MPI support on EGEE

Stephen Childs
Trinity College Dublin

www.eu-egee.org

Overview

- Resource broker and WMS
- Site configuration issues
- YAIM and Quattor modules
- Job submission
- Testing MPI

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)

RB/WMS configuration

- **gLite WMS allows jobwrappers to be edited**
  - Can remove hard-coded “mpirun” invocation
  - Needs to be done for each supported LRMS
  - /opt/glite/etc/templates/template.mpi.{pbs,lsf}.sh
- **Newer WMS should allow for “Normal” jobs with multiple nodes**
  - No hard-coded “mpirun”
  - Remove check for “pbs” or “lsf” jobmanagers
- **Don’t hard-code anything in WMS**
  - Maximum flexibility
  - How can users write wrappers that work everywhere? mpi-start

Site configuration

- **The recommended configuration**
  - Shared home filesystem between WNs (and possibly CE)
  - Use the “pbs” jobmanager not “lcgpbs” or edit GLUE information to publish “pbs” not “torque” (CE_BATCH_SYS)
  - Install mpi-start and MPI RPMs on WNs
  - Install “dummy” mpirun on WNs
  - Publish mpi-start availability and MPI versions in GLUE RTE
  - Set environment variables on WN describing MPI flavours

- **Can configure manually, or use Quattor or YAIM**

mpi-start

- **mpi-start provides flexibility and simplicity**
  - Allows site admins to “communicate” site config to job
  - Customisations for different types of site can be included
  - Hides details of MPI configuration from users

- **Site admin sets env. variables describing their site**
  - Location and version of MPI implementations
  - Support for passwordless ssh
  - Location of external mpiexecs (mainly for mpich/mpich2)

- **mpi-start running on WN reads these variables and configures MPI and scheduler appropriately**

YAIM

- **Module has dual aims:**
  - Configure Grid for cluster where MPI is already configured
    - Sysadmin tells YAIM details of installed MPIs
    - YAIM sets up Grid env. variables (WN) and GLUE (CE)
  - Add baseline MPI functionality in non-MPI cluster
    - Sysadmin just sets ENABLE_MPI
    - Install standard MPIs (mpich, mpich2, openmpi, mpiexec)
    - Set up Grid env. variables (WN) and GLUE (CE)

- **First version of org.glite.yaim.mpi committed to CVS**
  - Built for “modular” YAIM (v. 4)
  - Separate RPM to be installed
- **Ready for testing! (http://grid.ie/mpi/wiki/YaimConfig)**

YAIM meta-RPM

- **For installation of baseline MPI setup on WNs**
  - i2g-mpi-start
  - mpich
  - mpich2
  - openmpi
  - mpiexec (OSC)
  - …
- **Will hopefully be integrated into standard gLite release**

Quattor

- **Recommendations for MPI configuration are fully implemented in QWG templates**

File distribution

- Need mechanism for distributing (at least) the MPI binary to all nodes in MPI job. Options in order of preference/convenience:

1. Shared home filesystem
2. Passwordless-ssh between worker nodes
3. Use mpiexec to distribute files
4. Separate shared area
5. Small MPI “server” installed on all nodes (next slide)

- mpi-start can automatically distribute files (1, 2, 5 already supported; 3 & 4 coming soon)
- The availability of shared homes should be published by site in GLUE and described in environment variables

Distribution via MPI

1. **Use an MPI implementation that can start without needing ssh (e.g. mpiexec)**
2. **On first WN, tar up whole job directory**
3. **Use mpiexec to run ‘cat’ on all nodes in job**
   - Pipe in contents of source file
   - Configure mpiexec to copy stdin to all nodes
   - Pipe out to destination file
4. **Use mpiexec to untar file on all nodes**

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 # ($1)
export I2G_MPI_TYPE=$MPI_FLAVOUR  # ($2)
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
}
```

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**

- **OK=mpi-start**
- **WARN=no mpi-start**

Running nightly for dteam VO

Status

- **I2G's mpi-start has greatly eased configuration and job submission for EGEE MPI users**
  - Aim to deploy at all EGEE MPI sites
  - Meanwhile, mpi-start can easily be sent with the job
- **Deployment status (4/10/07)**
  - 14 sites with mpi-start installed
- **YAIM module being tested at a number of sites**
- **Adding new file distribution functionality to mpi-start**
  - Will bring in many more sites without shared FS or ssh

Conclusion

- **Recipes for site configuration and job submission on EGEE infrastructure exist and work**
- **We can now do “basic” MPI on EGEE, but …**
  - Interaction with schedulers
  - Compilers, licensing, …
  - Cross-site MPI
  - MPI debugging, tool support
- **EGEE should leverage I2G work on advanced topics**
- **All EGEE MPI documentation available via http://grid.ie/mpi/wiki**
- **Mailing list: project-eu-egee-tcg-mpi@cern.ch**