MPI on the grid: operational issues

Fokke Dijkstra
Donald Smits Center for Information Technology
University of Groningen
EGEE Conference Barcelona, 22 September 2009

Introduction

- **2nd MPI WG started 2009**
- **Questionnaires for users and site administrators**
- **Site administrator perspective**
  - Experiences with installing a site according to the guidelines from the previous WG
- **Note that some things are already fixed!**
- **Local site RUG-CIT**
  - Torque/maui
  - Yaim

YAIM

- **Special Yaim functions exist to configure for MPI**
  - Information system
  - Environment variables
  - Submit filter to adapt jobs
  - Dummy mpirun
- **Installation of packages is up to the site admin**

WMS support

- **MPICH jobtype**
  - Confuses implementation with jobtype
  - Broader parallel job support needed
- **Hardcoded call to mpirun in job script**
  - Still there!
  - MPICH mpirun can start a script
  - WG Suggestion to install wrapper script instead
    - Installed by YAIM
- **Only pbs and lsf supported by WMS**
  - Publish that you use pbs if using torque
- **MPICH jobtype abandoned, Normal jobtype can also be used!**

Information system

- **Available MPI implementations published in IS**
  - GlueHostApplicationSoftwareRunTimeEnvironment
    - MPICH, OPENMPI, etc.
    - MPICH-1.2.7, etc.
- **Environment variables set to point to location on file system**
- **Version requirements hard to use**
  - Current regexp matching in JDL makes this very hard
    - Exact or partial match only
- **Consistency not checked**
- **Information about interconnects and shared file system also published**
  - e.g. MPI-Infiniband
  - Needs to be consistent across sites

MPI packages

- **Lack of rpm packages for common MPI implementations in standard gLite repository (only MPICH-1)**
- **Inconsistent packaging (e.g. OpenMPI supplied in CentOS repository does not install into /opt)**
- **Support for mpiexec dropped because of switch to newer torque version**

- **This makes it hard for most site administrators to install MPI!**

- **MPI versions should also be compiled with support for local hardware and software (scheduler)**
  - E.g. OpenMPI can use torque scheduling daemons
  - Support for local interconnects hard to provide

mpi-start / mpiexec

- **mpi-start recommended as a tool to help users starting up MPI jobs**
  - Seems to work fine
  - Depends heavily on setting environment variables to correct values
- **Depends on mpiexec for mpich-1**
  - mpiexec support dropped from gLite!
  - Mpich-1 only available implementation in gLite repository
- **Is it still maintained?**

- **MPICH support dropped MPICH2 and OpenMPI added**
- **mpi-start still maintained**

Shared file system

- **Shared file system between WNs recommended**
  - Easier for applications
- **Existence has to be published**
- **Performance?**
  - NFS
  - Moving sequential jobs to local temporary directory
  - What, when a job does not need it for I/O?
  - Fast shared file system may be good for everyone
- **Configuration not supported by YAIM**
  - Is this a good idea anyway?

Starting up tasks using ssh

- **Most MPI implementations support ssh for starting up remote tasks**
- **Passwordless ssh set up by YAIM for torque job manager**
- **Site administrators may not like this**
  - For torque a pam module exists to prevent unwanted logins
  - Correct cpu time accounting problematic
  - Remote tasks not under control of scheduler
- **Gives user a lot of freedom to perform other parallel work without using MPI**
- **Some MPI implementations can use scheduler daemons**
  - OpenMPI (has to be compiled with support for this)
  - mpiexec for MPICH2

SAM tests

- **Sites that support MPI may still not work**
- **Lack of (enforced) SAM test for MPI**

- **What should be tested anyway?**
  - Consistency
    - Published information
    - Environment variables
    - Installed packages
  - Simple test program

Improved scheduling of cores

- **Use job requirement from the user**
  - Schedule full nodes
  - I don't care (site default applied)
    - Scheduling first available cores improves scheduling time dramatically on small sites
- **Sites have a different number of cores per node**
  - Always schedule complete nodes
  - Schedule just the requested cores
  - This probably should be another job requirement
- **Use of torque submit filter may break/remove site script**

Other problems

- **Topics not covered by the previous WG**
  - Support for other parallel job types
  - Scheduling of cores to the jobs
- **CPU time limits do not work with parallel jobs**
  - Best solution, only use wallclocktime
  - Other option make them very high
    - Some large VOs make requirements on CPU time (why?)
  - Local solution publish wallclock time limit as cpu time limit

My personal recommendations

- **Improve WMS**
  - Remove call to mpirun √
  - Remove requirement to publish pbs or lsf √
- **Improve repository**
  - Make rpms available of commonly used implementations
    - OpenMPI
    - MPICH2
  - mpiexec for MPICH2
  - Support mpi-start
  - Support for torque
  - Source RPMS for recompilation
- **Make use of SAM tests to check working MPI support**
- **Improve documentation**
  - General MPI
  - Guidelines for setting up:
    - Shared file system
    - Passwordless ssh
    - Recompiling MPI implementations with support for local hardware/software