Status of MPI support on the EGEE Grid

Stephen Childs
Trinity College Dublin

www.eu-egee.org

Overview

- Current issues and motivation
- New model for site config and job submission
- Testing MPI
- Conclusion and open issues

Issues

1. **Support in WMS and API is not flexible enough**
   - Incompatibilities with standard site configurations
   - Different model from “normal” MPI submission

2. **Lack of support for configuring sites for MPI**
   - Relatively few sites support it
   - Not sufficient information for discovery
   - HEP sites without MPI background can’t invest time

- **No huge technical obstacles, just small issues with a big impact on user experience**
- **EGEE TCG WG on MPI addressing issues in collaboration with int.eu.grid project**

Why should EGEE care about MPI?

- **Many application areas require MPI support**
  - Earth sciences, fusion, astrophysics, comp. chemistry…
  - Significant results can be obtained with 10s-100s of CPUs
- **Many clusters are MPI-ready**
  - Local use via direct submission
  - Shared filesystems, high-performance interconnects
- **But can't provide good access through the Grid**
- **Good MPI support will help to bring new users to EGEE**
  - As a useful infrastructure in itself
  - As a testbed prior to execution in high-end computing centres

Existing support (WMS)

- **MPICH job type in LCG RB and API (also gLite WMS)**
  - Allows multiple nodes to be requested
  - Sets number of nodes in Globus RSL
  - Wraps user binary in call to mpirun
  - Adds “MPICH” and no. cpus requirements to job ClassAd
- **Assumes (hard-codes) site configuration**
  - mpirun deprecated at many sites
- **Restricted set of jobmanagers**
  - Only “pbs” and “lsf” supported
  - Not “lcgpbs”, “torque”, “lcglsf”, etc.
  - Rules out > 80% of MPICH-configured sites!
- **“MPICH” is the wrong level of abstraction**
  - What about OpenMPI etc.?

lcgpbs	168
pbs	47
lcglsf	27
sge	14
pbspro	8
lcgcondor	6
lcgsge	3
lsf	3
condor	2
Total	278

Jobmanager types at sites advertising MPICH (9/5/07)

Existing support (Config)

- **If you support MPI, then publish the “MPICH” tag**
- **That’s it!**
- **Sites wanting to support MPI need to work out all the other details themselves**
  - High cost of entry
  - Only sites with existing MPI setup or strong demand will even try

Proposal

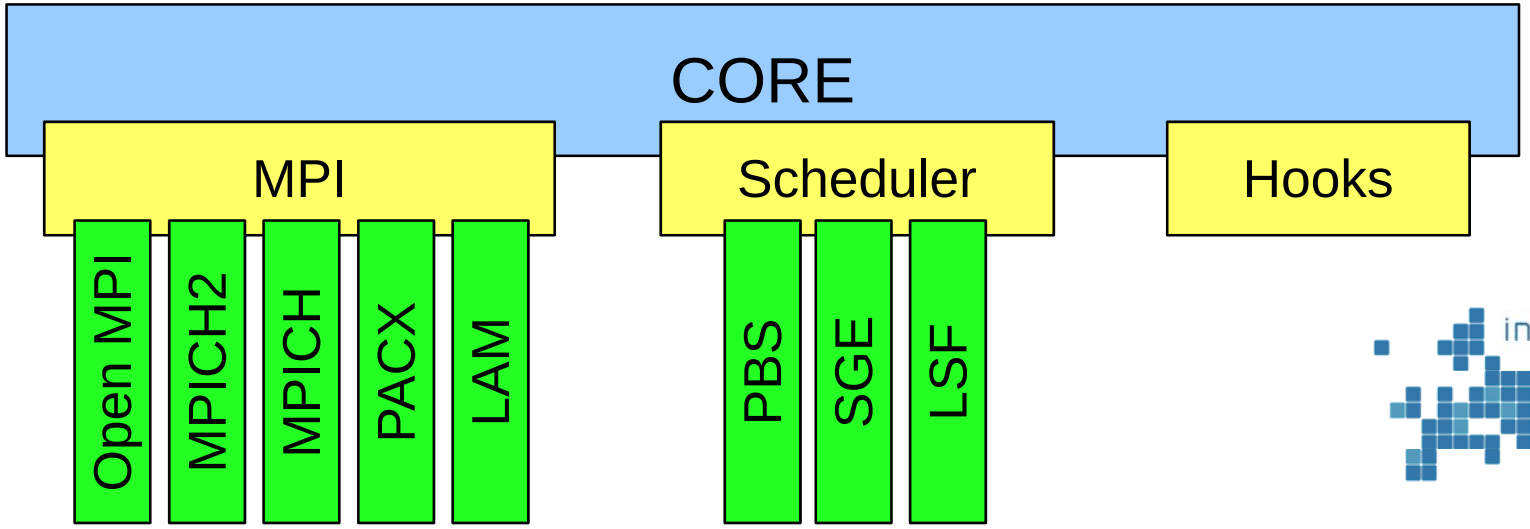
- **EGEE/int.eu.grid integration meeting (Dublin, Dec. 06) made the following recommendations:**
- **No special MPI job type in WMS**
  - Just jobs with multiple nodes
  - User submits wrapper script including MPI setup
  - Wrapper script can also include compile stage
  - Avoid hard-coded assumptions about site setup
- **Agreed configuration principles for MPI sites**
  - Info sys tags to advertise capabilities
  - Environment variables on WN so jobs can locate MPIs
- **Use mpi-start to reduce burden on user**
  - Details of MPI setup, jobmanagers, etc

Site configuration

- **Site admin provides info for job to set up MPI correctly**
- **Information providers**
  - Publish MPIs supported as software run-time environment tags
  - GlueHostApplicationSoftwareRunTimeEnvironment: OPENMPI
- **Environment variables**
  - Location of different MPIs (e.g. MPI_MPICH_PATH)
  - Version of different MPIs (e.g. MPI_MPICH_VERSION)
  - Shared home FS (MPI_SHARED_HOME)
- **mpi-start installed on CE and WNs**
- **Replace mpirun with dummy script (work-around)**

int.eu.grid's mpi-start

- Intermediate layer between WMS and batch systems
- Implemented as portable shell scripts
- Extensible via user hooks
- Takes care of MPI and scheduler setup
- User sets simple variables then invokes mpi-start



Job submission

JDL

```
JobType = "MPICH";
NodeNumber = 8;
Executable = "mpi-start-wrapper.sh";
Arguments = "mpi-test OPENMPI";
InputSandbox = {"mpi-start-wrapper.sh","mpi-hooks.sh","mpi-
test.c"};
Requirements = Member("OPENMPI",
other.GlueHostApplicationSoftwareRunTimeEnvironment);
```

wrapper

```
# Setup for mpi-start.
export I2G_MPI_APP=$MY_EXECUTABLE
export I2G_MPI_TYPE=$MPI_FLAVOUR
export I2G_MPI_PRE_RUN_HOOK=mpi-hooks.sh
export I2G_MPI_POST_RUN_HOOK=mpi-hooks.sh
# Invoke mpi-start.
$I2G_MPI_START
```

hooks

```
pre_run_hook () {
mpicc -o ${I2G_MPI_APP} ${I2G_MPI_APP}.c
}
```

Advantages

- **Flexibility**
  - No assumptions about site configuration
  - Site admins free to configure as they wish
  - Hooks allow for pre-compilation, post-processing
  - Users choose how to run MPI

- **Simplicity**
  - mpi-start hides details of MPI execution
  - Potential for standard wrapper scripts installed on UIs

Testing

- **Need reliable test of MPI functionality at sites**
  - Verify that published functionality works
  - For use by VOs to find suitable sites (e.g. FCR)
- **SAM sensor**
  - Needs own job to perform multi-node submits

No	RegionName	SiteName	NodeName	Status	dteam				
					mpi-inst	mpi	start	js-mpi	advert
1	SouthEasternEurope	AEGIS01-PHY-SCL	ce.phy.bg.ac.yu	OK	na	na	error	na	ok
2	NorthernEurope	BEgrid-KULeuven	kg-ce01.cc.kuleuven.ac.be	OK	na	na	error	na	ok
3	NorthernEurope	BEgrid-ULB-VUB	gridce.iihe.ac.be	OK	na	na	ok	na	ok
4	SouthEasternEurope	BG04-ACAD	ce02.grid.acad.bg	OK	na	na	error	na	ok
5	SouthWesternEurope	CNB-LCG2	mallarme.cnb.uam.es	OK	na	na	error	na	ok
6	France	GRIF	grid10.lal.in2p3.fr	WARN	ok	na	ok	na	ok
7	France	GRIF	ipnls2001.in2p3.fr	OK	na	na	ok	na	ok
8	France	GRIF	lpnce.in2p3.fr	OK	na	na	ok	na	ok
9	SouthEasternEurope	HG-03-AUTH	ce01.afroditi.hellasgrid.gr	OK	na	na	error	na	ok
10	France	IN2P3-IRES	sbgce1.in2p3.fr	OK	na	na	ok	na	ok
11	France	IN2P3-LAPP	lapp-ce01.in2p3.fr	OK	na	na	ok	na	ok
12	Italy	INAF-TRIESTE	grid001.oat.ts.astro.it	OK	na	na	error	na	ok
13	Italy	INFN-CAGLIARI	grid002.ca.infn.it	ERROR	na	na	error	na	ok
14	Italy	INFN-PADOVA	prod-ce-01.pd.infn.it	OK	na	na	error	na	ok
15	NorthernEurope	SARA-LISA	mu9.matrix.sara.nl	OK	na	na	error	na	ok
16	Italy	SNS-PISA	gridce.sns.it	OK	na	na	error	na	ok
17	SouthEasternEurope	TR-01-ULAKBIM	ce.ulakbim.gov.tr	OK	na	na	error	na	ok
18	SouthEasternEurope	TR-10-ULAKBIM	kalkan1.ulakbim.gov.tr	OK	na	na	error	na	ok
19	CentralEurope	TU-Kosice	ce.grid.tuke.sk	OK	na	na	error	na	ok
20	AsiaPacific	Taiwan-LCG2	quanta.grid.sinica.edu.tw	OK	na	na	error	na	ok
21	UK_Ireland	csTCDie	gridgate.cs.tcd.ie	OK	error	na	ok	na	ok

- **mpi-start**
- **Advertised MPIs**
- **MPI execution**
- **Shared FS**

Work in progress!

Status

- **Configuration fully implemented in Quattor templates**
  - Deployed at ~10 sites (France, Belgium, Ireland)
- **YAIM config function developed**
  - Need volunteers to test!
  - Will submit for certification once shown to work
- **WMS and user API**
  - Patch submitted for LCG RB and API
  - In negotiations with JRA1 to modify gLite WMS

Site status

MPICH with 2 nodes available	68
MPI-START with 2 nodes available	11
Sites passing MPI SAM tests	9

Open issues

- **Avoid over-complicated job submission while retaining flexibility**
  - Can the RB help without hard-coding?
- **Local batch system needs detailed requirements**
  - To efficiently schedule MPI and non-MPI jobs
  - To set up MPI more easily
  - Long-standing request – gLite CE should help
- **Cross-site MPI (build on int.eu.grid?)**

Get involved!

- **Applications people**
  - Submit MPI jobs with new config (select on MPI-START tag)
  - Send feedback to the mailing list

- **Site admins**
  - Configure your site for MPI with YAIM/Quattor
  - Send feedback to the mailing list

- **Web: http://www.grid.ie/mpi/wiki**
- **Mailing list: project-eu-egee-tcg-mpi@cern.ch**