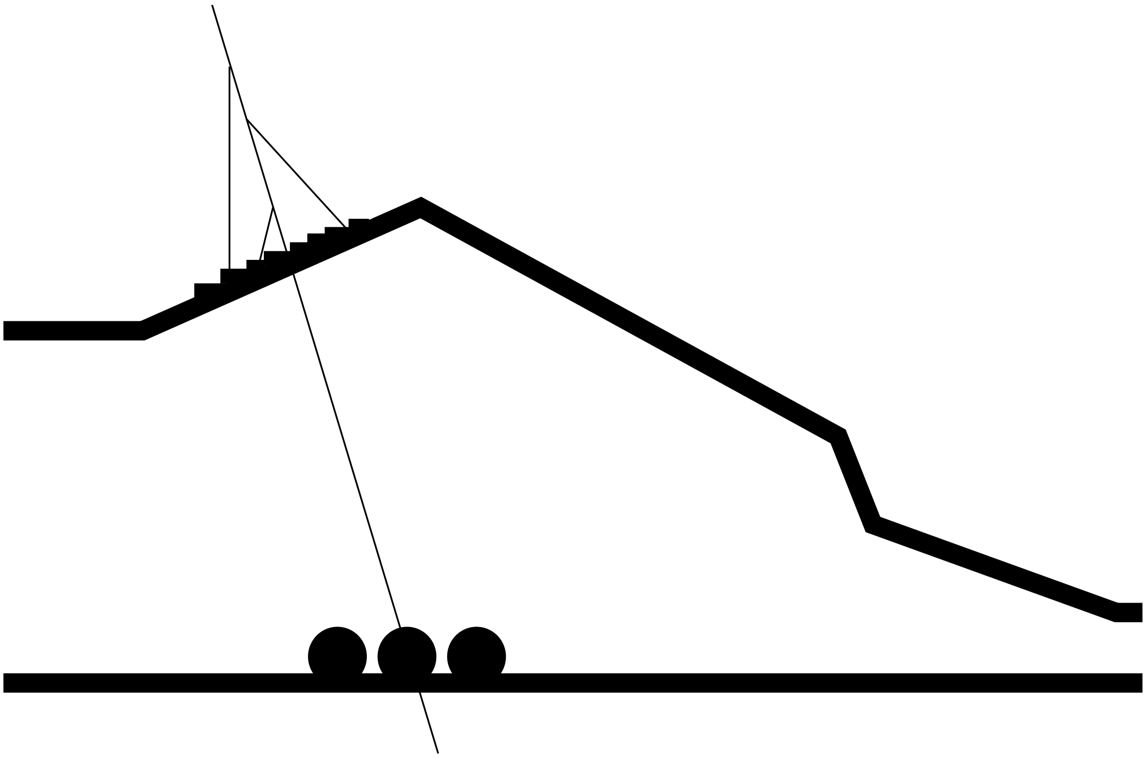
INFN/TC-98/07
26 Febbraio 1998



The LNGS AFS cell

A.D'Ambrosio, S.Parlati

INFN - Laboratori Nazionali del Gran Sasso

INFN - Istituto Nazionale di Fisica Nucleare
Laboratori Nazionali del Gran Sasso

INFN/TC-98/07
26 Febbraio 1998

The LNGS AFS cell

A.D'Ambrosio, S.Parlati

INFN - Laboratori Nazionali del Gran Sasso
S.S. 17bis Assergi (L'Aquila) 67010 Italy

Abstract

Since some years, the number of UNIX machines and people using them has been continuously growing at Gran Sasso Laboratories. The problem of software distribution and maintenance over different architectures, the needs of a common environment and an easy network access to LNGS files from any other INFN sites become soon evident. We tried to solve these questions using AFS (Andrew File System). Since July 1997 an LNGS AFS cell *lngs.infn.it* has been set up to be used by single users and experiments. In this preprint we present the results of this work, the problems we encountered and the future plans about AFS and other distributed filesystems.

1 Introduction

Since some years, the need of distributed filesystems integrated in many different UNIX platforms has been evident in almost every research institution. A distributed filesystem should in principle offer:

- distribution of programs, sources and documentation to a large number of machines;
- easy installation of common software (eg. CERNLIBS) or over many sites;
- an omogeneous environment to users over different UNIX architectures (eg. shared home directories);
- possibly shared network services as mail or news services.

One solution, adopted for instance by CERN, FNAL, DESY, is AFS, a ©Transarc Corporation product.

AFS is a product based on client-server model, as many other distributed filesystems, but it features a common naming, that hides to the user the physical location of the files. The AFS filesystem is organized in cells, collections of clients and servers, generally belonging to the same institution and managed independently from other cells.
Servers can be simple fileservers which just store files on disk or Authentication servers (also called DataBase servers) which manage user's password, file's ACLs and groups membership.
Data replication on fileservers and redundancy on the number of DB servers ensure the service continuity and reduce the risk of data loss.
The cache mechanism on client side allows best efficiency in file access; encryption mechanism, ACL and protection groups guarantee a good security over the network.

As will be explained in the following, INFN, together with CASPUR[1], is partecipating in AFS filesystem since 1995; the Gran Sasso Laboratory, after some tests as simple client site has now it's own AFS cell.

2 The choice of a separate cell

In 1995 INFN started the study of AFS filesystem, its possible application in HEP and it's configuration for the INFN sites. After some discussion it was decided to start with a single large cell, called *infn.it*[1] in which 3 sites act as authentication servers while a fileserver resides in almost every INFN site.
This situation seemed to be the simplest one and is still in use, but it's not excluded that will change in the future. Two INFN sections, Lecce and Pisa, preferred to have a separate cell for historical reasons.

[1] Consorzio per le Applicazioni di Supercalcolo per Universitá e Ricerca.

The main use of AFS in the INFN-wide cell is the distribution of software, particularly the CERN libraries: for each platform the complete ASIS[2] collection is mirrored from CERN to CNAF and from there to the other AFS fileserver. The software mirroring is completely done by CASPUR, as well as most of the systemistic support for AFS.

At the same time some tests on AFS have been performed at LNGS.
Some IBM AIX machines became AFS clients of the INFN cell, but because of incompatibility between local installed software and AFS-available ASIS programs, this test phase stopped in 1995.
In 1997, with the growing number of unix hosts, the use of AFS was re-examined.
The first and most important choice to be made was about the nature of our site: a single, independent cell or a fileserver of the national cell *infn.it*.
Each solution had advantages and disadvantages:

- Single cell:
  - \+ the local AFS managers have a higher freedom in managing AFS;
  - \+ the authentication and authorization processes do not depend from remote servers, so also in case of network failure, users can continue working on AFS;
  - \+ possibility to choose between local installed and AFS distributed CERN software;
  - – the complete installation and management of the cell and the mirroring of ASIS software is time and force consuming;
  - – CASPUR doesn't supply support for individual cells.
- Fileserver of *infn.it*:
  - \+ mimimal manpower requested for management;
  - \+ ASIS mirror performed by CASPUR;
  - – in case of network failure, AFS unavailable also on local files, because of loss of contact with the authentication servers;
  - – possible clash between experiment software and ASIS distribution.

One of the crucial point was the problem of authentication on an external Database server: since January '98, the lab is connected to the rest of the world through a single line and the possibilities that the line is off are not negligible: in the past we experienced network failures lasting even 24 hours! In these cases, no authentication process is possible on a remote server and if we choosed to be a fileserver of the INFN large cell, the whole AFS system would be unavailable to users (including user's files and directories stored locally). Taking into account that this situation actually happened in the past in the

[2] Application Software Installation Server; a CERN project to easily distribute and install CERN and public domain software.

INFN cell, causing problems to people working in AFS, we oriented ourselves to a local AFS cell. The possibility to have locally a Database server of the national cell was difficult, because it involved a new arrangement of the whole cell.

Another strong point in favour of a single cell was the problem of CERN vs. local software. Normally, the local software and ASIS are accessible by users in two different directories:

/usr/local -> /afs/<cell>/asis/<platform>/usr.local

/usr/local+ -> /<local_dir>

However, several local installed packages uses internal reference to libraries or accessory files residing in */usr/local*; since the AFS */usr/local* are *readonly* directories, it's quite hard to arrange the installation of local software. This problem, which may be negligible in many INFN sections, is very sensible at LNGS Laboratories, given the large number of non-EEC users and the use of analysis packages which rely little, if any, on the CERN software environment. Having a local cell, makes very easy to add and maintain non-ASIS software to the fileserver.[3]

For these reasons the 22nd of July 1997 the cell *lngs.infn.it* was created.

3 Cell configuration

AFS is designed to have data redundancy and to be machine fault tolerant: the database servers of a single cell can (and possibly should) be more than one (©Transarc recommends three[2]), while the volume replicas on more than one fileserver assure the continuity of the service.

Since the size of our cell is currently small, we decided however to use a single machine as (unique) DB server and fileserver.

3.1 The server

The ideal AFS server should have as much as possible these characteristics: a wide band network interface, a fast internal bus and high speed disks.
Our server, being the only one, should also be a very stable machine and possibly the same machine upon wich all our UNIX world depends.
These two requirements are present on RSGS02 that become our AFS server.
Its characteristic are:

- Type: IBM RS6000 - 580

[3] Since some months a new product ASISwsm allows using a writable */usr/local* AFS directory, but it's not supported by INFN-Caspur.

- Operating system: AIX 4.2 Server version (the upgrade to AIX4.2 was necessary to install the AFS server)

- Network interface: Ethernet - FDDI

- Buses: 8 slots MicroChannel, SCSI2, FW SCSI, diff SCSI2

- Disks: total 28 Gbytes - AFS 9 Gbytes

The server is set up also as AFS client.

This machine is also a NIS server and NFS server, exporting home directories and experiment's filesystem to all other central UNIX machines.

Due to security problems related to AFS, user accounts have been locked on the server and all unessential network services have been disabled.

3.2 The clients

All the LNGS central UNIX machines are client of the cell *lngs.infn.it* (see figure 1). They are: 5 Digital UNIX and 3 IBM-AIX; the AFS version is 3.4a.patches.03/97 for

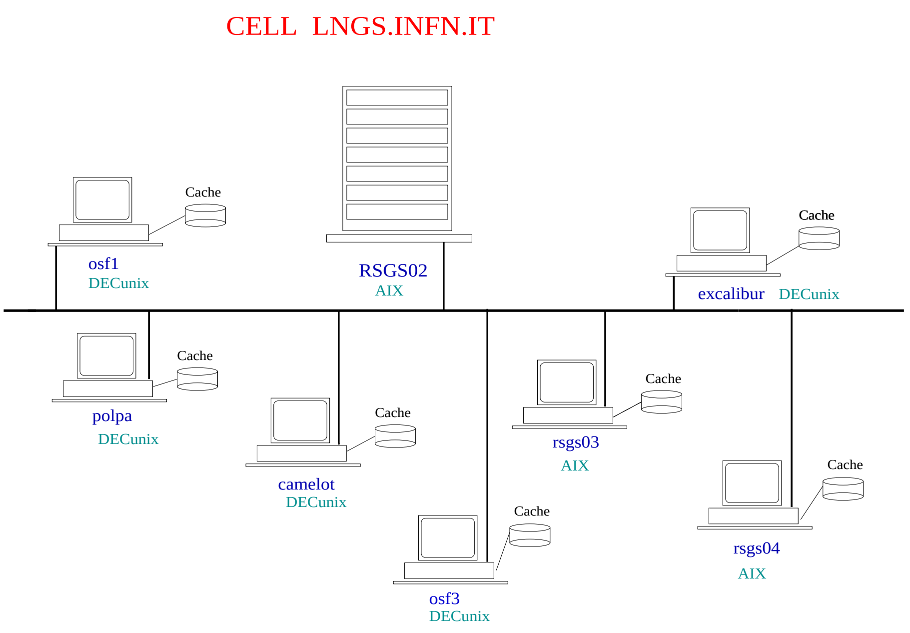


Figure 1: Structure of our cell.

both architectures.

The cache size of the clients ranges from 40 Mbytes to 80 Mbytes depending on type and machine utilization.

Table below shows an example of cache manager statistics from some clients:

	cache utilization (%)	cache misses vs cache hits (%)
camelot	94.81	0.90
excalibur	89.15	0.23
osf1	89.47	0.68
polpa	87.81	6.13
rsgs04	88.01	5.05
rsgs03	80.33	5.26

These statistics indicate that almost every client's request is found on cache, while just in a small fraction of cases (always less than 10%) a file has to be fetched from server. This is an indication that the cache size is well proportioned with respect to the user's needs.

We did not experience any problem during client script installation. The only operations really affected by AFS are OS upgrade, patch installation or installation of layered products on Digital UNIX machines: since AFS on this platform is included in the kernel and not dinamically loaded, AFS has to be uninstalled before any kernel change and then reinstalled.

4 AFS accounts

In order to use as much as possible the AFS capabilities, we decided to give an AFS account to each LNGS UNIX user.

For simplicity, users have on AFS the same UNIX username and UID.
In the */etc/passwd* file (shared via NIS) users store their UNIX password, in order to get into the system also when AFS authentication is not possible (server crashes or AFS login not available, see later).

A volume called *user.<username>* has been created for each user: local users are currently granted 40Mbytes, while external users have 10Mbytes available.

Experiment's public AFS accounts have been created as authentication-only accounts, that is without a own volume.
Currently about 60 AFS accounts are defined on our cell.

5 The LNGS AFS tree

Figure 2 shows the tree structure of our AFS cell.
About 5.5 Gbytes are used for ASIS software (Digital UNIX and AIX platforms + share), while about 1 and 0.5 Gbytes respectively are currently used for user's and experiment's volumes.

5.1 /afs/lngs.infn.it/system

In this directory all the AFS binaries for Digital UNIX and AIX platforms are kept, for the release currently in use and for the past releases.
The clients access this directory by the link:

/usr/afsws − > /afs/lngs.infn.it/system/3.4a-0397/@sys

5.2 /afs/lngs.infn.it/asis

It contains essentially the CERN program library, the complete TEXenvironment, the GNU software and X11 programs, for the two operating systems we have.
The ASIS collection is mirrored, at each new release, from the Roma1 AFS fileserver of the cell *infn.it*, which is the server nearest to LNGS (see figure 3).
Keeping all these packages on AFS represents a big improvement for the software management, because almost every program and library is ready as-is and requires no installation or revision or maintenance.
In this way all clients see the same products and the cache mechanism ensures a fast access to the most frequently utilized programs.

Two logical links are defined on the clients:

/cern − > /afs/lngs.infn.it/asis/@sys/cern[4]

/usr/local − > /afs/lngs.infn.it/asis/@sys/usr.local

in order to simplify the access and to make ASIS programs fully available to the system.
Locally installed packages are kept in

/usr/local+

when they require libraries or files in /usr/local, the same are put in the AFS server.
The system-wide variable PATH can be defined for all users specifying which package is found first.

Two prefix-less groups *infn-nodes* and *lngs-nodes* include respectively the IP addresses of all INFN local area networks and the LNGS one. These two groups have "lr" right on */asis* directories, ensuring the visibility of the subtree also to non-authenticated INFN users.

Before each new ASIS mirror, a BCK version of the asis volumes is done.

[4]the symbol @ in AFS environment translates the current Operating System name.

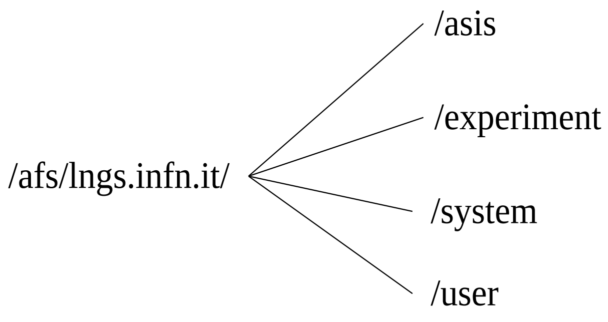


Figure 2: LNGS AFS tree.

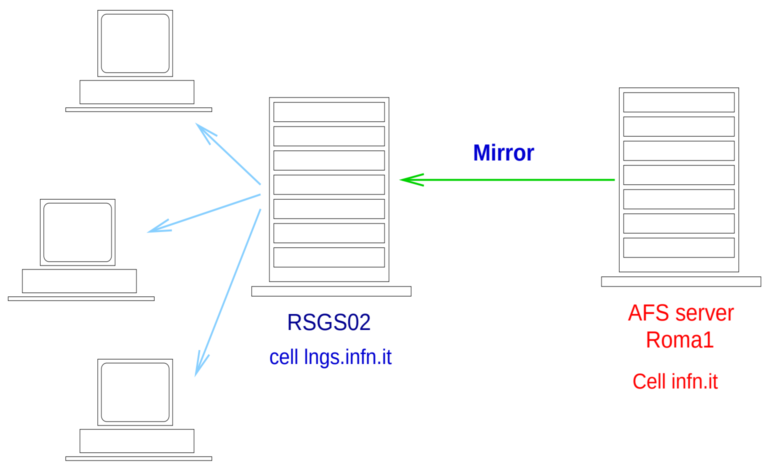


Figure 3: The ASIS software mirroring is done from the INFN cell.

5.3 /afs/lngs.infn.it/user

As said before, each AFS user has a volume: *user.<username>*.
Its mountpoint is /afs/lngs.infn.it/user/<username> and on the client side, this link is used:

/users −> /afs/lngs.infn.it/user

The group *system:anyuser* has been given by default the "lookup" right on user's top AFS directories, in order to guarantee the access to eventual *public* directories.
In principle, these AFS directories could be used as user home directories. In practice, a CDE problem prevented such a configuration as explained later.

Each night a BCK version of the user's volumes is created automatically, to allow the restore of accidentally deleted files.

5.4 /afs/lngs.infn.it/experiment

Our current strategy is to give each project a volume called *experiment.<experiment>*.
No need arised for other specific volumes such *experiment.<experiment>.software* even though this could be possible in future.
This volume mainly serves as software repository, documentation and images archive.

Experiment's small data file are also present but, because of limited client's cache size, we recommend not to store on AFS files larger than some Mbytes.

On client side, this directory is also reachable by the link:

/experiments −> /afs/lngs.infn.it/experiment/

Since at least two experiment's AFS groups are needed (for users and administrator), we choose to create two prefix-less AFS groups *<experiment>-people* and *<experiment>-admin*.

- *<experiment>-admin*: the group contains the administrators of the experiment AFS area; this group owns *<experiment>-people* in such a way that the members can easily change the collaboration list;
  this group has "all" right on */afs/lngs.infn.it/experiment/<experiment>*.
  The group is owned by *system:administrators*.

- *<experiment>-people*: this group include the list of people collaborating to the project; an experiment's public account is also member of this group; some IP address are member of this group too, when the machine belongs exclusively to the experiment and all its users are project's people.
  The group is owned by *<experiment>-admin*.
  The *<experiment>-people* right on the experiment volume are decided by the group administrators.

An interesting possibility to get easy access to AFS from machines wich are not AFS client, or by people without AFS account in our cell, through the use of an Internet browser is under test. From any machine, requesting the URL

ftp://<username>@host.address/afs/lngs.infn.it/experiment/<experiment>

(where host.address must be an AFS client with a web server enabled on it, and *<username>* could be an experiment's public account), a remote experiment's member gets AFS authentication on the client and can access or see project's software or documentation. This way the AFS file system can be "re-exported" via the FTP protocol.

6 Problems

We did not encountered any problem during the sw installation.

A surprising effect which arised after AFS was the following: working on Digital UNIX client machines, after a "su root" operation, all disks imported via NFS2 from the NFS server or from other sources became unaccessible (working on the host console seems not to be affected).
The messages received for instance with a "df" command were:

```
NFS2 fsstat failed for server rsgs02 : RPC: Authentication error
```

The same messages appeared after entering in a pagsh[5]. As a possible explanation, pagsh may have some problems in managing NFS authorizations. We found references to such a problem in the pagsh man pages together with a suggestion on how to solve it. The proposed remedy didn't work.
So, the only way for root users to access NFS imported filesystems is working at the console on AFS clients Digital UNIX machines.

Another problem concerned the user authentication at login time.
Almost all clients use CDE as desktop enviroment. The CDE authentication program is dtlogin which is not AFS-compliant; for this reason, users opening a Xsession (on the console or from an Xterminal) should authenticate in AFS later with "klog"; this implies that user's home directories cannot be AFS directories, as long as CDE is the desktop manager.

A simple way to overcome this problem, is to use a modified version of xdm (xdm-afs) which includes AFS authetication.
This should allow authenticated login from any source (including X-session from Xterminals) and thus home directories in AFS.
But in this case, since xdm-afs must be called inside a pagsh, the root problem became even worst, since the NFS problems appeared also at the console.
Since this is a serious limitations for system management, we decide to come back to

[5]Process Authentication Group shell: is a shell used in AFS environment for security purposes.

CDE's dtlogin and leave user's home directories on local filesystems.

7 Future work and conclusions

Some aspects of the AFS cell management are still under way, and should be taken into account in the next months.

The most important is the backup strategy: at the moment all user's and experiment's directories are simply packed in gzip-tar file and taken on disk. As soon as possible, a tape coordinator machine shoud be configured (probably the same server) and a tape backup strategy should be defined.

The security aspect of AFS has to be better understood, especially in the area of integration with other security mechanisms.

Even though presently Digital UNIX and AIX are the only platforms present in our cell, we should consider the possibility to provide AFS service for other operating systems (Linux, SunOS, HP), both as AFS binaries and as ASIS collection. This will require considerably more disk space.

Depending on the long term plans for the UNIX system at LNGS, it could be possible that fileserver and database functionality will be separated in two AFS servers; this because the disk space requirements of new experiments could require a new fileserver strategy that will include AFS volumes, while the authentication functionality may be better integrated with other network services.

Recent news from ©Transarc, encourage using AFS in the future, since AFS will not only be supported for current platforms, but will be developed and improved mainly on the management aspects.
In parallel to AFS works, we plan to investigate the potentiality of DCE/DFS as a solution to the current AFS weakness and as an unifying environment for the unix world.

Our experience indicate that AFS represents a very useful tool for software distribution and maintenance (ASIS is the best example); single users and experiments can take great advantage from the possibilities AFS offers about cooperative work (ACL and group protection) and about network available filesystems.
Despite the advantage of a single distributed file system, however, it's very improbable to extend the reach of the cell outside the lab, given the license problems which may arise with international collaborations. In this sense, some different groupware tools may be needed.
The choice of a separate cell revealed to be a correct one, since we didn't experience AFS failures even when the wide area network was unreachable.
Time spent to maintain AFS software and mirror asis collection, even though every-day care (user's assistance, account creation, volume backup...) is nowadays minimal.

8 Aknowledgements

We would like to thank A.Maslennikov (CASPUR) and R.Gomezel (INFN-TS) for their help and their very useful “AFS system administrator” course.

References

[1] Proposta di adozione e uso di AFS nell'INFN.
http://www.roma1.infn.it/AFS/proposta_afs.html.

[2] AFS System Administrator's Guide.
©Transarc Corporation

A Available Documentation

The ©Transarc AFS user's manual is available on-line in postscript format at the page:

http://osf1.lngs.infn.it/docs/afs/afs.user.ps

while slides (PS or html format) of the AFS course given at Gran Sasso in October 1997 can be found at:

http://osf1.lngs.infn.it/docs/afs/info_afs

The above documents can be also found in the LNGS Library.