Rome 28-30 September 2009, Università Roma Tre
European AFS meeting 2009
http://www.dia.uniroma3.it/~afscon09

AFS in a GRID context

G. Bracco, S.Migliori, S. Podda, P. D'Angelo A. Santoro, A. Rocchi, C. Sciò
ENEA FIM, C.R. ENEA Frascati
V. E. Fermi 45 Frascati ROMA (Italy)
bracco@enea.it

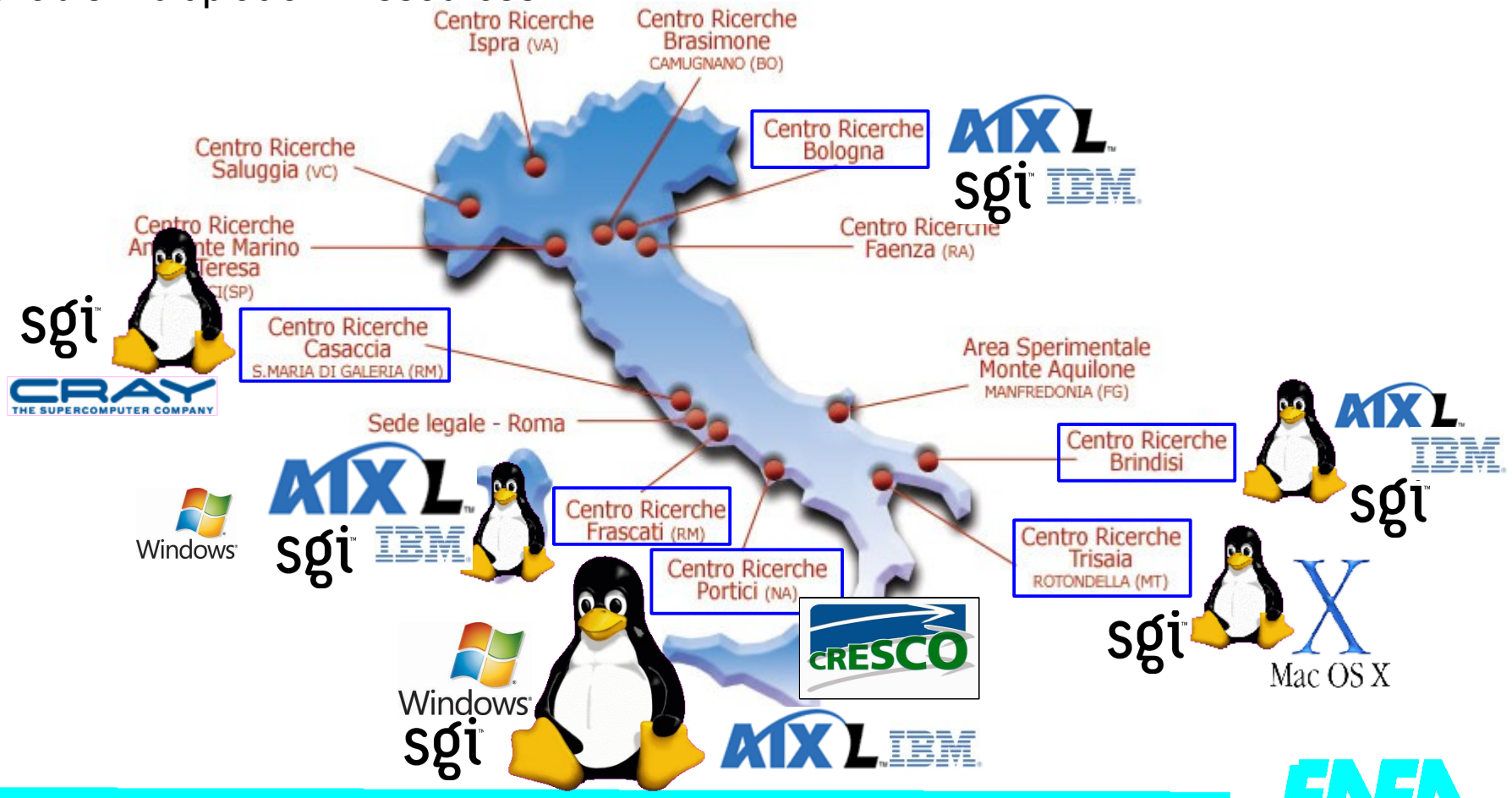
Motivation

Illustrate the utilization of AFS in a GRID framework, showing architectural aspects, advantages and issues from experience in ENEA-GRID, a GRID infrastructure operating since 1999

- ENEA-GRID and its computational resources, CRESCO HPC system, interoperability with other grids
- AFS and ENEA-GRID
- GRID related AFS pros/cons
- Conclusions

ENEA-GRID www.eneagrid.enea.it

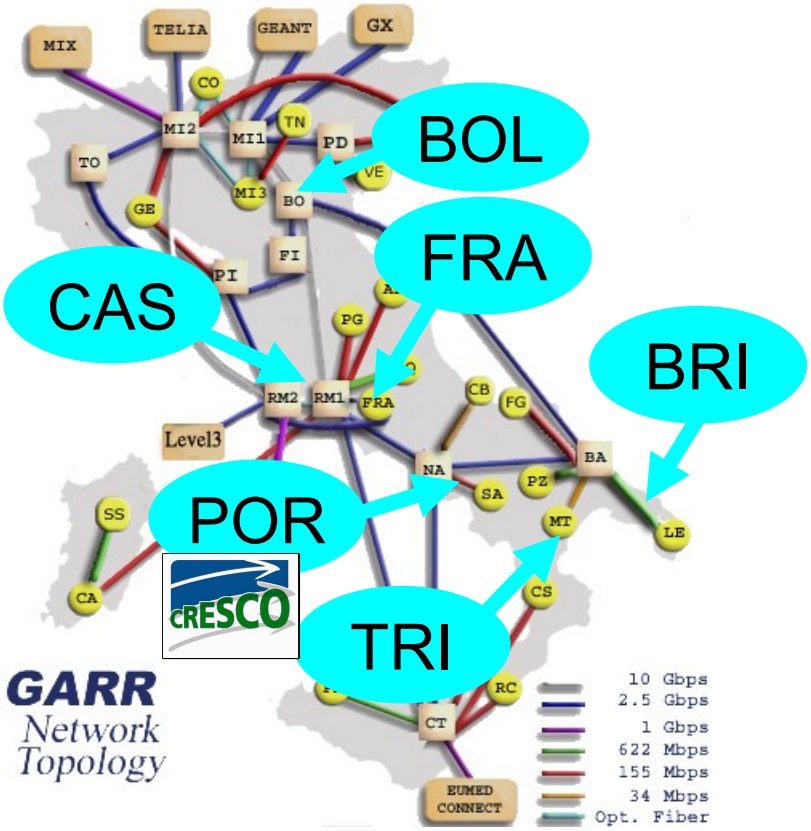
ENEA, the Italian Agency for New Technologies, Energy and Sustainable Development, has 12 research sites, 6 of them with computer centers. All ENEA main computational resources are integrated in the ENEA-GRID infrastructure which provides to ENEA researchers and their collaborators an easy access to the available multiplatform resources.



ENEA GRID and the Network

WAN connection is provided by GARR, the Italian Academic and Research Network Consortium: for ENEA
9 PoP, 18-2000 Mbps

Brindisi	150 Mb/s
Bologna	30 Mb/s
Casaccia	200 Mb/s
Frascati	1000 Mb/s
Portici	2000 Mb/s
Trisaia	18 Mb/s
Palermo	
Pisa	
Roma Sede	



ENEA-GRID Architecture

GRID functionalities (authentication, authorization, resource and data discovery, sharing & management) are provided by mature multiplatform components (ENEA-GRID Middleware):

Distributed File System : OpenAFS/Kerberos 5 integrated

WAN resource manager: LSF Multicluster [www.platform.com]

User Interface: Java & Citrix Technologies, now also FreeNX

These production ready components have permitted to integrate reliably over the years (ENEA-GRID started in 1999) the state of the art computational resources.

ENEA participates in GRID projects (Datagrid, EGEE,EGEE-II & III, BEINGRID, Italian PON projects, GRISU..) focusing on interoperability solutions with other middleware (gLite, Unicore): a gateway implementation method (SPAGO, Shared Proxy Approach for GRID Objects) has been developed and applied to interoperability with gLite based grids.

ENEA-GRID computational resources

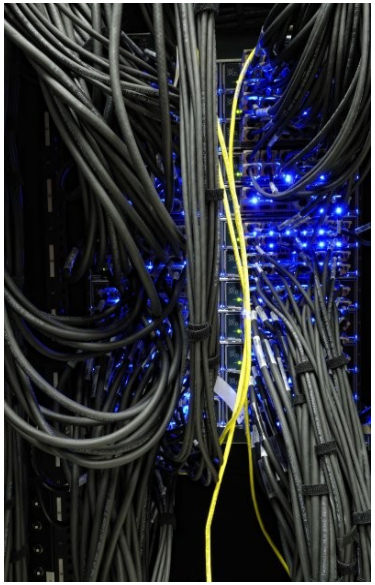
Hardware

- Most relevant: CRESCO HPC system, located in Portici (NA) rank #125 in Top500 June/2008 (rank #2 in Italy) 17.1 Tflops, 300 hosts, 2720 cores, InfiniBand 4xDDR

- Others: ~100 hosts ~650 cpu
  - AIX: IBM SP5 256 cpu (12 p575 1.5GHz, 16 cpu + 1 p595 1.9 Ghz, 64 cpu, 1.5 Tflops); SP4, 96 cpu
  - SGI Altix 350 (IA64) 32 cpu & Onyx
  - Cray XD1 24 cpu
  - Linux clusters 32/x86_64; Apple cluster; Windows servers....

Software: commercial codes (fluent, ansys, abaqus..); computational environment (Matlab, IDL,..), research codes (CPMD, MCNP.....)

CRESCO HPC system (1)



CRESCO (Computational Research Center for Complex Systems) is an ENEA Project, co-funded by the Italian Ministry of University and Research (MUR) in the framework of PON 2000-2006 call 1575. In operation since spring 2008.

www.cresco.enea.it

CRESCO HPC system (2)

A general purpose facility based on leading multicore x86_64 technology. Three sections:

Section 1: 672 cores: large memory requirement

42 fat nodes IBM x3850/x3950-M2, 4 Xeon Quad-Core Tigerton E7330 processors (2.4GHz/1066MHz/6MB L2), 32 GB RAM (4 extra-fat nodes with 64 GB RAM, 1 coupled node (x2) 32 cores /128 GB.

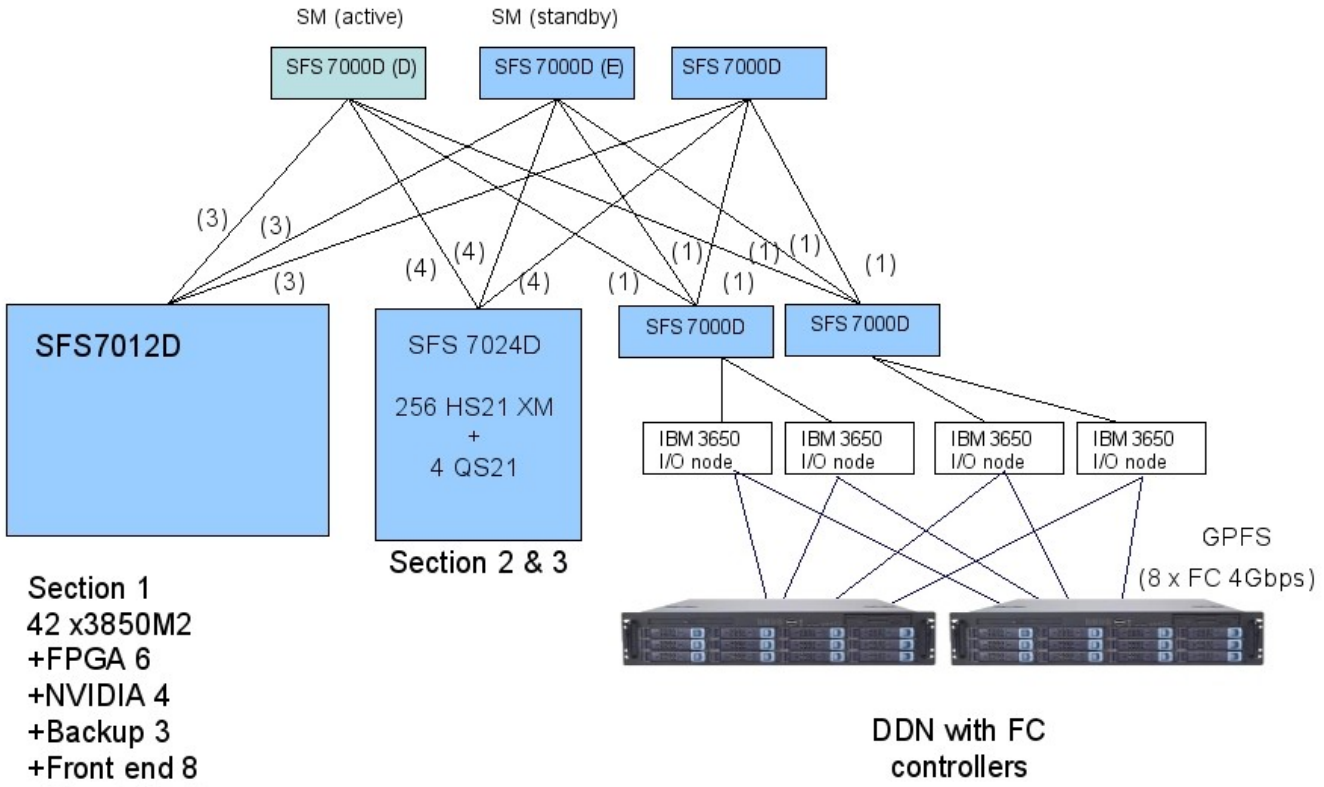
Section 2: 2048 cores : high scalability

256 blades IBM HS21, Xeon Quad-Core Clovertown E5345 processors (2.33GHz/1333MHz/8MB L2), 16 GB RAM

Experimental section: cell, FPGA, GPU

- 4 blades IBM QS21, 2 Cell BE Processors 3.2 Ghz
- 6 nodes IBM x3755, 4 sockets AMD Dualcore 8222, FPGA VIRTEX5 LX330
- 1 node IBM x 3755, 4 sockets AMD Dualcore 8222, NVIDIA Quadro FX 4700 X2 video card

CRESCO Infiniband Network



CRESCO Storage:

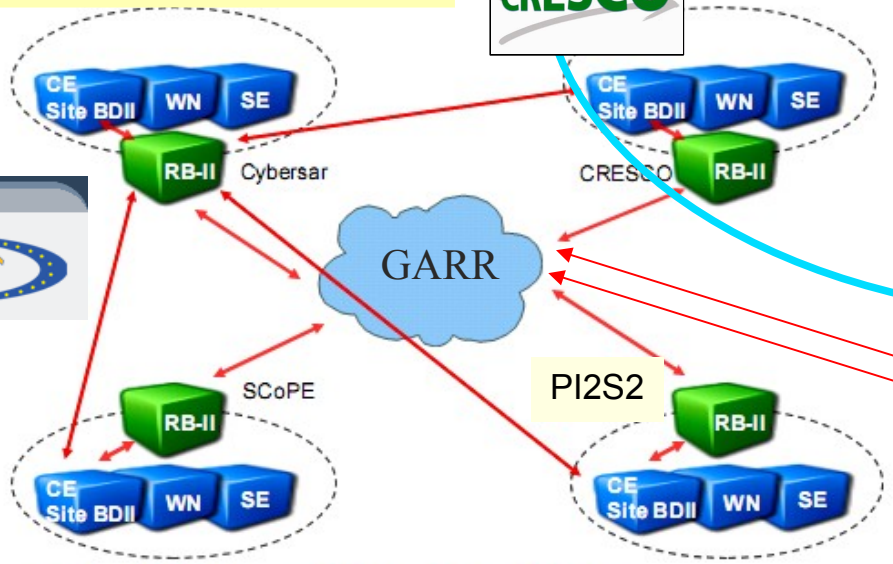
DDN S2A9550

180 TB raw; 120 TB GPFS

ENEA-GRID interoperability in production

IGI & EGEE: Italian and European GRID

GRISU: Southern Italy GRID



www.grisu-org.it

flusso delle informazioni del Sistema Informativo Grid

ENEA-GRID and Research

- Requirements
  - Computing power: High Performance / High Throughput Computing
  - Systems at the state of art of computational performance
  - Reliable and stable user environment
  - Tools for collaborative activities
  - User support (access, infrastructure integration, resource availability and monitoring)
- ENEA-GRID & Research: ENEA Examples
  - Computational Chemistry
  - Nuclear Fusion, plasma stability
  - Climate/Weather/Ocean Simulations
  - Pollutant Athmospheric Diffusion
  - Combustion Simulation
  - ....

ENEA-GRID & Industry

- Additional specific requirements/issues
  - Access to proprietary codes “Certification issues” (license management)
  - Reduction in simulation time, also at the price of weak problem scalability
  - Synergy between proprietary and open source codes (e.g. OpenFoam vs Fluent)
  - Customization of access control to codes and data
  - Security and traceability, intellectual property protection
  - SLA, charging model,....
- Some examples with the following companies
  - AVIO
  - AnsaldoEnergia/AnsaldoRicerche
  - AAPS Informatica
  - CETMA
  - ...

AFS and GRID

AFS was born much earlier the GRID paradigm was introduced (1995 by Foster and Kesselman)

- But many of the AFS features were already GRID like
  - Born on the WAN with a Global name space
  - Strong authentication
  - Data location transparency for the user
  - Powerful PTS group management
  - Multi-platform
- When GRID infrastructures started to develop
  - AFS was a closed/proprietary software
  - Some performances issues
- Not a problem for ENEA when ENEA-GRID was started
  - AFS was multi-platform and production ready

AFS and ENEA-GRID

- History
  - 1999 initial Transarc installation
  - Collaboration with CASPUR since 2004
  - OpenAFS migration 2006/may
  - Kerberos 5 (MIT) migration 2007/february
- Features
  - enea.it cell, servers and architecture
  - User and Data layout
  - Management Utilities
  - LSF integration

enea.it AFS cell

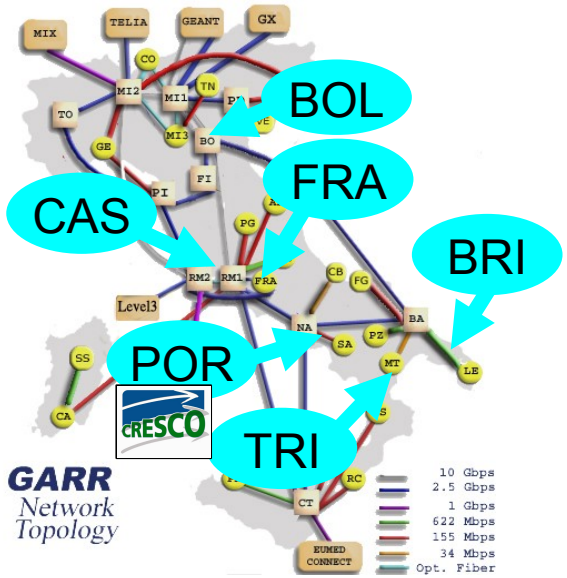
- **Dbservers, (scientific linux 4.x, openafs 1.4.0)**
  - 3 standard dbservers located in one site (Frascati) and listed in the public available CellServDB at grancentral.org.
  - 5 “clone” dbservers in the other sites (2 in Portici, where CRESCO HPC is located)
  - All dbservers run fakeka and are secondary KDC of the Realm ENEA.IT, primary KDC is in Frascati.
- **Fileservers (SL 4x-5x, AIX 5.3, Solaris 9, openafs 1.4.1,1.4.4,1.4.10)**
  - 11 fileservers: at least two in the main sites, Portici and Frascati
  - Total allocated space: 44 TB Total used space 11.2 TB
- **Clients**
  - SL 4x, SL 5x, CentOS 5.x, AIX 5.1-5.3, IRIX 6.5, MacOSX 10.4.11
  - Computational Working nodes: CellServDB contains only local dbservers+1 standard on for sites where only 1 dbserver is available
  - Vlserver and fileserver preferences set automatically by a cron script

enea.it user & data setup (1)

- User propagation using AFS: /etc/passwd is managed using a hourly cron job and user data in AFS space

- AFS space is organized following the ENEA sites

  - /afs/enea.it ./bol
    - » ./bri
    - » ./cas
    - » ./fra
    - » ./por
    - » ./tri



enea.it user & data setup (2)

- User space: dual entry to user HOME , global and site entries
  - /afs/enea.it/user/b/bracco < global path

  is the same area as
  - /afs/enea.it/fra/bracco < local path : HOME on Frascati site

    - » ~/private
    - » ~/public
    - » ~/public_html
    - » ~/rem/por user volume on Portici site
    - » ~/rem/bri user volume on Brindisi site
    - » ~/PFS/por link to user GPFS in Portici site

user space on remote sites is provided directly from the user HOME

enea.it user & data setup (3)

- Data space for software and projects (organized in 2 level volumes: project volume, subproject volumes)
  - /afs/enea.it/software
    - » ./maxima
    - » ./maxima/html > http://www.afs.enea/software/maxima
  - /afs/enea.it/projects
    - » ~/eneagrid
    - » ~/eneagrid/html > http://www.afs.enea/project/eneagrid
    - » ~/eneagrid/docs
    - » ...

project and software volumes have standard name p.site.nickname.subx while for users the standard approach is used: e.g user.bracco

enea.it user & data setup (4)

- Also Project/Software area have also global and local access paths
  - /afs/enea.it/projects
    - » ~/eneagrid
  - /afs/enea.it/fra
    - » ~/arcproj/eneagrid
    - » ~/user
    - » ~/remote < a site based entry for user remote volumes

- Backup volumes mount point have only a global access path
  - /afs/enea.it/backup
    - » ./user/b/bracco
    - » /user/bracco.rem/fra > remote volume backup
    - » ./arcproj/eneagrid
    - » ./arcproj/eneagrid.html

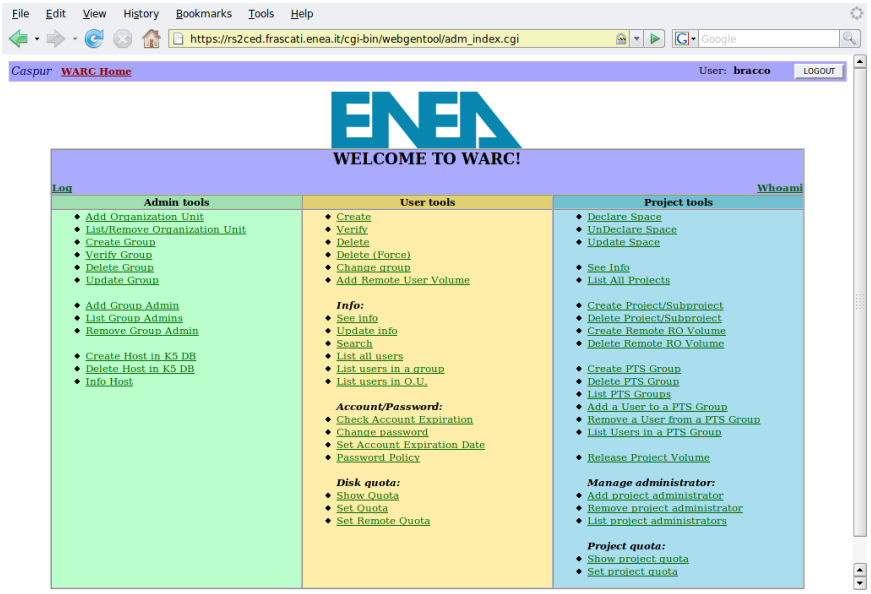
enea.it user & data setup (5)

- Projects are also used to provide to the user a scalable data space for code results
  - In ENEA-GRID it has been chosen to limit the volume size to about 50 GB for an easy management
  - Many subproject volumes can be allocated to provide a large data space to selected users
    - That is pratical when user needs some hundreds GB
    - It becomes cumbersome for the user when many TB are required
      - In that case the user must solve the problem of fitting his data into fixed size AFS volumes

.

Management Utilities

- Users and projects are maintained using a dedicated application (WARC) developed by CASPUR in collaboration with ENEA; WARC provides also management delegation to project administrators and site administrators



WARC

- User management
- Project creation
- Project management
  - PTS groups
  - Release RO volumes

- Historical analysis and routine checkup of enea.it cell is performed using AMACA application: see Alessio Rocchi presentation on Tuesday 29/9

LSF integration/ Interoperability issues

- Token management is the key point for LSF integration
  - The integration is performed using gettok/puttok routines in the version provided by Platform which required some patch to run correctly with the tokens generated by kinit afslog procedure.
  - blaunch integration for mpi jobs is in production
  - gettok/puttok available also for windows (result of a collaboration between ENEA and Salerno University in the framework of CRESCO project)
- Token management has been also an issue in participating to other GRID projects (for example EGEE) where authentication is based on X509 certificate, with extended features (e.g. Virtual Organization support)
  - gssklog/gssklogd is the tool used to convert X509 to AFS token
  - In the framework of ENEA participation to EGEE, VO support has been added to gssklog/gssklogd and that work was presented at AFS BPW in Ann Arbor, 2006.

Grid related AFS pros/cons (1)

- The main advantage of AFS is the transparent access to applications and data, together with good security
  - Advantage for the user: no explicit data transfer required
  - Advantage for the administrator: easy deployment of applications and data
- Issues arise essentially from performance problems
  - It is well known that AFS performances over WAN are limited by rx protocol [e.g. see H.Reuter presentation at last European AFS Workshop, Graz, 2008] to ~4 MB/s for rtt=10 ms
  - If performance matters, then the user must be aware of the data localization
  - In ENEA-GRID we have tried to solve that by providing local and remote volumes to users, but of course some of the AFS elegance is lost in the process

Grid related AFS pros/cons (2)

- For users, data transfer over WAN can be several times faster using standard methods (scp) than by direct AFS trasfer
  - over LAN (CRESCO) AFS (memory cache) trasfers at 50 MB/s
- What about reliability?
  - AFS is very reliable but obviously it can fail
  - More often failures depends on the network and that induces a reliability issue for the local resources due to remote network problems
- The effect for the user sometimes can be minimized:
  - By a proper configuration of read-only copies
  - By the availability of clone dbservers and proper cell definition on the clients

Grid related AFS pros/cons (3)

- The PTS group in AFS can be exploited to provided VO (Virtual Organization) functionality for a GRID structure. In ENEA-GRID PTS groups:
  - are used define the ACL for access to software/and data (standard)
  - are connected to defined Projects/Software AFS space
  - are used to enable access to services by controlling for example the access to dedicated Web pages or portals
  - can be defined by non-admin users, by means of WARC utility, as also the user membership. In fact the WARC utility permits to delegate these operations to Project Administrators
- These features permits to organize the cooperative activity of group of users with functionalities similar to the one provided by other VO implementation

Conclusions

- AFS is one of the main component of ENEA-GRID and at the moment there are non alternatives for the features it provides
- Some of the GRID features in ENEA-GRID are based on AFS (together with some utilities): data location transparency, data security, Virtual Organization management (authorization, collaboration)
- The impact of the low WAN performances of AFS is relevant and any improvement on this issue can have important effect on the AFS usability in GRID context