GRID computing with MPI

BELNET - BEgrid

UCL - INGI – 14th December 2010

Agenda

- A word on BELNET
- GRID in Europe
- BELNET and GRID
- How it works
  - Virtual Organisations
  - Authentication
  - Searching for resources
  - Sending a job
- Message Passing Interface – MPI
  - MPI principles and API
  - GRID and MPI
- References

BELNET
The network of knowledge

BELNET

- BELNET is the Belgian National Research and Education Network (NREN)
  - Provide network connectivity to education world
    - Universities
    - Research centers
    - High schools
    - …
  - Offer services
    - « base services »
    - GRID
    - …
  - CERTs – Computer Emergency Response Team
    - BELNET CERT
    - National CERT (CERT.be)

GRID in Europe

- GRID pushed by EU commission and CERN for LHC needs
  - *205 173 cores*
  - PB of storage
  - *33* countries
  - *73*M€ in 4 years
  - *13 000* researchers
- In Belgium mainly used by IIHE and UCL for CMS (High-energy physic)
- Resources and authentication distributed worldwide

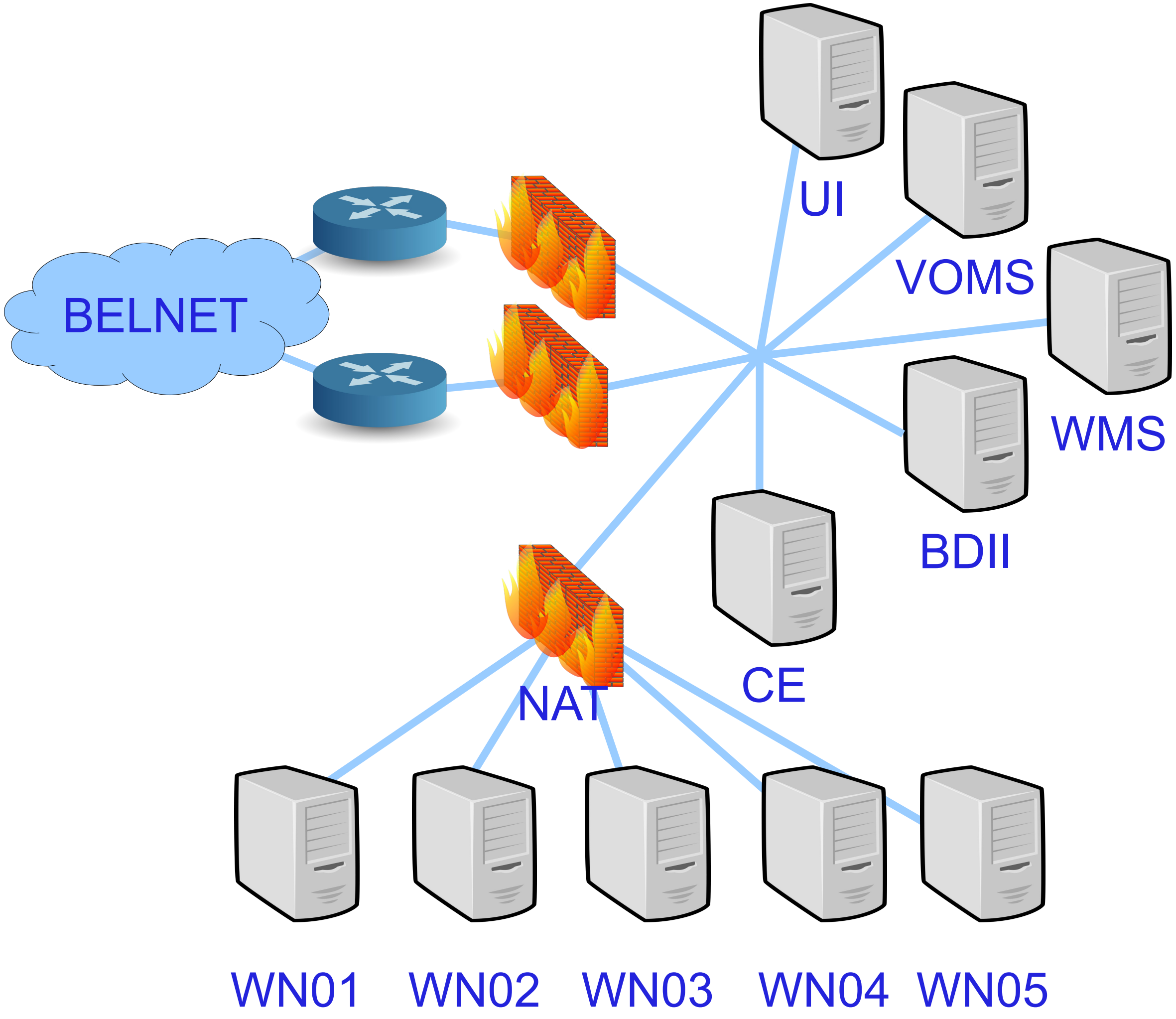
« Central » GRID services

- Belgian « virtual organisations » (*VO*) permission management (*VOMS*)
- Work Management Systems (*WMS*)
- Information Services (*BDII*)
- Monitoring
- GRID Security
  - In collaboration with CERTs
- Support and training
- Bring resources to the GRID ;-)

BEgrid

- BEgrid is a collaboration between belgian universities to participate in the GRID
- BELNET acts as coordinator
- Round *1000* cores available
  - + *2000* cores given by the Netherlands (SARA)

Connection to BELNET



.begrid.be
193.190.113.128/26

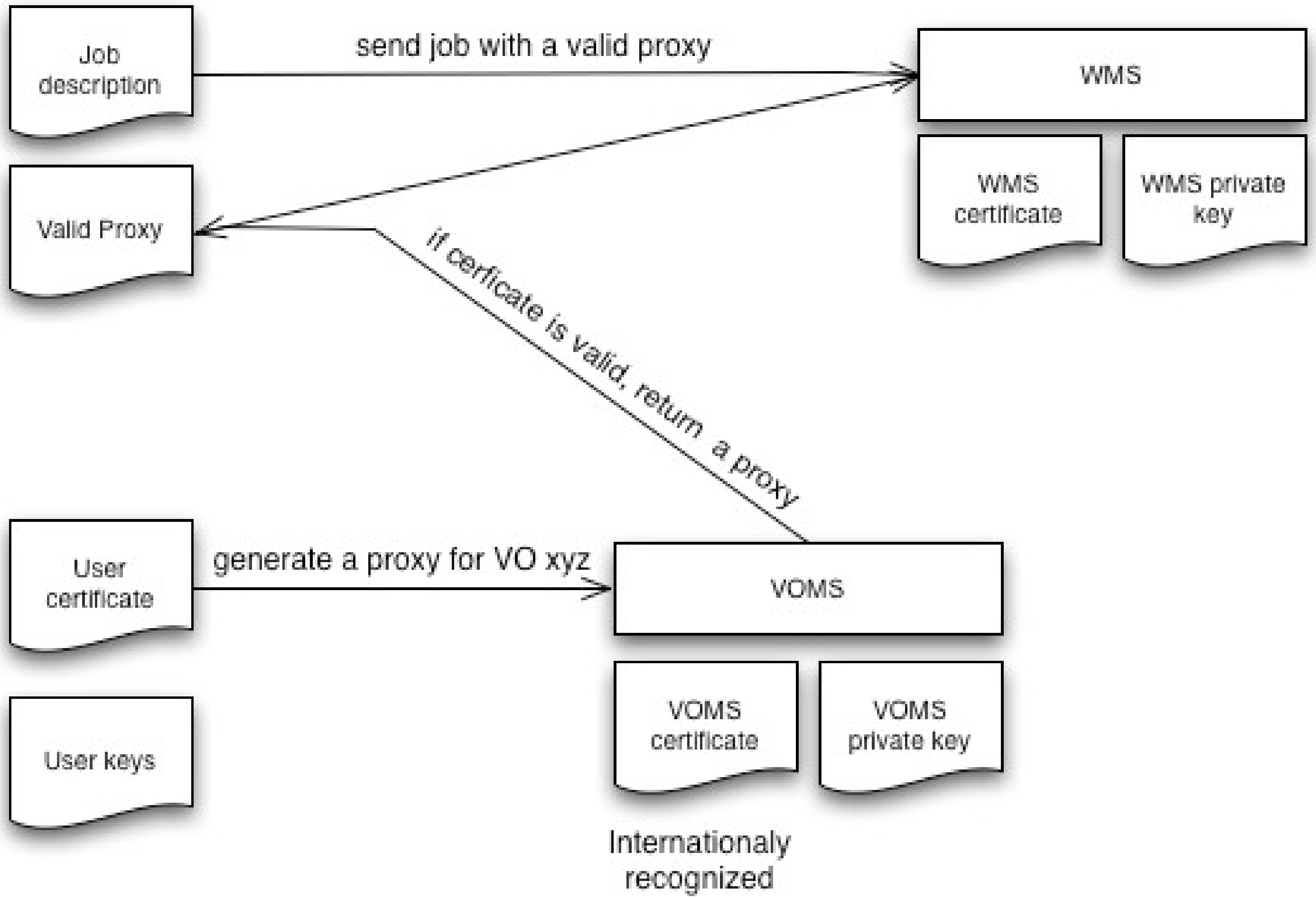
Virtual Organisations

- GRID is not dedicated to a single research field / experiment
- Principe of Virtual Organisations (VO)
- For each VO
  - Each site decide to allow to use resources
  - Priority at site level
  - Access control to softwares, datas...
- Membership to some VO is controlled on VOMS servers (Virtual Organisation Management System)
  - For BELNET: 4 local VOs
    - betest
    - beapps
    - becms and becms-t2

Security principles

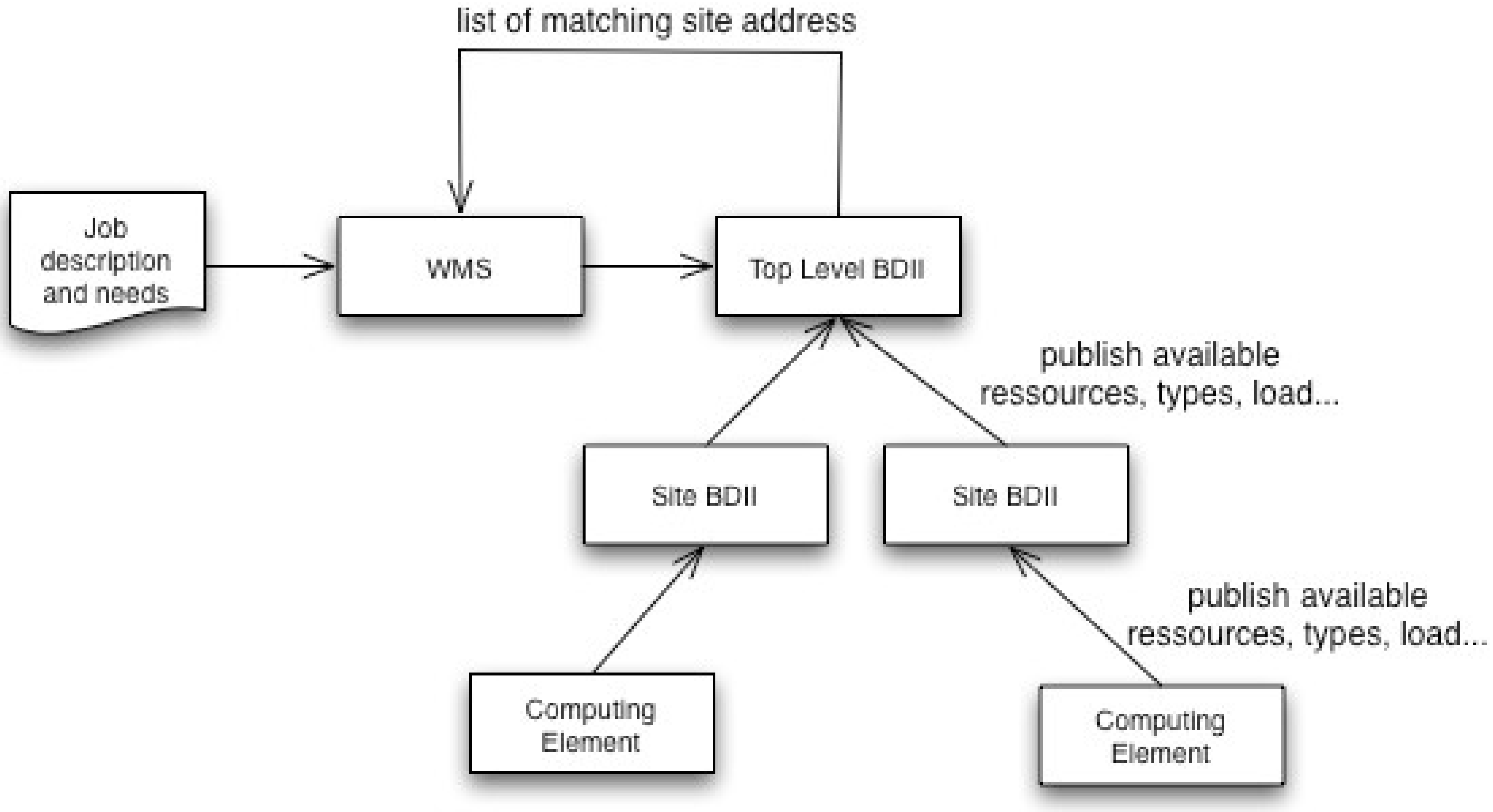
- Authentication:
  - Each server has a certificate valid for 2 years
  - Each user has a certificate valid for 1 year
  - Each certificate is signed by a recognized registration authority (RA)
    - BELNET is a recognized RA
- Authorisation:
  - Each job is send with a certificate signed by a VOMS server
    - Valid for maximum 24 hours
    - Per default, 8 hours on BELNET Virtual Organisations

Security principles (2)

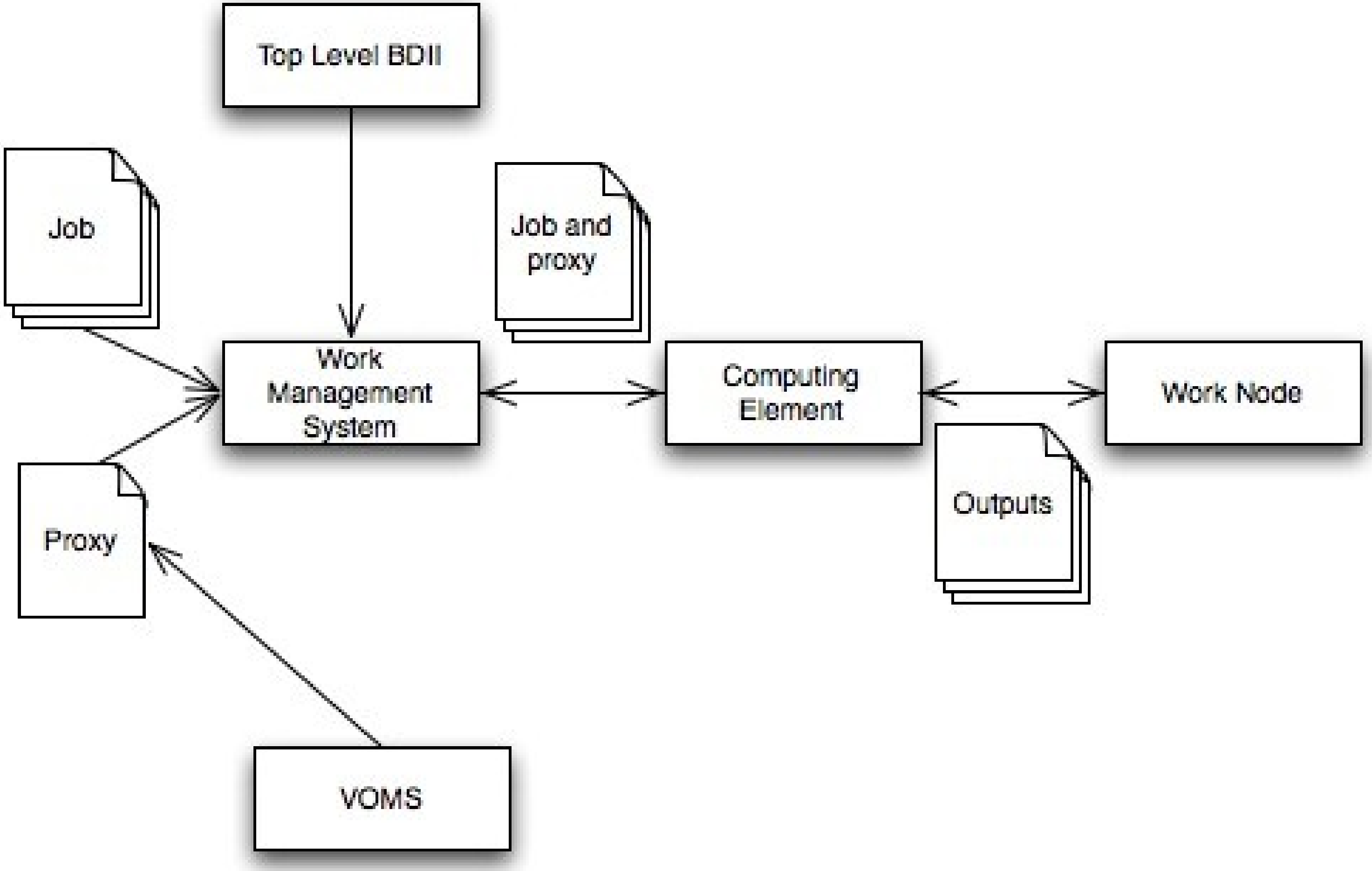


Each server is able to check identity of his pair

Searching for resources


list of matching site address
Job description and needs
WMS
Top Level BDII
publish available ressources, types, load...
Site BDII
Site BDII
publish available ressources, types, load...
Computing Element
Computing Element

Global overview



A simple job

- Job described using the Job Description Language (JDL)
- « Minimum » set of parameters:
  - Command / binary to execute
  - Input parameters
  - Output filename
  - Error filename
  - Sandbox(es)

```
Executable = "/bin/echo";
Arguments = "Hello World";
Stdoutput = "message.txt";
StdError = "error.txt";
OutputSandbox = {"message.txt","error.txt"};
```

Software stack

- Operating Systems – RedHat based
  - ScientificLinux 5.x – Computing resources
  - CentOS 5.x
- Middleware
  - gLite 3.1 and gLite 3.2 (www.glite.org)
- Central management of OS and softwares
  - Quattor (www.quattor.org)
- Regristration Authority
  - OpenTrust (www.opentrust.com)
- Firewalls
  - pfSense (www.pfsense.org)

Requesting access

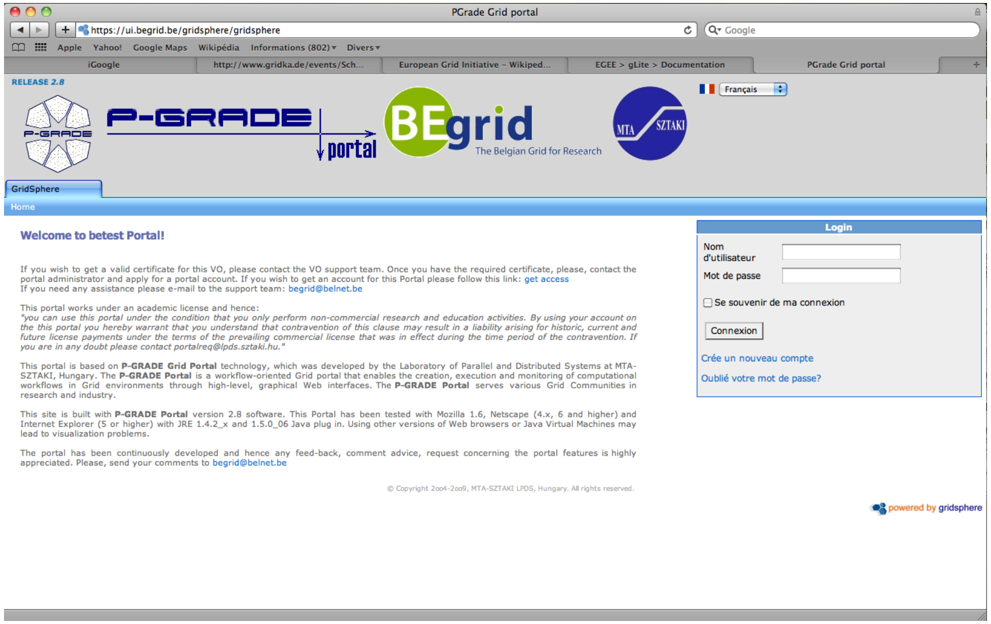
1. Request a user certificate
   - At BELNET (https://gridra.belnet.be)
   - At TERENA
   - Any recognized institution
2. Request VO membership
   - Depends on the VO
   - https://voms.begrid.be:8443/vomses/
     - For betest, beapps, becms and becms-t2
3. Request an account on a user interface
   - At BELNET, send a mail to begrid@belnet.be with your public SSH key for ui.begrid.be

« Defaults » applications

- Compilers / Interpreters
  - gcc – C, C++, Fortran...
  - Perl
  - Python
  - MPI compiler
- Computing environments
  - Octave
  - R
- VO specifics applications
- Other applications
  - Based on needs
  - Based on licenses
  - If needed, request!

Job control environnent

- Command line interface
  - Typically a SSH on a machine with tools installed
  - ui.begrid.be for instance – not a computing node!
- Web interfaces



Welcome to betest Portal!

If you wish to get a valid certificate for this VO, please contact the VO support team. Once you have the required certificate, please, contact the portal administrator and apply for a portal account. If you wish to get an account for this Portal please follow this link: get access
If you need any assistance please e-mail to the support team: begrid@belnet.be

This portal works under an academic license and hence:
"you can use this portal under the condition that you only perform non-commercial research and education activities. By using your account on the this portal you hereby warrant that you understand that contravention of this clause may result in a liability arising for historic, current and future license payments under the terms of the prevailing commercial license that was in effect during the time period of the contravention. If you are in any doubt please contact portalreq@lpds.sztaki.hu."

This portal is based on **P-GRADE Grid Portal** technology, which was developed by the Laboratory of Parallel and Distributed Systems at MTA-SZTAKI, Hungary. The **P-GRADE Portal** is a workflow-oriented Grid portal that enables the creation, execution and monitoring of computational workflows in Grid environments through high-level, graphical Web interfaces. The **P-GRADE Portal** serves various Grid Communities in research and industry.

This site is built with **P-GRADE Portal** version 2.8 software. This Portal has been tested with Mozilla 1.6, Netscape (4.x, 6 and higher) and Internet Explorer (5 or higher) with JRE 1.4.2_x and 1.5.0_06 Java plug in. Using other versions of Web browsers or Java Virtual Machines may lead to visualization problems.

The portal has been continuously developed and hence any feed-back, comment advice, request concerning the portal features is highly appreciated. Please, send your comments to begrid@belnet.be

Nom d'utilisateur
Mot de passe

□ Se souvenir de ma connexion

CLI tools – execute a job

- Create a valid proxy
  - voms-proxy-init --voms betest
- Send a job using the proxy
  - glite-wms-job-submit -a myjob.jdl
    - Output an address on the WMS
- Get job status
  - glite-wms-job-status wms_job_address
- Retrieve result
  - glite-wms-job-output -a --dir /path/you/want wms_job_address
- Get matching ressources
  - glite-wms-job-list-match -a myjob.jdl

Message Passing Interface - MPI

- Language-independent
- Message passing programming
- MPI Goals
  - High performance
  - Scalability
  - Portability
- Two major versions used
  - MPI-1.2
  - MPI-2
- Often one process per processor / core

MPI-1 and MPI-2

- MPI-1 (1994)
  - Point-to-point communication
  - Collective operations
  - Process groups and topologies
  - Communication contexts
  - Datatype management
- MPI-2 (1997)
  - Dynamic process management
  - File I/O
  - One sided communications
  - Extension of collective operations

MPI concepts: Communicator

- Connect groups of processes in the MPI session
- In each session
  - Each process receive an independent ID
  - Communicator create a topology
- Communication is
  - Single group intra-communication
  - Bilateral inter-communication
- Could be partitioned

MPI concepts: Point-to-point

- MPI_SEND
  - Allows one process to send a message to another process
- Blocking or not blocking communication
- « Ready-send »
  - A send could only be made when the matching receive request has been done

MPI concepts: Collectives

- Communication among all processes in a process group
- MPI_Bcast
  - One process send a message to all other processes in the process group
- MPI_Reduce
  - Take data from all processes
  - Perform some operation
  - Store result
- MPI_Alltoall
  - *n* items exchanged
  - *n*th node receive the *n*th item from each node

Inspired from Wikipedia MPI page (http://www.wikipedia.org)

MPI concepts: Datatypes

- Pre-defined MPI types for standard types
  - MPI_INT for int
  - MPI_CHAR for char
  - MPI_DOUBLE for double
  - ...

MPI concepts: Datatypes

- Pre-defined MPI types for standard types
  - MPI_INT for int
  - MPI_CHAR for char
  - MPI_DOUBLE for double
  - ...

GRID and MPI

- To simplify « GRIDification » of MPI Jobs
  - MPI-Start
    - Portable
    - Permit to enable debug
    - Interface to run MPI job
      - MPI command invisible to user
    - Allow to run MPI job without change to GRID middleware
  - Wrapper
    - Set environment for MPI-Start
    - Call MPI-Start
    - Submitted with the job
  - Hooks
    - Handle compilation
    - Submitted with the job

Start wrapper

```
#!/bin/bash

MY_EXECUTABLE=`pwd`/$1
MPI_FLAVOR=$2

MPI_FLAVOR_LOWER=`echo $MPI_FLAVOR | tr '[:upper:]' '[:lower:]'`

eval MPI_PATH=`printenv MPI_${MPI_FLAVOR}_PATH`

eval I2G_${MPI_FLAVOR}_PREFIX=$MPI_PATH
export I2G_${MPI_FLAVOR}_PREFIX

touch $MY_EXECUTABLE

export I2G_MPI_APPLICATION=$MY_EXECUTABLE
export I2G_MPI_APPLICATION_ARGS=
export I2G_MPI_TYPE=$MPI_FLAVOR_LOWER
export I2G_MPI_PRE_RUN_HOOK=mpi-hooks.sh
export I2G_MPI_POST_RUN_HOOK=mpi-hooks.sh

export I2G_MPI_START_VERBOSE=1
#export I2G_MPI_START_DEBUG=1

$I2G_MPI_START
```

Start wrapper

```
export I2G_MPI_APPLICATION
```

- Executable

```
export I2G_MPI_APPLICATION_ARGS
```

- Parameters to give to the executable

```
export I2G_MPI_TYPE
```

- MPI implementation (*OpenMPI* for BEgrid)

```
export I2G_MPI_PRE_RUN_HOOK
export I2G_MPI_POST_RUN_HOOK
```

- Path to hooks

```
export I2G_MPI_START_VERBOSE=1
export I2G_MPI_START_DEBUG=1
```

- Enable verbose / debug modes

Hooks

```
#!/bin/sh

pre_run_hook () {

  echo "Compiling ${I2G_MPI_APPLICATION}"
  cmd="mpicc ${MPI_MPICC_OPTS} -o ${I2G_MPI_APPLICATION} ${I2G_MPI_APPLICATION}.c"
  echo $cmd
  $cmd
  if [ ! $? -eq 0 ]; then
    echo "Error compiling program.  Exiting..."
    exit 1
  Fi

  echo "Successfully compiled ${I2G_MPI_APPLICATION}"
  return 0
}

post_run_hook () {
  echo "Executing post hook."
  echo "Finished the post hook."

  return 0
}
```

C job – Hello World

```c
#include "mpi.h"
#include <stdio.h>
int main(int argc, char *argv[]) {

  int numprocs;  /* Number of processors */
  int procnum;   /* Processor number */

  /* Initialize MPI */
  MPI_Init(&argc, &argv);

  /* Find this processor number */
  MPI_Comm_rank(MPI_COMM_WORLD, &procnum);

  /* Find the number of processors */
  MPI_Comm_size(MPI_COMM_WORLD, &numprocs);
  printf ("Hello world! from processor %d out of %d\n", procnum, numprocs);

  /* Shut down MPI */
  MPI_Finalize();
  return 0;
}
```

Job description

```
JobType = "MPICH";
NodeNumber = 16;
Executable = "mpi-start-wrapper.sh";
Arguments = "mpi-test OPENMPI";
StdOutput = "mpi-test.out";
StdError = "mpi-test.err";
InputSandbox = {"mpi-start-wrapper.sh","mpi-hooks.sh","mpi-test.c"};
OutputSandbox = {"mpi-test.err","mpi-test.out"};
Requirements =
  Member("MPI-START", other.GlueHostApplicationSoftwareRunTimeEnvironment)
  && Member("OPENMPI", other.GlueHostApplicationSoftwareRunTimeEnvironment)
  ;
```

- **Ask for a site with MPI-START and OPENMPI**
- **Job Type set to MPI**
- **16 cores claimed**

Ressources

- BELNET trainings
- Hands-on
- BEgrid website
  - http://www.begrid.be
- BEgrid WIKI
  - http://quattorrepository.begrid.be
  - Only available from R&E institutions
- GridCafé
  - http://www.gridcafe.org

Ressources - MPI

- The Message Passing Interface (MPI) Standard
  - http://www.mcs.anl.gov/research/projects/mpi/
- MPI Documents
  - http://www.mpi-forum.org/docs/docs.html
- Open MPI documentation
  - http://www.open-mpi.org/doc/
- EGEE and EGI projects documentation
  - http://www.eu-egee.org/fileadmin/documents/UseCases/MPIJobs.html
  - https://quattorrepository.begrid.be/trac/centralised-begrid-v6/wiki/MPI_on_the_grid
- Wikipedia

Questions and answers

Thanks!

?

Feel free to contact us at:
begrid@belnet.be
http://www.begrid.be

Now... do it yourself :-D

- Compute the value of π using MPI

$$\pi \approx \int_{-1/2}^{1/2} \frac{4}{1+x^2}\,dx$$

Hint:

- Use MPI_Broadcast and MPI_Reduce calls
- Compute by summing rectangles
- Help: http://www.open-mpi.org/doc/v1.5/