Deployment and Management of Grid Services for a data intensive complex infrastructure

Álvaro Fernández (IFIC – Valencia)

5th User Forum

Uppsala, April 2010

www.eu-egee.org

e-infrastructure

Contents

- Motivation for the talk
- Proposed Solution
- Computing and Storage Hardware
- Configuration and Management
- Developments
- Conclusions and Future

Motivation

- **Several user communities require access to computational and storage resources in efficient manner.**
  - National Grid Initiative applications
  - Grid-CSIC
  - Atlas Tier2 and Tier3
  - Local users
- **Different kinds of applications and data access patterns**
  - Parallel and sequential Applications
  - Public and private data. Need for sharing policies
  - File size varies, sequential and random access

Proposed solution

- To use grid technologies that we have been working with, and Hardware and Software configurations that alow to grow in the future.
- Storage and Network Hardware based on a NAS
- Lustre as a parallel backend filesystem
- Storm as a Storage Resource Manager for our disk based storage
- Grid Services based on glite
- Development of plugin, Tools and High level applications for user commodity

Computing Hardware

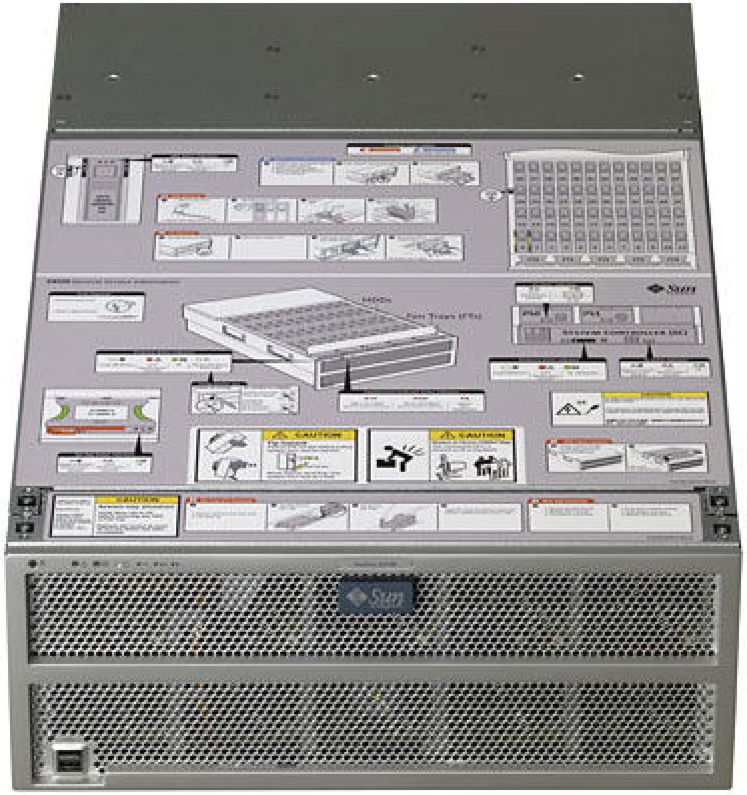
- Grid-CSIC resources
- Sequential: 106 nodes
  - 2xQuad Core Xeon E5420 @2.50GHz
  - 16 Gbytes Ram
  - 2xHD SAS 134 BG RAID 0
- Parallel: 48 nodes
  - Same as before with Infiniband Mellanox Technologies MT25418 DDR 4x

- Atlas VO resources
- 32 + 19 nodes HP DL160
  - 2xQuad Core Xeon E5472 @3.0GHz
  - 16 GB RAM
  - 2xHD SAS 134 BG RAID 0

- Total cores : 1704

Storage Hardware

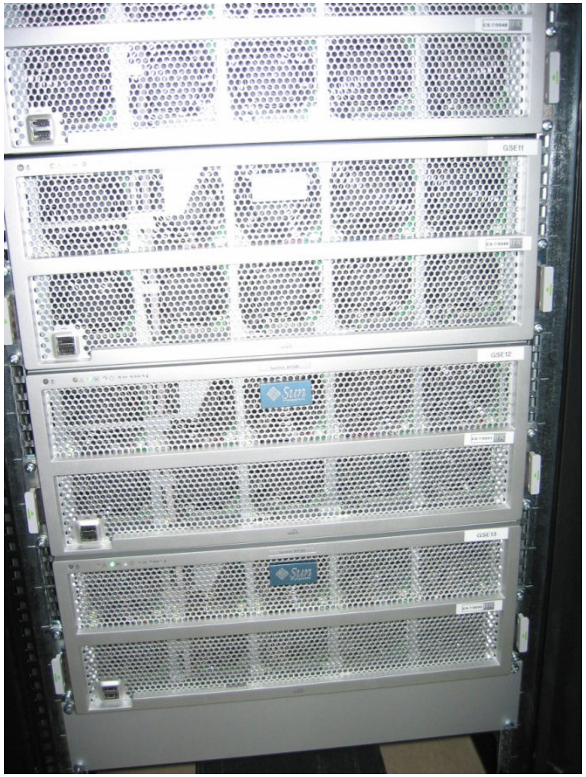
- Tape library for long time storage (with legacy castor sw)
- Disk servers to store online data (as part of Tier-2 requirements)
  - Fast access and always available



- Sun X4500/X4540 server
  - 48 disks (500Gb/1Tb), with 2 disks for OS
  - Redundant power supply
  - 4x1Gb Ethernet ports
  - Scientific Linux IFIC v4.4 with Lustre kernel
  - Ext3 system (volumes until 18/36 Tib)

Storage servers configuration

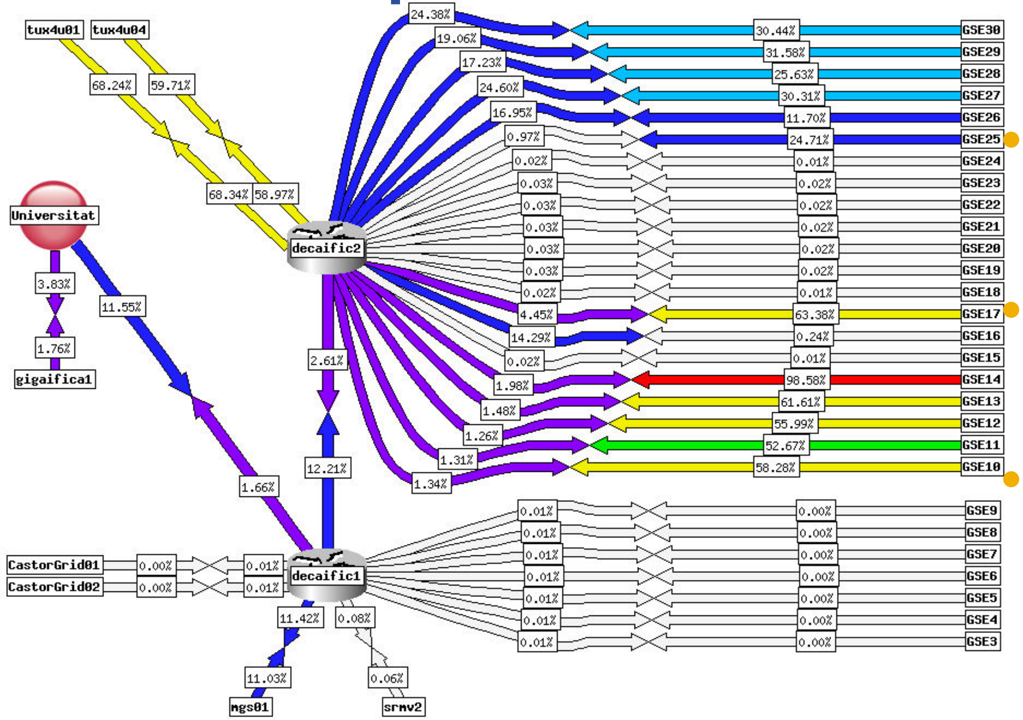
- **IFIC tested best configuration:**
  - Disks in raid 5 (5x8 + 1x6). Usage ratio of 80%
  - Best performance with 1 raid per disk controller
  - Bonnie++ aggregated test results:
    - Write: 444,585 KB/s
    - Read: 1,772,488 KB/s
- **Setup:**
  - We have 21 servers providing a raw capacity of around 646 TB



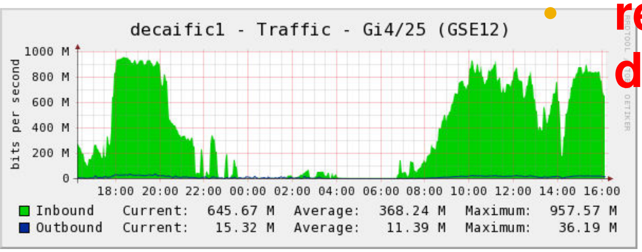
Network

- Data Network based on gigabyte ethernet.
- 10GB uplink to the backbone network

Created: Apr 08 2010 16:00:33



- Cisco 4500 – core centre infrastructure.
- Cisco 6500 – scientific computing infrastructure
- Data servers with 1GB connection. Channel bonding tests were made. We aggregate posibly 2 channels in future.
- WNs and GridFTP servers with 1GB

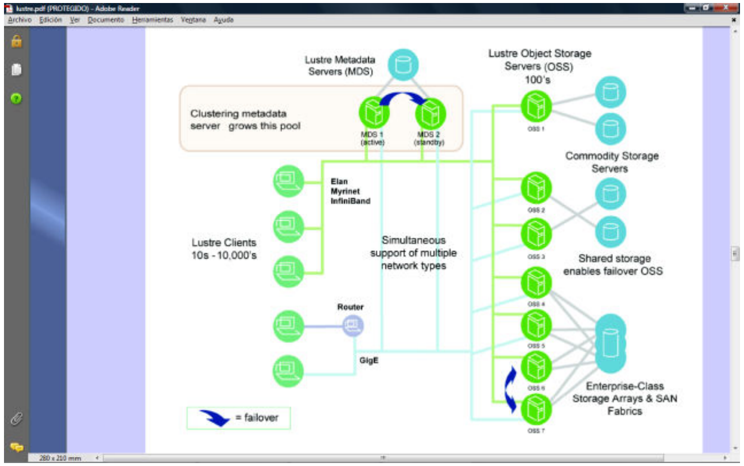


```
🟩 Inbound   Current:  645.67 M  Average:  368.24 M  Maximum:  957.57 M
■ Outbound  Current:   15.32 M  Average:   11.39 M  Maximum:   36.19 M
```

- reach 1Gbit per data server

Lustre installation

- **We chose Lustre** because is a **Petabyte** scale **posix** filesystem, with **ACL** and **quotas** support. **Opensource** and supported by SUN/Oracle. In our installation with a single namespace (/lustre/ific.uv.es):
- **MGS: Management Server** (1 per cluster) + **MDS: Metadata Server**. (1 per filesystem)
  - 2xQuadCore Intel(R) Xeon(R) CPU
  - 32 GB.
  - Planned to add HA
  - Currently 4 Filesystems (MDT)

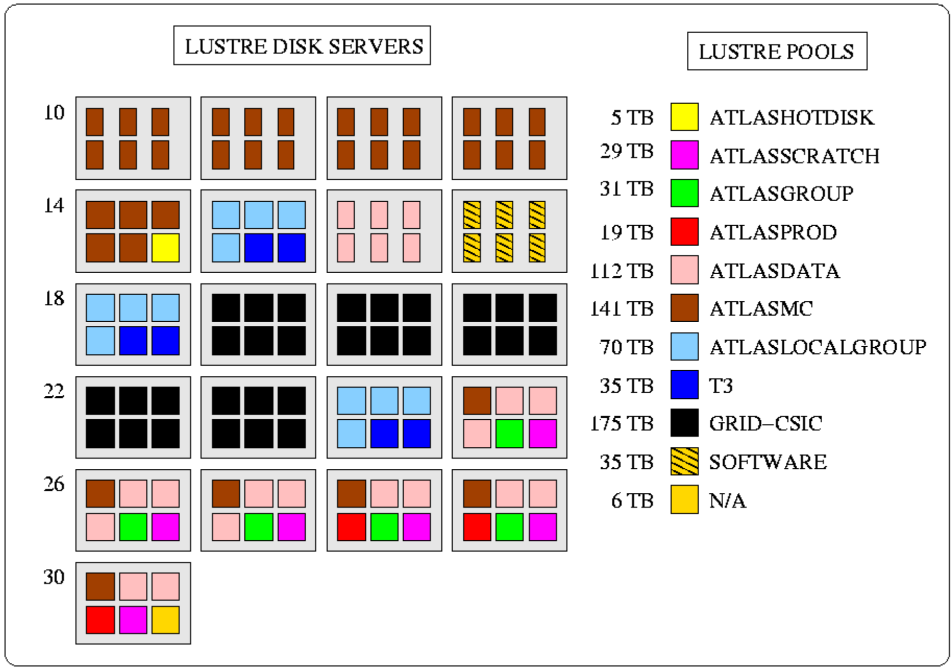

lustre.pdf (PROTEGIDO) - Adobe Reader
Archivo Edición Ver Documento Herramientas Ventana Ayuda
Lustre Metadata Servers (MDS)
Clustering metadata server grows this pool
MDS 1 (active)
MDS 2 (standby)
Lustre Object Storage Servers (OSS) 100's
OSS 1
OSS 2
OSS 3
OSS 4
OSS 5
OSS 6
OSS 7
Commodity Storage Servers
Shared storage enables failover OSS
Enterprise-Class Storage Arrays & SAN Fabrics
Elan Myrinet InfiniBand
Lustre Clients 10s - 10,000's
Simultaneous support of multiple network types
Router
GigE
= failover
280 x 210 mm

- **OSS: Object Storage Server**. (Several per cluster guaranteeing scalability) 1 in every SUN X4500.
  - 6 OSTs per server (raid-5 5x8 + 1x6)
- **Clients**. (SL5 kernel with lustre patches), mainly User Interfaces (UI) and Worker Nodes (WN) with configurations in RW and RO mode

Filesystems and Pools

- With the Lustre release 1.8.1, we have added pool capabilities to the installation.
- Pools Allow us to partition the HW inside a given Filesystem
  - Better data management
  - Assign determined OSTs to a application/group of users
  - Can separate heterogeneous disks in the future

- **4 Filesystems with various pools:**
  - /lustre/ific.uv.es Read Only on WNs and UI. RW on GridFTP + SRM
  - /lustre/ific.uv.es/sw. Software: ReadWrite on WNs, UI
  - /lustre/ific.uv.es/grid/atlas/t3 Space for T3 users: ReadWrite on WNs and UI
  - xxx.ific.uv.es@tcp:/homefs on /rhome type lustre. Shared Home for users and mpi applications: ReadWrite on WNs and UI



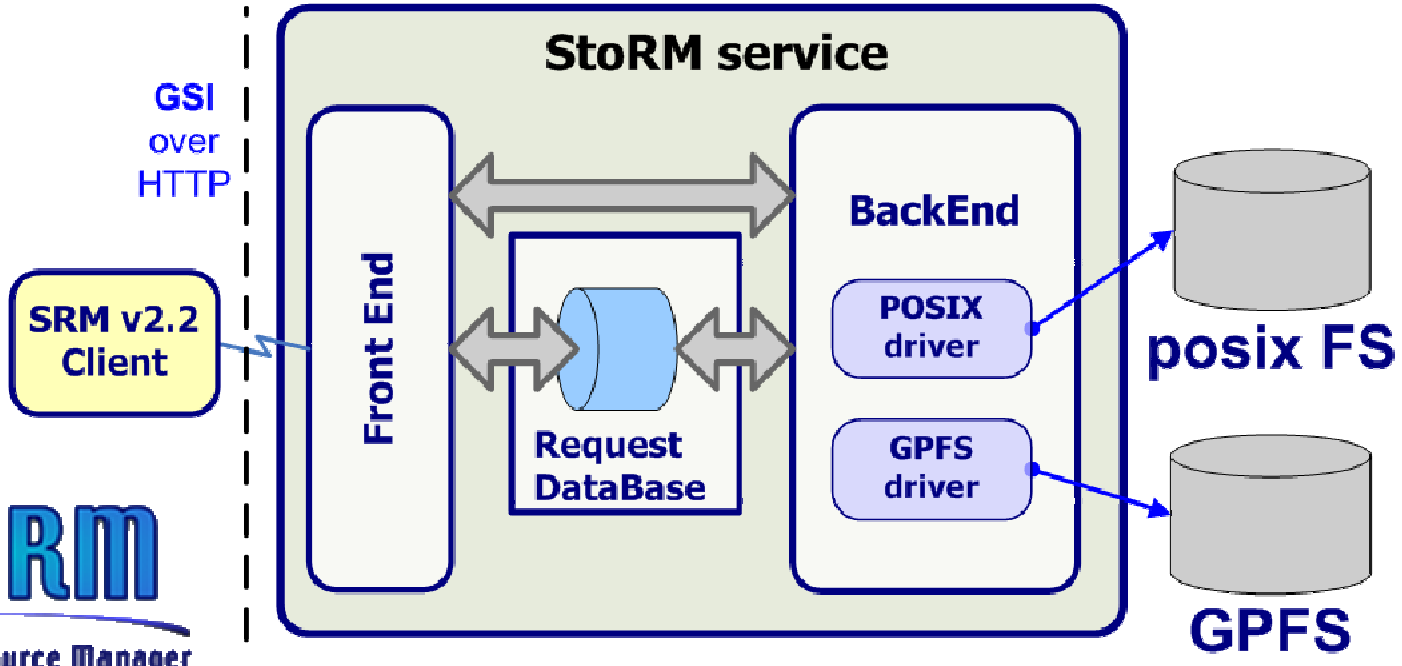
5 TB	ATLASHOTDISK
29 TB	ATLASSCRATCH
31 TB	ATLASGROUP
19 TB	ATLASPROD
112 TB	ATLASDATA
141 TB	ATLASMC
70 TB	ATLASLOCALGROUP
35 TB	T3
175 TB	GRID–CSIC
35 TB	SOFTWARE
6 TB	N/A

Lustre management

- **Quotas**
  - Applyed per Filesystem
  - For the Users in the Tier3 space
  - For the Shared Home Directory
  - Not being applied in general VO pools/spaces because all data is common for the VO
- **ACLs**
  - Applied in the diferent VO pools/spaces to allow policies to be implemented (ie: only some managers can write, all users can read)
  - In the User and shared home directories
  - In the SW directories
  - This to be respected by the higher Grid Middleware Layers (STORM part of the presentations)

StoRM as an SRM server

- Interface of the storage network with the GRID.
- For disk based Storage Elements
- Implementing SRMv2.2 specification.
- Allows accessing Lustre filesystem with the posix interface.

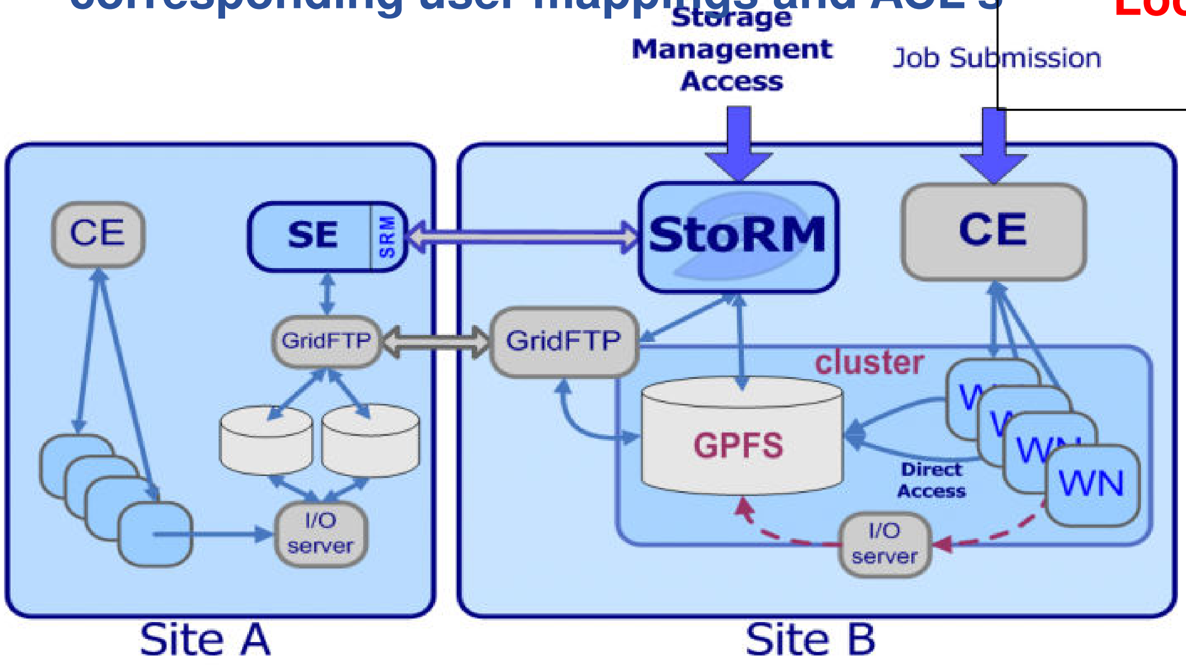


Enabling Grids for E-sciencE

Storm local implementation & policies

- Access like a local file system, so it can create and control all the data available in disk with a SRM interface. (BE + FE Machine)
- Coordinate data transfers, real data streams are transferred with a gridFTPserver in another physical machine. (Pool of gridfts)
- Enforce authorization policies defined by the site and the VO. (users are mapped to common vo account)
- We Developed Authorization plugin to respect local file system with the corresponding user mappings and ACL's

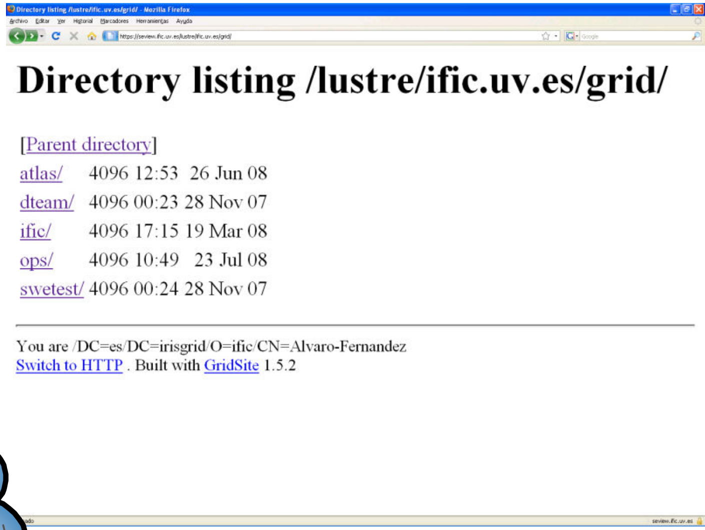
LocalAuthorizationSource available > Storm 1.4



- # file: atlasproddisk
- # owner: storm
- # group: storm
- user::rwx
- group::r-x
- group:atlas:r-x
- group:atlp:rwx
- mask::rwx other::---
- default:user::rwx
- default:group::r-x
- default:group:atlas:r-x
- default:mask::rwx default:other::---

Other Services: SEView

- Besides posix CLI:
- IFIC users additionally can access data through a WEB interface.
- Ubiquitous and X.509 secured easy read-only access.
- Developed with GridSite

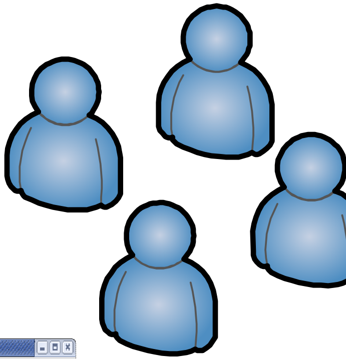

Directory listing /lustre/ific.uv.es/grid/ - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
https://seview.ific.uv.es/lustre/ific.uv.es/grid/
Google

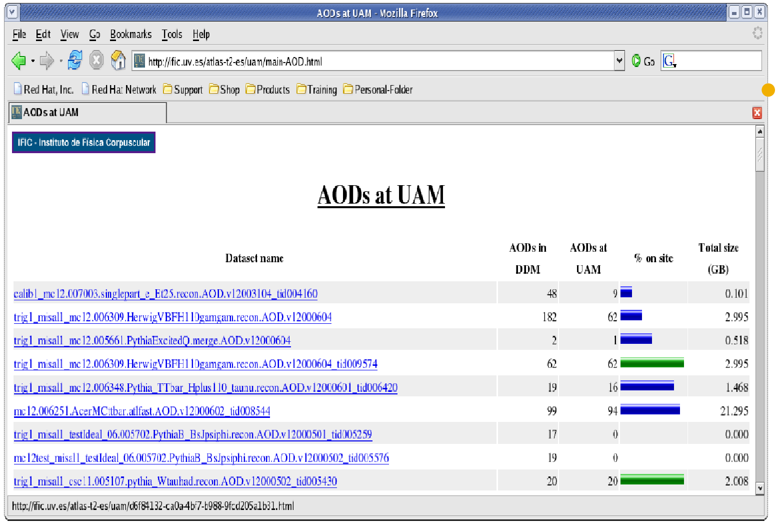
Directory listing /lustre/ific.uv.es/grid/

[Parent directory]

atlas/	4096	12:53	26 Jun 08
dteam/	4096	00:23	28 Nov 07
ific/	4096	17:15	19 Mar 08
ops/	4096	10:49	23 Jul 08
swetest/	4096	00:24	28 Nov 07

You are /DC=es/DC=irisgrid/O=ific/CN=Alvaro-Fernandez
Switch to HTTP . Built with GridSite 1.5.2




AODs at UAM - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://ific.uv.es/atlas-t2-es/uam/main-AOD.html
Go
Red Hat, Inc. Red Hat Network Support Shop Products Training Personal-Folder
AODs at UAM
IFIC - Instituto de Física Corpuscular

AODs at UAM

Dataset name	AODs in DDM	AODs at UAM	% on site	Total size (GB)
calib1_mc12.007003.singlepart_e_Et25.recon.AOD.v12003104_tid004160	48	9		0.101
trig1_misal1_mc12.006309.HerwigVBFH110gamgam.recon.AOD.v12000604	182	62		2.995
trig1_misal1_mc12.005661.PythiaExcitedQ.merge.AOD.v12000604	2	1		0.518
trig1_misal1_mc12.006309.HerwigVBFH110gamgam.recon.AOD.v12000604_tid009574	62	62		2.995
trig1_misal1_mc12.006348.Pythia_TTbar_Hplus110_taunu.recon.AOD.v12000601_tid006420	19	16		1.468
mc12.006251.AcerMCttbar.atlfast.AOD.v12000602_tid008544	99	94		21.295
trig1_misal1_testIdeal_06.005702.PythiaB_BsJpsiphi.recon.AOD.v12000501_tid005259	17	0		0.000
mc12test_misal1_testIdeal_06.005702.PythiaB_BsJpsiphi.recon.AOD.v12000502_tid005576	19	0		0.000
trig1_misal1_csc11.005107.pythia_Wtauhad.recon.AOD.v12000502_tid005430	20	20		2.008

http://ific.uv.es/atlas-t2-es/uam/d6f84132-ca0a-4bf7-b988-9fcd205a1b31.html

WEB Interface to check datasets (Atlas groups of files) at the Distributed Tier-2

Conclusions & Future

- **Presented a data management approach for multi-VO in a computer centre based on grid technologies and mature software.**
  - Tested well in stress tests (Step09,UAT)
  - more challenges to come, prepared to manage problems
- **Lustre gives good performance and scalability to grown next year until Petabyte scale.**
- **Storm presents our disk-based storage to the grid.**
  - Performed very well, and suits our necessities.
  - We would like it to continue not to being a complex system
- **In the future, deploy an implementation for HSM and Lustre:**
  - Follow CEA/Oracle work on this, these days: http://wiki.lustre.org/index.php/Lustre_User_Group_2010)

Final words

More information:

https://twiki.ific.uv.es/twiki/bin/view/Atlas/LustreStoRM

https://twiki.ific.uv.es/twiki/bin/view/ECiencia/WebHome

Alvaro.Fernandez@ific.uv.es

Thanks for your attention !