ATLAS TIER3

WP: Distributed storage (Lustre & GPFS)

https://twiki.cern.ch/twiki/bin/view/Atlas/LustreTier3
Index

0. Members .................................................................................................................................................. 3
1. Introduction............................................................................................................................................... 3
   1.1 Lustre components in a basic cluster................................................................................................. 3
      1.1.1 Minimal Setup............................................................................................................................. 4
      1.1.2 Base version ............................................................................................................................... 5
      1.1.3 Existing Installs and contacts ..................................................................................................... 5
   1.2 GPFS components in a basic cluster ................................................................................................. 5
      1.2.1 Minimal Setup............................................................................................................................. 7
      1.2.2 Base version ............................................................................................................................... 7
      1.2.3 Existing installs and contacts ..................................................................................................... 7
2. Lustre first configuration recommendation ....................................................................................... 7
   2.1 General ................................................................................................................................................ 7
   2.2 Metadirectory ....................................................................................................................................... 8
   2.3 Disk Servers ...................................................................................................................................... 9
   2.4 Clients ................................................................................................................................................ 9
3. GPFS first configuration recommendation .......................................................................................... 10
   3.1 General ............................................................................................................................................... 10
   3.2 Metadirectory ................................................................................................................................... 10
   3.3 Disk Servers ................................................................................................................................... 11
   3.4 Clients ............................................................................................................................................... 12
Appendix A: Lustre + Storm Install (Valencia case) ................................................................................. 13
   A.1 Lustre Configuration and Setup ....................................................................................................... 13
      A.1.1 Software Installation ................................................................................................................ 13
      A.1.2 Setup ......................................................................................................................................... 13
      A.1.3 Nodes setup ............................................................................................................................. 13
   A.2 StoRM Configuration and Setup ................................................................................................... 14
      A.2.1 Current services ........................................................................................................................ 14
      A.2.2 OS, Required packages & Storm Installation ............................................................................. 14
      A.2.3 Configuration ............................................................................................................................ 15
      A.2.4 ATLAS authorization configuration ......................................................................................... 17
      A.2.5 Dinamic Info Provider .............................................................................................................. 18
      A.2.6 Configuration issues ............................................................................................................... 18
Appendix B: GPFS + Storm Install (UK case) ............................................................................................ 19
   B.1 Installing GPFS .................................................................................................................................. 19
      B.1.1 Install RPMS ............................................................................................................................. 19
      B.1.2 Build source .............................................................................................................................. 20
      B.1.3 Create GPFS Cluster ............................................................................................................... 20
      B.1.4 Create a Network Shared Disk (NSD) ..................................................................................... 20
      B.1.5 Create GPFS filesystem ........................................................................................................... 21
   B.2 Adding GPFS to storm ..................................................................................................................... 21
      B.2.1 Mount GPFS space onto StoRM node ...................................................................................... 21
      B.2.3 Configure StoRM to use GPFS drivers ................................................................................... 21
      B.2.4 Testing ...................................................................................................................................... 23
Appendix C: Inventory/Survey form ......................................................................................................... 23
0. Members
Santiago González de la Hoz (IFIC-Valencia)

LUSTRE:
- UAM-MADRID Tier 2 (Juan Jose Pardo and Miguel Gila)
- LIP-COIMBRA Tier 2 (Miguel Oliveira)
- BONN-Physikalisches Institut (Simon Nderitu)
- DALLAS-Southern Methodist University (Justin Ross)
- IFIC-VALENCIA (Javier Sánchez, Álvaro Fernández, Alejandro Lamas)
- ISRAEL T2/T3 Federation-Weizmann Institute, Tel Aviv University, The Technion (Lorne Levinson and Pierre Choukroun)
- DESY (Yves Kemp and Martin Gasthuber)
- U. OKLAHOMA (Horst Severini)
- Wuerzburg (Andreas.Redelbach)

GPFS:
- Edinburgh (Wahid Bhimji)
- Italian sites (Gianpaolo Carlino and Fulvio Galeazzi)

1. Introduction

In this document, current usage in ATLAS Tier2 and Tier2 sites for Lustre a GPFS distributed storage is given. The aim of this document is to determine and make available best practices guidelines and suggestions for deployment Lustre and GPFS at all Tier3 sites.

1.1 Lustre components in a basic cluster

A Lustre file system consists of the following basic components (see figure1).

- **Metadata Server (MDS)** - The MDS server makes metadata stored in one or more MDTs available to Lustre clients. Each MDS manages the names and directories in the Lustre file system(s) and provides network request handling for one or more local MDTs.

- **Metadata Target (MDT)** - The MDT stores metadata (such as filenames, directories, permissions and file layout) on an MDS. Each file system has one MDT. An MDT on a shared storage target can be available to many MDSs, although only one should actually use it. If an active MDS fails, a passive MDS can serve the MDT and make it available to clients. This is referred to as MDS failover.

- **Object Storage Servers (OSS)**: The OSS provides file I/O service, and network request handling for one or more local OSTs. Typically, an OSS serves between 2 and 8 OSTs, up to 8 TB each2. The MDT, OSTs and Lustre clients can run concurrently (in any mixture) on a single node. However, a typical configuration is an MDT on a
dedicated node, two or more OSTs on each OSS node, and a client on each of a large number of compute nodes.

- **Object Storage Target (OST):** The OST stores file data (chunks of user files) as data objects on one or more OSSs. A single Lustre file system can have multiple OSTs, each serving a subset of file data. There is not necessarily a 1:1 correspondence between a file and an OST. To optimize performance, a file may be spread over many OSTs. A Logical Object Volume (LOV) manages file striping across many OSTs.

- **Lustre clients:** Lustre clients are computational, visualization or desktop nodes that run Lustre software that allows them to mount the Lustre file system. The Lustre client software consists of an interface between the Linux Virtual File System and the Lustre servers. Each target has a client counterpart: Metadata Client (MDC), Object Storage Client (OSC), and a Management Client (MGC). A group of OSCs are wrapped into a single LOV. Working in concert, the OSCs provide transparent access to the file system.

Clients, which mount the Lustre file system, see a single, coherent, synchronized namespace at all times. Different clients can write to different parts of the same file at the same time, while other clients can read from the file.

More information for this section can be found on [http://wiki.lustre.org/index.php/Lustre_Documentation](http://wiki.lustre.org/index.php/Lustre_Documentation)

**1.1.1 Minimal Setup**

- 1 x MDS (+MGS): metadata server (+ config server). The MDS RAM memory would be at least 32 GB using a connectivity of 1 Gbps. The kind of MDS processors should be the standard server configuration nowadays, i.e 2 x quad core.
- 1 x OSS (wich servers one or more OST). Sites should use a kind of disk redundancy, for instance RAID5 or RAID6. Perhaps RAID6 is better if you have a lot of disk. The
disk server should have at least 16 GB of RAM memory and 2 x quad core processors. For the connectivity we think that NAS storage topology is the easier and cheaper option for a Tier3 with a connectivity of 1 Gbps.
- 1 client. With a connectivity at least 1 Gbps and with read-write access.

(more details in section 2)

1.1.2 Base version

Lustre: 1.8.1.1
OS: SLC 5

1.1.3 Existing Installs and contacts

IFIC-Valencia

Contacts:
Javier.Sanchez@ific.uv.es, Alejandro.Lamas@ific.uv.es, Alvaro.Fernander@ific.uv.es

Installation:
thttps://twiki.ific.uv.es/twiki/bin/view/Atlas/LustreStoRM

DALLAS-Southern Methodist University

Contacts:
ross@mail.smu.edu

Installation:

ISRAEL T2/T3 Federation-Weizmann Institute, Tel Aviv University, The Technion

Contacts:
Lorne.Levinson@weizmann.ac.il, pierre.choukroun@weizmann.ac.il

Installation:
thttp://ific.uv.es/~sgonzale/Lustre_experience_of_the_Israel_Tier-2.pdf

1.2 GPFS components in a basic cluster

GPFS Clients can access the storage via the Network Shared Disk (NSD) protocol. This is illustrated in the figure below. Worker nodes will be NSD clients and, as with Lustre, will see and access a coherent, synchronised namespace.
NSD (I/O) Servers. The disks themselves are attached to servers either via a SAN or directly. GPFS allows up to 8 servers per disk and having more than one server per disk will provide additional redundancy. Note that servers after the first one will be contacted only in case of failure of the previous ones: hence, load sharing between servers serving the same set of NSDs is achieved by shuffling the list of servers for each NSD.

Metadata
Unlike Lustre, GPFS allows all nodes of the cluster to perform metadata operations. Therefore a separate metadata server is not required and so is not a single point of failure. However it can be useful to separate metadata and have it served by faster disks if the data stored is in small files or if there are numerous metadata operations (for example experiment software areas). Alternatives would be to not store the software areas on the GPFS filesystem or to export them via Cluster NFS (CNFS).

GPFS File system
A GPFS cluster can contain up to 256 filesystems with no practical limit on the size of the filesystem. The block size of the filesystem is configurable and for serving large data files a block size of at least 1MB is recommended. Storage pools allow for tiering storage and the namespace can also be partitioned into filesets to set quotas.


1.2.1 Minimal Setup

In addition to clients, one would expect
- At least 3 data NSD servers (to allow a quorum) and that these had at least 4 GB of RAM with 1 Gb networking.
- The servers would usually have RAID 6 configured in a multiple of 4 (eg. 4+1P, 8+1P or 8 + 2P etc.) for GPFS write performance.

(more details in section 3)

1.2.2 Base version

GPFS: 3.3
OS: SLC5

1.2.3 Existing installs and contacts

Edinburgh

Contacts:
Wahid Bhimji wbhimji@staffmail.ed.ac.uk

Installation:
https://www.gridpp.ac.uk/wiki/Storm_GPFS_Install

2. Lustre first configuration recommendation

2.1 General

Every site has an associated Tier2, is using Grid tools (is “a la Grid”) and 50% of the sites are sharing Lustre with their Tier2.

The number of users using the Tier3 is less than 50 people, and the Lumber of groups less than 4.
2.2 Metadirectory

All the sites are using the same machine for MGS and MDT, so the recommendation would be use the same machine to simply the installation but with different disks for MGS and MDT.

IFIC is the only site using HA in the metadirectory server. We think the is not necessary at Tier3 but Lustre plans for 2.2 version is to include a clustered Meta Data, so probably is worth to wait for it. On the other hand we are assuming we are doing some kind of redundancy (i.e RAID1) and it would be advise to do MGS and MDS back up, this is the case in Coimbra and in Valencia.

The metadirectory server should fine work with connectivity at least 1 Gbps, 32GB of memory to assure a good functionality with the number of file systems used and with the number of files in each file system, and 2 x quad core processors (8 cores in the MDS), which is the standard server configuration nowadays. AT IFIC-Valencia with 16 GB of RAM memory we had problems with the ATLAS software area. Some time the memory was full and the machine performance was a bit low. For that reason we think that at least 32 GB we will never seen any problems.

The number of systems (fs) would be tied to different aspects:

- Quotas in the file system. Because the quota is applied to the whole file system, sometimes several fs are needed to have different policies.

- It is sometimes useful to have several fs to split hardware between different usages, although "pools" can be used to do this.

- Maybe a recommendation can try to explain those aspects and include some setup examples, for instance, split space between scratch space (without quotas) and long term (with quotas).

Only a few sites are using Lustre pool features, for instance IFIC-Valencia. The reason to use that is using "pool feature" to isolate ATLAS space tokens in such a way that every token has a dedicated OSTs with a limited disk space within the same file system without using quotas. There is a one to one relation between space token and pool. Maybe this use can be extrapolated to T3. On the other hand limiting disk space can be done via StoRM, so no need for the disk pool features but a Tier3 could not to have a SRM.

The recommendation would be to setup at least a default pool. OSTs would be assigned to it. In the future the site would decide to include new hardware in the existing pool or create a new one. This has a little setup costs and would save the site to do it later.

It would be advise to use quotas and it would be interesting give general ATLAS recommendation for the quotas management and setups. On the other hand, maybe is not necessary to use users ACLs because in a Tier3 more or less the same people (users) are working on the same analysis and they should write in the same directories. But for instance, if you would like that only a few users access to one directory you will have to
use it. For collocated Tier2/Tier3 this is a must so that Tier3 users can access directly (read) the Tier2 grid storage area.

2.3 Disk Servers

Sites are using at this moment less than 1 PB of data. Lustre is working fine for PB of data, so we should not have problems with that.

The number of disk servers depends on the on the disk space you would like to have. Looking into the form (site configuration) at least sites have 10. The same about the number of OSTs by disk server, looking into the sites configuration at least sites has 6. The number of disks by OST depends on the site configuration (see the form).

About the redundancy sites are using RAID5-RAID6. So the recommendation would be use RAID5-6 redundancy. Maybe RAID6 is better if you have a lot of disk because you are losing free space and is slower than RAID5. For instance, for 12 disks is OK RAID6.

In general, for those last four questions recommendation would be made about how many disks per RAID group to be made, how many partitions (OSTS) per RAID group.

- ie, try to answer questions like:

   It is better to have a single 12 disks RAID 6 or two 6 disks RAID 5? (Having in mind T3 possible use).

About the back up policy and/or recommendation will depend about the size and we will have to discuss about. For instance, how we can do back up about 10 TB of data in a Tier3. In a Tier2, actually, we do not have a back up system and/or policy.

The disk server has at least 16 GB of RAM memory are 2 x quadcore processors and they are not using another kind of connectivity. DESY is using fibre channel because they have SAN (storage topology). For a Tier3 the easier is to have NAS storage topology. A SAN infrastructure is probably more expensive for a Tier3, so the recommendation would be to choose a NAS one.

The disk server should have at least a connectivity of 1 Gbps. Here, maybe we should to provide a ratio Space/BW. For instance, IFIC has a ratio of 30TB/1Gbps. In general only one network interface is enough. There are sites, like DESY, which have Fiber Channel with InfiniBand.

2.4 Clients

In general the sites have more than 400 between cores and nodes with access to Lustre. We have to think how to evaluate the load, this means try to evaluate the load when more than 400 cores are accessing to Lustre while an usual ATLAS analysis job is running.
We have done some test in this direction using the HammerCloud, for instance: http://gangarobot.cern.ch/hc/1307/test/. And the CPU/wall time was very good while more than 200 jobs was running in different cores accessing the same data in Lustre.

Actually the clients accessing to Lustre could be a WN, UI, SE, SRM, gridftp, virtual machines, etc...and the connectivity of this clients recommended would be at least 1 Gbps with read-write access. For instance, in the case of IFIC-Valencia, our UI can read-write in the tier3 area, the SRM and Gridftp can read-write in the Tier2 areas and the WN can read-write in the software area. Whatever client can read-only in whatever area.

3. GPFS first configuration recommendation

3.1 General

Only Edinburgh is sharing resources with associated Tier2 (UKI-SCOTGRID-ECDF): all other sites are independent from any Tier2.

All sites are in GRID. It would be worth clarifying what is meant by “we are in GRID”. Roma3, Trieste and Edinburgh do have a Localgroup space-token in ToA: Milano does not (at least, is not registered in ToA), Bristol is mainly a CMS site. So it would be recommended for Grid sites providing SMR access to data using StorRM or Bestman currently.

All sites are, or will be, using Storm as a SRM. Storm backend and frontend on a single host for all sites (Roma3 has them on a virtual host, but will move to physical). Only one gridftp server for Bristol, Roma3, Trieste: 2 gridftp for Edinburgh (Atlas disk share 20 TB out of 160 total), 4 virtual-ip balanced server for Milano (100% Atlas, 240 TB). The recommendation would be to use multiple Gridftp servers if > 50 TB provided and accessed remotely. We can also say that 1 gridftp server every 50 TB of disk space (served with SRM).

The groups and the users using the Tier3 are very different in each site. Bristol accepts only GRID jobs, for Atlas but has some local CMS users. Edinburgh is the largest site, having 100s of local users (though only around 30 from HEP community including grid users). The three Italian sites have between 30 and 50 local users, belonging to a small number of groups (ranging from 1 to 8).

3.2 Metadirectory

The Matadirectory configuration is very varied. Bristol and Edinburgh have a separate server for metadata for home directories (as well as data and metadata
for applications). Milano has only dataAndMetadata NSDs. Roma3 has a few dataAndMetadata NSDs for all metadata of “data” partitions, all other NSDs are dataOnly. Trieste has separate metadata NSD servers (to be verified). Metadata servers are useful for data that is in small files and/or many metadata operations – (if servers are not high performing in terms of IOPS.). The most troublesome area is the software area,. It makes sense to somehow separate metadata and let them be served by faster disks, but really we don’t know how to quantify the statement: for example, at Roma3 metadata is held on SATA disks, but metadata resides on just two NSDs.

Bristol, Edinburgh and Trieste are using SAS disks and all the sites have 2 replicas of the metadata. We are using separate filesets (for quotas). Milano explicitly states it is using only system storage pool: Roma3 is mapping all ATLAS space-tokens to separate storage pool. So the recommendation would be to use separate filesets. If site also hosts a consistent fraction of data outside GRID, it could make sense to put such data on separate storage pools. Not 100% sure about the best philosophy for NSD-raidset association, whether full separation or full mix is optimal. However, another thing to do would be to configure the primary server for GRID data as the secondary server for local data, and vice versa, to optimize usage of diskserver bandwidth.

All the sites are using ACL’s in order to make Storm disk space accessible to local users. So the recommendation would be yes. ACLs are required for StoRM. Some examples: https://www.gridpp.ac.uk/wiki/Storm_GPFS_Install

3.3 Disk Servers

The capacity in GPFS is around 200 TB in average; there are sites with more than 200 TB, for instance Bristol (400) and sites with less than 200 TB, for instance Roma3 (100). The fraction dedicate to ATLAS is very variable: Milano 100%, Edinburgh and Roma3 40% and Bristol and Trieste is marginal, few TB.

All the sites are doing any kind of redundancy. The recommendation would be RAID6 (or RAID 5). The disk used for each RAID volume is extremely varied in all the sites. We think we should not make any recommendation here. Or may be a very mild one like: “We should do any kind or benchmark test comparisons and find this compromise to work fine for us”.

The block size used most of time for the sites is 1 MB, so we should recommend 1 MB, so that is recommended here.

Roma3, Trieste and (part of) Edinburgh use a SAN with 4, 5, 8 disk servers respectively. Bristol and Milano are using DAS (9 servers). Bristol and Trieste have 1 filesystem and all others have 3-6 main ones. The size of the individual NDS is around 4-8 TN. The recommendation would be for each RAIDset I create 5-6 NSDs such that I have a bit of flexibility in giving space to users, and enlarging/shrinking disk areas.
The features of the disk servers would be at least 4 GB RAM, up to ~20 (Edinburgh). In many sites servers have 4-8 cores, in some cases from Xeon CPUs (no fancy Intel 5k, real Xeon with HT). Though GPFS license charges are often per core so that may motivate getting less cores in each server.

The connectivity for the disk servers is varied among the sites. Bristo and Trieste are using 10 Gb, all other 1 Gb in trunking (Edinburgh 4 x 1 GB). So, the recommendation would be more than 1 Gb. Only Milano is using Fiber Channel.

The quorum nodes the sites have is any number from 2 to 8 (possibly counting quorum nodes across different filesystems, in most cases). The recommendation would be greater than 2. Or we can say greater than 3 for quorum needs, and greater than 2 if using tie-breaker disk.

For data replication we should think about a policy. At this moment we don’t use data replication with the exception of home directories (Edinburgh) and possibly for hotdisk (Roma3).

GPFS is not used for ATLAS software areas, with the exception of Roma 3(using CNFS) and Milano (CNFS, but moving to NFS). The recommendation would be not using GPFS currently for software area, though a possibility is using CNFS which Roma are happy with, because it is easy to setup and provide server redundancy. Only possible issue is coexistence of data server and CNFS server on same host: manual says it is envisaged to work like this, but Milano had some troubles.

Normally we don not use CNFS, with exception of Roma3 (home dirs and software) and Milano (software, but going to dismiss it).

### 3.4 Clients

The sites have between 200 to 700 clients accessing to GPFS. The kind of clients accessing to GPFS are “all”, WN, UI, etc. including CE.

The connectivity used by the clients is 1 Gb, so the recommendation would be greater than 1 Gb.

All sites are defining more than one cluster, except Trieste. The recommendation would be define separate clusters for servers, clients. For example, it would make sense to define a cluster for each blade system. Otherwise, due to quorum mechanism, there may be unwanted long-range interactions.

The cluster has read-write access with the exception of Milano. This issue was put here just because there are some software areas being mounted read-only, with the batch system directing sgm jobs only to read-write nodes. Since all sites are using some flavor of NFS, it does not tell us much.
Appendix A: Lustre +Storm Install (Valencia case)

A.1 Lustre Configuration and Setup

A.1.1 Software Installation.

O.S:

Headnode and OSS with Scientific Linux IFIC 4.4 (x86_64)
Client with Scientific Linux IFIC 3.0.6 (i386)

Updated software in sli44 machines:

tcl-2.2.39-1.1.x86_64.rpm  libtcl-2.2.39-1.1.x86_64.rpm  attr-2.4.28-1.2.x86_64.rpm  libattr-2.4.28-1.2.x86_64.rpm  dump-0.4b41-1.x86_64.rpm  rmt-0.4b41-1.x86_64.rpm  e2fsprogs-1.39.cfs2-0.x86_64.rpm (the e2fsprogs version must be equal or higher than 1.38)

Updated/Installed/Removed software in sli306 machine:

rpm -Uhv libgcc-3.4.6-3.i386.rpm  rpm -ihv libstdc++-3.4.6-3.i386.rpm  rpm -ihv db4-4.2.52-7.1.i386.rpm  rpm -Uhv e2fsprogs-1.39.cfs2-0.i386.rpm  rpm -e modutils  rpm -ihv module-init-tools-3.1-0.pre5.3.2.i386.rpm

Lustre: kernel, modules and lustre utilities.

kernel-lustre-smp-2.6.9-42.0.10.EL_lustre_1.6.0.1.x86_64.rpm  lustre-modules-1.6.0.1-2.6.9-42.0.10.EL_lustre_1.6.0.1smp.x86_64.rpm  lustre-1.6.0.1-2.6.9-42.0.10.EL_lustre_1.6.0.1smp.x86_64.rpm

A.1.2 Setup

Sincronize nodes with ntp. Disable selinux until create a proper profile

A.1.3 Nodes setup

Headnode: will host MGS + MDT

MGS

mkfs.lustre --mgs /dev/sda5  mkdir -p /mnt/lustre/mgs  mount -t lustre /dev/sda5 /mnt/lustre/mgs

MDT

mkfs.lustre --fsname=ificfs --mdt --mgsnode=wn181@tcp0 --mountfsopptions=acl /dev/sda6  mkdir -p /mnt/lustre/mdt-ificfs  mount /dev/sda6 /mnt/lustre/mdt-ificfs

Look for the partition labels with e2label and add the lines to fstab for next mounts.

LABEL=MGS /mnt/lustre/mgs lustre defaults,_netdev 0 0
LABEL=ificfs-MDT0000 /mnt/lustre/mdt-ificfs lustre defaults,_netdev 0 0

OSS:

mkfs.lustre --fsname=ificfs --ost --mgsnode=wn181@tcp0 /dev/sda5
mkdir -p /mnt/lustre/ost0-ificfs mount -t lustre /dev/sda5
/mnt/lustre/ost0-ificfs/

fstab entry:
LABEL=ificfs-OST0000 /mnt/lustre/ost0-ificfs lustre
defaults,_netdev 0 0

Client:

mkdir -p /lustre/ific.uv.es/ mount -t lustre wn181.ific.uv.es:/ificfs
/lustre/ific.uv.es/

fstab entry:
wn181@tcp:/ificfs /lustre/ific.uv.es lustre defaults,_netdev 0 0

A.2 StoRM Configuration and Setup

A.2.1 Current services

The current installation in production is based on Storm release 1.4.0. We are currently evaluating version 1.5 to be installed first in the pre-production services: [http://igrelease.forge.cnaf.infn.it/doku.php?id=doc:updates:ig3_1_32:infngrid-update31-61](http://igrelease.forge.cnaf.infn.it/doku.php?id=doc:updates:ig3_1_32:infngrid-update31-61)

A.2.2 OS, Required packages & Storm Installation

O.S: Storm is currently released for Scientific Linux 4, and we are using SLI 4.7, which is the Cern Version with some IFIC packages and configurations for our institute.

Relases for SL5 is promised by the developers, and although there are some users that have installed it over a SL5 there are a high dependency to legacy packages that makes it nonsense for us to install it in this moment.

Repositories:

We are using yum for the os, base and middleware repositories.

```
-rw-r--r-- 1 root root 256 Nov 13 12:03 ific-extra.repo
-rw-r--r-- 1 root root 585 Nov 13 12:03 glite-generic.repo
-rw-r--r-- 1 root root 362 Nov 13 12:03 dag.repo
-rw-r--r-- 1 root root 279 Nov 13 12:03 lcg-CA.repo
-rw-r--r-- 1 root root 630 Nov 13 12:03 jpackage.repo
-rw-r--r-- 1 root root 330 Nov 13 12:03 ig.repo
-rw-r--r-- 1 root root 350 Nov 13 12:03 ific-update-rpms.repo
```
NTP: Install and configure ntp for node date synchronization Java: Installed Java 1.5.0.15 and exclude jre from updates (/etc/yum.conf).

```bash
java-1.5.0-sun-devel-1.5.0.15-1jpp
java-1.5.0-sun-1.5.0.15-1jpp
```

If there are dependency problems in the installation you can remove the packages that require jre, in our SLI installation:

```bash
mozilla-plugin, ificdummy openoffice* jere susefax susefax-cover
```

Install CAs and Storm packages:

```bash
yum install lcg-Ca
```

we are using another node/s for GridFTP so we don't install GridFTP here although this is a possible scenario for smaller sites. Depending on access you can also install frontend and backend in different nodes, or even having several frontends for larger sites.

Other:

```bash
copy grid certificates /etc/grid-security
```

In production we create the user "storm" (with id 666:666) prior to yaim configuration to maintain the permissions to access the data in /Lustre.

A.2.3 Configuration

We are using the ig-yaim release provided by INFN which builds in top of generic gLite releases. Following instructions as indicated in [http://igrelease.forge.cnaf.infn.it/doku.php?id=doc:guides:install-3_1](http://igrelease.forge.cnaf.infn.it/doku.php?id=doc:guides:install-3_1)

Yaim packages:

```bash
Ig-yaim-storm-4.0.8-1
glite-yaim-core-4.0.8-7
ig-yaim-4.0.8-3_1
```

Prepare the configuration files: Yaim is able to maintain a site config file, and a per-service file to maintain the variables belonging to this service. We include here the specific Storm variables in the file services/ig-se_storm_backend (*check below some notes about not included parameters)

```bash
# Storm backend configuration variables

# ig : necesario para igyaim-config
NTP_HOSTS_IP=147.156.1.1
STORM_HOST=$SE_SRМ_HOST
```
STORM_PORT=8443
STORM_USER=storm

STORM_GRIDFTP_POOL_LIST=(tux4u01.ific.uv.es)

# Database settings.
# Host for database connection. COMPULSORY
STORM_DB_HOST=localhost
# User for database connection. COMPULSORY
STORM_DB_USER=storm

# Protocol support.
# If set to 'FALSE', the following variables prevent the corresponding protocol
# to be published by the StoRM gip. OPTIONAL - Available values: [true|false] - Default value: true
#STORM_INFO_FILE_SUPPORT=true
#STORM_INFO_GRIDFTP_SUPPORT=true
#STORM_INFO_RFIO_SUPPORT=true
#STORM_INFO_ROOT_SUPPORT=true
STORM_INFO_RFIO_SUPPORT=false
STORM_INFO_ROOT_SUPPORT=false

STORM_DEFAULT_ROOT=/lustre/ific.uv.es/grid

ATLAS_STORAGE_AREAS="atlas atlasdatadisk atlasmdisk atlaseduser
atlasgroupdisk atlashotdisk atlascalgroupdisk atlashprodisk
atlasscratchdisk atlassuserdisk"
STORM_STORAGEAREA_LIST="$ATLAS_STORAGE_AREAS dteam ops swetest ific
opcsic gcscic ngies opsngi genngi formgni grid4build nanodev mosfet
archist blast slgrid photonics qcomp frodock odthpiv filogen timones
ghase turbulencia meteo"

STORM_ATLASDATADISK_VONAME="atlas"
STORM_ATLASDATADISK_ROOT=$STORM_DEFAULT_ROOT/atlas/atlasdatadisk
STORM_ATLASDATADISK_TOKEN="ATLASDATADISK"
STORM_ATLASDATADISK_ACCESSPOINT=$STORM_ATLASDATADISK_ROOT

(the same for the other spacketoken)

About the defined Storage areas, are defined in the STORM_STORAGEAREA_LIST variable and configured in the STORM_<VO_NAME>_* related variables. But only the ones referring to ATLAS tokens are included here since the others are similar are ATLAS are in the scope of this wiki.

Configuration files issue:

the infn release installs by default several services that we don't need. We remove the configuration for GRIDEYE and NFS ( commenting/removing corresponding entry in yaim file: /opt/glite/yaim/node-info.d/ig-se_storm_backend):

/opt/glite/yaim/node-info.d/ig-se_storm_backend:
#config_fmon_client
#config_nfs_sw_dir_server

Configure the services:

/opt/glite/yaim/bin/ig_yaim -c -s site-info-pps-ific-SL4.def -n ig_SE_storm_backend -n ig_SE_storm_frontend
A.2.4 ATLAS authorization configuration

To allow a proper authorization configuration and being able access to data, we are using a special authorization plugin developed by us, and that now it is included in the main storm sources. This means you can use it, but it is not very well documented.

Modify the plugin to be used, and related variables in the storm.properties file:

```
/opt/storm/backend/etc/storm.properties:

authorization.sources = LocalAuthorizationSource
# -----------------------
# nuevas directrices
# -----------------------
# poner el permiso de escritura en directorios nuevos
# lo necesita LocalAuthorizationSource
directory.writeperm = true
```

In our srm and gridftp configuration, all atlas users all mapped to a single user. Check the mapping files:

```
/etc/grid-security/voms-grid-mapfile:

"/atlas/Role=lcgadmin" atls000
"/atlas/Role=lcgadmin/Capability=NULL" atls000
"/atlas/Role=production" atlp000
"/atlas/Role=production/Capability=NULL" atlp000
"/atlas/Role=software" atls000
"/atlas/Role=software/Capability=NULL" atls000
"/atlas/Role=pilot" atlu000
"/atlas/Role=pilot/Capability=NULL" atlu000
"/atlas" atlu000
"/atlas/Role=NULL" atlu000
"/atlas/Role=NULL/Capability=NULL" atlu000
```

```
file /opt/edg/etc/edg-mkgridmap.conf:

# ATLAS
# Map VO members (prd)
group vomss://voms.cern.ch:8443/voms/atlas/?/atlas/Role=production
atlp000
# Map VO members (sgm)
group vomss://voms.cern.ch:8443/voms/atlas/?/atlas/Role=lcgadmin
atls000
# Map VO members (root Group)
group vomss://voms.cern.ch:8443/voms/atlas/?/atlas atlu000
```

The idea is that with this configuration the plugin respects the ACLs on disk to authorize or not permission to access the files, according to ATLAS policies. You can check the ACLs that we have defined in the following spaces of the atlas pool in /lustre/ific.uv.es/grid/atlas:

```
# file: ..
# owner: root
# group: root
user::rwx
group::r-x
other::r-x
```
See in the previous box for example the atlasproddisk space, where common atlas users can access and read the files, but only production users (atlp mapped GID) can write. The access ACLs are created for this policy, and the default acl and mask to respect it in the created files and subdirectories.

A.2.5 Dinamic Info Provider

The initial default configuration for the available and free space in disk checks the disk, but it does not take into account our pool configuration for the different spaces, specially for ATLAS vo. It does not update correctly the information when it is runnint, in the Information system and in the internal Storm database.

To be able to publish information correctly we developed a script to update this data frequently, that you can find here: space-tokens-dynamicinfo_v2.tgz (https://twiki.ific.uv.es/twiki/bin/view/Atlas/LustreStoRM)

A.2.6 Configuration issues

- if you are not using the default port, check the configuration file since yaim is not changing it properly
  /opt/storm/backend/etc/storm.properties:
  storm.service.port=8444
  fe.port = 8444

- On previous version configuration was not understanding correctly FQAN at voms proxies, like this:
  /opt/srmv2storm/var/log/srmv2storm.log:

  06/14 18:15:46 26838,0 Rm: UserDN=/C=ES/O=DATAGRID-ES/O=IFIC/CN=Alvaro Fernandez Casani 06/14 18:15:46 26838,0 Rm: Number of FQANs: 0
• in this case, check that the yaim configuration creates the correct directories containing the voms servers certificates (or for newer versions the DN of the voms server):

```
[root@srnmv2 storm]# ll /etc/grid-security/vomsdir/
total 432
drwxr-xr-x  2 root root 4096 Nov 26 15:12 atlas
-rw-r--r--  1 root root 4620 Jan  19
       15:41 ccclgsvomsli01.in2p3.fr.1413.pem
-rw-r--r--  1 root root 1440 Dec  2 16:09 cert-voms01.cnaf.infn.it.pem
-rw-r--r--  1 root root 1440 Dec  2 16:09 cert-voms01.cnaf.infn.it.pem.1
-rw-r--r--  1 root root 1424 Dec  2 16:09 cert-voms01.cnaf.infn.it.pem.2
```

• Then correct interpretation of FQANs is accomplished:

```
02/18 13:13:08   495,0 Mkdir:
UserDN=/DC=es/DC=irisgrid/O=ific/CN=Alvaro-Fernandez
02/18 13:13:08   495,0 Mkdir: Number of FQANs: 2
02/18 13:13:08   495,0 Mkdir: FQAN[0]:
/atlas/Role=NULL/Capability=NULL
02/18 13:13:08   495,0 Mkdir: FQAN[1]:
/atlas/lcg1/Role=NULL/Capability=NULL
```

• If there are problems with java in reconfiguration, then it gives problem locating java and starting storm-backend server. Solution:

```
[root@ccc01 yaim]# alternatives --install /usr/bin/java java
/usr/java/jdk1.5.0_07/bin/java 1
[root@ccc01 yaim]# alternatives --config java
```

There are 2 programs which provide 'java'.

```
Selection Command
-----------------------------
*+ 1 /usr/share/java/libgcj-java-placeholder.sh
  2 /usr/java/jdk1.5.0_07/bin/java
```

Enter to keep the current selection[+], or type selection number: 2

**Appendix B: GPFS + Storm Install (UK case)**

**B.1 Installing GPFS**

**B.1.1 Install RPMS**

Start by installing the base release and then any updates on top of that.

Run install script and accept licence terms - requires an X connection unless --text-only is used

```
./gpfs_install-3.3.0-0_x86_64 --text-only
```

Install RPMS

```
cd /usr/lpp/mmfs/3.3/ rpm -Uvh gpfs.docs-3.3.0-0.noarch.rpm rpm -Uvh gpfs.gpl-3.3.0-0.noarch.rpm rpm -Uvh gpfs.gui-3.3.0-0.x86_64.rpm rpm -Uvh gpfs.msg.en_US-3.3.0-0.noarch.rpm
```
If they are update RPMS install those

cd /root/3.3.0-1 rpm -Uvh gpfs*rpm cd /usr/lpp/mmfs/

If necessary change /etc/redhat-release to make sure it claims to be Red Hat Enterprise Linux rather than Scientific Linux

B.1.2 Build source

cd /usr/lpp/mmfs/src
make Autoconfig
make World
make InstallImages

If necessary change /etc/redhat-release to make sure it claims to be Red Hat Enterprise Linux rather than Scientific Linux

B.1.3 Create GPFS Cluster

Ensure disk servers can access each other as root (via ssh or whatever methods selected) without password and that they can talk to each other on all relevant ports.

Add mmfs command dir to PATH
export PATH=$PATH:/usr/lpp/mmfs/bin

Create a GPFS cluster

mmcrcluster -N pool1.glite.ecdf.ed.ac.uk:manager-quorum,pool2.glite.ecdf.ed.ac.uk:manager-quorum \ -p pool1.glite.ecdf.ed.ac.uk -s pool2.glite.ecdf.ed.ac.uk -r /usr/bin/ssh -R /usr/bin/scp -C gridpp.ecdf.ed.ac.uk

-p and -s indicate the primary and secondary servers -r -R the remote shell and copy commands -C gives the cluster name (not resolved)

Say that you have a licence for the server nodes you will be using

mmchlicense server -N pool1,pool2

Startup cluster

mmstartup -a

B.1.4 Create a Network Shared Disk (NSD)

For each disk make a file sdk.dsc of the form

DiskName:ServerList::DiskUsage:FailureGroup:DesiredName:StoragePool

E.g.

sdk:pool1.glite.ecdf.ed.ac.uk::dataAndMetadata:7:cad7vd03vol2

The failure group (a number from -1 to 4000) is so that when gpfs creates replicas in can avoid putting them on the devices in the same failure group.

Then create the NSD with this file

mmcrnsd -F sdk.dsc

Add other servers as needed
B.1.5 Create GPFS filesystem

```
mmcrfs fs_test cab7vd03vol2:::dataAndMetadata:7::: -A no -B 512K -D posix -E no -j cluster -k all -K whenpossible -m 1 -M 2 -n 256 -T /exports/fs_test
```

fs_test is the name of the filesystem. cab7vd03vol2:::dataAndMetadata:7::: could be replaced with -F and the desc file above which will have been rewritten by mmcrnsd to contain this

-A no : Don't mount automatically
-B 512k Block size. Here we recommend using 1MB or greater
-D posix (Unless you want to support nfsv4)
-E no - Don't always update mtime values
-j cluster . After round robinning across disks in a filesystem, GPFS can "cluster" try to keep adjacent blocks together or scatter randomly.
-k all - support a range of acls
-K whenpossible - Whether replication is enforced
-m -M : minimum and maximum number of metadata replicas (Ideally set to more than one to assist in recoveries)
-n : nodes accessing filesystem
-T : mountpoint

Now mount filesystems:

```
mmmount all
```

B.2 Adding GPFS to storm

B.2.1 Mount GPFS space onto StoRM node

Follow steps above to install GPFS onto the storm node.

On one of the existing nodes in the GPFS cluster do
```
mmaddnode se2.glite.ecdf.ed.ac.uk
```

On the StoRM node do
```
mmchlicense server -N se2
```

Check the state
```
mmgetstate -a
```

Startup
```
mmstartup -N se2
```

Mount filesystem,
```
mmmount fs_test
```

B.2.3 Configure StoRM to use GPFS drivers

Add either through yaim or directly into the namespace.xml file.
For YAIM set

```bash
STORM_ATLAS_FSTAB=gpfs
STORM_ATLAS_FSDRIVER=gpfs
STORM_ATLAS_SPCDRIVER=gpfs
```

for each VO. Then rerun Yaim.

Note: Rome case modified init.d/storm-backend like the following to fix two issues:
- "storm-backend status" was returning "Storm dead" due to using the wrong ps command
- They added a part to save stderr and stdout which would otherwise be overwritten at each Storm restart.

To define the spacetokens I just had to (in services/ig-se_storm_backend) set:

```bash
STORM_ATLASLOCALGROUPDISK_VONAME=atlas
# **Optional** variables
STORM_ATLASLOCALGROUPDISK_ACCESSPOINT=/atlaslocalgroupdisk
STORM_ATLASLOCALGROUPDISK_FSTYPE=gpfs
STORM_ATLASLOCALGROUPDISK_GRIDFTP_POOL_LIST=${MY_STORM_ST_GRIDFTP_POOL}
STORM_ATLASLOCALGROUPDISK_QUOTA=true
STORM_ATLASLOCALGROUPDISK_QUOTA_DEVICE=/dev/gpfsdev_data02
# this is the GPFS device hosting spacetoken filesets
STORM_ATLASLOCALGROUPDISK_QUOTA_FILESET=atlaslocalgroupdisk
STORM_ATLASLOCALGROUPDISK_ROOT=${MY_STORM_ST_ATLAS_ROOT}/atlaslocalgroupdisk
# this is the physical mount point
STORM_ATLASLOCALGROUPDISK_ROOT_HOST=${MY_STORM_ST_ROOT_HOST}
STORM_ATLASLOCALGROUPDISK_RFIO_HOST=${MY_STORM_ST_RFIO_HOST}
STORM_ATLASLOCALGROUPDISK_STORAGECLASS=T0D1
STORM_ATLASLOCALGROUPDISK_TOKEN=ATLASLOCALGROUPDISK
# STORM_ATLASLOCALGROUPDISK_ACLMODE=aot
```

and also one need to set the ACL for local access like this:

```
[root@storm-01 ~]# mmgetacl -d /storage/data02/atlaslocalgroupdisk/
#owner:storm
#group:storm
user::rwxc
group::-----
other::-----
mask::rwx-
#group:atlas:r-x-
#group:prdatlas:r-x-
#group:sgmatlas:r-x-
#group:pilatlas:r-x-
```
group:atlasrm3:r-x-  <--- this is the group for local users, with this addition they can do (read-only) POSIX access to LOCALGROUP files

**B.2.4 Testing**

Copy file in with lcg-cp

```
  lcg-cp -D srmv2 -b -v --vo atlas file:/phys/linux/wbhimji/bob srm://se2.glite.ed.ac.uk:8444/srm/managerv2?SFN=/atlas/brian2
```

Watch it magically appear in your GPFS space.

Retrieved from "https://www.gridpp.ac.uk:443/wiki/Storm_GPFS_Install"

Note: Note that, after Roma3 reinstalled Storm 1.5, they found they had to set `security.disable.mapping=true` in the frontend configuration file. Otherwise transfers would complete OK but FTS would think they crashed, and would start a new one... rapidly filling up my storage space. The problem may have been fixed in a recent/future release.

**Appendix C: Inventory/Survey form**

To have a real overview of technologies, configuration (Hardware and Software) at various sites using Lustre File System and GPFS, an inventory/survey form for each one was created.

The information can be found for lustre on:

- http://spreadsheets.google.com/ccc?key=0AjDZJgYDLICadFVFQkFFczdORDY2bC1raTRkd21hN1E&hl=en&pli=1#gid=0

and for GPFS on:

- http://spreadsheets.google.com/gform?key=0AqcCwHr39RA6dGdiMU5aajNvYnNSRktoOWhSQ3V5aWc&hl=en&gridId=0#chart