EGEE-III MPI WG activities

Roberto Alfieri - INFN Parma

CCR Workshop 2010 **Napoli, January 25-26-27 2010**

Information Society and Media

MPI-WG activities

WG Start-up: 02/2009. Mandate:

- complete the work of the previous WG (07-08, http://www.grid.ie/mpi/wiki/):

recommend (rather then to solve) a method for deploying MPI support that will work for both users and site administrators .

- investigate how to get the allocated cores to be all on the same physical machine, or packed into as few physical machines as possible

Pre-final version of the MPIWG-recommendation-1.doc (24/12/09) includes:

- Survey for users and system administrator
- Recommendations, including multiple cores allocation

MPI-WG results will be presented at the EGEE User Forum in April 2010.

Since 09/2009 (EGEE'09 Conference) is active the MPI-Task Force with an operational mandate.

MPI-WG Survey per utenti

Questionario realizzato e diffuso dal MPI-WG di Egee-III nel maggio 2009.
Abbiamo chiesto agli utenti di inviare i dati anche al Cnaf per poter analizzare i dati nazionali.

Dati rilevanti:

Tutti interessati al calcolo parallelo, ma pochi usano MPI/Grid

Ambiti scientifici: fisica 35%, chimica 23%, astronomia 18%, ..

Strumenti software: mpi-2 54%, openMP 20% , mpi-1 13% ..

- Interesse in architetture multi-core

Risorse utilizzate per HPC: locale 60% , altro (grid, provider comm.) 40%

- Difficile trovare documentazione per l'uso di MPI in grid (100%)

dati presentati a Palau – Maggio 2009

MPI-WG Survey per amministratori

Questionario realizzato e diffuso dal MPI-WG di Egee-III nel Maggio 2009

Abbiamo chiesto agli utenti di inviare i dati anche al Cnaf per poter analizzare i dati nazionali.

Dati rilevanti:

I siti **non hanno infrastrutture di rete** adeguate: Ethernet 70%, Infiniband 23%

L'implementazione MPI maggiormente supportata è MPICH e ciò è in contrasto con le esigenze degli utenti che richiedono un maggiore supporto a MPI-2

Il processo d'installazione e configurazione risulta abbastanza semplice ma sarebbe necessaria una documentazione più accurata

Scarso utilizzo di Job MPI rispetto al numero complessivo: < 10%

dati presentati a Palau – Maggio 2009

Using multiple cores within a node

A new attribute in the JDL, SMPGranularity, should be introduced in order to allow users to specify how the cores can be distributed for the allocation. Example:

SMPGranularity = 8;

(Match-making: other.GlueHostArchitectureSMPSize >= SMPGranularity;)

For multi-threaded applications the users would have the ability to reserve whole nodes with a boolean attribute:

WholeNodes = True;

The introduction of these new attributes would affect some gLite Middleware components (WMS and Computing Element).

Recommendations: JobType

Users require multiple MPI flavours (MPICH-2, openMPI, ..), hence JobType =MPICH should be deprecated.

Use “Normal” tag for MPI jobs and specify the CPUnumber. (feb 2009)

The site can state the support for parallel jobs by publishing the keyword “Parallel” in the GlueHostApplicationSoftwareRunTimeEnvironment

Example:

```
Type = "Job" ;
JobType = "Normal" ;
CpuNumber = 8 ;
MPIGranularity = 4 ;
```

Recommendations: MPI-start

Mpi-start (developed by HLRS) has been recommended by the previous WG as a way to start MPI jobs.

Mpi-start is included in gLite since version 3.1 (feb 2008).

The main advantage of MPI-Start is the possibility to detect and use site-specific configuration features, like:

- **batch scheduler** (SGE, PBS, LSF)
- **MPI implementations** (openMPI, MPICH, MPICH2, PACX-MPI, LAM)
- **File distribution** (if the home isn't shared)
- **Workflow control** (user's pre/post execution scripts)

Recommendations: Information system

MPI-related informations have to be published in the Information System and exported on the WNs system environment (needed by MPI-start).

MPI_<flavour>_VERSION=”x.y.z”

MPI_<flavour> (for those who don't care about the version)

MPI_<flavour>_PATH=”/opt/mpi/<flavour>_<version> (env var. for MPI-start)

MPI_SHARED_HOME=”yes”

MPI_SSH_HOST_BASED_AUTH="yes"

May be other info such as:

MPI-<network-type>

Recommendations: MPI Packages

Distribution of:

MPI-start and pre-compiled packages of MPICH-2 and OpenMPI, possibly with support for Torque.

All packages installed in /opt/mpi/[flavour]-[version]

Source RPMS, including instructions on how to modify the configuration (different compilers, special network interconnections, different batch systems, etc)

Recommendations: SAM TESTS

Check the following items:

- MPI flavour and version published and the corresponding environment variables and verify the correct location.
- the home is shared and verify it.

Actual MPI SAM tests:

- Searching for published MPI tags: MPI-start, MPICH, MPICH2, OPENMPI
- Verify ENV variables
- submit “pingtest” with 2 CPUs using the supported flavours

Recommendations: Home File System and Job startup

Shared file system between WNs is recommended

- avoid files distribution
- ease output transferring

File distribution is supported by MPI-start but it is not deeply tested.

Remote Start-up of MPI job can be achieved via password-less SSH

Recommendations: Cpu Time Limits and Accounting of parallel jobs

Most of the sites have a WC time limit similar to CPU time limit.

lcg-info --list-ce --attrs MaxWCTime,MaxCPUTime -vo theophys

Typical for sequential jobs. Parallel jobs rapidly face the CPU time limit.

Recommended solution:

Do not set the CPU-time limit or set the CPU time limit very high
(in case of job with requirements on CPU time limit)

An **MPI accounting** portal was developed at Cesga for the in.eu.grid project to show **number of CPUs used by a job**, **MPI efficiency,** etc.

This portal could be useful for a future EGEE accounting portal.

Recommendations: MPI SUPPORT

There is an urgent need to establish structural support for MPI.

This task will be covered by a “Task Force” composed by
Isabel Campos, John Walsh and few representatives of the WG.

questions?