Integration of heterogeneous computational resources in EGEE infrastructure: a live demo

A. Santoro, G. Bracco , S. Migliori, S. Podda, A. Quintiliani, A. Rocchi, C. Sciò *

ENEA-FIM, ENEA C.R. Frascati, 00044 Frascati (Roma) Italy, (*) Esse3Esse

Summary

The SPAGO (Shared Proxy Approach for Grid Objects) architecture enables the EGEE user to submit jobs not necessarily based on the x86 or x86_64 Linux architectures, thus allowing a wider array of scientific software to be run on the EGEE Grid and a wider segment of the research community to participate in the project. It also provides a simple way for local resource managers to join the EGEE infrastructure and the procedure shown in this demo further reduces the complexity involved in implementing the SPAGO approach.

This fact can widen significantly the penetration of gLite middle-ware outside its traditional domain of the distributed and capacity focused computation. For example the world of the High Performance Computing, which often requires dedicated system software, can find in SPAGO the easy way to join the large EGEE community. SPAGO will be used to connect ENEA CRESCO HPC system (#125 in top500/2008) to EGEE infrastructure.

The aim of this demo is to show how a computational platform not supported by gLite (such as AIX, Altix, Irix or MACOS) may still be used as a gLite Worker Node, and thus be integrated inside EGEE by employing the above mentioned SPAGO methodology.

All the machines required to support the demo (consisting of both the gLite infrastructure machines and the non-standard Worker Nodes) reside on the ENEA-GRID infrastructure. Specifically, the demo will make use of two shared filesystems (NFS and AFS), Worker Nodes belonging to five different architectures (AIX, linux, altix, cray, IRIX), and one resource manager systems (LSF), plus five Computing Elements, two gLite worker nodes (which will act as proxies) and a machine acting as BDII. All those resources are integrated into the ENEA-GRID infrastructure which offers a uniform interface to access all of them.

The case of a multiplattform user application (POLY-SPAN) which takes advantage of the infrastructure is also shown.

FOCUS OF THE DEMO

1) We show how a Worker Node whose architecture and operating system are not explicitly supported by gLite can still be integrated into EGEE. The demo summarizes the steps to integrate a generic UNIX machine into the grid and job submission will be demostrated to AIX, Altix, IRIX, Cray(Suse) and MacOSX worker nodes.

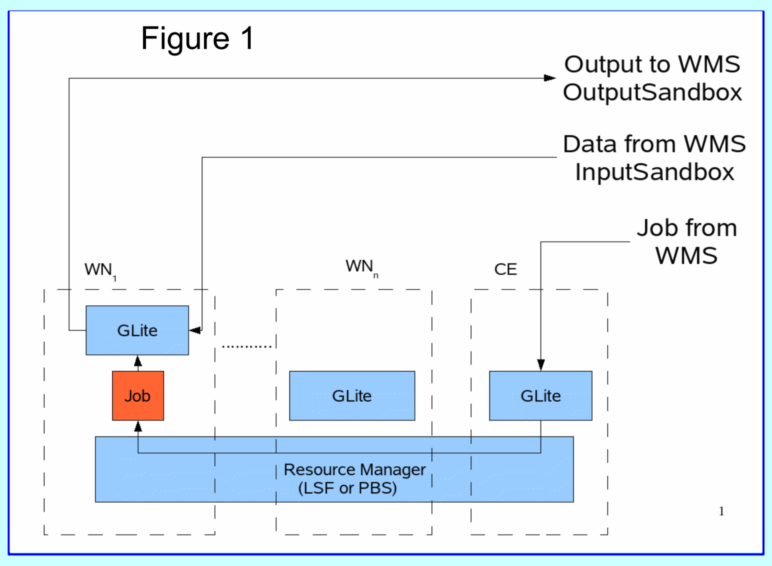
2) We show how jobs submitted by users for a specific, non standard platform are automatically redirected to the proper Worker Nodes.

3) We present an user application [POLY-SPAN] compatible with many different platforms not supported by gLite, and will show how it can run on the non-standard worker nodes presented above.

The *SPAGO* approach

The Computing Element (CE) used in a standard gLite installation and its relation with the Worker Nodes (WN) and the rest of the EGEE GRID is shown in the Figure 1.
When the Workload Management Service (WMS) sends the job to the CE, the gLite software on the CE employs the resource manager (LSF for ENEA-INFO) to schedule jobs for the various Worker Nodes.
When the job is dispatched to the proper worker node (WN_1), but before it is actually executed, the worker node employs the gLite software installed on itself to setup the job environment (it loads from the WMS storage the files needed to run, known as the InputSandbox). Analogously, after the job execution the Worker Node employs gLite software to store on the WMS storage the output of the computation (the OutputSandbox).
The problem is that this architecture is based on the assumption underlying the EGEE design that all the machines, CE and WN alike, employ the same architecture. In the current version of gLite (3.1) the software is written for intel-compatible hardware running Scientific Linux.

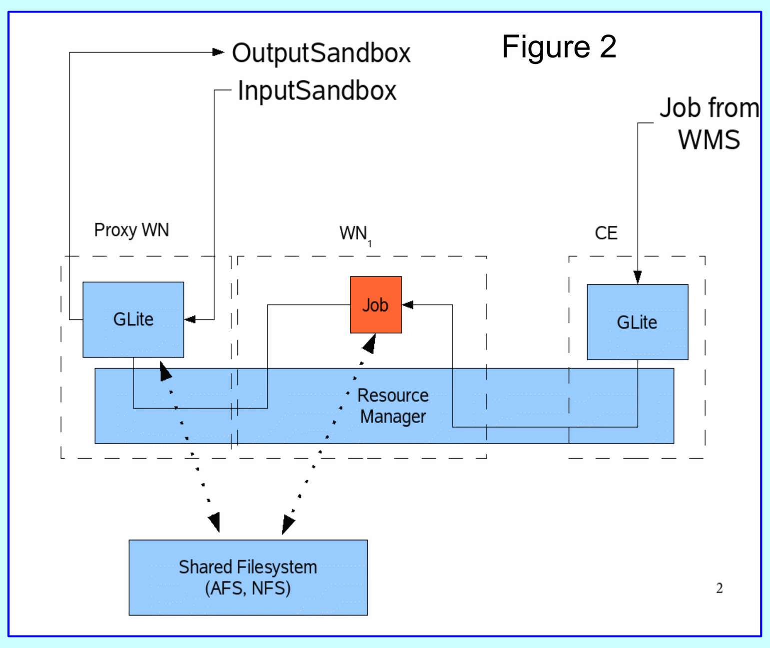
CE & WN layout for the standard site



The SPAGO approach: no middleware on WN

The basic design principle of the ENEA-INFO gateway to EGEE is outlined in Figure 2 and it exploits the possibility to use a shared file system. When the CE receives a job from the WMS, the gLite software on the CE employs the resource manager to schedule jobs for the various Worker Nodes, as in the standard gLite architecture.
However the worker node is not capable to run the gLite software that recovers the InputSandbox. To solve this problem the LSF configuration has been modified so that **any attempt to execute gLite software on a Worker Node actually executes the command on a specific machine, labeled Proxy Worker Node** which is able to run standard gLite.
By redirecting the gLite command to the Proxy WN, the command is executed, and the InputSandbox is downloaded into the working directory of the Proxy WN.
The working directory of each grid user is maintained into the shared filesystem, and is shared among all the Worker Nodes and the Proxy WN, thus downloading a file into the working directory of the Proxy WN makes it available to all the other Worker Nodes as well. Now the job on the WN_1 can run since its InputSandbox has been correctly downloaded into its working directory. When the job generates output files the OutputSandbox is sent back to the WMS storage by using the same method.
In the above architecture, the Proxy WN may become a bottleneck since its task is to perform requests coming from many Worker Nodes. In that case **a pool of Proxy WN** can be allocated to distribute the load equally among them.

CE & WN layout for SPAGO Architecture



Tested implementations

Shared Filesystems	*Resource dispatchers*	*Worker Nodes Architecture*
• NFS • AFS (requires additional modification of the CE due to authentication issues) • GPFS (in progress)	• LSF Multicluster (v 6.2 and 7.0) • Script SSH • PBS (under investigation)	• Non-standard LINUX • AIX 5.3 (in production) • IRIX64 6.5 • ALTIX 350 RH 3, 32 cpu • CRAY XD1 (Suse 9) 24 cpu • MacOSX 10.4

Modifications on CE

YAIM: config_nfs_sw_dir_server, config_nfs_sw_dir_client, config_users
Gatekeeper: lsf.pm, cleanup-grid-accounts.sh
Information system: lcg-info-dynamic-lsf

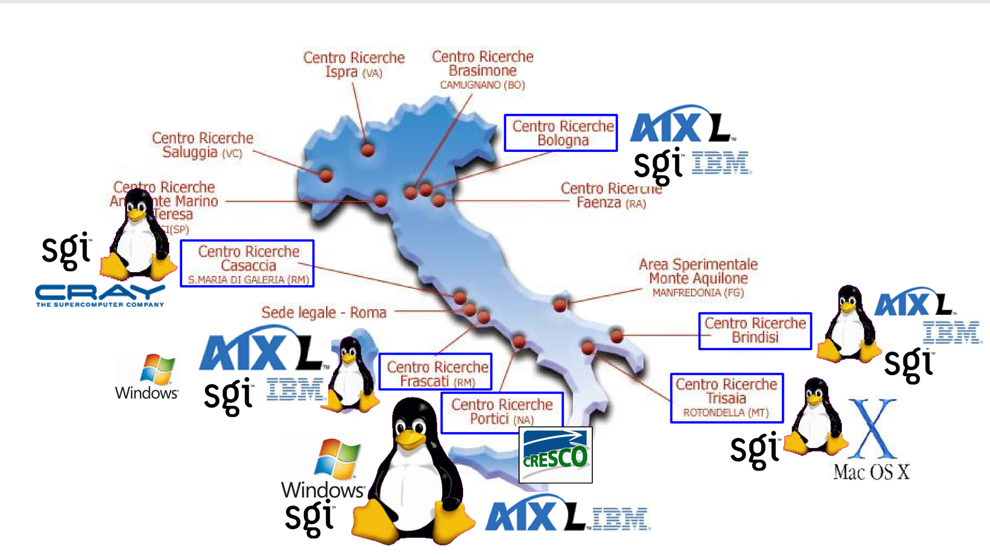
Modifications on WN

Worker nodes: the commands that shoyuld have been executed on the WN have been substituted by wrappers on the shared filesystem that invoke a remote execution on the Proxy Worker Node.

ENEA

[Italian National Agency for New Technologies, Energy and Environment]

12 Research sites and a Central Computer and Network Service (ENEA-INFO) with 6 computer centres managing multi-platform resources for serial & parallel computation and graphical post processing.



ENEA GRID architecture

GRID functionalities (unique authentication, authorization, resource access and resource discovery) are provided using "mature", multi-platform components:

Distributed File System: OpenAFS
Resource Manager: LSF Multicluster [www.platform.com]
Unified user interface: Java & Citrix Technologies

These components constitute the ENEA-GRID Middleware.

http://www.eneagrid.enea.it

OpenAFS
• user homes, software and data distribution
• integration with LSF
• user authentication/authorization, Kerberos V

ENEA GRID

ENEA-GRID computational resources:

• Hardware: ~400 hosts and ~3400 cpu : IBM SP; SGI Altix & Onyx; Linux clusters 32/ia64/x86_64; Apple cluster; Windows servers. Most relevant resources:

• CRESCO 2700 cpu mostly dual Xeon Clovertown 4 core
• IBM SP5 256 cpu; 3+ frames of IBM SP4 105 cpu

ENEA GRID mission [started 1999]:

• provide a unified user environment and an homogeneous access method for all ENEA researchers, irrespective of their location.

• optimize the utilization of the available resources

CRESCO HPC Centre

www.cresco.enea.it

CRESCO (Computational Research Center for Complex Systems) is an ENEA Project, co-funded by the Italian Ministry of University and Research (MIUR). The project is functionally built around a HPC platform and 3 scientific thematic laboratories:

• the Computing Science Laboratory, hosting activities on HW and SW design, GRID technology and HPC platform management

The HPC system consist of a ~2700 cores (x86_64) resource (~17.1 Tflops HPL Benchmark #125 top500/2008), InfiniBand connected with a 120 TB storage area (GPFS). A fraction of the resource, part of ENEA-GRID, will be made available to EGEE GRID using gLite middle-ware through the SPAGO approach.

• the Computational Systems Biology Laboratory, with activities in the Life Science domain, ranging from the "post-omic" sciences (genomics, interactomics, metabolomics) to Systems Biology;

• the Complex Networks Systems Laboratory, hosting activities on complex technological infrastructures, for the analysis of Large National Critical Infrastructures.

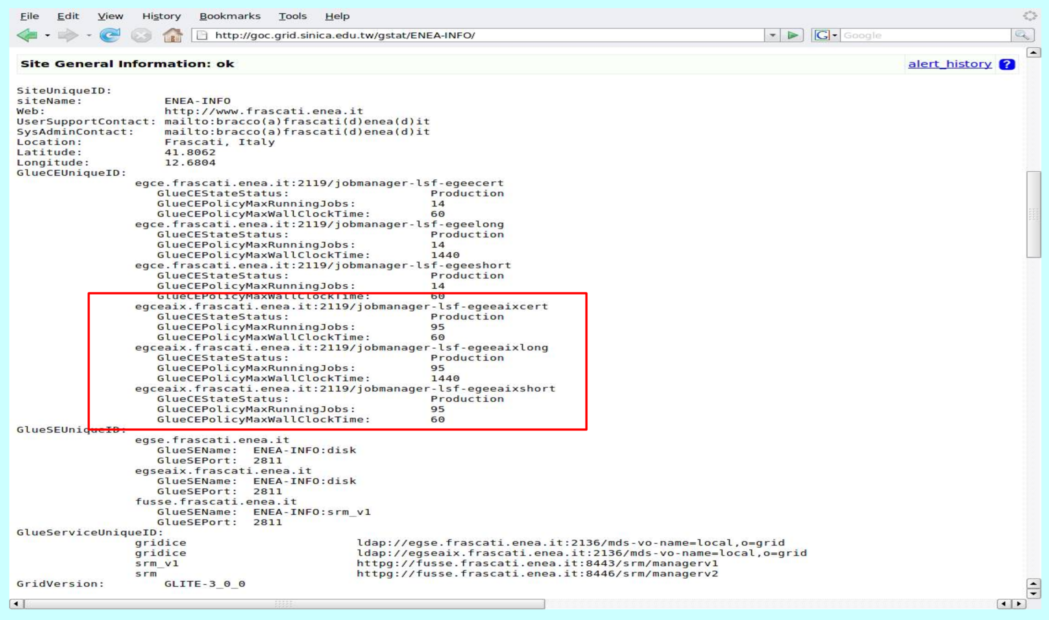
The Issues of SPAGO Approach

The gateway implementation has some limitations, due to the unavailability of the middleware on the Worker Nodes. The Worker Node API are not available and also the monitoring is partially implemented.

As a result, RGMA is not available as also the Worker Node GRIDICE components. A work around solution can be found for GRIDICE, by collecting the required information directly using a dedicated script on the information collecting machine, by means of native LSF commands.

SPAGO in EGEE Production GRID

GOC/GSTAT page with AIX WN information



The activity has been supported by the ENEA-GRID and CRESCO TEAM

P. D'Angelo, D. Giammattei, M. De Rosa, S. Pierattini, G. Furini, R. Guadagni, F. Simoni, A. Perozziello, A. De Gaetano, S. Pecoraro, A. Funel, S. Raia, G. Aprea, U. Ferrara, F. Prota, D. Novi, G. Guarnieri