

CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Extending towards an Heterogeneous Environment and
Improvements in Performance of MC-GPFS for Production

Deliverable ID: DEISA-SA2-4A
Due date: April, 30th, 2006
Actual delivery date: May 17, 2006
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	2
List of Figures	2
List of Tables	2
1. Introduction	1
1.1 Executive Summary	1
1.2 References and Applicable Documents	1
1.3 Document Amendment Procedure	1
1.4 List of Acronyms and Abbreviations	2
2. Overview on Multi-Cluster GPFS for DEISA	3
3. Configuration Work at the AIX Sites	3
3.1 Disk Space Extension	3
3.2 MC-GPFS Test-Suite	3
3.3 Integration of the CSC site with its own GPFS	3
3.4 Integration of the ECMWF site with its own GPFS	4
3.5 Test Systems for Integration of the non-AIX sites	4
3.6 Configuration of the AIX Site Production Systems	4
4. Configuration Work at the non-AIX MC-GPFS sites	5
4.1 Preparation of Integration of the SARA site	5
4.2 Preparation of Integration of the BSC site	6
5. Showcase for the Supercomputing SC05 in Seattle	7
6. Current and Planned Configuration	8
6.1 Overview on Current Configuration and General Problems	8
6.2 General Scalability and Reliability	8
7. Work for Future Configurations and Roadmap	9
7.1 Integration of the non-core Sites in Production	9
7.2 Upgrade of the MC-GPFS software	9
7.3 Future Configuration of the MC-GPFS throughout DEISA	10

List of Figures

Figure 1: CSC Production Configuration	5
Figure 2: Current Test Configuration between SARA and RZG	6
Figure 3: Current Test Configuration between BSC and RZG	7
Figure 4: GPFS provision throughout DEISA	10

List of Tables

Table 1: Current hardware configuration of the core sites providing MC-GPFS	4
---	---

1. Introduction

This deliverable was formerly intended to be D-SA2-5A. Since the new partner LRZ will start operation with its huge SGI Altix System in the mid of this year, priority was given to the integration of this system into the existing, homogeneous MC-GPFS environment of AIX systems. Thus the integration of the SGI Altix System at SARA was also brought forward. Last but not least, this is the first step to the integration of the large and important Linux PPC System at BSC which will also benefit from a common file system.

The former target of the timetable was the implementation of the Hierarchical Storage Management (HSM) feature. Since furthermore the availability of disk space is currently not a demanding problem this has been delayed to month 30 and will be reported in October 2006 as D-SA2-5A.

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), not only on all the AIX-computers participating in DEISA, but also integrating other architectures, like SGI-Altix systems and PPC-Linux systems. This document, "Extending towards an Heterogeneous Environment and Improvements in Performance of Multiple-Cluster-GPFS", is the fourth SA2 deliverable, describing the availability of the MC-GPFS on the non AIX sites (SARA, BSC) and the work on improving the performance between the four "core" sites (IDRIS, RZG, CINECA and FZJ). Furthermore, the number of core sites is practically extended to five, since CSC now provides its own GPFS file system and is fully integrated into the MC-GPFS. The last section describes work in progress and some future aspects.

1.2 *References and Applicable Documents*

- [1] DEISA home-page: <http://www.deisa.org>
- [2] Deliverable D-SA2-2A
- [3] Deliverable D-SA2-3A
- [4] Deliverable D-SA2-3B
- [5] Deliverable D-SA2-4B
- [6] Acronyms and Abbreviations:
<http://cgi.snafu.de/ohei/user-cgi-bin/veramain-e.cgi>
- [7] 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
HSM	Hierarchical Storage Management
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 setup and configuration of these four core sites has been documented in the previous deliverables D-SA2-2A [2] and D-SA2-3A [3].

The integration of CSC as a site not providing its own file system was also described in the latest deliverable. Thus, DEISA offered already a distributed European wide file system spanning the four core sites and CSC that are currently connected by the 1Gbit/s DEISA network.

This configuration, which included disk space only on the four core sites, is now extended by the fifth AIX-system CSC with its own disk space. Furthermore the comfortable use of the transparent file access through the MC-GPFS showed that the extension to a heterogeneous environment including non-AIX sites would provide enormous advantages. So the current main work is done on providing a MC-GPFS for SGI-Altix systems, which are already running in SARA, and will soon be running at the LRZ. Also the integration of the powerful PPC-Linux system at BSC would make DEISA more useful to the scientists, doing their calculations on the DEISA systems.

3. Configuration Work at the AIX Sites

3.1 *Disk Space Extension*

Compared to the last deliverable almost all sites have enlarged their disk space provided as MC-GPFS to DEISA. Details can be found in the table summarizing the current configuration of the AIX sites below (section 3.6).

3.2 *MC-GPFS Test-Suite*

In order to be able to get an overview on status of the ever more complex MC-GPFS in the DEISA infrastructure, a set of tests is developed, which provides information on the status and the current performance of the MC-GPFS. This test suite is currently only used by RZG to run tests on all sites and will be provided to the other sites for individual use soon. The major problem in this is that a permanent testing affects the performance of real applications. On the other hand, a rare testing does not provide useful near-time status information. However, the tools are quite useful to determine the current status of the MC-GPFS and to check the effects of changes made to the configuration. A WEB display of the later regularly sampled data is under preparation.

3.3 *Integration of the CSC site with its own GPFS*

Since the CSC site in principal is like all the other AIX core sites based on an AIX-system internally connected with an HPS-Switch and externally connected with 1Gbit/s to the DEISA network, it was quite natural not only to include the CPUs into the DEISA network, but also to setup its own local GPFS. After buying additional disk

hardware, CSC now can provide its own GPFS disk space of 2TB. A detailed configuration can be found in Figure 1 below.

3.4 *Integration of the ECMWF site with its own GPFS*

At the very end of the reporting period all external problems, like the missing 1 GBit/s connection of ECMWF with the DEISA network, could be solved and a first test configuration of the GPFS file system could be mounted from ECMWF in the production environment of the homogeneous sites. This initial configuration consists only of 2 nodes with a total of about 1TByte of disk storage, but will be extended, if it proves to be stable. In the next deliverable a detailed configuration will be included.

3.5 *Test Systems for Integration of the non-AIX sites*

Although the MC-GPFS will be available for other systems than AIX the provision of disk space can not be done on all other non-AIX systems. Therefore, RZG configured an additional disk space of 0.25TB and 0.5TB to provide several file systems to be used for the testing of the integration of these systems. One of these file systems is used for the porting activities of MC-GPFS to the SGI-Altix system, done at SARA. The other test file system is used for testing of the integration of the PPC-Linux system located at BSC, which will be integrated with its own MC-GPFS after the successful testing period.

3.6 *Configuration of the AIX Site Production Systems*

Each of the six AIX sites offer now a locally configured MC-GPFS, which is exported to the other AIX sites. Thus, strictly speaking DEISA does not use a single MC-GPFS over all these sites, but five MC-GPFS interwoven in such a way that they appear to be a single, shared file system. They are mounted as `"/deisa/<site>"`. Below this directory 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 [7]).

The current situation is summarised 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	5 TB	480 Power5 (1.9 GHz)	1152 GB
CSC	2	2 TB	416 Power4 (1.1 GHz)	672 GB
ECMWF	2	1 TB	48 Power4 (1.9 GHz)	48 GB
FZJ	2	4 TB	1288 Power4 (1.7 GHz)	5152 GB
IDRIS	2	2 TB	1024 Power4 (1.3 GHz)	3136 GB
RZG	2	7 TB	928 Power4 (1.3 GHz)	2368 GB

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

The detailed configurations of the four AIX production systems of CINECA, FZJ, IDRIS and RZG can be found in the last deliverable. The current configuration of CSC is displayed the following figure:

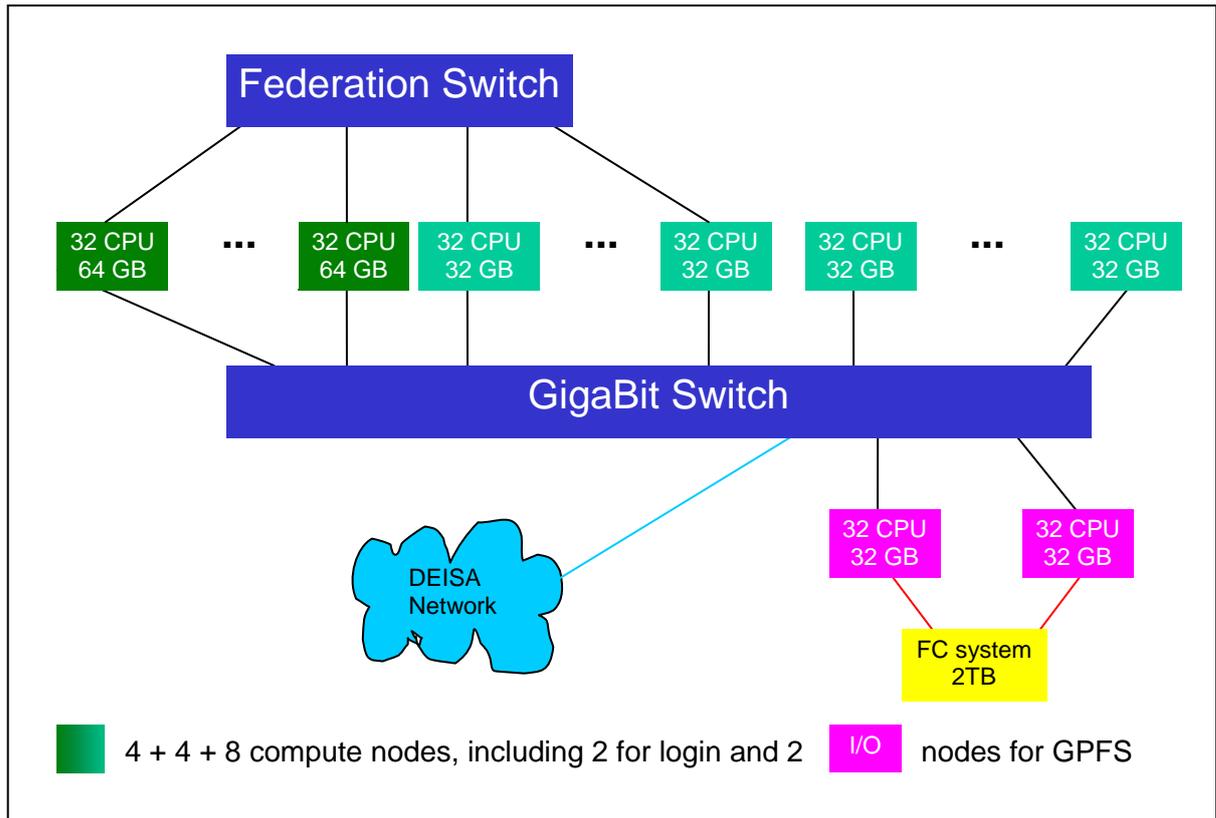


Figure 1: CSC Production Configuration

4. Configuration Work at the non-AIX MC-GPFS sites

4.1 Preparation of Integration of the SARA site

In order to make the file access very transparent for the users of DEISA, it is essential that as many different systems as possible can share the data via MC-GPFS. Thus a contract with IBM was signed concerning the porting of MC-GPFS to a SGI-Altix system.

It was clear from the beginning, that the SGI-Altix itself would never export its own file system as a file server, but only work as a client. So the disk space has to be provided externally. To start as quickly as possible for the first porting phase RZG provides a small MC-GPFS on a test machine connected to the DEISA network (see above 3.5).

Additionally, for a good local performance it is required to have disk space located at the same site as the compute system. For this reason SARA has bought additional servers, based on Intel architecture running Linux. These four servers currently provide about 800GByte of disk space. At the end of the year additional disk systems, which are part of a procurement currently setup at SARA, will extend this configuration. This file system then will be integrated into the existing MC-GPFS production infrastructure of DEISA.

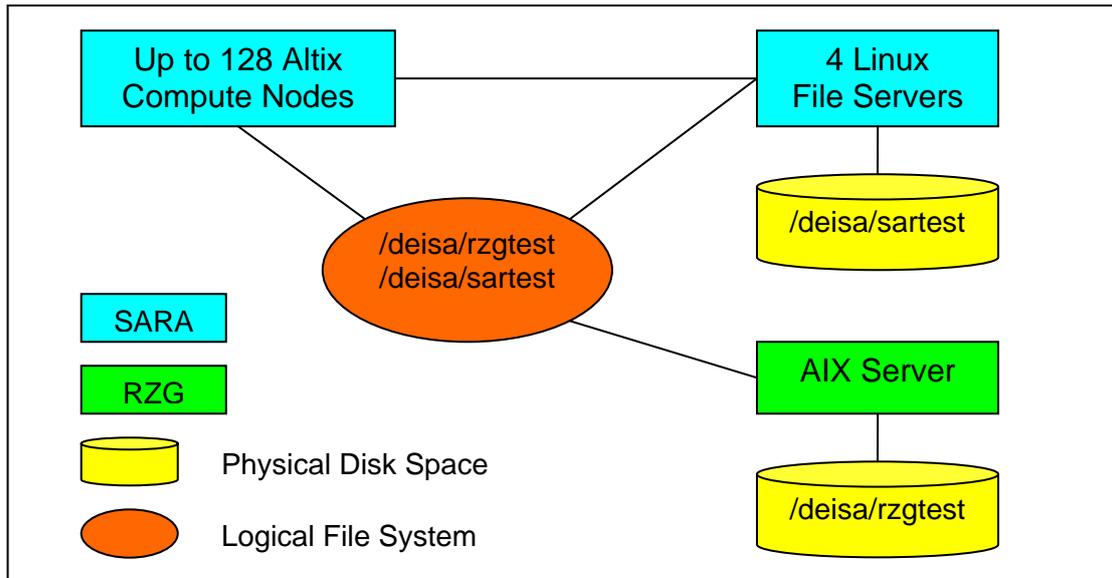


Figure 2: Current Test Configuration between SARA and RZG

4.2 Preparation of Integration of the BSC site

The integration of the BSC site is a big challenge, since it consists of thousands of individual systems. The MC-GPFS requires the possibility of communication between all systems. Thus, the number of partners for communication increases rapidly compared to the situation, where large single systems of the AIX machines have been coupled. Thus, a rather conservative plan for the integration of the BSC system has been defined, which tries to catch all problems without causing problems on the production systems.

For this reason, test systems at BSC and RZG are defined and configured which are used for the first phases, before the real production systems are included. The detailed list of actions looks as follows:

- Testing the network between BSC and RZG
- Exporting GPFS from BSC-Test to RZG-Test
- Exporting GPFS from RZG-Test to BSC-Test
- Exporting GPFS from RZG-Test to BSC Production
- Exporting GPFS from BSC to RZG-Test
- Exporting GPFS from BSC to RZG
- Exporting GPFS from BSC to all DEISA
- Exporting GPFS from RZG to BSC
- Exporting GPFS from all DEISA to BSC

The testing of the network should provide good initial parameters for the configuration of the setup parameters for general networking as well as for the MC-GPFS parameters. In each of the following steps the functionality and performance is then tested and possibly some of the parameters and configuration is adopted until the full integration of BSC into the MC-GPFS infrastructure of DEISA is achieved.

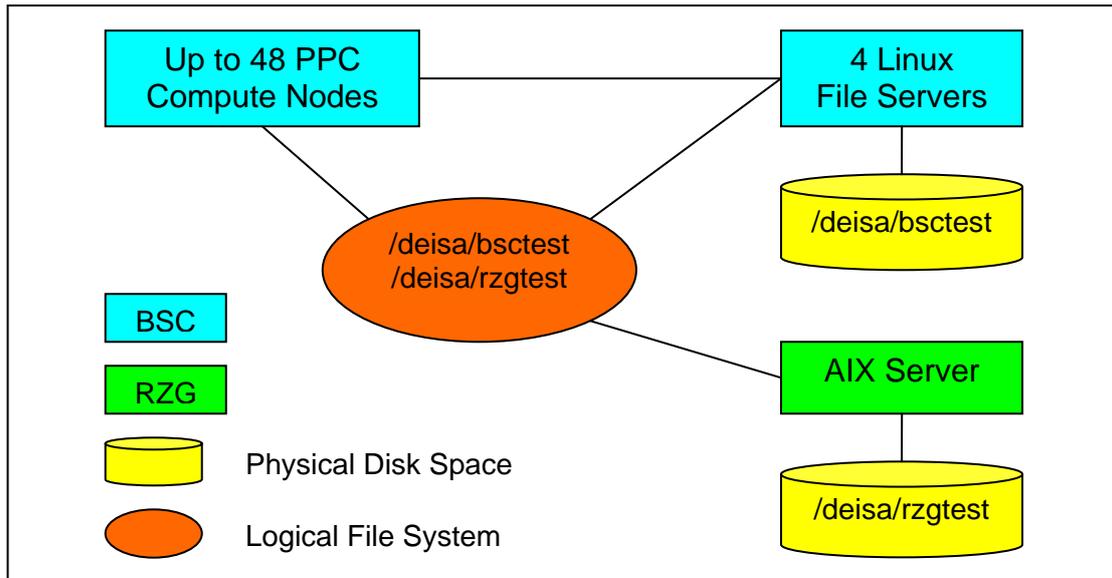


Figure 3: Current Test Configuration between BSC and RZG

5. Showcase for the Supercomputing SC05 in Seattle

As announced in the last deliverable, a showcase was prepared for the SC05 in Seattle in November 2005. The demonstration was performed live in Seattle and proved with real applications (ROSETTA, GADGET, ENZO and TORB), running in DEISA or TeraGrid and using the MC-GPFS file systems of DEISA and TeraGrid, that a world wide transparent data access is possible with MC-GPFS.

In preparation of the showcase, all DEISA sites participating in the showcase had to reconfigure the MC-GPFS in order to allow for connection with the TeraGrid MC-GPFS. The reconfiguration was restricted to a minimum, in order not to perturb the currently running productive system. Due to high round trip times the performance was not very high, and no optimisation was done, since the productive systems should not be interrupted by reconfiguration work for just a few days of the show. More details about this showcase can be found in the article "Exploring the hyper-grid idea with grand challenge applications: The DEISA-TeraGrid interoperability demonstration" to appear in the IEEE proceedings of the SC05.

Boiled down, during the Supercomputing SC05 in Seattle the following live performances have been taken place:

- 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

Both the TORB and GADGET demonstrations have been performed at the DEISA booth, which was hosted by the Dutch booth. The TORB code run at RZG and used the DEISA file system for the input data and the TeraGrid for the output data, which after the short demo run was displayed on a computer screen at the SC05 running a

visualisation program on a TeraGrid computer. The GADGET program computed data at RZG, wrote it to the TeraGrid MC-GPFS. The data then was immediately displayed from a TeraGrid machine on a display at the SC05, with automatically updates of newly produced data.

By this demonstrations it was shown, that it is possible to access data transparently on a world wide range. This global visibility and usability of data, without the need of any user driven transfer using any copy tools eases the possibility of doing calculations anywhere around the globe and using these data afterwards on any other place. Thus a real global file system was achieved and the way DEISA tried to address this issue was proven right.

6. Current and Planned Configuration

6.1 Overview on Current Configuration and General Problems

Currently the five AIX sites CINECA, CSC, FZJ, IDRIS and RZG are connected to the MC-GPFS in production. These systems are also coupled with the MC-LoadLeveler, a batch queuing software, so that these five production systems almost look like one huge system to the user. Although no jobs can currently share resources of more than one site at one time, a DEISA user has in principle access to all systems and can easily access the data located on the MC-GPFS transparently at each site. Thus a huge system consisting of more than 4000 CPUs is now connected together with the MC-GPFS.

With the DECI projects running on the AIX sites accessing data transparently using the MC-GPFS it is proven that the concept of the MC-GPFS is working in real production. However, some issues, which cannot simply be addressed, are seen now. The main organisational issue arises from the fact, that from the point of a DEISA user there is in fact one big system, but for the administration there are five units. So maintenance at a site A affects jobs running at a remote site B, when accessing data from the MC-GPFS located at the site A. Currently this situation is tried to be circumvented by avoiding such situations, so not starting jobs of another site, if this site is planning a maintenance. But for the long term other concepts, best included into the job setup, have to be considered.

6.2 General Scalability and Reliability

The real production use of the MC-GPFS showed that the performance is very dependant on the extreme stability of the underlying network. During the last few month GEANT did a major reconfiguration of the network infrastructure, which raised the round trip times by factors. Before and after the reconfiguration the round trip times between the original four core sites are around 20ms. During the reconfiguration phase this times raised up to 100ms. Furthermore the routing went very different ways. This caused a performance degradation up to a factor of five. Thus the configuration parameters of the MC-GPFS were no longer adequate. This required some adjustments of the network parameters, which only could be done during maintenance since these network parameters become active when establishing the connection, which is done with the first communication after starting up the single MC-GPFS.

Furthermore it was discovered that there is a difference in performance between nodes, directly connected by Gigabit Ethernet, and nodes routed from the internal Federation Switch via a Gigabit-adaptor to the DEISA network. This routing is done by an also internal machine connected to the DEISA network, whose nominal

bandwidth should be sufficient. Thus, this behaviour is not understood yet. For example a job running at FZJ using the file system located at RZG achieves with a single stream of file-I/O an average rate of about 400 Mbit/s. With multiple streams achieved by one application or by more than one application the connection can be saturated to the 1Gbit/s limit currently available. In the opposite direction, if the Federation Switch is involved, the rate for a single stream dropped significantly, while it was comparable if performed on a machine directly connected to the DEISA network. But only the performance of MC-GPFS was affected while pure network test showed that the network was capable of providing the full 1Gbit/s performance. This problem was investigated together with IBM. The upgrade of the hardware to the new 10Gbit/s adapters resolved this issue. This behaviour can be explained under the assumption that the Federation Switch does not react adequately on flow control requests related to GPFS, since the Federation Switch itself is assuming a bandwidth capable to handle all GPFS-requests. Thus actually the problem is related to the drivers of the Federation Switch or the Operating System itself. The 10Gbit/s adapters match the bandwidth of the Federation Switch and therefore the problems described above should be solved by an upgrade of the GigaBit adapters.

The stability of MC-GPFS is in principle as good as the reliability of the underlying network, which is extremely high. But as mentioned above any reboots or maintenance actions at the single sites can affect the whole system concerning productive jobs. But the successful running of the DECI projects proves the stability and availability of the MC-GPFS as an outstanding distributed file system.

7. Work for Future Configurations and Roadmap

7.1 *Integration of the non-core Sites in Production*

In the last deliverable it was considered that the sites hosting non AIX systems, which cannot provide MC-GPFS file systems directly are provided with remote disk space from one or more of the AIX sites only. With the improvements of the MC-GPFS it is now possible to include a MC-GPFS hosted at BSC using PPC based Linux-Servers. When the testing of this configuration proves to be a stable solution, implementation can take place. Similarly, the local MC-GPFS Intel-Linux-Servers installed for testing reasons at SARA could be integrated directly, thus each site would have local MC-GPFS disk space.

This approach may be preferable compared to the availability of remote disk space. The maintenance arguments above is even worse, when a remote site hosts disk space for other sites, since during maintenance of the site hosting the remote disk space, all jobs from all sites using this remote disk space are affected. This makes maintenance even more complicated. So local disk space is much preferable and it is considered for the heterogeneous sites to provide their local MC-GPFS disk space based on some small Linux Boxes.

7.2 *Upgrade of the MC-GPFS software*

The Multi-Cluster functionality of the current version of the MC-GPFS was built into the existing GPFS. With the productive use of this version it became clear that some design improvements are required especially when coupling the MC-GPFS over a Wide Area Network. So the next version will provide lots of improvements for better a cooperation between different clusters, including stability and performance. This new version will be installed as soon as possible. But there are some major restrictions in doing so. Currently it is required that all AIX systems are upgraded to AIX 5.3. The new version is incompatible to the old version in a way that all sites would have to

upgrade at the same time. This is simply an operational impossible task. Thus, DEISA will engage with IBM to relieve this restriction. Furthermore, the integration of the heterogeneous sites requires that these MC-GPFS clusters can also be upgraded or stay on the lower release level. Testing and eventually upgrading will be one of the main task for the next two years of the DEISA project for the SA2 group.

7.3 Future Configuration of the MC-GPFS throughout DEISA

The following figure 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. For the AIX systems the file servers are an integrated part of the production system, for all others the file servers are separated from the production system but connected to the local network. Since EPCC/HPCx is not integrated into the MC-GPFS RZG hosts the home and data directories for the DECI users from EPCC/HPCx on the RZG-MC-GPFS.

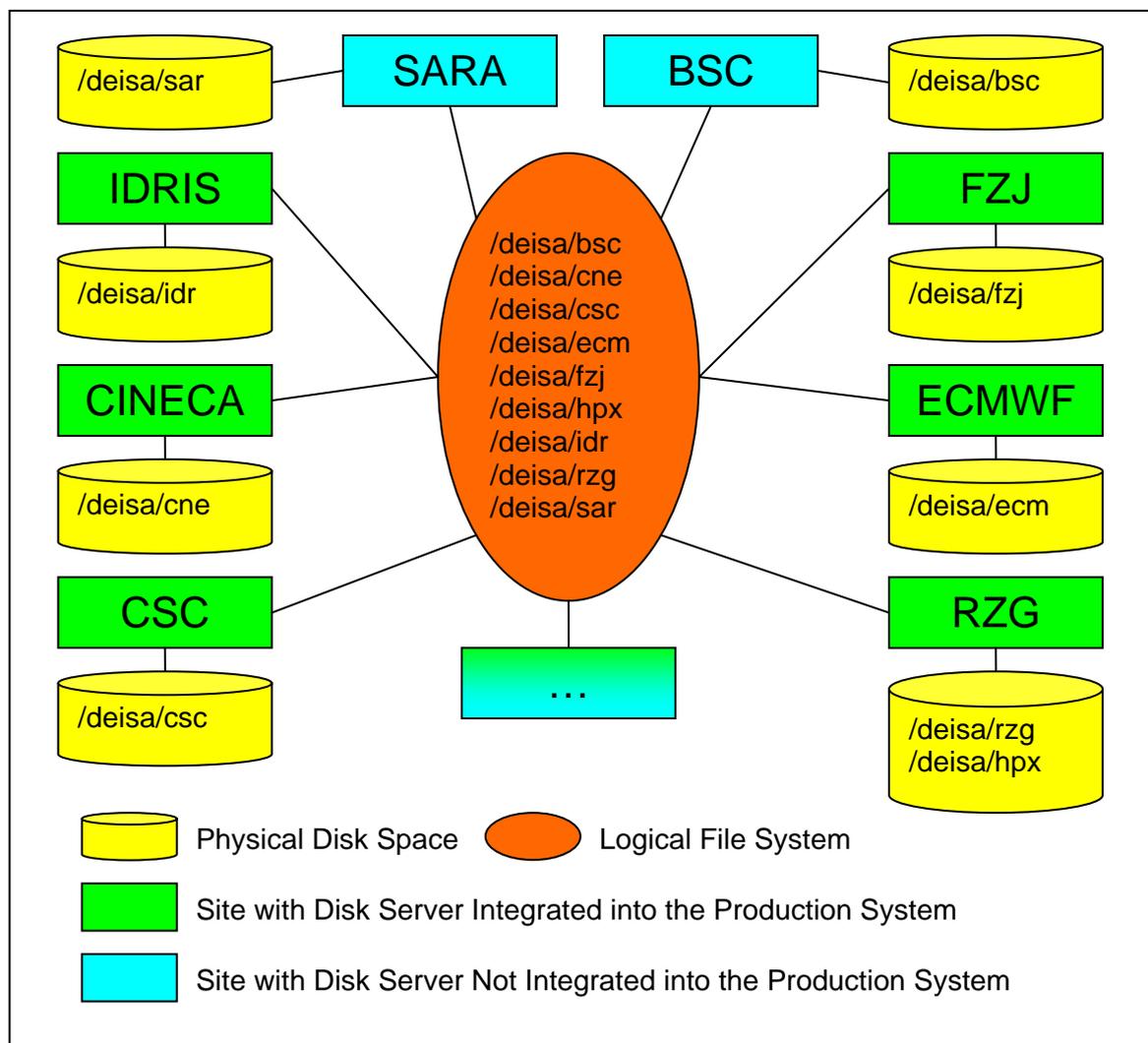


Figure 4: GPFS provision throughout DEISA