



CONTRACT NUMBER 508830

DEISA
DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Integration of SGI Altix System and
Hierarchical Storage Management for MC-GPFS

Deliverable ID: DEISA-SA2-6A
Due date: April, 30th, 2007
Actual delivery date: May, 25th, 2007
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	2
1.4	List of Acronyms and Abbreviations	2
2.	Evolution of Multi-Cluster GPFS for DEISA	3
2.1	Initial state and Planning	3
2.2	General Development.....	3
2.3	Hierarchical Storage Management	3
2.4	Integration of SGI Altix System at LRZ.....	4
2.5	Integration of AIX System at ECMWF	4
2.6	Integration of SARA.....	5
2.7	Integration of BSC	5
2.8	Integration of new Cray-system at CSC	5
3.	Work for Future Configurations and Roadmap	6
3.1	Configuration of the DEISA wide MC-GPFS	6
3.2	Future Configuration of the MC-GPFS throughout DEISA	6

List of Tables

Table 1:	Hardware configuration of the DEISA sites integrated with MC-GPFS.....	6
----------	--	---

1. Introduction

This deliverable summarizes the work done in the last half year (November 2006 – April 2007). The highlight is the integration of the SGI Altix system at LRZ, which is one of the most powerful computers in the world. After an upgrade to new processors in March 2007, which also implied changes in the MC-GPFS, more than 60 TFlop/s have been added to the DEISA cluster integrated by the transparent MC-GPFS.

The other main work was done in investigating the possible upgrade of the PPC-system at BSC. Beside the internal complexity due to the huge number of individual nodes in the BSC system this work was heavily influenced by the upgrading of the networking infrastructure to 10 Gbit/s. Thus, the production system could not yet be integrated, but it is planned to integrate the PPC-Linux-system in the last year of DEISA.

A next major work concerned the functionality of a hierarchical storage management. This feature was added on parts of the MC-GPFS of RZG. Furthermore the GPFS Test Suite was extended to better check the configurations of the MC-GPFS and also to deliver some performance data. This test suite is used occasionally for problem determination only, but it is intended to run on regular basis thus providing some sort of monitoring for the file system.

Last but not least, decisions about new procurements at SARA and CSC influenced the future planning. It was decided not to spend further work into the re-integration of SARA's old Altix, which will be switched off soon. Instead, the new Power-based Linux-system at SARA should be integrated as soon as possible. In addition, investigations started on the possible integration of the new Cray-system, which is installed at CSC.

Beside these efforts for the new systems available in DEISA, HLRS is working on opportunities to integrate the NEC SX-8 vector-based system into the DEISA-wide MC-GPFS environment.

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 SGI Altix System and Hierarchical Storage Management", is the sixth SA2 deliverable, describing the availability of the MC-GPFS on the non AIX sites LRZ in production and the implementation of a hierarchical Storage Management feature at RZG. Furthermore, a test suite is described and the planning for the integration of the BSC system in production and the new systems at SARA and CSC is explained.

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-SA2-5A
- [6] Deliverable D-SA2-3B
- [7] Deliverable D-SA2-4B

- [8] Acronyms and Abbreviations:
<http://cgi.snafu.de/ohei/user-cgi-bin/veramain-e.cgi>
- [9] 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. Afterwards, the document is 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	Central Processing 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 two years. In DEISA, this Grid file system provides transparent access to data just like with a local file system. The setup and configuration of these sites have been documented in the previous deliverables [2], [3], [4], and [5].

As described in the last deliverable [5], serious problems lead to the decision to upgrade MC-GPFS to a new version. This version was incompatible with the version on which the integration of SARA's Altix and the BSC-PPC-system was prepared. Thus, both systems could no longer be integrated in production after the upgrade. An upgrade of the MC-GPFS software on both of these systems would have been required, but this is not a simple straight forward task.

In addition, new challenging developments occurred, with SARA and CSC deciding for buying new system, both with an architecture or combination of hardware and operating system new to DEISA. The integration of these systems is of high importance, since the success to do so, proves the right way of DEISA, deciding for a common transparent file system as the comfortable data backbone for the users.

2.2 *General Development*

During the last half year and also in the remaining last year of DEISA, the underlying network infrastructure undergoes major changes. The initial 1 Gbit/s connectivity is upgraded to 10 Gbit/s. This has a major impact on the configuration of the networking parameters relevant for MC-GPFS. In principle, the new network speed stressed the existing file server structure to their limits. While MC-GPFS in the old structure was limited by the interconnecting network infrastructure, now, the bottleneck has been moved to other parts of the installed systems. Only latency effects are sometimes limiting the performance from the network side. With the 10 Gbit/s network connection in place, there is now room for the installation of future HPC Systems. These ongoing network upgrades require careful observation of the MC-GPFS and its configuration. For this reason, the GPFS Test Suite was extended to better check for such parameters and also to provide some sort of performance measurements. Currently, the test suite is available for the AIX systems and used occasionally. It will be adapted for the other integrated non-AIX systems. Furthermore, an integration of this software into some general system monitoring software, like the already used INCA, is considered.

2.3 *Hierarchical Storage Management*

Although, currently the online disk space on the DEISA-wide MC-GPFS is sufficient for the productive use, in the future a need for a hierarchical storage management is quite probable. Thus, for future development, one part of the MC-GPFS located on the file servers of RZG has been reconfigured to move data from online disk space to tapes and allow retrieval of that migrated data transparently by simply accessing the file in the file system.

This feature is at the moment only available on that specific part of the DEISA-wide file system. All users of RZG get a directory there automatically during account creation. To provide this feature for all DEISA users, each user not originating from RZG, will get a directory with the migrating feature enabled on the RZG part of MC-

GPFS on request. They can access this data transparently using the path to their directory created on the RZG part of the globally visible file system in the same way as they access the data on their other MC-GPFS directories. Since the data may be offline there is a need for pre-staging mechanisms for batch processing together with SA3.

2.4 Integration of SGI Altix System at LRZ

Since the new version of MC-GPFS is incompatible with the old version, the upgrade of MC-GPFS was stopping the integration of the Altix at SARA with their old version of the software. Therefore, a porting of the new version was required. The original port at SARA had been done for the version 2.3 for Red Hat Linux with a 2.4 kernel. It was decided, that the porting of the new version of MC-GPFS was best to be done already for a Linux kernel version of 2.6, which was installed with SLES 10 in the Altix machine at LRZ. Thus the porting was done at LRZ. Later, after a planned upgrade of the Operating System of the Altix at SARA, this port could then also be used on that machine.

So, already during the acceptance phase of the LRZ Altix, the porting of MC-GPFS started on a small test partition. In principle, it worked the same way as for the former porting at the Altix at SARA. The test cluster at RZG was used as counterpart for the development. Due to the near neighbourhood of RZG to LRZ, RZG also provided the file server disk space for the MC-GPFS of LRZ, since the port only covered a client for accessing MC-GPFS.

After a successful setup of MC-GPFS between the LRZ test partition and the RZG test machine, other test machines at IDRIS and FZJ have been connected. Even the initial testing environment of ECMWF was already integrated (see next section) Thus, the WAN functionality was tested and the adaptation of the networking parameters for MC-GPFS had been performed. All these tests have proven the possibility of a stable integration of the SGI-Altix at LRZ.

Next, the ported MC-GPFS client was installed on the production partitions of the LRZ machine. After that, all the test clusters of FZJ, IDRIS and RZG have been mounted by the 16 production partitions of the SGI-Altix at LRZ in January 2007.

Finally, in February 2007, the LRZ machine was the fully integrated into the existing DEISA common file system production infrastructure as the first non-AIX machine using the new enhanced GPFS version.

In March 2007, the Altix machine at LRZ was upgraded from old Madison processors to newer and faster Montecito processors. After the upgrade, MC-GPFS did no longer work. A detailed analysis showed, that the slightly different internal architecture of the new processors lead to a different thread handling. After the adaptation to these new conditions, the MC-GPFS client again could be started. Thus, by end of March the Altix at LRZ is the biggest and fastest machine integrated in the DEISA wide global file system, delivering more than 60 TFlop/s.

2.5 Integration of AIX System at ECMWF

As already explained in the last deliverable [4], the integration of ECMWF was a quite straight forward task after ECMWF had been connected to the DEISA network with a dedicated 1 Gbit/s link.

First, the ECMWF test machine had been connected to the RZG-test machine and later to IDRIS, FZJ and even the LRZ in order to check the parameters and find out about potential problems.

Soon after these successfully finished tests, ECMWF could be integrated into the production environment of MC-GPFS in DEISA.

2.6 Integration of SARA

Conforming to the initial planning of the DEISA common file system infrastructure, the SGI-Altix of SARA could be integrated soon after the EU Review Meeting in June 2006. Then, the necessary upgrade of the MC-GPFS version lead to the situation, that the system could not longer be integrated due to the incompatibility of the two version of the software. A port of the new version was done on the LRZ machine for a newer kernel version. Thus, a re-integration of SARA's Altix would only be possible after an upgrade of the Operating System.

In the meantime, SARA decided a procurement for a new machine. The new system is based on Power processors running Linux. This machine will be delivered mid of 2007. At the same time, the Altix system will be switched off. Thus, it did not make sense to put much effort in the re-integration of the Altix system, which will disappear soon. Therefore, no further work was done for the Altix system, but it is planned to integrate the new Power-based system into the DEISA wide MC-GPFS as soon as possible.

2.7 Integration of BSC

The same problem as for SARA is in principle also valid for the BSC system. Everything was prepared for an integration of that system into the MC-GPFS of DEISA. But the incompatible version upgrade of MC-GPFS stopped the progress.

Although, the new version of MC-GPFS is available for PPC-Linux, a simple installation was considered to be too risky. The reason is the work which had been invested to make the old version stable and reliable locally. Thus, intensive compatibility and stability test are need before any upgrade to be performed at BSC. This work is still in progress.

In addition, BSC had undergone a major upgrade of its system and furthermore, the network connection has been switched to 10 Gbit/s. The latter one introduced some problems, which are meanwhile solved. So, the further steps of testing, described in the last deliverable in detail, will proceed. It is still planned to integrate the BSC system into the MC-GPFS infrastructure in the last year of the DEISA project.

2.8 Integration of new Cray-system at CSC

Not only SARA, but also CSC had a procurement, in which they decided for a hardware new to DEISA. For the Cray-architecture, CSC has decided for, no GPFS-software is available yet. Thus, a new porting of the software will be required. Therefore investigations and discussions with IBM have been initialized in order to be able, to also integrate this system into the MC-GPFS of DEISA.

3. Work for Future Configurations and Roadmap

3.1 Configuration of the DEISA wide MC-GPFS

Only six sites offer a locally configured MC-GPFS, which is exported to the other 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 [9]).

Due to the near neighbourhood, LRZ does not provide its own disk space, but uses the servers of RZG. Also for users of sites, not yet integrated into the MC-GPFS infrastructure, the file servers at RZG provide the respective home and data directories.

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	Number of Fileservers	Storage TByte	Compute CPUs (Type, GHz)	Memory GB	TFlops
CINECA	2	5	480 (Power5, 1.9)	1152	2.6
CSC	2	2	416 (Power4, 1.1)	672	2.2
ECMWF	2	1	2640 (Power5+, 1.9)	2250	20.1
FZJ	2	4	1288 (Power4, 1.7)	5152	8.9
IDRIS	2	2	1024 (Power4, 1.3)	3136	6.7
LRZ	0	0	9728 (Montecito, 1.6)	39064	62.3
RZG	2	7	928 (Power4, 1.3)	2368	4.6

Table 1: Current hardware configuration of the DEISA sites integrated with MC-GPFS

3.2 Future Configuration of the MC-GPFS throughout DEISA

As already mentioned above, SA2 will make all efforts to integrate all DEISA sites into the MC-GPFS. Actually, all architectures promised to be integrated are integrated. This covers the AIX-systems and the SGI-Altix at LRZ. The PPC-system at BSC is still under testing and we are still confident to be able to integrate it.

Since HPCx will be connected to the 10 Gbit/s network, it is considered to integrate the AIX system there also into the MC-GPFS. This should actually be a straight forward task, to happen as soon as the dedicated network connection is available. This is expected for June 2007.

Additionally, all efforts will be made to integrate the new Power-based system at SARA, which substitutes the Altix, as soon as possible. And last but not least the integration of the completely new Cray-architecture at CSC is considered with top priority.