems participating in DEISA. The MC-GPFS is a highly parallel, high-performance global file system providing transparent access to data, and designed for the use within wide area networks (WAN). Within DEISA, each core-site provides disk space for a file system, which is managed locally. This file system can be exported by means of GPFS to all the other sites, thus giving it the global dimension. If the WAN is not available, local access is still possible. This functionality is in principle also provided by common distributed file systems such as NFSv3. However, the performance and availability of NFSv3 are significantly worse than that of GPFS. GPFS provides the global access that serial and parallel applications need with excellent performance and availability characteristics.

Concerning security, MC-GPFS requires a trusted host environment over all sites involved. This means that all administrators of all nodes act benevolent and no external person gains administrative access to the network or one of the node participating therein.

The MC-GPFS between the four core-sites (IDRIS, RZG, CINECA and FZJ) is installed in a test environment which resembles the future production environment. The test systems are already connected to the dedicated wide area DEISA network, which is a trusted host environment as specified above.

2.2 GPFS (General Parallel File System)

Basically, GPFS is a parallel disk file system. GPFS allows users shared access to files that may span multiple disk drives on multiple nodes. Parallel applications can simultaneously access the same files, or different files from any node in the GPFS node environment.

GPFS provides several advantages when used on AIX- or a Linux-Cluster.

Firstly, on the server-side, a file is spread over multiple servers in parallel. This increases the system performance, because the client uses multiple network streams to access a file. As an important side-effect, all server-disk usage is automatically balanced.

Spreading the file system over multiple servers also enhances its availability. If a single network connection breaks down or the network load is very high, the GPFS client looks on its own for an alternative network path to reach the server. Beside this, if a server crashes, its functionality can be taken over by one of the other servers, if the disks can be seen by that server, thus keeping the file system available to the client nodes.

GPFS is also parallel on the client side. It allows simultaneous access of multiple processes or applications on all nodes. The file consistency is assured by a sophisticated token management no matter on which network path the clients access the file.

Beside these parallel features, GPFS is a journaling file system, which guarantees a fast recovery in case of a failure.

The last important advantage is that the GPFS can be extended and reconfigured while being in production. There is no need to shut down the file system when adding new nodes or server-disks. Also, once the GPFS is configured initially it can be reconfigured to increase the throughput.

GPFS Components

Figure 1 gives an overview of the components of a local GPFS.

On the client side (compute nodes), the GPFS is realised in the GPFS-layer and the VSD-layer. When an application on a compute node wants to access a file, the OS passes this request on to the GPFS-layer. This one forwards the request to the VSD (Virtual Shared Disk). The VSD of the compute node contacts via Ethernet the VSDs of the relevant servers which in turn perform the actual disk-i/o and return the requested data to the compute node via the network.

In MC-GPFS, the VSD are assisted by the so-called NSD (Network Shared Disk), which manages remote file access.

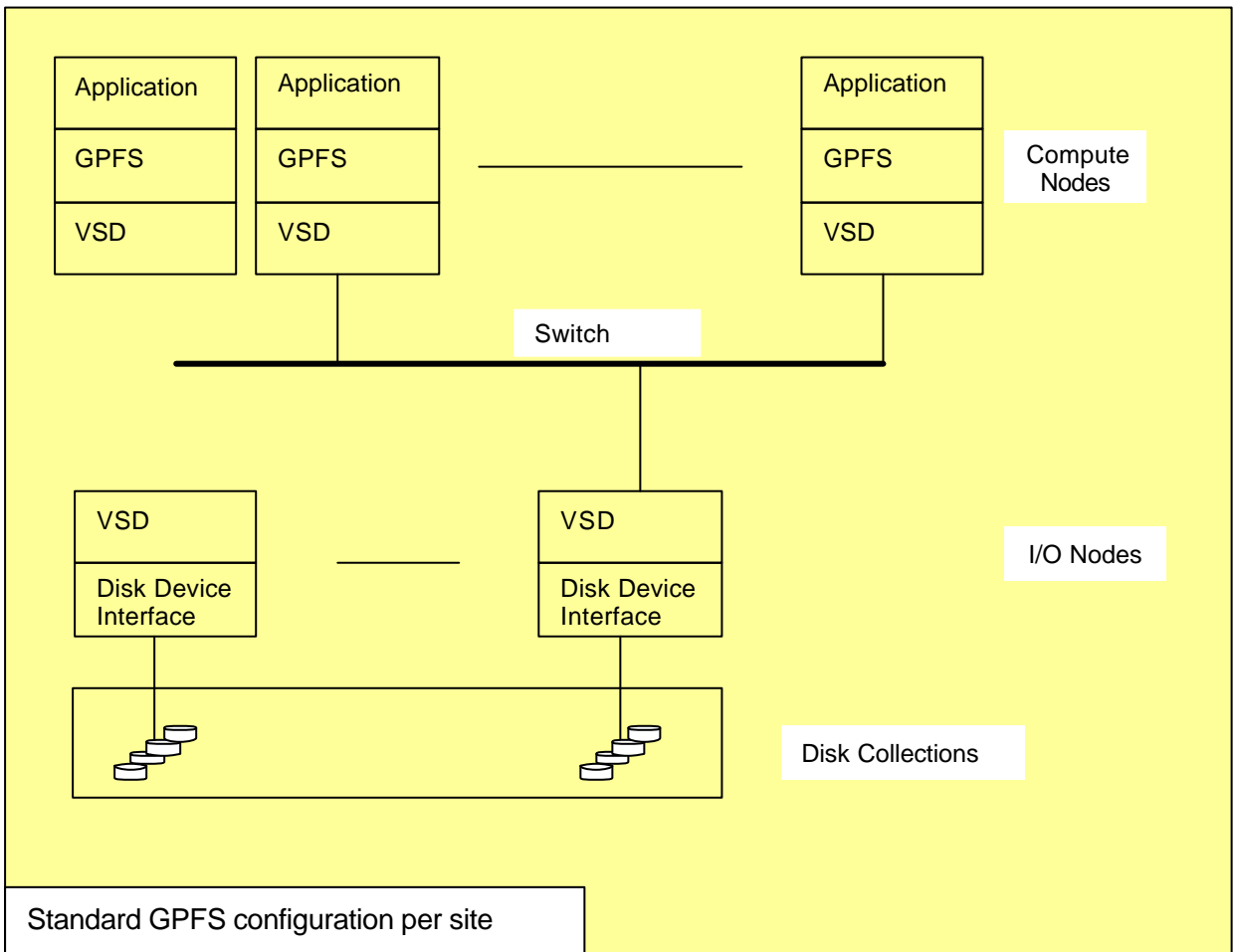


Figure 1: Simplified overview of the GPFS components.

2.3 GPFS Solution in a Local Environment

GPFS is supported on AIX and Linux systems and even in a mixed environment. The data traffic can flow either directly via a Storage Area Network (SAN), based on FibreChannel (FC) from each CPU node or via network connection to I/O nodes which have the direct disk access (see figure 2).

The network connection can be a high performance switch as the Federation Switch - typically used in the Regatta systems - or a Myrinet / Infiniband switch, often used in Linux-Clusters. But even standard TCP/IP network like Gigabit Ethernet is supported.

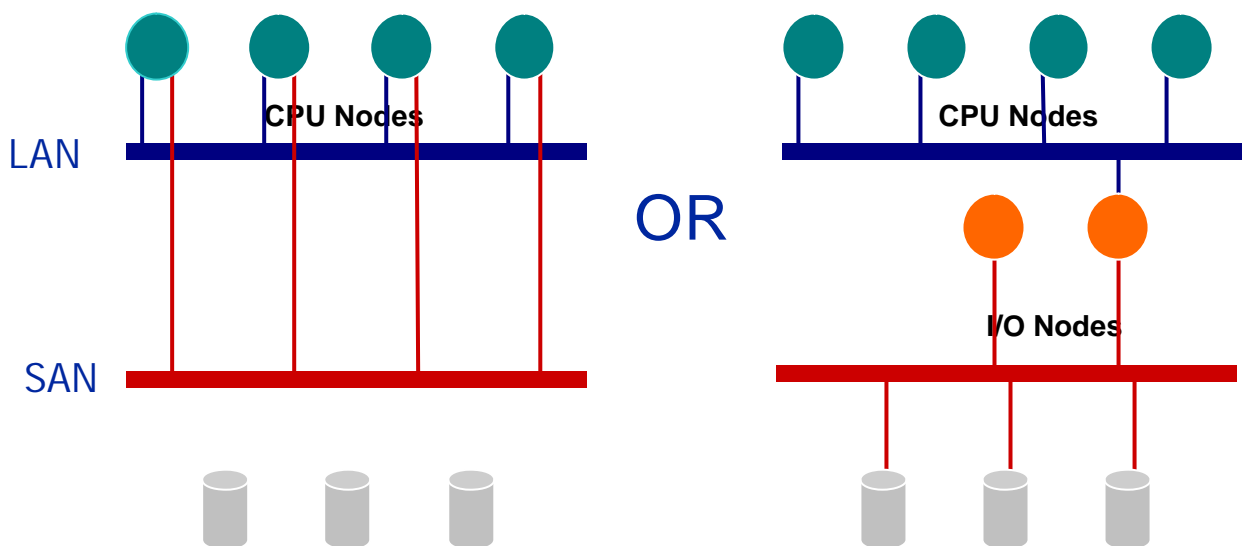


Figure 2: GPFS Cluster configuration in a LAN

Extending this configuration to a WAN (see figure 3) environment to have a single GPFS over more than one site, as planned for DEISA, raises the following difficulties:

- ? Loss of reliability
The performance of GPFS is achieved by delegating file system internal administrative task – like locking of files – to client nodes. This means that no locality of data access can be guaranteed. Thus the stability of the network connection between the different sites is essential for any access to data. No local work at all may be possible if the network connection to one site is interrupted. Even if the data is local the mechanism of quorum nodes, which controls the access to the file system at all, may stop the local access if the network to the other sites is down.
- ? Single administration
The GPFS-software has to be identical on all participating nodes independently of the site, thus no different levels of the software are allowed at the different sites. The configuration is global for all sites and could be changed on one site with impact on all other sites. Thus theoretically each site has to know the exact

- HW-configuration of every site. Furthermore each site could have complete access to all facilities on any other site.
- ? Common user administration
 Since it is one file system over more than one site and there is no UID-mapping in this version of GPFS identical, unique UIDs have to be provided for each user. This effectively means that all sites no longer can administrate the users by themselves but every user is globally administered for all sites.
 - ? Security
 All data and even meta data traffic is not encrypted.

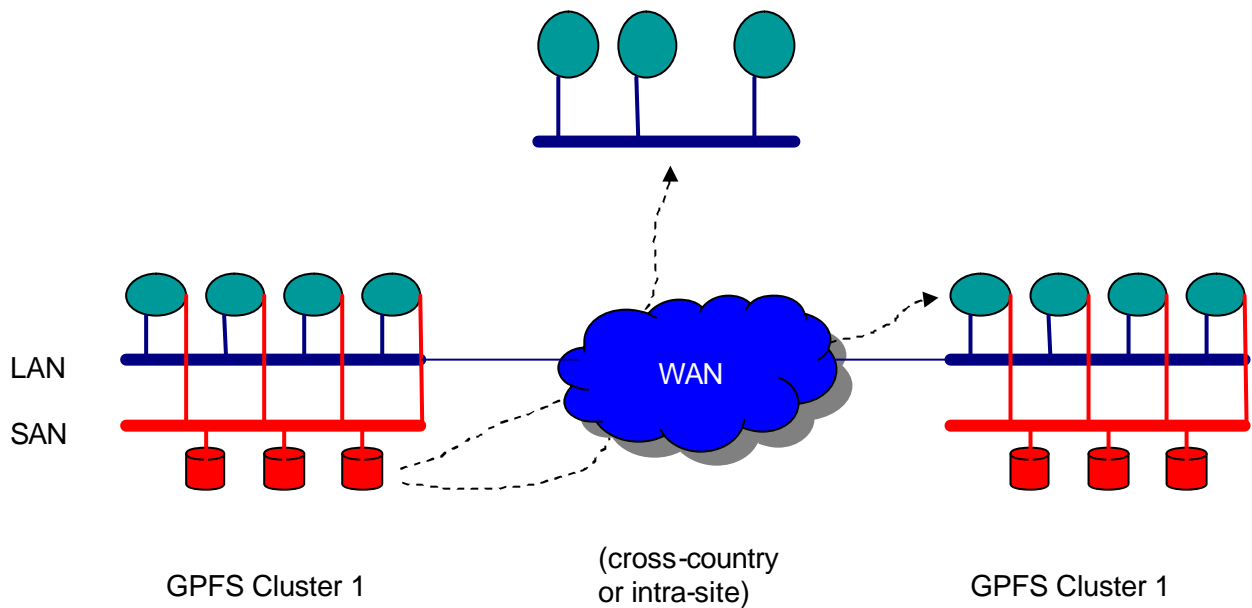


Figure 3: Single GPFS Super-Cluster configuration (no MC-GPFS)

All these restrictions would couple sites together so strongly, that they lose their independence. Therefore the described GPFS scenario is not a solution for a European scaled file system but for a local site configuration only.

2.4 MC-GPFS: a Solution for a European Wide Environment

The aim of MC -GPFS is to keep the benefits especially the highly performing access of the local GPFS and extend it for the requirements of a WAN environment (see figure 4).

This is achieved by defining two different views of GPFS – local and remote –:

Each file system belongs to one “owning” cluster, responsible for

- ? administration
- ? lock management
- ? recovery

“Remote” nodes can mount this GPFS and

- ? request locks
- ? access data and metadata directly over the WAN
- ? but do not participate in quorum or administration.

To provide this functionality the following new features were implemented:

Data Access

The disk access is done by the new NSD layer similar to the local VSD layer.
The Cluster manager detects node failures and drives recovery.

Administration

The administration and configuration of an “owning” cluster GPFS is done locally and the local parameters are advertised to the remote sites.

Security

A trusted configuration between sites is still required, but the communication and the mount of the “remote” file system is secured by the SSL protocol.
Different mechanisms of UID mapping for file access can be provided.

GPFS Protocol

The protocol is extended for remote access to provide scaling by limiting protocol traffic to “active” nodes.
Failure detection is implemented using GPFS disk leasing/heartbeating outside that set of remote machines.

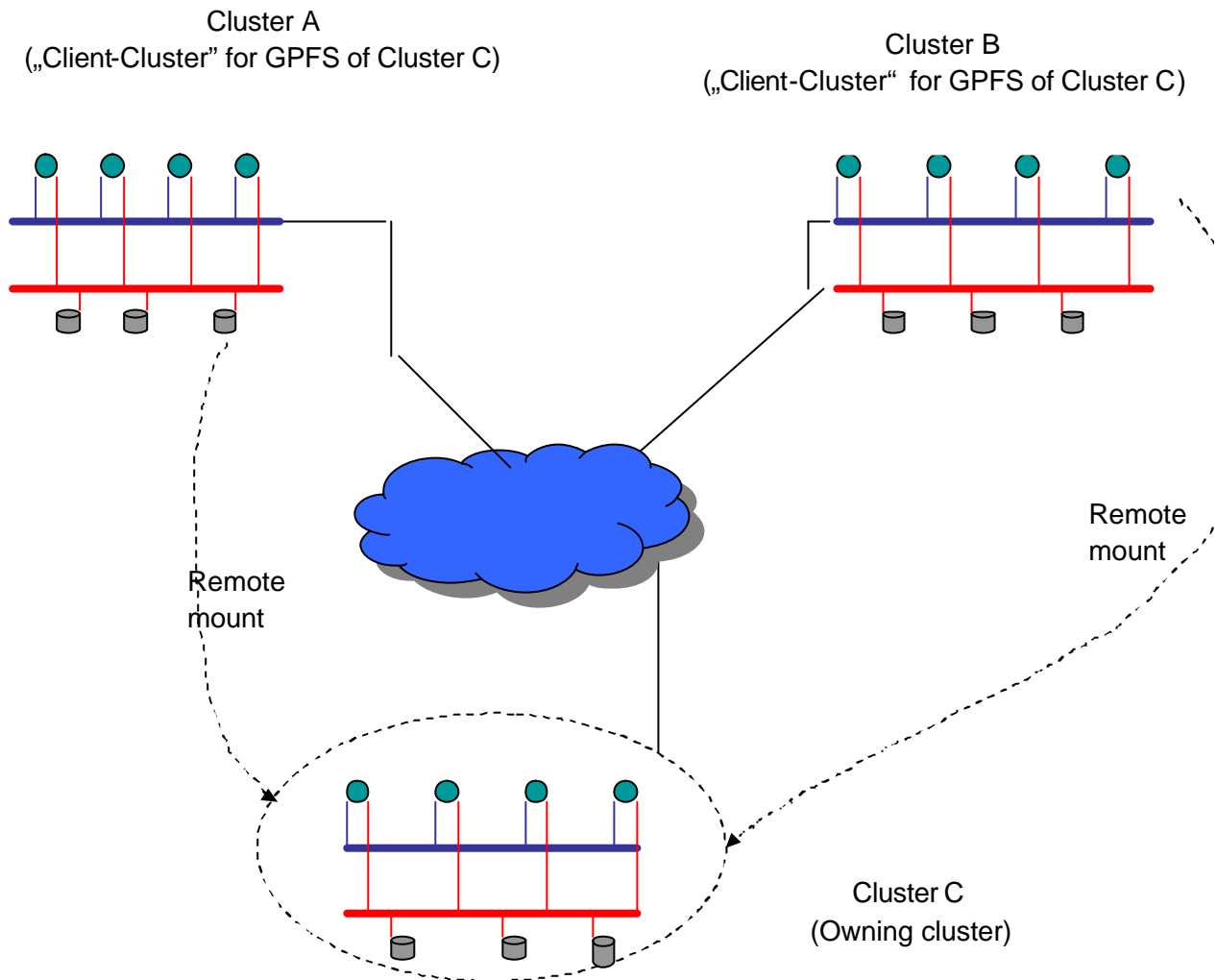


Figure 4: MC-GPFS configuration with remote mounts

2.5 Setup of Dedicated Test Systems

Separate test systems have been provided, installed and configured at each core site in order not to interfere with the production systems.

Test systems of the four core-sites

IDRIS has three p655 nodes, two are equipped with 4 CPUs and one with 8 CPUs, dedicated to the DEISA test-environment. These machines are connected internally with an HPS. Externally all three machines are connected to the DEISA-network with a 1Gigabit Ethernet-interface.

RZG has one full p690 node divided into four LPARs having 8 CPUs each dedicated to the DEISA test-environment. The four LPARs are connected internally and externally with the DEISA-network via a 1 Gigabit ethernet-interface.

FZJ has four p690 LPARs consisting of 2 CPUs each dedicated to the DEISA test-environment. The four machines are connected internally and externally with the DEISA-network via a 1Gigabit ethernet-interface.

CINECA has two p690 LPARs with 8 CPUs each dedicated to the DEISA test environment. The two machines are connected internally and externally with the DEISA-network via a 1Gigabit ethernet-interface.

Dedicated DEISA network

Currently the test environment is using the already installed 1 Gbit/s DEISA wide area network (more information in SA1).

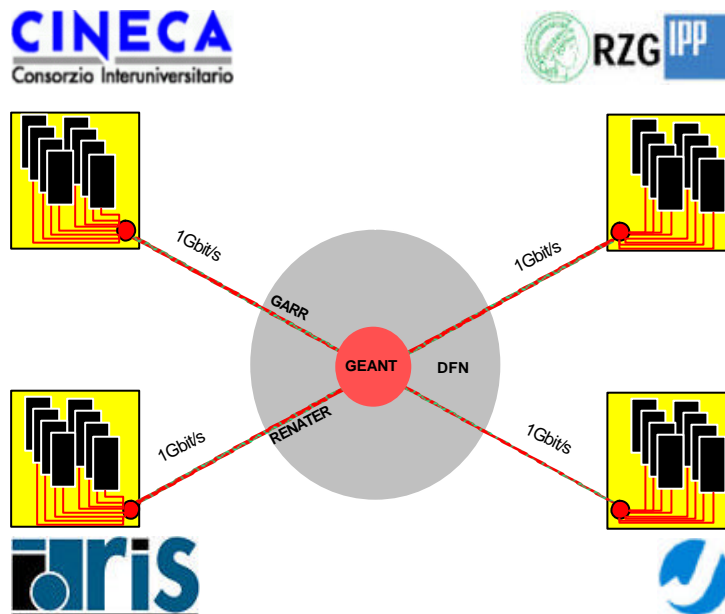


Figure 5: Testbed Configuration of the four Core Sites

Software

The MC-GPFS (2.3 beta release) requires at least three nodes at each site providing a file-system. These nodes have to be connected to the DEISA-network and be able to communicate with all nodes at each site. The version of the used AIX has to be 5.2 ML 03 or higher. Furthermore, it is recommended to be the same version on all the nodes. During the beta testing phase the GPFS Software is required to be at the same level on each site. In the installed beta-version of GPFS provides neither security nor UID or GID mapping features.

The beta software is updated regularly including fixes and improvements.

2.6 Installation

In the beginning of September 2004 all GPFS specialists of the four core sites and ECMWF met at IDRIS for the first installation process. During this week an installation of the first beta release of MC-GPFS at IDRIS and RZG was achieved. A file system was set up on both sites using two of the three nodes dedicated to the DEISA test environment. They could be mounted locally and remotely. Access to data (read and write) as well as locking functionality was provided by the software. The locking functionality was tested with special programs.

After these successful tests CINECA was included as a client site (not providing its own file system) by using only one machine. This configuration was also working and showed the same access and locking functionality as the two-site configuration.

Finally different failure scenarios – local and wide are network disruptions, system failure – were tested.

Extension to all 4 sites

Meanwhile a MC-GPFS file system is set up at each site and mounted by all other sites.

The sites provide the following GPFS disk configuration:

IDRIS provides a Cluster GPFS striped over two internal disks of the two 4 processor nodes with a capacity of 144 GByte each.

At RZG all four LPARs act as servers for 864 GByte disk space. The four machines are grouped in two pairs and connected to the disks in a way that one machine can take over the disks of the other partner automatically in the case of a failure.

At FZJ three partitions act as server each supporting 6 internal 72 GByte disks, thus providing 1.2 TByte.

CINECA has two file servers supporting SAN connected disk space (400 GByte) with fail-over functionality.

2.7 Testing Results and Conclusions

After the installation of the MC-GPFS at IDRIS and RZG tests showed that the access (local and remote) to files is distributed evenly over the file servers. This confirms the parallelism required for high performance access locally and over the WAN connection (DEISA dedicated network). Up to now no measurements have been performed, but all single components are working in the expected way. Extrapolation to the real production system is therefore possible, thus predicting full bandwidth access to remote file-systems, limited only by the WAN, which is currently a 1 Gbit/s connection.

The file/directory locking features which are essential for data integrity were intensively tested between local and remote sites. All tests were passed successfully.

While some requirements and limitations of GPFS 2.3 are well documented, better solutions seem nevertheless desirable in a number of cases.

Remote Data Access via the new NSD in MC-GPFS

It was observed that all NSD traffic from a remote node to a disk flows directly to the server controlling that particular disk, but not to other nodes that would then get to the disk through VSD. This was considered the expected and desired behaviour to prevent unnecessary traffic within the owning cluster's local network. Different behaviour was observed only as a result of an NSD configuration no longer matching a modified VSD configuration. Obviously, all nodes mounting a file system must thus be able to communicate at least with the owning cluster's contact as well as manager and NSD server nodes. It is understood that, currently, other GPFS nodes do also communicate directly with each other, in particular when handing over file locks. It is assumed that this concept aims at improving scalability by minimizing load on the file system manager.

The file/directory locking tests included the setting of a lock on either the local or a remote site. It was proven that the access to that file/directory could be gained on any site only after removing the lock on the site holding the lock. And it was seen that the information about the release of the lock was propagated almost instantaneously to the other sites.

Communication Restrictions and Future Requirements in the WAN

The failure tests, including loss of communication between sites or shutting down servers, showed some behaviour of MC -GPFS which may be undesirable for a production environment. Within a single, local cluster, it is feasible to ensure necessary connectivity between any two nodes at all times. Thus, shutting down the file system if such connectivity is lost, is considered an acceptable solution. However, with MC -GPFS, the network involved is much more complex and problems are bound to occur more frequently. An experiment was performed to determine the behaviour if multiple remote clusters access the same file system. It was found that, if the IDRIS file system is mounted at both RZG and CINECA, file locking may involve communication not only between RZG and IDRIS and between CINECA and IDRIS, but also between RZG and CINECA. Similarly, if any other cluster local to IDRIS were to access the IDRIS DEISA GPFS file system, it would require full connectivity also to other DEISA sites. This is considered very undesirable and much more restrictive than expected. It would be strongly appreciated if the file system manager got the ability to proxy requests on behalf of clients in case direct communication between clients is not possible. Direct communication could still be preferred as a default for scalability reasons.

In the same context, it was found that a remote node mounting a file system, and hence virtually joining the cluster, needs the ability to talk to all nodes already using that file system. It was found that, if CINECA is trying to mount the IDRIS file system while unable to communicate directly with RZG, this does in fact not prevent CINECA from obtaining the mount, but rather causes RZG to lose its existing mount. This ability of a badly configured site to 'steal' an existing mount is clearly unacceptable. In conjunction with the '-A dynamic' mount option, it might even result in oscillating mounts, which would cause tremendous confusion for users and administrators trying to understand the situation.

UID Mapping Implementation Requirements

It is understood that the need to trust the kernel on all clients accessing a GPFS file system results from the fact that all nodes communicate directly with the kernel part controlling the disks. The absence of a high-level server-software which could check the clients authenticity is the base of GPFS scalability. A consequence of this is that all access control is left to individual local or remote nodes and that UID mapping is also planned to be left to remote nodes rather than then owning cluster. From a security perspective, however, a concept giving the owning cluster significantly more control would be preferable. For instance, IDRIS might want to grant a remote site access on behalf of DEISA users, but prevent it from messing with data of IDRIS users not involved in the DEISA project.

Further Security Considerations

While it may be impossible to enforce such restrictions for nodes that have physical access to disks, remote nodes will typically communicate with disks via the NSD protocol and NSD servers within the owning cluster. The encryption and mutual authentication is going to be implemented in this NSD. It seems that additional restrictions could be enforced at the same point if appropriate provisions were built into the protocol. While performance and scalability are the main focus when using GPFS locally, security

considerations become more important for remote access. At the same time, remote performance is anyway more likely to be limited by network constraints than by the processing capabilities of NSD servers.

2.8 Test MC-GPFS Scenario

In the MC-GPFS test environment each of the four core sites configures and administrates its own part of the MC-GPFS, but all sites have a transparent view on this file system, e.g. a common namespace/directory structure can be provided. Each site can decide on which sort of Hardware – highly performing or inexpensive – its part of the MC-GPFS is located. Furthermore each site is able to extend its part of the MC-GPFS according to local needs, but all other sites see the changes to the file system without any need for reconfiguring their systems.

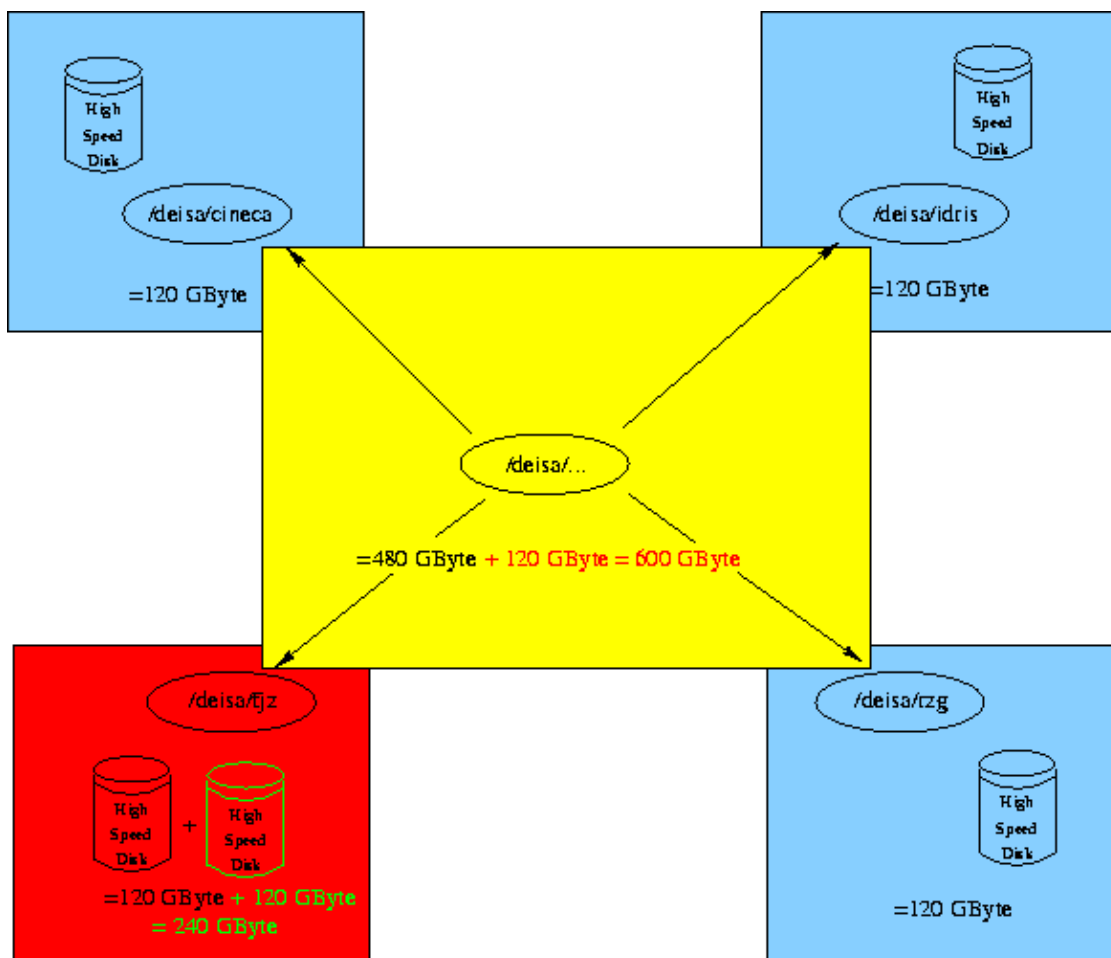


Figure 6: Common File System Structure at the Four Core Sites

2.9 Application Tests

With code ORB (further details in JRA3), a full production code, it was shown that the configuration supports the remote data access through an application. The check-pointed program was started at RZG on the local MC-GPFS. After one calculation-cycle it created a restart file on the RZG-GPFS. The program was then restarted at IDRIS resuming computation from the check point reading in the restart file from the remote Cluster GPFS. Then the computation completed successfully at IDRIS.

2.10 Future Investigations

Performance Tests

After the good results of the network team – nearly peak performance of the raw network-traffic between the single sites – further detailed performance measurements of the MC-GPFS will be performed. Especially the multi-stream feature of MC-GPFS has to be evaluated.

Reliability Test

It is required to perform further tests on the stability of the MC-GPFS configuration when high I/O rates put heavy load on the file systems. Furthermore network failure and recovery has to be tested intensively.

Security

The new security features which are currently delivered – uid/gid-mapping, host identification during the mount command – require further testing.

Preparation for the Production Environment

Because the introduction of MC-GPFS on the production systems at all sites implies the same release level when the GPFS software is finally delivered at GA (general availability) status a detailed upgrade plan has to be set up. Especially the current production GPFS file systems have to be supported by the new release and to be upgraded without any data corruption.