



CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Performance Improvements of AFS in a Wide-Area Network and a
HPC-Environment
Deliverable ID: D-SA2-4B

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 Content

Table of Content.....	2
List of Figures.....	2
List of Tables.....	2
1. Introduction.....	3
1.1 Executive Summary.....	3
1.2 References and Applicable Documents	3
1.3 Document Amendment Procedure	4
1.4 List of Acronyms and Abbreviations	5
2. Description and Configuration	5
2.1 Introduction.....	5
2.2 Advances in Performance Improvement in the WAN Environment.....	5
2.3 Improvements in the LAN Environment.....	6
2.4 Future Developments	7

List of Figures

Figure 1 Fast file access on a common GPFS between AFS-Server and Client.

List of Tables

1. Introduction

1.1 *Executive Summary*

The Service Activity 2 within the DEISA project deals with the connectivity of all DEISA-sites on the file system level. Two strategies are pursued in parallel:

- Deploying IBM's GPFS. See deliverable D-SA2-4A [2].
- Implementing a distributed file system structure for heterogeneous environments.

This document discusses the research into performance improvements of the AFS in wide area networks.

1.2 *References and Applicable Documents*

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

1.3 *Document Amendment Procedure*

Not applicable.

1.4 List of Acronyms and Abbreviations

AFS	Andrew File System, used in the open-source implementation OpenAFS
AIX	Advanced Interactive eXecutive (IBM's derivative of UNIX OS)
Altix	Multi-Processor compute node from SGI
AMD	CPU-vendor
DEISA	Distributed European Infrastructure for Supercomputing Applications
DFN	Deutsches Forschungs-Netz, German NREN
FS	File System
GPFS	General Parallel File System, proprietary FS from IBM.
IBM	Computer Manufacturer
Intel	CPU-vendor
Itanium	64bit based CPU manufactured by Intel.
IPP	Institut für Plasmaphysik
IPP-HGW	Institut für Plasmaphysik, Section Greifswald
Linux	Free, open source UNIX version
MC-GPFS	Multi Cluster GPFS
MR-AFS	Multi-Resident AFS; enhanced AFS, providing hierarchical storage management
NREN	National Research and Education Network
Opteron	64bit-CPU manufactured by AMD
OS	Operating System
RX	Network traffic protocol, based on the same RFC as TCP, but uses UDP
SGI	Computer Vendor
TCP	Transmission Control Protocol, a network protocol
UNIX	Operating system family
UDP	User Datagram Protocol, a network protocol
WAN	Wide area network

2. Description and Configuration

2.1 Introduction

Due to the availability of MC-GPFS on all major computer-platforms used within the DEISA-project, the work on AFS has been reduced and priority is given to MC-GPFS. Thus, the AFS-clients on the DEISA sites other than RZG have not been changed in the last 18 months. Similarly, there was no need to update the server software. Nevertheless, work has been done on improving the performance in both LAN and WAN environments. The sites involved in this development were RZG and the IPP-HGW in Greifswald, Germany, which is partly under the administration of RZG. Both sites have much experience with AFS which is important for this kind of improvement and testing, and have a geographical distance of several hundreds of kilometres.

The role of AFS as a fall-back solution for MC-GPFS has to be defined carefully. There are three situations which require attention:

- MC-GPFS fails remotely, but works locally, the dedicated WAN network is intact.
- MC-GPFS fails both remotely and locally.
- MC-GPFS fails remotely, but works locally, the dedicated WAN network is down.

In the worst and very unlikely case of complete MC-GPFS failure at a site, the data of the user are inaccessible and a full shutdown of the site in order to fix this problem with the help of IBM is expected.

Thus the presumption is valid that when AFS will step in as a fallback solution, the local GPFS is still accessible.

From the user's point of view, the file paths under which the data reside should not change and should be the same regardless if it accesses the files locally or remotely. The only thing allowed to change is the number of files accessible.

Important user-data will have to be copied then into the AFS-specific regions of the GPFS, and then be exported via the AFS-Server both to local and remote computing nodes.

In order to keep the user-data consistent, the direct use of the local GPFS has to be prohibited, and AFS has to be used also locally.

Besides some improvements in the WAN discussed in section 2.2, performance improvements of AFS within a local HPC-cluster are described in section 2.3.

Last but not least, the interruptions for the user should be taken into account, which will be examined in section 2.4

2.2 Advances in Performance Improvement in the WAN Environment

In the WAN environment, further investigations into the fine-tuning of network parameters were carried out. The network connectivity between RZG and IPP-HGW is at the moment not comparable with the connectivity between DEISA sites. In DEISA the sites are connected using a premium service of the NRENs which is close to a dedicated point to point connection, whereas RZG and IPP-HGW are connected by the German NREN DFN without any special treatment.

This situation will change later this year, when RZG and IPP-HGW will establish a direct connection.

However, it is necessary to test both of these connection types between RZG and IPP-HGW in order to optimise AFS as fallback for the MC-GPFS with a simultaneous breakdown of the dedicated DEISA network and without such a network-failure.

AFS uses the so-called RX-Protocol on top of UDP. This RX-protocol is based on the same RFC as TCP, but is more efficient on the server-side when dealing with many clients.

The performance of network traffic on a WAN is governed by latency and packet-loss, it is likely that the RX-window-size and the number of parallel streams play a role in improvements of the performance. Both these aspects are taken into account for MC-GPFS and GridFTP, but implemented differently.

As described in deliverable D-2A-2B, the maximum window-size for AFS per call has been increased from 32 to 128 packets resulting in the expected increase of throughput by a factor of four. The feasible maximum is at 256 packets, which results in another improvement factor of 1.5. Higher window-sizes require non-trivial changes to the source-code. The window size governs the number of packets which are in transit, which means no ACK-packet has to be returned to the sender for this number of packets. The two main differences between the dedicated network and the normal internet connection are the round-trip time and the number of packet-losses per time.

The value mentioned above, however, has been done with an underlying dedicated network. When the point to point connection between RZG and IPP-HGW is established, then a thorough investigation to higher window sizes beyond 128 packets can be carried out. The tests involve deep changes in the AFS-clients and likely the UDP buffer sizes of the OS, so that the usage of the AFS-client at CINECA for these tests has not been taken into consideration.

The implementation of parallel streams into AFS is still an objective for the future, but is not currently implemented.

2.3 Improvements in the LAN Environment

As mentioned in the introduction, in case of a problem there is reason to assume that the local GPFS is still accessible, but the remote is not. In such a case, all file access of the user has to be done through the AFS-interface in order to use the same file paths locally and remote.

In order to optimise AFS-performance in the LAN environment, an extension to the AFS-client has been developed which allows the use of a local cluster file system as a basis for the AFS. Like this, the AFS-server just passes the meta-data of a given file to the client and the client uses the underlying cluster file system, here the GPFS, to obtain the actual data. Thus the data is accessible with nearly native GPFS-speed. A schematic view on this is shown in figure 1.

