

CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Availability and Performance of Multiple-Cluster-GPFS
(MC-GPFS)

Deliverable ID: DEISA-SA2-2A
Due date: April, 30, 2005
Actual delivery date: May 15, 2005
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)	

Document Keywords and Abstract

Keywords:	DEISA, HPC, Grid, GPFS
Abstract:	This document reports about the operation, performance and availability of the first official release of MC-GPFS for the AIX-5.2 operating system, installed at the four core sites FZJ, CINECA, RZG and IDRIS.

Table of Contents

Table of Contents	1
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	3
1.4 List of Acronyms and Abbreviations	3
2. Multi-Cluster GPFS for DEISA	5
2.1 Introduction	5
2.2 Benchmarks	6
2.3 Local implementation and performance of MC-GPFS	7
2.4 Network architecture and performance of MC-GPFS	9
2.5 Performance Results	10
2.6 Changes in the Setup of the Dedicated Test Systems	13
2.7 Preparation of the Migration of the Production Systems	14
2.8 Integration of Heterogeneous Architectures	16
APPENDIX A: NETWORK DEBUGGING	16

List of Figures

- Figure 1Common File System Structure at the four core sites
- Figure 2Hardware Configuration of MC -GPFS Fileservers at RZG.
- Figure 3RoundRobin Striping of a single fileserver.
- Figure 4Random striping of a single fileserver.
- Figure 5Balanced Random Striping of a single fileserver.
- Figure 6Schematic view of the network-connections between clients and servers.
- Figure 7Local disk performance at all core sites.
- Figure 8Remote file access performance from CINECA.
- Figure 9Remote file access performance from FZJ.
- Figure 10Remote file access performance from IDRIS.
- Figure 11Remote file access performance from RZG.
- Figure 12Future Configuration of the Regatta system at the RZG in May 2005.
- Figure 13Network throughput measurement from RZG to IDRIS.
- Figure 14Repeated network throughput measurement from RZG to IDRIS.

List of Tables

- Table 1 Current hardware configuration of the different sites providing MC -GPFS.

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, "Availability and Performance of MC-GPFS", is the second SA2 deliverable, describing the availability and the performance of the MC-GPFS 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-SA2-2A
- [3] Iozone: <http://www.iozone.org>
- [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)
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

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. Multi-Cluster GPFS for DEISA

2.1 Introduction

The four core-sites of DEISA, facilitating IBM's HPC systems, share a Multi-Cluster General Parallel File System (MC-GPFS) to be used as a Grid File System. The basic technology of MC-GPFS and the proof of functionality of a pre-release has already been described in the last deliverable D-SA2-2A.

This deliverable goes one step further and discusses the realisation of the production release, i.e. its configuration and interaction with the underlying network infrastructure and last but not least its performance.

Configuration

Each of the core-sites has a local configured MC-GPFS, which is exported to the other core-sites. Thus, strictly speaking DEISA does not use a single MC-GPFS over all core-sites, but many, which are interwoven in such a way that they appear to be a single, shared file system. However, each user uses only the branch of his/her "home" GPFS. The figure 1 below describes this situation. Whenever one site increases its file space, the total virtual file space of the DEISA project is increased, but only the users of the site increasing the disk space benefit from this, because users of other sites usually do not use this branch of the common MC -GPFS tree.

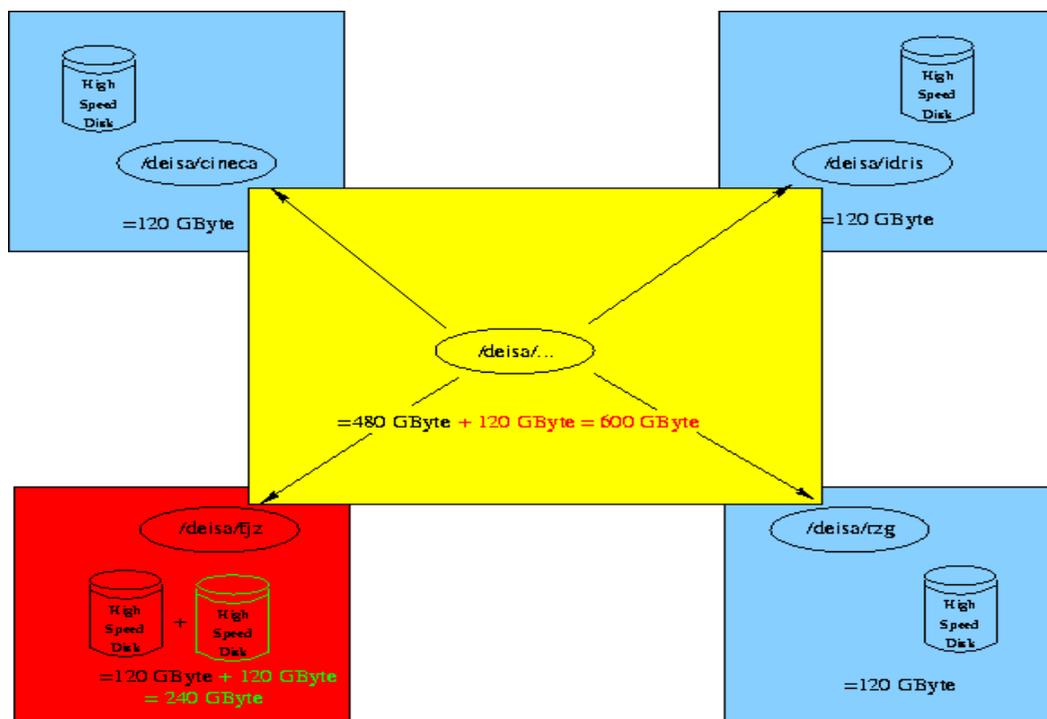


Figure 1: Common File System Structure at the four core sites, numbers are fictitious

Performance

It is necessary for redundancy and performance that each site uses a number of file servers to provide its MC-GPFS. Table 1 gives an overview of the current situation.

Site	Number of Fileservers	Provided Storage
CINECA	3	1086 GByte
FZJ	3	1230 GByte
IDRIS	3	1355 GByte
RZG	2	6966 GByte

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

During the tests, nearly the full 1 Gbit/s bandwidth of the network could be satisfied. The availability of the MC -GPFS was during this test phase very high, outages were always due to reconfiguration and maintenance at the different DEISA-sites. The performance of the MC-GPFS is governed by various issues. First of all, the network has to be stable and the packet loss has to be extremely small. Secondly, the network topology needs to be optimized for high-speed traffic. Thereafter, the locally attached storage has been configured and redundantly attached to the file servers. Moreover, the software configuration of the different single MC-GPFS were harmonised with all core-sites. As an additional test, a local machine at RZG was configured to be a remote client, so that it acts exactly like a client at IDRIS or any other core site, without the WAN. Like this, the effect of the WAN could be estimated. The realisation of the MC -GPFS infrastructure was test-driven, i.e. each step has to stand specific tests so that problems or bottlenecks could be identified at an early stage. The tests mentioned above were purely focussed on peak-performance and stability, thus only a small number of users had accessed only few files. After testing the basics of the MC-GPFS, a real application was successfully run by a member of JRA3, making use of the full network speed. However it has to be said, that the tuning of the file system was based on the sequential file access solely. Tuning to random access patterns would require good statistics and therefore too much time without the guarantee of a significant improvement. Thus, the last test is to show resilience only, not performance under heavy random access. The subsequent sections focus on the different implementation levels mentioned above, discussing the realisation and test-results in detail.

2.2 Benchmarks

Network benchmark

The basic network tests were done by members of SA1 using among others a tool called "netperf", which can measure the bandwidth of a TCP or UDP stream between different hosts. Since MC -GPFS is based on TCP, it is essential that as few packets as possible get lost in transmission. After the first disappointing results at the GPFS-level from RZG to IDRIS, the network connection between these two sites was tested thoroughly and a

problem in one of the routers was found. The details about this search are given in Appendix A.

File I/O benchmark

The file system tests were done with the well-known I/O-benchmark "iozone" [3]. It is a very flexible benchmark which tests various file-access patterns from one or more hosts simultaneously on one or more files.

The real file access pattern within the DEISA-production lies between two extremes: On the one hand, few large files are accessed sequentially, this may happen when a computational job is being started at a remote site and wants to read its input or write its output into the MC-GPFS. On the other hand, many small files are accessed simultaneously, which happens using some other applications.

These two extremes were simulated by different usage of iozone.

In order to get the peak-performance, only the sequential read and write tests of large files were run. The test-files were always large enough to eliminate caching effects on the client side.

Two different accesses were used:

- ? One client reads/writes a single file.
- ? Multiple clients read/write a single file.

The other two cases, one client writes multiple files and multiple clients write more than one file show the same behaviour as the two test cases above and are thus ignored here.

For testing the resilience, many instances of iozone are started on many clients accessing many small files simultaneously.

While the benchmark is running, the node is monitored by "iostat" and/or "nmon" to see the limiting factor of the test. In the optimal case, the performance is only limited by the network speed.

Application I/O Benchmark

The highest level of benchmark is of course a real application, started by a real user. The ORB code solves equations pertinent to the study of transport related instabilities and turbulences on fusion relevant plasmas.

Initiated at the CRPP, the ORB code has been substantially enhanced at IPP. It is a massively parallel Monte-Carlo code, simulating at least 100 million particles. This code will be available to all users within the DEISA production environment. This code, like many others, stores information between two successive runs in so called "restart" files. For "Grand Challenge" runs, which are aimed at within DEISA, these files are of order of tens of GBytes. Thus, it is absolutely vital for DEISA to guarantee fast file access between the remote sites, otherwise too much CPU-time would be wasted in I/O operations.

2.3 Local implementation and performance of MC-GPFS

The local implementation depends heavily on the available hardware, which differs from site to site. Thus, only the implementation at RZG is described in figure 2.

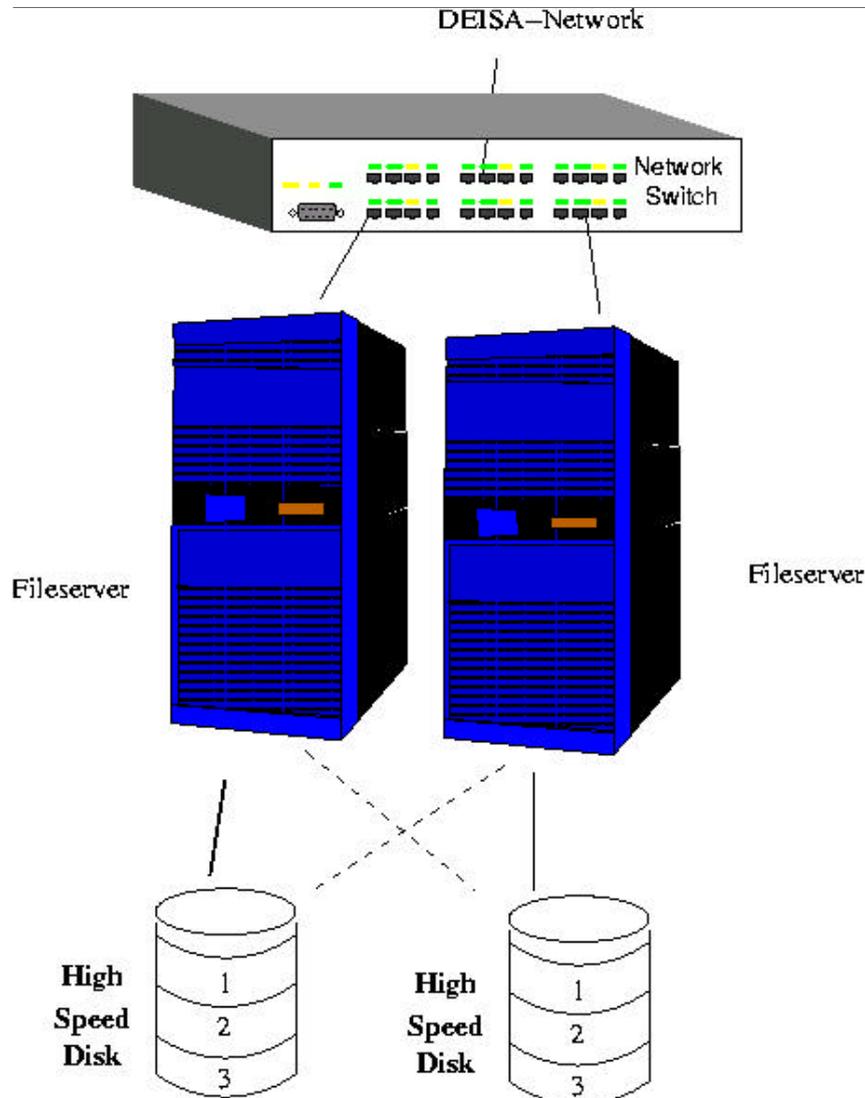


Figure 2: Hardware Configuration of MC-GPFS Fileservers at RZG. Two file servers see both storage systems. Each storage system consists of three RAID 5 systems.

The hardware-redundancy shown in figure 2 is necessary for providing high performance and guaranteeing availability. At the RZG, two file servers are twin tailed-attached to two storage systems which consists of three RAID 5 systems.

Using more than one fileserver and storage system enhances the performance by a technique called "striping".

Enhancing performance

Striping enhances the flow of information by using multiple streams of data, which can be processed in parallel.

Here, there are three levels of striping, one on the network layer – half the data goes to fileserver one, the other half to fileserver 2 – and two levels of striping on the disk level. Firstly, the file system is not written contiguously on one disk, but distributed in "stripes" across the three RAID 5 systems. The second level of striping is the distribution of data

across the individual disks in a RAID 5 system. RAID 5 is a common technology and allows high throughput combined with some fail-safety on commodity hardware (here: SATA-disks).

The actual implementation of the disk and the striping is described in more details in Appendix B.

Guaranteeing Availability

Replication of data

Besides using multiple disk- or RAID systems for striping, MC-GPFS offers redundancy on the data level via “replication”. It allows up to two copies for data and for metadata on different storage systems. That means, while storing data, GPFS writes it in up to three different locations, so when one RAID fails, the data can still be retrieved from one of the two other RAIDs. This implies of course a performance-degradation and moreover decreases the maximum file size GPFS is able to handle.

At the RZG, the data replication is turned off by default, the failover capacity of the RAID 5 – systems is considered safe enough, based on the experience of other exported file systems at the RZG. The much more sensitive metadata, however is replicated within the GPFS, since a failure in the metadata is much more dangerous than in the data.

Replication of fileserver: automatic failover capability

However, the twin tailed server configuration provides redundancy on the fileserver-level. Whenever one of the servers goes down, the other takes over the disks and exports its data as well. This is marked in the figure 2 above by the dashed lines.

Failure Groups

Failure groups are a collection of disks which share a common single point of failure. A single point of failure is one or more nodes exporting the disks. In this case, the disks are connected to two servers. Thus, only if both servers fail, then this group of disks is not accessible anymore.

Quorum

In order to protect the GPFS in case of massive failures, the whole GPFS is only available when more than 50% plus one of all so-called “quorum” nodes is reachable, although special definitions are allowed.

It makes sense to only specify fileserver as “quorum” nodes since clients are usually rebooted more often and should not be able to bring down the GPFS although all servers are still up and running.

2.4 Network architecture and performance of MC-GPFS

Figure 3 shows a schematic view of the network streams between the clients and the servers. Each client opens its own connection to each fileserver. Thus, there are always (Number of clients) x (Number of servers), in the example $3 \times 2 = 6$, TCP-streams. The number of files opened by the client does not affect this. GPFS handles this on the protocol-layer inside the TCP-stream. The advantage of having more than one TCP-stream from each fileserver is embedded in the TCP-protocol. The TCP-protocol

provides an intrinsic flow control to adjust automatically the speed of the connection. Whenever a packet gets lost, the speed of the connection is automatically reduced. This is not a problem if only very few packets get lost, but if too many packets are missing, the impact is severe, slowing down the network rate by a significant amount. Thus, in practice it is better to have multiple slower streams than one high speed stream. In total, the slower streams outperform the singular high speed stream.

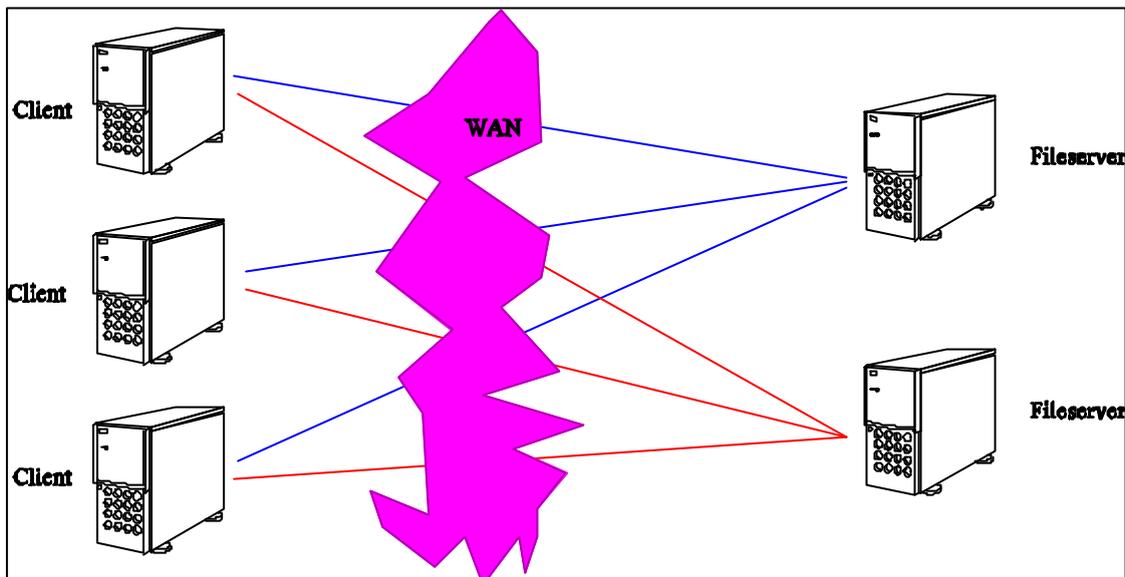


Figure 6: Schematic view of the network-connections between the clients and servers

2.5 Performance Results

This section discusses the results of the I/O-benchmark “io-zone” both locally and remotely. The results vary from site to site, since each site has a different configuration. First, the local disk performance has been tested. This tests, except the one marked RZG*, are run on the file servers of each site, that means the nodes have been servers and clients at the same time. Thus the performance measured here is a mixture between the speed from one node to the storage system as well as the internal network speed. Since all file servers on all sites are connected by IBM’s fast Federation Switch, the two speeds are comparable. In other words, the network interconnect should not slow down the speed of the GPFS. The site marked “RZG*” is a special case: one node which is locally connected at the RZG via a 1Gbit/s Ethernet are configured as remote clients. Thus, from the GPFS point of view they are treated as if they are remote, but without the latency of the WAN.

Like this, one can see the difference cause by the WAN.

As one can easily see, most speeds are above the network speed of 1 Gbit/sec. Thus, it is ensured that the local disks are almost always fast enough for the dedicated network.

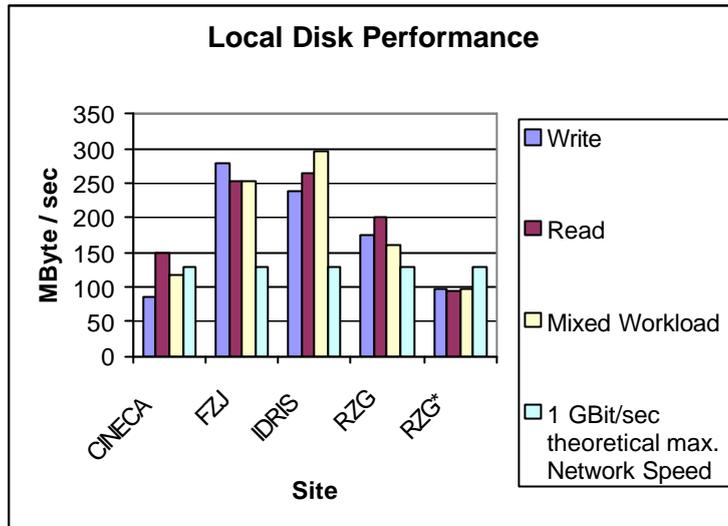


Figure 7: Local disk performance at each site. The site marked as “RZG*” denotes a client physically located at the RZG, but mounted the RZG file servers as remote.

The special “local-remote” case “RZG*” shows that the 1 Gbit/s network is nearly filled up by a single client. Thus, the overhead of mounting a remote site as opposed to a local site within the GPFS is negligible. This means that the GPFS has principally the same performance locally and remote as long as the network is the same speed.

The next four graphs show the I/O-performance of each site to each other site. The site on which the test is run is given in the graph-title. The number of clients used at each site is given in the figure-caption.

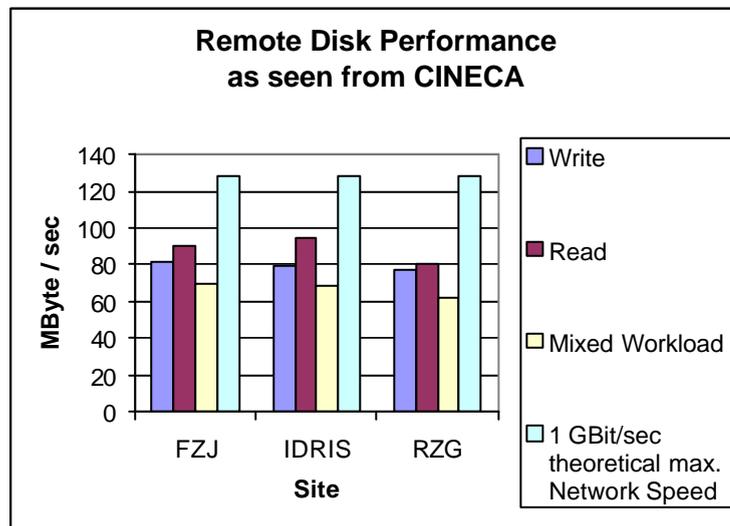


Figure 8: File access performance from 2 Clients at CINECA to the other three sites

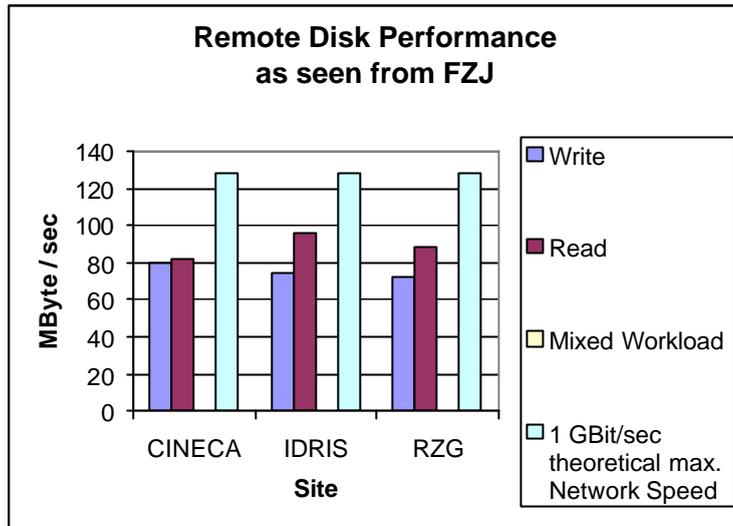


Figure 9: File access performance from 4 Clients at FZJ to the other three sites

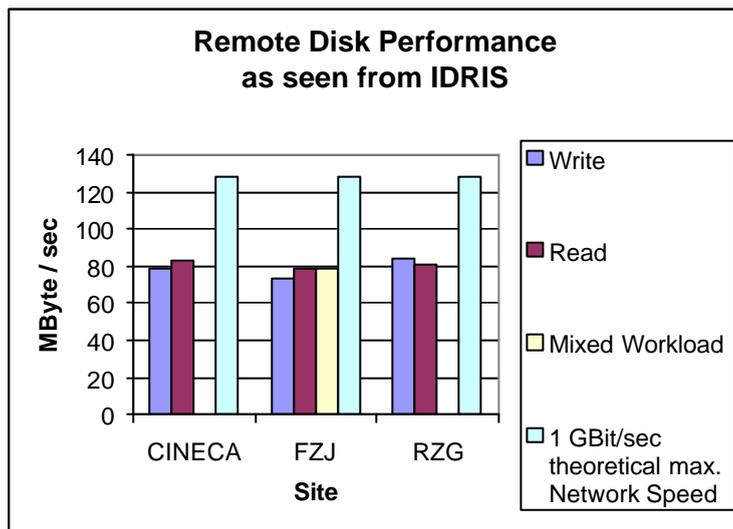


Figure 10: File access performance from 2 Clients at IDRIS to the other three sites

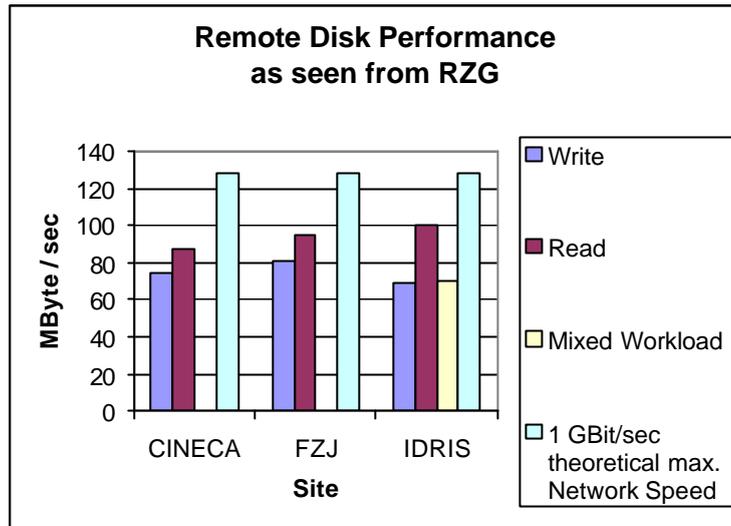


Figure 11: File access performance from 2 Clients at RZG to the other three sites

As one can see, the performance over the network is not as good as the local performance. This is natural, since for the local tests, the network speed is higher and the network latency is significantly lower than for the remote tests. Some of the remote tests, fill up the complete wire, some do not perform that well. The reasons for this are subject to the ongoing investigation. On average, however the file performance is acceptable. One must always bear in mind that only very few clients and thus network streams were involved in these measurements. In the real production systems, when tens of clients are accessing the GPFS the performance of a single network stream is not as significant as it is in these test. Thus, the assumption is justified that in the production system the GPFS fills up the network speed to the theoretical limit.

File access by a real application

The ORB-code, explained above, was running already on three (FZJ, IDRIS, RZG) of the four core-sites with remote file access. The file access times seen with the application were in agreement with the results given in the previous section. The details of these runs are given in the deliverable of JRA3.

2.6 Changes in the Setup of the Dedicated Test Systems

In order to test different I/O-configurations, e.g. for the later use in the production system, the dedicated some reconfigurations have been made to test systems.

RZG has connected 7TB of new FC-SATA disk space to two servers. These servers have in the meantime already been integrated into the Federation Switch Communication Network, thus providing the best possible I/O performance. Each system can takeover the disks of the other machine in order to guarantee an uninterrupted access to the data in the case that one machine is down for maintenance. Furthermore

each machine is directly connected to the dedicated DEISA-network with one 1Gbit/s-Ethernet uplink. This is sufficient to saturate the disks also with a remote I/O operation.

The two other machines could not be connected to the Federation Switch. They only have a 1Gbit/s connection. These machines are currently used to locally emulate a remote cluster, thus being able to test the influence of the DEISA-WAN communication in comparison to a local 1Gbit/s network. Additionally these machines are used to test the MC-functionality of the LoadLeveler by being able to move jobs between the two local clusters as LoadLeveler-Administrator (required by SA3).

Similarly, IDRIS provides FC-disk space of 1TB with two servers also connected to the Federation Switch. Additional test-file systems are connected to other machines and exported to the DEISA test environment.

The configuration of CINECA has not undergone major changes. It provides now also about 1TB on SSA-disks.

The same holds for FZJ, which offers also about 1TB on internal disks distributed over the four LPARs.

2.7 Preparation of the Migration of the Production Systems

IDRIS prepared a complete new test-cluster similar to the DEISA-test-cluster which will be used to test the migration of the production system to the state required by DEISA. This includes starting from a software and network environment equal to the current production system - but not part of it - the upgrade of the software, e.g. the GPFS from the current production version 2.1 to the new required MC-GPFS version 2.3. Then, the address-configuration of the machines and the GPFS has to be modified to be able to be part of the dedicated DEISA network.

Although there is a paper with instructions from IBM, this is a very critical step, because the data on the production system must not be corrupted by this migration process. Thus intense tests are required to be sure that the migration of the real production system is hassle free. It is intended to upgrade GPFS of the productions systems in April and to switch to the DEISA-network also in April or May. After these two steps the IDRIS production system will be fully available to DEISA.

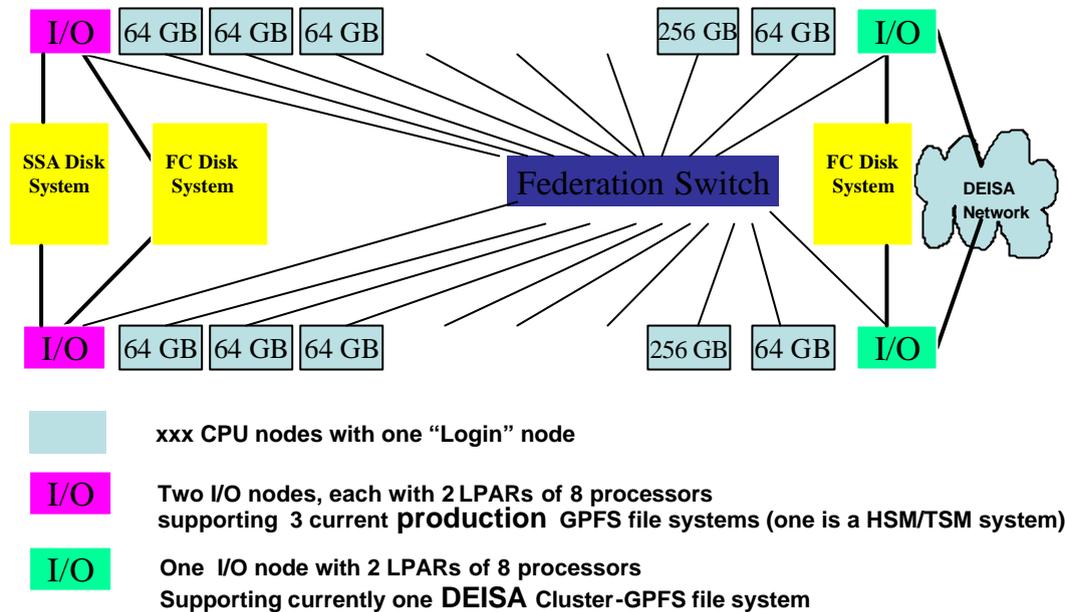


Figure 12: Future Configuration of the Regatta system at the RZG in May 2005

The above described modification of the RZG test systems is also the first step of the migration of the RZG production system to be part of DEISA. The two I/O-servers with 7TB of disk space will later be the DEISA-I/O-servers of the production system. The same steps as in IDRIS are required in RZG.

The GPFS software has to be upgraded and the production network including the GPFS has to be migrated to the DEISA-address-range. The tests performed at IDRIS will help to do the migration similarly. At RZG there is a software-add-on product on top of GPFS, the HSM/TSM functionality, which is able to move files transparently from disk to tape. Unfortunately, the support of this product for the DEISA required version of GPFS will be available not earlier than May.

Since only one version of GPFS can run in a cluster at one time, RZG cannot migrate the production system earlier than this product is available. But RZG will migrate immediately after that. Thus RZG should have moved the production system to DEISA at the end of May.

CINECA currently uses no HSM/TSM functionality and will upgrade, after IDRIS has proved the migration process, as soon as the production requirements will allow the upgrade and reconfiguration of the production system.

FZJ is very heavily relying on the HSM/TSM functionality in their production cluster, since the one and only GPFS uses this feature. The migration process of the FZJ production system to be fully integrated into the DEISA setup, will take some more time and will take place, after the HSM/TSM functionality has been certified by RZG. This will be done when the production requirements allow the upgrade and reconfiguration of the production system.

Security

In order to improve security in the whole cluster all GPFS filesystems are mounted with the "nosuid" option turned on. This is required not only for the remote GPFS filesystems but also for the local ones, which are exported. Thus, the potential compromise of one site cannot propagate to other sites via the GPFS. Additionally the feature of UID-mapping was already successfully tested for the root user, thus restricting super-user-rights to the local site owning the GPFS filesystem.

2.8 Integration of Heterogeneous Architectures

While IBM offers GPFS for both AIX on pSeries and Linux (PowerPC, IA32, and AMD64), the DEISA consortium wants to extend the use of GPFS also to other Hardware. Therefore a port of the MC-GPFS-client for the ALTIX SGI-machine, based on IA64 processors, at SARA is required. Meanwhile there have been already some successful tests for this GPFS-client.

Thus it will be possible to fully integrate SARA into the GPFS of the AIX-systems of the four core sites, so that programs on the ALTIX machine will be able to read and write data into that global file system of DEISA.

For the new associate of DEISA Barcelona Supercomputer Centre (BSC) facilitating a strong PowerPC-cluster with a mixed AIX/Linux GPFS environment some additional work on a "hierarchical" GPFS has to be done to be able to join DEISA's MC-GPFS as a potential 6th member.

Appendix A: Network debugging

A report from the network engineer at the RZG

During the past months we had been monitoring the network connections between the DEISA sites by short TCP throughput measurements with the analysis tool "iperf". The results showed constant quality and acceptable rates around 750 Mbps. To avoid interference with other testing activities these throughput measurements ran for only ten seconds each.

When tests with GPFS begun (using larger amounts of data) these throughput values could not be achieved by far. Running iperf for longer times with a resolution of one second showed that the throughput dropped every 61 seconds when TCP traffic was sent from RZG. Since the recovery time is much longer than 61 seconds the mean transmission rate was very poor (see figure 9).

This effect could be observed when sending data from any DEISA machine at RZG. Each showed this 61 seconds schedule but because the events occurred at different wall clock times on different machines we first arrived at the (wrong) conclusion that this effect was caused by the end systems. We therefore closely monitored the activity. We even stopped all subsystems and processes on one of the test systems but without avail.

Closer examination of the output of "tcpdump" finally showed that during these events a whole RTTs worth of data was retransmitted, although no data packet was really lost.

Therefore no new data was transmitted for one RTT and TCP flow control was forced to do a so-called "slow-start". Retransmission was obviously triggered by too many packets arriving out of order at the receiver side (apparent by DUP-ACKs, duplicate Acknowledgement Packets, followed by ACKs with again increasing sequence number during short intervals).

In the past there had been rumours that some brands of routers caused problems by delivering packets out of order. So we removed our Foundry "BigIron 15000" from the DEISA network and connected an IBM pSeries directly: the drops in transmission rate every 61 seconds were gone. Our Foundry vendor recommended using the latest software release, which seemed to fix the problem. The software upgrade of course required a reboot of the router. So we don't know yet whether the new software really fixed the problem permanently, or it will occur again after some time. Figure 14, shows that the disruptions have disappeared. The still visible drops in performance are correlated with the automated short tests between the DEISA sites.

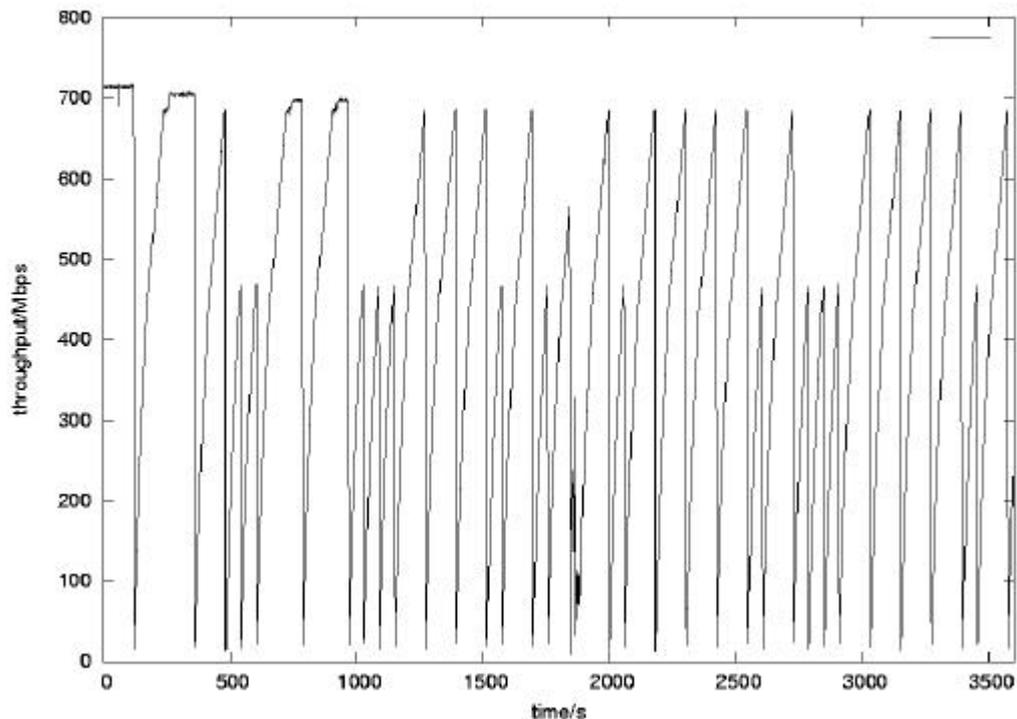


Figure 13: Fine resolution throughput measurement from RZG to IDRIS before firmware upgrade of central router at RZG.

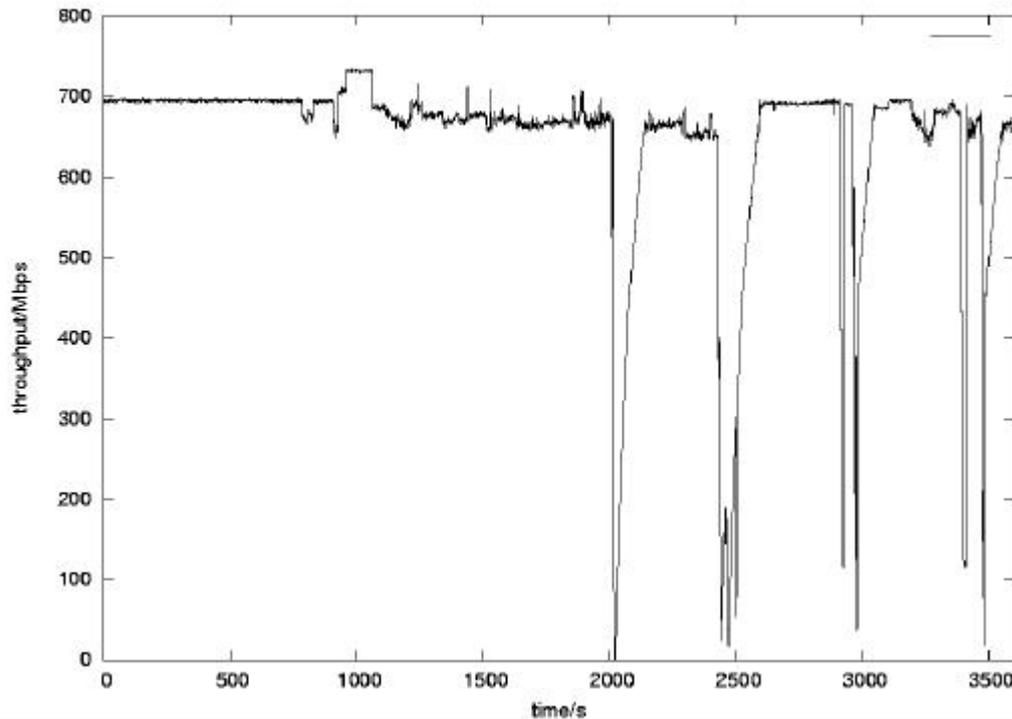


Figure 14: Fine resolution throughput measurement from RZG to IDRIS after firmware upgrade at central router at RZG. The drops in performance still visible in this figure are correlated with the automated short tests between the DEISA sites.

Appendix B: MC-GPFS internals

In this section, the striping available in GPFS file systems is explained in deeper detail. As stated in section 2.3, the file system is not written contiguously on one disk, but distributed in “stripes” across the three RAID 5 systems. RAID 5 is a common technology and allows high throughput combined with some fail-safety on commodity hardware (here: SATA-disks).

For example, the first data-block resides on RAID 1, the second block on RAID 2, number three on RAID 3, number four on RAID 1 again and so on. Like this, the three RAIDs embodying one storage-system are intrinsically parallelized without the user noticing it. MC-GPFS offers three different algorithms to distribute its data onto the different RAID systems:

? RoundRobin:

Data and MetaData-Blocks are written on each disk, looping through the disks in always the same order. For sequential access this delivers good performance and has a perfect load balance on the disks.

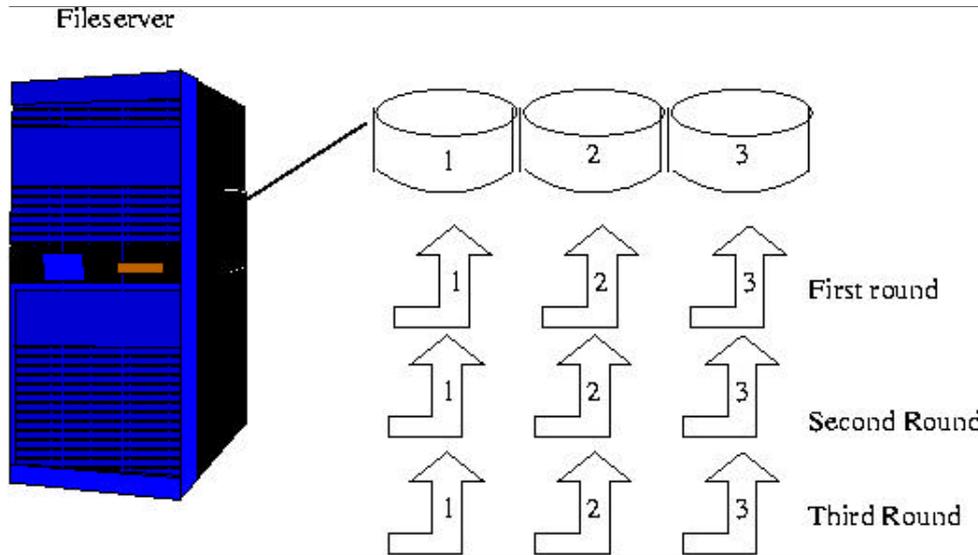


Figure 3: RoundRobin Striping of a single fileserver

- ? Random
The Blocks are distributed randomly across all disks. Load balancing on the disks is not guaranteed and performance may vary for sequential file-access.

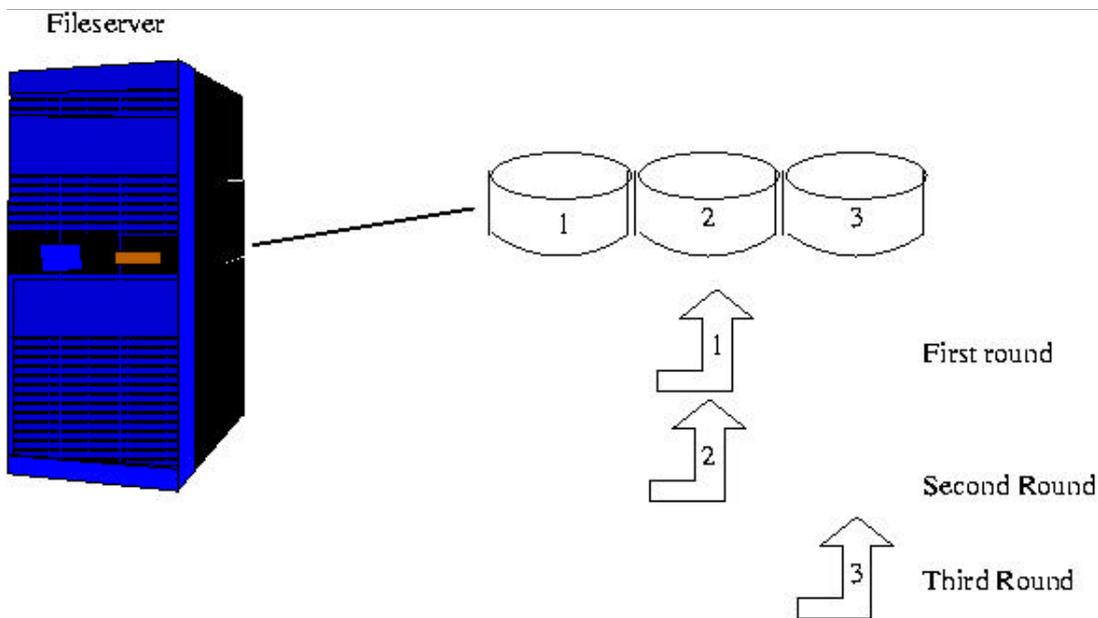


Figure 4: Random striping, load balancing is not guaranteed.

- ? BalancedRandom
Similar to RoundRobin, but with the difference that the disk-order in which the data is written changes from loop to loop.

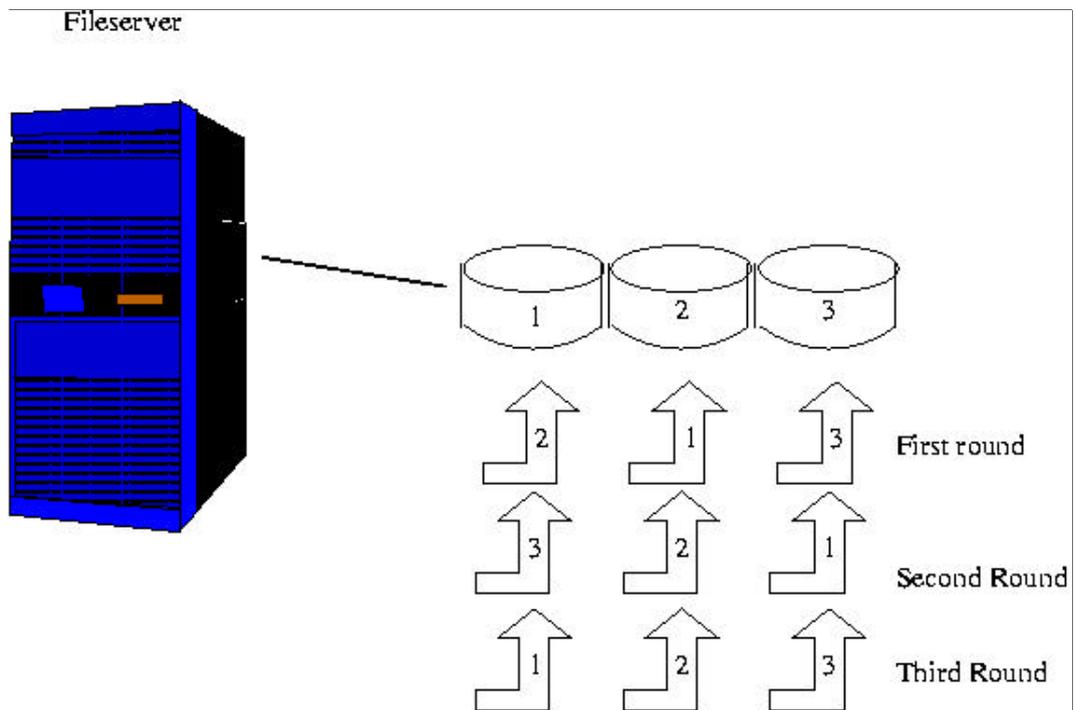


Figure 5: Balanced Random, the order in which the disks are accessed is random in each round, but each disk is accessed exactly once in that round.

At the RZG, the RoundRobin algorithm has been chosen, because it promises the highest performance due to its simplicity and guaranteed load-balancing.