1.MPI SURVEY REPORT

The questionnaire is devoted to identify the current status of the computational resources made available by those sites that are part of the Italian end EGI infrastructure, focusing the attention on the resources devoted for parallel calculations and, where not available, on the intention to enable such resources for parallel calculations.

A number of 29 questionnaires have been collected on a total of 57 production sites in the IT domain. The questionnaire is representative of the IT Site domain and during the data analysis procedure the 29 collected questionnaires are the sample of the current study.

All the sites involved are in PRODUCTION state. The resource-related statistics are distributed as follows:

TOTAL IT PRODUCTION SITES:	57
TOTAL SAMPLE:	29

O.S.	%
SL4.5	4
SL4.7	11
SL5.5	81
CENTOS 5	4

Middleware	%
gLite 3.1	27
gLite 3.2	69
EMI	4

Node	%
LGC-CE	37
CREAM	63

2. MPI SUPPORT

The IGI Sites declaring to support MPI are 16 (55% of the sample) with a total of 3511 cores. The 87% of these resources are shared among different projects. The direct consequence of that is a reduction of the availability of MPI resources in the domain.

An amount of 19% of the Sites supporting MPI declare to hide the correct MPI TAGs due to the following reasons:
- Configuration problems; Nagios failures
- Nagios failures due to multiple supporting libraries (supporting one library the problem does not appear)

The 46% of the IGI Sites declaring to not support MPI declare as well that they are planning to support MPI in the next future. Those Sites who are not planning to support MPI motivated the decision with the following sentence:

- Inadequate hardware site resources

The MPI-related statistics are distributed as follows:

Supported Library	Distribution*(%)
MPICH	20
MPICH2	7
MPICH1-2	60
MVAPICH1-2	33
OPENMPI	47
* Sites may support more than one MPI Library	

3. DOCUMENTATION

The sample has been queried about the official documentation[1] already available for Admins and Users involving the following procedures description:
- MPI-installation
- MPI-configuration
- MPI-getting started

Analysing the information collected from the sample, the Admins seem to use the available documentation but the absence of a "Very Good" or "Excellent" rate indicates that it is unsatisfactory for people that already had experience with MPI in the three aspect described above. A deep restyling is needed and appreciate specially after the introduction of the new MPI attributes. The statistics are distributed as follows:

MPI-Installation	Distribution (%)
Poor	10
Fair	24
Good	45
Very Good	--
Excellent	--
No Experience Yet	3
Not Answered	18

MPI-Configuration	Distribution (%)
Poor	7
Fair	27
Good	45
Very Good	--
Excellent	--
No Experience Yet	3
Not Answered	18

MPI-getting started	Distribution (%)
Poor	--
Fair	32
Good	45
Very Good	--
Excellent	--
No Experience Yet	3
Not Answered	20

[1] http://www.eu-emi.eu/products/-/asset_publisher/z2MT/content/glite-mpi

4. CUSTOMER CARE

The sample has been queried for their customer experience (and the service received) described as follow:
- MPI support as Admin
- MPI support as a User

A value between 60% and 70% of the Admins seem to have no relation with the MPI support in EGI. On the other hand, who of them had contact with it seem to appreciate the effort spent by the people involved in MPI (at the moment we cannot define a proper MPI support team) but also in this case, the absence of a "Very Good" or "Excellent" rate indicate that this effort is unsatisfactory and greater attention to the customer is needed. The statistics are distributed as follow:

MPI support as Admin	Distribution (%)
Poor	--
Fair	14
Good	24
Very Good	--
Excellent	--
No Experience Yet	34
Not Answered	28

MPI support as a User	Distribution (%)
Poor	--
Fair	7
Good	17
Very Good	--
Excellent	--
No Experience Yet	38
Not Answered	38

5. CONCLUSIONS

The aim of the survey is to report to the actual status of MPI in the IT domain of those sites that support EGI. The survey pointed out the need to solve some open problems here resumed:
- Restyling of the available documentation;
- Resolution of the most common configuration problems and probe failures
- Revision of the actual Nagios probes
- Implementation of new probes able to identify
    - MPI resources effectively available in a site [2]
    - Max allowed MPI resources for a single calculation [3]
- Work out of the MPI related "known issues" in order to distribute a stable version of the package(s) [4,5,6]
- Compilers: some applications have compilation problems with the standard compilers in SL5.X (gcc 4.1)
- Impossibility to handle more instances of the same MPI library

[2] It is possible populating the SUB-CLUSTER TAGs in the site-info.def (GLUE1.3). The WMS is not able to distinguish between different sub-clusters of the same cluster (by IGI_operations_team).

[3] It is possible by setting the batch property concerning the VO group specifications. The GLUE1.3 has not such settings implemented, moreover Site Admins probably do not want to publish their internal policies (by IGI_operations_team).

[4] VT MPI within EGI- Task 5: https://wiki.egi.eu/wiki/VT_MPI_within_EGI

[5] Workaround by UNINA-SCOPE (IT) 14/03/2011
Link to the italian version of the document: http://trac.scope.unina.it/attachment/wiki/TR/SIS_Baco_MAUI.pdf
· Get the RPM source (es.): maui-3.2.6p21-snap.1234905291.5.el5.src.rpm
· Install it with the command: #rpm -ivh maui-3.2.6p21-snap.1234905291.5.el5.src.rpm
· cd rpmbuild/SPECS
· run the command: #rpmbuild -bs maui-xxxx.spec
· Open the file: /usr/lib/rpm/redhat/macros and modify the line 182 as follow:
o %__global_cflags -O2 -g -pipe -Wall -Wp,-D_FORTIFY_SOURCE=2 -fexceptions -fstack-protector --param=ssp-buffer-size=4
with
o %__global_cflags -O2 -g -pipe -Wall -Wp,-D_FORTIFY_SOURCE=0 -fexceptions -fstack-protector --param=ssp-buffer-size=4
· run the command: #rpmbuild --rebuild rpmbuild/SRPM/maui-3.2.6p21-snap.1234905291.5.el5.src.rpm

[6] Software Release announcement by IGI (17/02/2012)
There are available new services in IGI Release, based on middleware services developed by EMI & IGI: * MPI v. 1.0.0 - providing updated MAUI, v. 3.3-4, solving issues like: https://ggus.eu/ws/ticket_info.php?ticket=57828 https://ggus.eu/ws/ticket_info.php?ticket=67870

APPENDIX 1 (Test MPI attributes)

As from the last MPI-VT meeting related to new Nagios probes (17/02/2012 @ Universe), NGI_IT has tested via both the WMS and direct submission to CEs the following MPI attributes:
- SMPGranularity
- HostNumber
- WholeNodes

The attributes have been used in the JDL specifying the number of CPU used in the calculation. In all the job submissions the following attribute's values have been used:

```
CPUNumber = 8;
SMPGranularity = 4;
HostNumber = 2;
WholeNodes = true;
```

The above configurations require a CE with 8 CPU and two nodes with SMP>=4 as from the MPI user guides.[7,8]

[7] https://wiki.infn.it/strutture/pi/datacenter/cluster_gruppo_iv/csn4cluster/job_paralleli;
[8] http://www.eu-emi.eu/products/-/asset_publisher/z2MT/content/glite-mpi

A1.1 WMS LIST-MATCH

The tests have been performed using the wms-multi @cnaf.infn.it with the gLite3.2 middleware installed.
The glite-wms-job-list-match command has been used to monitor the available resources
No errors have been registered using the specified attributes in the JDL

A1.2 WMS Submission

In all the job submissions the following configuration has been used in the JDL:

```
CPUNumber = 8;
SMPGranularity = 4;
HostNumber = 2;
WholeNodes = true;
```

In all the tests CREM CEs have been used. The job performs the parallel execution of an HelloWorld file that write the name of the Host used for each process.
From the results it is clear that a minimum of
- Nodes with SMP>=4
- Hosts>=2

are guaranteed from the batch system of the CE but this does not means that the processes will be equally distributed among the Hosts.

Examples

m3pec.u-bordeaux1.fr

```
=[START]=================================================================
[0] Machine Name = wn10.m3pec.u-bordeaux1.fr
[1] Machine Name = wn10.m3pec.u-bordeaux1.fr
[2] Machine Name = wn13.m3pec.u-bordeaux1.fr
[3] Machine Name = wn13.m3pec.u-bordeaux1.fr
[4] Machine Name = wn13.m3pec.u-bordeaux1.fr
[5] Machine Name = wn13.m3pec.u-bordeaux1.fr
```

```
[6] Machine Name = wn25.m3pec.u-bordeaux1.fr
[7] Machine Name = wn25.m3pec.u-bordeaux1.fr
=[FINISHED]==========================================================
```

scope.unina.it

```
=[START]=============================================================
[0] Machine Name = wn181.scope.unina.it
[1] Machine Name = wn181.scope.unina.it
[2] Machine Name = wn181.scope.unina.it
[3] Machine Name = wn181.scope.unina.it
[4] Machine Name = wn181.scope.unina.it
[5] Machine Name = wn181.scope.unina.it
[6] Machine Name = wn179.scope.unina.it
[7] Machine Name = wn179.scope.unina.it
=[FINISHED]==========================================================
```

A1.3 DIRECT CREAM Submission

In all the job submissions the following configuration has been used in the JDL:

```
CPUNumber = 8;
SMPGranularity = 4;
HostNumber = 2;
WholeNodes = true;
```

In all the tests CREM CEs have been used. The job performs the parallel execution of an HelloWorld file that write the name of the Host used for each process. No submission errors have been registered.
The tests bring to the same results of A1.2.
From the results it is clear that a minimum of
- Nodes with SMP>=4
- Hosts>=2

are guaranteed from the batch system of the CE but this <u>does not</u> means that the processes will be equally distributed among the Hosts.

Examples

m3pec.u-bordeaux1.fr

```
=[START]=============================================================
[0] Machine Name = wn13.m3pec.u-bordeaux1.fr
[1] Machine Name = wn13.m3pec.u-bordeaux1.fr
[2] Machine Name = wn13.m3pec.u-bordeaux1.fr
[3] Machine Name = wn13.m3pec.u-bordeaux1.fr
[4] Machine Name = wn25.m3pec.u-bordeaux1.fr
[5] Machine Name = wn25.m3pec.u-bordeaux1.fr
[6] Machine Name = wn25.m3pec.u-bordeaux1.fr
[7] Machine Name = wn25.m3pec.u-bordeaux1.fr
=[FINISHED]===========================================================
```

scope.unina.it

```
=[START]=============================================================
[0] Machine Name = wn166.scope.unina.it
[1] Machine Name = wn166.scope.unina.it
[2] Machine Name = wn166.scope.unina.it
[3] Machine Name = wn166.scope.unina.it
[4] Machine Name = wn166.scope.unina.it
```

```
[5] Machine Name = wn166.scope.unina.it
[6] Machine Name = wn166.scope.unina.it
[7] Machine Name = wn157.scope.unina.it
=[FINISHED]==========================================================
```