

CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Integration of Altix and PPC Systems and
Improvements in MC-GPFS for Production

Deliverable ID: DEISA-SA2-5A
Due date: October, 31st, 2006
Actual delivery date: November 24, 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

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	1
2.	Evolution of Multi-Cluster GPFS for DEISA	2
2.1	Initial state and Planning	2
2.2	Software Problem and Decision for Solution	3
2.3	Upgrade of the Software and further Integration	3
3.	Configuration Work at the AIX Sites	4
3.1	Disk Space	4
3.2	MC-GPFS Test-Suite.....	4
3.3	Integration of the ECMWF site with its own GPFS	4
3.4	Test Systems for Integration of the non-AIX sites	4
3.5	Configuration of the AIX Site Production Systems	4
4.	Configuration Work at the non-AIX MC-GPFS sites	5
4.1	Configuration Work at the SARA Site	5
4.2	Configuration Work at the LRZ Site.....	5
4.3	Configuration Work at the BSC Site	6
5.	Showcase for the EU review meeting in Paris, June 2006	6
6.	Work for Future Configurations and Roadmap	7
6.1	Integration of the non-AIX Sites in Production.....	7
6.2	Future Configuration of the MC-GPFS throughout DEISA	7

List of Figures

Figure 1: GPFS provision throughout DEISA.....	8
--	---

List of Tables

Table 1: Current hardware configuration of the AIX sites providing MC-GPFS	5
--	---

1. Introduction

This deliverable summarizes the work done in the last half year. Beside the extension of the homogeneous MC-GPFS cluster to six AIX sites only the SGI Altix system from LRZ could be introduced as first non-AIX architecture. The integration of the SGI Altix system from SARA and the PPC system from BSC had to be delayed due to the need of an upgrade of the MC-GPFS software. This upgrade was required due to stability reasons in the productive homogeneous AIX environment. Nevertheless the complete cluster including all AIX systems (except HPCx, which will follow in early 2007) and all non-AIX systems (except the NEC-vector machine at HLRS) will be in production approximately at the end of the year 2006.

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, "Integration of Altix and PPC systems and Improvements in MC-GPFS", is the fifth SA2 deliverable, describing the availability of the MC-GPFS on the non AIX sites (LRZ, SARA, and BSC) and the work on improving the stability between the AIX sites (CINECA, CSC, ECMWF, FZJ, IDRIS, and RZG). It furthermore discusses the serious problems faced in the last half year and describes how they were addressed and solved.

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-SA2-4A
- [5] Deliverable D-SA2-3B
- [6] Deliverable D-SA2-4B
- [7] Acronyms and Abbreviations:
<http://cgi.snafu.de/ohei/user-cgi-bin/veramain-e.cgi>
- [8] 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)
Altix	Multi-Processor compute node from SGI

BlueGene	Specialised HPC System from IBM
CPU	Computing Processor Unit
DEC	DEISA Executive Committee
DEISA	Distributed European Infrastructure for Supercomputing Applications
FC	Fibre Channel (disk-connection protocol)
GA	General Availability
GID	Group IDentification (UNIX Group)
GPFS	General Parallel File System
HPC	High Performance Computing
HSM	Hierarchical Storage Management
IBM	International Business Machines (Computer Manufacturer)
I/O	Input/Output
IP	Internet Protocol
LAN	Local Area Network
Linux	Free UNIX-like Operating System
MC-GPFS	Multi-Cluster GPFS
OS	Operating System
PPC	PowerPC (a CPU chip type)
RAID	Redundant Array of Independent Disks
SAN	Storage Area Network
SGI	Silicon Graphics Incorporated (Computer Manufacturer)
TCP	Transmission Control Protocol
UID	User IDentity (UNIX User)
UNIX	Most used HPC Operating System
WAN	Wide Area Network

2. Evolution of Multi-Cluster GPFS for DEISA

2.1 *Initial state and Planning*

The five DEISA sites CINECA, CSC, FZJ, IDRIS and RZG, all operating IBM HPC systems, share the Multi-Cluster General Parallel File System (MC-GPFS) thus providing a grid file system since more than a year. 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 [2], [3] and [4].

During the review meeting in Paris in June 2006 a common MC-GPFS of the AIX system at RZG connected with the SGI-Altix system at SARA and the PPC system at BSC was shown in operation.

2.2 Software Problem and Decision for Solution

It was intended to integrate the Altix and PPC systems soon after that into the productive grid file system, but the installation of a new HPC machine (BlueGene) at FZJ using the same file servers as DEISA introduced some unexpected instabilities. Deeper investigations together with IBM showed a fundamental lack of communication facilities in the currently used version of the MC-GPFS software.

The severe problem in the communication protocol of MC-GPFS showed up, when FZJ started their new BlueGene and connected it to their file servers, which provide GPFS locally as well as for DEISA. Although no file system was shared between the BlueGene and any DEISA system, connection attempts between the single machines have been logged. After integrating the connection addresses of the BlueGene into the DEISA range, the connections could be satisfied. However, due to the huge number of clients and the frequent reboots of single nodes of the BlueGene cluster, the number of reconfiguration requests made to the DEISA cluster lead sometimes to spontaneous umounts of the DEISA file system.

This behaviour was unacceptable. For a short term solution FZJ abandoned the use of GPFS from its BlueGene cluster. However, in the long run, an upgrade of the MC-GPFS to the next release which includes solutions to all these problems was the only way to go. In order to solve the previously mentioned problems, major changes in the communication protocol of MC-GPFS were required. This resulted in an incompatibility between the new version and the old one. This means, that only one version of the software may be installed and a common GPFS is only possible within an environment with either version of the MC-GPFS release.

The new version provides a lot of features improving WAN functionality and a stable operation could not be surely promised with the old version, when enlarging the MC-GPFS by more and more remote sites. Furthermore the operation of the BlueGene system at FZJ was heavily degraded due to the lack of the local GPFS.

Facing all these complex issues after long discussions it was decided to delay the extension of the existing grid by new sites to after an upgrade of the software. Providing stability in the already productive DEISA environment is of higher importance and is more desirable than the availability of MC-GPFS on more DEISA systems.

2.3 Upgrade of the Software and further Integration

The upgrade itself required a strictly planned time management, since the new release was incompatible with the old one. Thus no common Grid File System would have been possible again until all sites would have performed that software upgrade. Immediately after the holiday season this upgrade took place mainly within two days.

Since the PPC system at BSC uses a specially tuned version of the old version of the MC-GPFS software, intense test are still required with the new version, before the PPC system can be integrated.

For the SGI Altix system the new version of the software had to be ported again. This new version requires also a new version of the Operating System running on the SGI Altix systems. Thus the integration of the system of SARA is delayed until SARA has upgraded its machine also with a new Operating System. The SGI Altix at LRZ will be integrated after its public availability is announced.

After the availability of the network connection with ECMWF, some testing has been performed. In succession to these successful tests the ECMWF cluster could be integrated into the homogeneous AIX environment after the installation of special firewall router hardware.

Thus, by the end of the year the MC-GPFS of DEISA will consist of six AIX-sites (CINECA, CSC, ECMWF, FZJ, IDRIS and RZG) as well as the SGI-Altix at LRZ. Soon in the next year the SGI-Altix system of SARA will be integrated after the upgrade of the Operating system. The PPC system of BSC will be added after successfully testing and subsequently installing the new version of MC-GPFS.

3. Configuration Work at the AIX Sites

3.1 Disk Space

The disk space currently provided in the MC-GPFS is sufficient for the projects and other computations performed in DEISA. Thus no extensions had been setup. Since LRZ is close-by RZG, LRZ is using file servers and disk space from RZG. A detailed configuration can be found in a table below (section 3.5).

3.2 MC-GPFS Test-Suite

The test suite, initially used at RZG only, has been improved and rewritten for better portability. The test suite now not only checks for connectivity and transfer rates between sites, but also for local configuration parameters and changes of the configuration. The results can be viewed with a web browser from a web server located at RZG. Critical changes can be mailed to the responsible persons to take action.

For testing the performance at each site machines are selected randomly. Thus the overhead of data transfers for testing purposes should be minimized in order to reduce the influence on running jobs using the infrastructure.

3.3 Integration of the ECMWF site with its own GPFS

The testing between the test environments of ECMWF and RZG were done successfully with the old version of MC-GPFS. Before integrating the productive system, the upgrade of the MC-GPFS software was a requirement, too. Therefore, the system at ECMWF was also upgraded to the new MC-GPFS version at the beginning of October. The fileservers will then been integrated into the productive MC-GPFS cluster after a short testing period at the end of November, when the new firewall router hardware has been installed and configured.

3.4 Test Systems for Integration of the non-AIX sites

In the beginning of the reporting period, RZG still provided the test disk server used for porting the MC-GPFS software to the SGI Altix system at SARA. When it became clear, that it was required to do also a port for the new MC-GPFS version, RZG upgraded its test environment, to provide a test disk server for the SGI-Altix at LRZ, who provided a 64-node SGI-Altix (for more details about this porting see below).

3.5 Configuration of the AIX Site Production Systems

Each of the six AIX sites offers 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 six 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 [8]).

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	5.0 TB	480 Power5 (1.9 GHz)	1152 GB
CSC	2	2.0 TB	416 Power4 (1.1 GHz)	672 GB
ECMWF	2	0.7 TB	2432 Power5+ (1.9 GHz)	2250 GB
FZJ	2	4.0 TB	1288 Power4 (1.7 GHz)	5152 GB
IDRIS	2	2.0 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 AIX sites providing MC-GPFS

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

4.1 Configuration Work at the SARA Site

After successfully porting MC-GPFS to the SGI-Altix system, its functionality was proven in a live demonstration at the EU-review meeting in Paris in June 2006. It was intended to integrate the SARA site into the productive MC-GPFS configuration immediately after some final tests had been finished.

All partitions of the SGI-Altix machine have been prepared with the MC-GPFS software, keys for the secure communication have been generated and exchanged throughout the AIX-sites forming the MC-GPFS cluster at that time.

Shortly after everything was prepared to integrate the SARA system into the DEISA MC-GPFS, the aforementioned problems with the BlueGene Cluster appeared. Considering the deep root of that problem, a decision has been made to upgrade the GPFS-software on the core-sites before extending it to non-AIX sites.

After this principal decision, the integration of the SGI Altix system was delayed after a new version of MC-GPFS would be available for SGI Altix, too. The new version of MC-GPFS requires also a new release of the Operating System. SARA plans to upgrade its OS at the end of the year. After that the new version of MC-GPFS can be installed and SARA can be integrated immediately.

4.2 Configuration Work at the LRZ Site

Porting the new version of MC-GPFS for a SGI-Altix system required a different Operating Environment. Such a system could be provided by LRZ. So porting of the new release of MC-GPFS took place on a 64-node SGI-Altix at LRZ using the test file servers of RZG for data exchange.

Due to the short distance between LRZ and RZG this communications was almost like in a local network. So the performance of the data transfer achieved the physical disk-I/O limits of the test servers at RZG.

After the successful upgrade of the MC-GPFS software in the productive GPFS the next step is to integrate LRZ immediately after they open their SGI-Altix system to the public, which will be in November 2006. Since RZG is connected with LRZ comparable to a local network, LRZ will not provide its own disk servers, but use the disk servers provided by RZG.

4.3 Configuration Work at the BSC Site

Similar to SARA also the PPC Linux system at BSC was already integrated into the DEISA test cluster, which was used during the live demo at the EU-review in Paris in June 2006. The same considerations about stability and usability hold also for BSC. So it was decided to integrate BSC only after the upgrade to the new version of MC-GPFS.

For the PPC Linux system at BSC this new version of software is already available as natively supported release. But the system at BSC is a quite complex one, and currently runs with a specially adopted and tuned old release of MC-GPFS. So an upgrade to the new version is not as easy as it was with the AIX systems.

So first the new release has to be tested locally at BSC and then within DEISA before the PPC Linux system can be integrated. So the whole test sequence described in the last deliverable has to be applied again:

- Testing the network between BSC and RZG (firewall configuration)
- 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.

5. Showcase for the EU review meeting in Paris, June 2006

During the EU review meeting in Paris in June 2006 a live demonstration was presented, which coupled the AIX at RZG, the SGI Altix at SARA and the PPC Linux system at BSC. All sites provided disk space. In this showcase a job running at SARA used the data from the part of the GPFS located at RZG as input and wrote its output data to the GPFS hosted at BSC. Although this was done in a test environment with an Altix client still in the beta stadium, the demonstration again proved the concept of a Grid File System with transparent data access as a user

friendly, easy understandable solution for the HPC community. Thus MC-GPFS is a real added value provided by DEISA to its users.

6. Work for Future Configurations and Roadmap

6.1 *Integration of the non-AIX Sites in Production*

There are no fundamental restrictions concerning the integration of the missing non-AIX sites BSC and SARA. For both architectures, PPC Linux and SGI Altix the MC-GPFS software is available. The only obstacles come from the current productive use of the machines. The installation of the new release of MC-GPFS, which is required for the integration into the existing productive MC-GPFS, requires some major maintenance actions. This work will interrupt the local usability of the machine and also the use for the DECI projects within DEISA.

For BSC this means that first the new release of the MC-GPFS software has to be tested in a test configuration. If this proves to be stable and reliable the whole productive PPC Linux cluster will be upgraded. This upgrade is intended to happen around the end of the year 2006.

Concerning SARA the MC-GPFS software is already tested sufficiently, since it is in productive use at LRZ already. But the installation of the MC-GPFS software requires the upgrade of the operating system of the SGI Altix system. This has to be done in strong cooperation with SGI, which provides a special version of the Linux operating system for the Altix architecture. Currently it is planned to do the upgrade of the operating system and the installation of the MC-GPFS compatible with the DEISA productive GPFS also approximately at the end of the year 2006.

6.2 *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 most of the others the file servers are separated from the production system but connected to the local network. LRZ will use the disk servers provided by RZG because the distance is so short, that LRZ and RZG almost share a local network. Since EPCC/HPCx is not yet integrated into the MC-GPFS RZG hosts the home and data directories for the DECI users from EPCC/HPCx on the RZG-MC-GPFS.

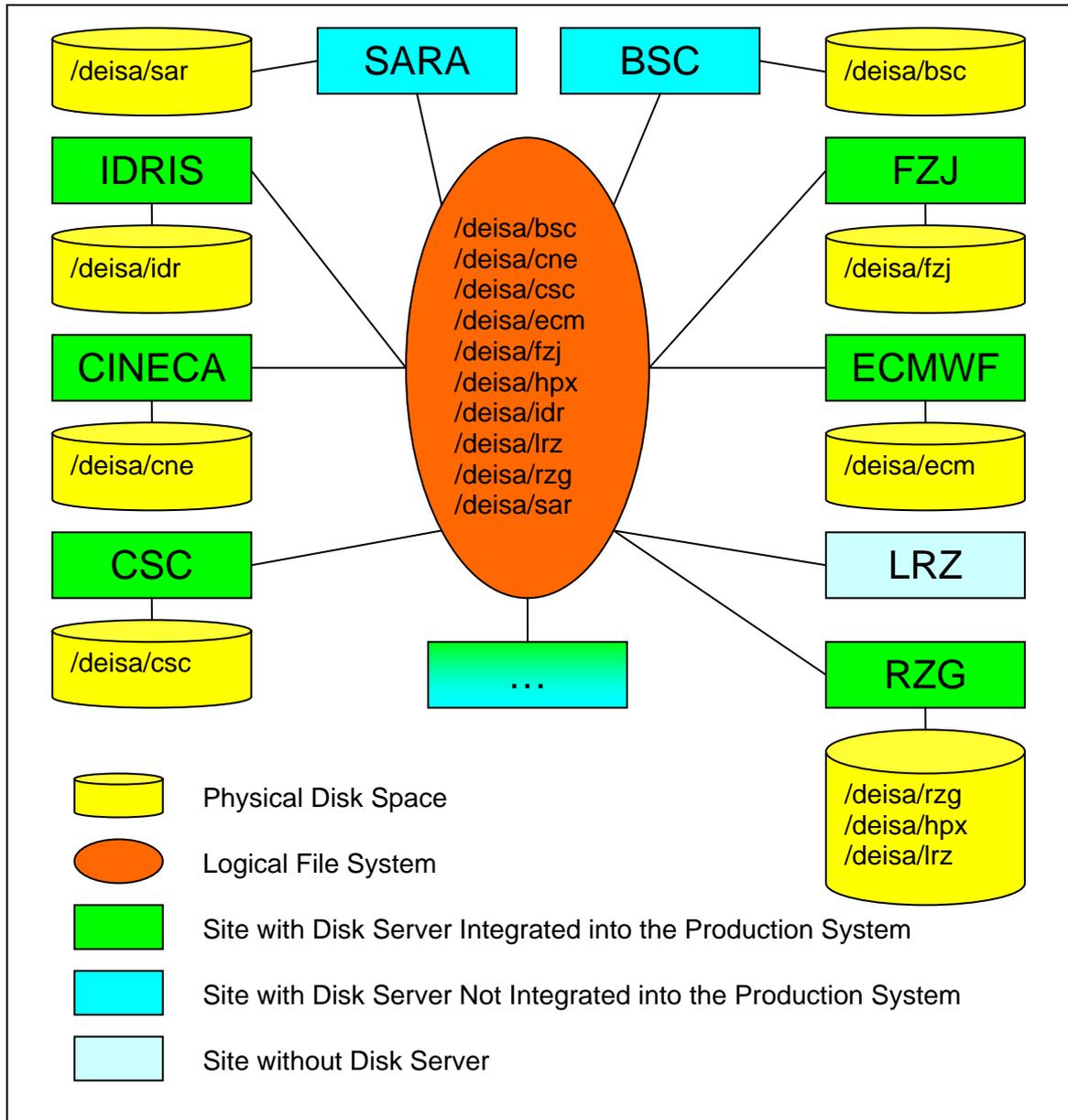


Figure 1: GPFS provision throughout DEISA