MPI Support

John Walsh (TCD)

CERN – JRA1/SA3 All hands meeting

Dec 15th/16th 2009

Background

- **MPI-WG (2nd Working Group)**
  - Working Group examined MPI support in gLite
    - Survey of Site Admins/Users
  - Delivered interim draft report on state of MPI (EGEE09)
    - Critical of current state of MPI support
  - Full report expected Jan'10
- **MPI-WG hosted EGEE09 session**
  - User communities reported bad performance and support
    - ESR (Earth Sciences)
      - *Only 7 of 26 sites usable/stable*
    - CompChem
      - *51% success for 8proc jobs,*
      - *~21% success for 16-64 processors*
  - Results are discouraging uptake/investment by new users
- **EGEE-III mid-term review critical of level of support**

MPI-TF

- **Established after EGEE09**
  - Based on request for dedicated support from users and admins
  - Oversight by TMB
  - Joint leadership (Isabel Campos/John Walsh)
  - Dedicated effort across activities (JRA1, NA4, SA1,SA3)
- **Remit**
  - Implement short term feasible recommendations of MPI-WG
  - 7 month life span
    - Note: long term strategy defined by MPI-WG
  - Engage with Users and Sites
  - Seek improvements at current sites
  - Encourage deployment at sites

Objectives

- **Improvements/Updates in documentation (NA4)**
  - User examples and use cases (NA4), localisation & training
  - Installation and support procedures (SA1/SA3)
  - Unify under central location (NA4)
- **MPI_utils (SA3)**
  - Update outdated requirements
  - Update installation/configuration
  - Fix long outstanding bugs
  - Test/Certify components
- **MPI-START (JRA1)**
  - Enables multiple MPI distribution to be installed/supported
- **SAM testing (SA1)**
- **Produce "Knowledge-base" (ALL)**
- **Interact with community (users/admins)**

MPI TF Composition

- **SA1**
  - Isabel Campos (IFCA, ES)
- **SA3**
  - John Walsh (TCD, IRL) – Also SA1,
  - Dennis van Dok (NIKHEF, NL)
- **JRA1**
  - Enol Fernandez (IFCA,ES)
- **NA4**
  - Jeroen Engelberts (SARA, NL)
  - Others to be appointed from User communities

- **Jeroen/Dennis are bridge to MPI-WG**
  - MPI-TF is a subgroup of MPI-WG

SAM testing & Site Interactions

- **SAM testing**
  - Implemented by Karolis Eigelis, Konstantin Skaburskas)
  - Validation of sites in progress (test several times per day)
  - Manual review of results (JW/IC/EF + others at IFCA)
  - 105 CE tested (based on publishing MPI-START tag)
    - ~50 CE pass for all supported MPI distributions at site
  - Tests need refinement (e.g E
  - Failing sites in process of being individually examined
    - Tickets raised/follow-up in GGUS
  - Results being used to build knowledge-base of problems and their resolution
  - Currently tests are **NOT CRITICAL**
    - Need solid quorum of sites to pass tests successfully
    - Need complete knowledge base
    - Need release of updated components

MPI_utils

- **New version in certification (patch 3092/3225)**
  - Updates to glite.yaim.mpi
  - Updates to meta-package
  - Fixes mpiexec/torque dependency

- **HOWEVER, SOME ISSUES**
  - OS support for MPI distributions
    - OpenMPI not compiled with torque support (SL4/SL5)
    - MPICH2 available for SL4/SL5 (from EPEL)
      - *SL4 version has bad run-time dependency of unavailable RPM*
      - *SL5 OK*
  - Fixes requested to remove explicit dependency on any MPI distribution
  - Error found in logic of function (MPI_SHARED_HOME tag)

MPI-START

- **Code review of mpi-start (JW/EF)**
  - Weaknesses in error reporting identified and fixed
  - Improved file distribution mechanism
- **New version of MPI-START produced (0.60)**
  - Not yet in certification

Batch Support

- **Systems**
  - Support for Torque/Maui and LSF
    - Well supported (LCG-CE/CREAM-CE)
  - SGE support for MPI in progress (Alvaro Simon Garcia)
    - Design stage (Dec 2009)
    - Implementation LCG-CE/CREAM-CE (Feb'10)
    - Testing LCG-CE/CREAM-CE (Mar'10)
  - Condor
    - Unknown, limited support for MPI under Condor

Issues

- **Large range of testing environments**
  - CE (LCG-CE + CREAM-CE)
  - Batch system (Condor(?), LSF, Torque, SGE)
  - Shared homes vs. non-shared homes
  - Interconnect support
  - 32/64 bit support
  - Multiple MPI implementations (can only test most common)
    - Supporting interconnect X, Y & Z
  - Potentially 2x4x2x2xNxM scenarios!!!!
    - Pragmatism dictates only common scenarios feasible
- **MPI-TF needs to cover a lot of ground quickly!**
- **Support for generic parallel jobs(?)**

Events/Meetings/Misc

- **IC/JW have met F2F several times since EGEE'09**
- **Produced input for EGI-INSPIRE and EMI proposals**
- **Abstract submitted to EGEE User Forum**

- **Intensive 2/3 day (Santander - Jan'10)**
  - Members of MPI-TF and others
  - Goal is to produce major upgrade (release Feb'10)
  - Cover all objectives of MPI-TF

Questions?