Support for MPI in the EGI Infrastructure

Community Forum 2012

Álvaro Simón (CESGA)

e-infrastructure

Outline

- Introduction
  - Virtual Team framework
  - MPI Virtual Team
- MPI support within EGI era
  - Documentation
  - LRMS and MPI support
  - Accounting
  - Monitoring
  - MPI user communities and VOs
- Future work

Introduction

- Virtual Team framework:
  - EGI VTs are short living projects that focus on well defined, non-operational activities around the production infrastructure.
  - Enables NGI and EGI.eu members to initiate an participate in different activities.
  - VTs focus around new communities and sustainability.
  - The VT project framework was developed within the NA2 activity of the EGI-InSPIRE project.
  - but it is also open for NGIs, countries and regions that are not involved in NA2 or in EGI-InSPIRE.
  - Virtual Teams have members from multiple NGIs and carry out a targeted action within the following activity

Introduction

- MPI VT it was created in November 2011
  - It's a short living project (end date: 31/May/2012)
  - MPI VT EGI Wiki: https://wiki.egi.eu/wiki/VT_MPI_within_EGI
  - MPI VT mailing list: vt-mpi@mailman.egi.eu

- MPI VT Objectives:
  - Collaborate with user communities and projects that use MPI resources.
  - Promote MPI deployment.
  - Enhance MPI experience and support for EGI users.

MPI support within EGI era

- Documentation:
  - MPI documentation was fragmented in different places.
  - EGI MPI documentation was reviewed and extended by MPI VT members.
  - MPI Administrator guide: https://wiki.egi.eu/wiki/MAN03
  - MPI VT members are working to provide a new MPI user guide.
  - It will be available in a single endpoint.

MPI support within EGI era

- LRMS and MPI support:
  - MPI VT members are working to detect current LRMS MPI issues.
  - At this moment CREAM GE EMI release fully supports MPI. Tested by MPI VT members.
  - It was found an Issue with Torque / Maui MPI support
    - GGUS tickets #57828 & #67870
    - Not supported by EMI but by a third party.
    - EPEL includes torque server / client packages and an external EMI repo includes MAUI updates.
    - This issue will be fixed in next EMI2 release.

MPI support within EGI era

- Parallel jobs accounting:
  - APEL is not ready to registry MPI usage records yet.
    - This work is on progress and tracked by JRA1.4 task and EMI.
    - New APEL with MPI capabilities will be ready at the end of summer.
  - The only way that we have at this moment to identify MPI jobs is checking job efficiency >1 (as the ratio of CPU time to wall time).
  - MPI VT members are in contact with EMI developers to describe the parallel usage record for the different batch systems.
  - New APEL with MPI capabilities probably will be ready at the end of summer.

MPI support within EGI era

- Monitoring:
  - Current SAM nagios probes does not detect if MPI framework is really working fine or not (submits just an “hello world”).
  - A new set of nagios probes specifications were created by MPI members to improbe MPI monitoring: https://wiki.egi.eu/wiki/VT_MPI_within_EGI:Nagios
  - The new Nagios probes specifications are being reviewed by MPI VT.
  - After this review the new probes will be developed by SA3.

MPI support within EGI era

- Monitoring:
  - The new nagios probes include these tests.
    - EnvSanityCheck checks:
      - The information published by the (MPI or Parallel) service.
      - If site has the GlueCEPolicyMaxSlotsPerJob variable set to a reasonable value (not 0 nor 1 nor 999999999) for the queue where the MPI job is supposed to run.
      - Site publishes reasonable GlueCEPolicyMaxCPUTime and GlueCEPolicyMaxWallClockTime values (not 0 nor 999999999).
      - Timeout 120s.

MPI support within EGI era

- Monitoring:
  - The new nagios probes include these tests.
    - This probe requires two slots in different machines.
      - (JobType="Normal"; CpuNumber = 2; NodeNumber=2)
    - SimpleJob checks:
      - The MPI functionality with a minimum set of resources.
      - MPI-START is able to find the type of scheduler, determine if the environment for the MPI flavour, is able to distribute the application binaries, etc.
      - Timeout: 120s.

MPI support within EGI era

- Monitoring:
  - The new nagios probes include these tests.
    - This probe requires 4 slots with 2 instances running in different dedicated machines
      - (JobType="Normal"; CpuNumber = 4; NodeNumber=2; SMPGranularity=2; WholeNodes=True)
    - ComplexJob checks:
      - The MPI functionality and if the recommendation from the EGEE MPI WG are being implemented.
      - MPI-START is able to find the type of scheduler, determine if the environment for the MPI flavour, is able to distribute the application binaries, etc.
      - Timeout: 120s.

MPI support within EGI era

- Monitoring:
  - MPI VT request the creation of a new GOCDB Service Type
    - GOCDB acting like an ITIL CMDB: all configuration items should be registered
    - Release the dependency from the information system
    - Allows accurate infrastructure information about infrastructure status

MPI support within EGI era

- MPI user communities and VOs:
  - MPI VO bring together sites and users interested in MPI.
  - This VO is NOT intended for everyday use by all users wishing to use MPI.
  - The main reason for its establishment is to collect experience that will be later adopted by regular VOs.

MPI support within EGI era

- MPI user communities and VOs:
  - A new VOMS service was created and is ready to include new users: https://egee.cesnet.cz/mpi/registration/prihlaska_priprav.php
  - MPI VO configuration is available in the MPI VT wiki page.
  - At this moment CESNET and CESGA are supporting this VO for testing purposes.

MPI support within EGI era

- MPI user communities and VOs:
  - Home Page: https://www.metacentrum.cz/en/VO/MPI/index.html
  - Registration: https://egee.cesnet.cz/mpi/registration/prihlaska_priprav.php
  - Mailing List: mpi-kickstart@metacentrum.cz
  - Environment Settings:
    - VO_MPI_VOMS_SERVERS="'vomss://voms1.egee.cesnet.cz:8443/voms/mpi?/mpi'"
    - VO_MPI_QUEUES=""
    - VO_MPI_SW_DIR="$VO_SW_DIR/mpi"
    - VO_MPI_DEFAULT_SE=""

Future Work

- Continue to unify and simplify MPI documentation
- Inspect how the IS schema can adjust to specific LRMS policies, and enforce the correct values
  - May involve interaction with LRMS IP developpers
- Tested “new” probes functionality
  - Should be developed by SA3
- Include new sites to support VO MPI
  - For testing purposes.

Questions

- Questions?