Towards a unified framework for Grid parallel applications

Enol Fernández (CSIC)

(gLite) MPI PT

DoW Objectives

EMI will consolidate the existing efforts targeting the support of parallel jobs in 3 steps. The highest priority is given to providing a common interface to multi-core jobs on all resources; when this is achieved multi-node execution on interconnected clusters will be addressed; and finally special scenarios like advanced topologies, FPGAs, GPGPUs, and massively multi-node jobs will be investigated for use on high-end resource types.

JRA1.1.2 Tasks

Subtask	Name	Owner	Due
A11.1	define a proposal for a parallel execution framework within EMI	MPI TF	M18
A12.1	implementation of the proposal for a parallel execution framework within EMI	MPI TF	M32
A13.1	enable capabilities to support multi-core, multi-node execution in ARC	Arc CE	M36
A13.2	enable capabilities to support multi-core, multi-node execution in gLite	gLite JM	M36
A13.3	enable capabilities to support multi-core, multi-node execution in UNICORE	UNICORE *	M36
A13.4	enable capabilities to support cross-middleware multi-core, multi-node execution	MPI TF	M36

Common Execution Framework

- EMI-ES will provide a common interface for *submitting* the jobs
- MPI-Start can provide a common interface for *executing* the jobs
  - It does not impose any changes in the CEs
  - It is already ready for integration (see presentation in Vilnius)
  - Has a established user base (compchem, theophys, biomed) and EGI support
  - Extensible and open to new parallel applications (Open MP, gromacs)

MPI-Start

- MPI-Start was initially developed in int.eu.grid for working within a gLite environment
- But
  - It is independent of the middleware
  - It just interacts with the batch system and the MPI implementation
- Objectives:
  - Provide an easy way for users to run their parallel (MPI) jobs in heterogeneous environments
  - Provide an easy way for site admins to support the execution of parallel (MPI) jobs.

gLite MPI

- The gLite MPI is a meta-package with
  - MPI-Start
  - yaim plugins for helping admin to configure
- In the CE
  - Configures the Information system to publish the supported MPI implementations
  - Configures the torque submit filter to allow submission of parallel applications
- In the WN:
  - Configures the environment for MPI-Start usage

ARC Parallel Job Revisited

```
&(jobName="mpi-start")
(count="16")
(runtimeenvironment="ENV/MPI-START")
(executable="/usr/bin/mpi-start")
(arguments="-t openmpi hello-ompi.exe")
(inputfiles=("hello-ompi.exe"))
(stdout="std.out")
(stderr="std.err")
(gmlog="gmlog")
(wallTime="10 minutes")
(memory="1024")
```

gLite Parallel Job Revisited

```
JobType          = "Normal";
CpuNumber        = 16;
Executable       = "/usr/bin/mpi-start";
Arguments        = "-t openmpi hello-ompi.exe";
InputSandbox     = {"hello-ompi.exe"}
StdOutput        = "std.out";
StdError         = "std.err";
OutputSandbox    = {"std.out","std.err"};
Requirements     =
    Member("MPI-START",
           other.GlueHostApplicationSoftwareRunTimeEnvironment)
 && Member("OPENMPI",
           other.GlueHostApplicationSoftwareRunTimeEnvironment);
```

UNICORE Parallel Job Revisited

```
{
  Executable: "./hello-ompi.exe",
  Imports: [
    {From: "/myfiles/hello.mpi", To: "hello-ompi.exe" },
  ],
  Resources:{ CPUs: 16, },
  Execution environment: {
    Name: mpi-start,
    Arguments: { mpi-type: openmpi,  },
  },
}
```

~~MPI~~ Parallel Jobs Task Force

- Same “experience” for all middleware
  - EMI-ES + mpi-start
- Go beyond MPI
  - multi-node execution on interconnected clusters (some experience with PACX-MPI)
  - GPUs, FPGAs
- We need ARC, gLite, UNICORE teams involved!

Status and progress of (gLite) MPI-* for Y2

Enol Fernández (CSIC)

gLite MPI PT

gLite-MPI in EMI-1

- gLite MPI successfully released as part of EMI-1, including:
  - Yaim plugins x.x.x
  - MPI-Start 1.0.4
- MPI-Start 1.0.4 maintains backward compatibility and:
  - Build according to EMI (Fedora) guidelines
  - Bug fixes
  - Support for new schedulers and new execution environments

MPI-Start in EMI-1

- MPI-Start 1.0.4 released as part of EMI-1
- Includes:
  - Scheduler plugins for (Sun/Oracle/Univa) GE, PBS/Torque, LSF, Slurm & Condor
  - Execution plugins for Open MPI, MPICH, MPICH2, LAM, PACX-MPI
  - Hooks for OpenMP, Marmot, MPI Trace
- Use of command line parameters (instead of environment variables)
- Updated documentation
- Linux FHS compliant

Towards EMI-2

- Yaim plugin review:
  - Interaction with glite-CLUSTER
- MPI-Start evolution:
  - More options for processes placement:
    - Per core, per socket, per node
  - MPI Processor and memory affinity
  - Better OpenMP support
  - Support for non Linux OS (MacOS, BSD)

Thank you

EMI is partially funded by the European Commission under Grant Agreement INFSO-RI-261611