MPI Accounting

Enol Fernández (CSIC)

MPI TF

MPI & Parallel jobs

- MPI is not the only kind of parallel jobs, moreover there is more than one MPI implementation
  - E.g. OpenMP
  - EMI-ES includes a “Parallel Environment” with several types already defined:
    - *MPI,GridMPI,IntelMPI,LAM-MPI,MPICH1 … PVM*
- Is Application type/Parallel Environment accountable?

Relevant fields in EMI UR

- 11.7 WallDuration
  - WallClock time elapsed during the job execution. basically it EndTime-StartTime no matter on how many cores, processors, nodes, sites the user job ran on.

- 11.12 Host
  - It should contain all the hosts involved in the MPI execution.

Relevant fields in EMI UR

- 11.8 CpuDuration
  - **Aggregated** CPU time consumed by the job.
  - Passwordless SSH startup mechanism may cause this to be wrongly accounted by the batch system
  - Sites should use a startup mechanism integrated with the batch system:
    - OSC MPIEXEC for Torque + MPICH
    - Open MPI with tight integration for SGE (included in OS distributed version) and Torque **(not included)**
    - Further investigation needed for MPICH2

Relevant fields in EMI UR

- D.5. **NodeCount**: Number of nodes used (Positive integer)
- D.6. **Processors**: Number of processors used (Positive integer)
  - Is this the number of slots? Number of cores?
- Resource Usage:
  - D.1. **Network**, D.2. **Disk,** D.3. **Memory,** D.4. **Swap**

Other information

- Total Number of processes/threads?
- Can we get per node (or even per core) information?
  - E.g. non aggregated CPU utilisation, memory usage, network usage
  - This would allow to detect when an application is using the whole node for other reasons than CPU (e.g. memory)
  - Is it possible to get it from batch system?
  - LLView (FZJ) created some extensions to UR in order to publish such information
- Network topology