MPI on the Grid

Camille Coti
ICL Friday Lunch
Friday, February 13th 2009
The QosCosGrid project

» Middleware partners

- Institut National de Recherche en Informatique et en Automatique
  France
- Poznan Supercomputing and Networking Centre
  Poland
- Israel Institute of Technology
  Israel
- Platform Computing SA
  France

» Application partners

- University of Ulster (Coordinator)
  Northern Ireland
- Collegium Budapest
  Hungary
- AITIA International Inc.
  Hungary
- Universitat Pompeu Fabra
  Spain

- Cranfield University
  United Kingdom
- University of Amsterdam
  The Netherlands
- University of Queensland
  Australia
### QosCosGrid Use-Cases

#### Major Source of complexity ->

<table>
<thead>
<tr>
<th>Number of entities</th>
<th>Structure of entities</th>
<th>Interaction Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic / Evolutionary Algorithms Toolbox</td>
<td>Gene Regulatory Networks</td>
<td>Parameter Estimation in Highly Coupled ODEs</td>
</tr>
<tr>
<td>Stellar Dynamics and Evolution</td>
<td>Protein Folding and P-P Interactions</td>
<td>Gene Regulation (spatial model)</td>
</tr>
<tr>
<td>Ecological Evolution Models</td>
<td>Business Supply Chain Dynamics</td>
<td>Social Influence on Discrete Choices</td>
</tr>
</tbody>
</table>

- **Inter-Model Parallelisation**
  - (Parameter space search, Model fitting, Sensitivity analysis)

- **Intra-Model Parallelisation**
  - (Spatial segregation, Multi-scale modeling, etc.)

- **Inter- and Intra-model Parallelisation (agent-based models)**

- **Use-cases using MPI**

  - 3 categories / communication patterns
    - Evolutionary computation
    - Living simulations
    - Coevolutionary agent models
The QosCosGrid architecture
Grid = cluster of clusters
Grid-aware MPI applications for QosCosGrid

» The programmer describes:
» Communication schemes
» Associated resource requirements
» In companion file -> jobProfile
» The scheduler
» Allocates resources wrt this companion file
MPI for Grids

» Grid-enabled MPI middleware
  » Inter-cluster connectivity
  » Efficient collective operations
» Grid-enabled MPI application
  » Hierarchical master-worker
  » Virtual topology
QCG-OMPI Design

» New communication drivers
  » OMPI: new BTL
  » ORTE: new OOB

» Extended runtime environment
Connectivity

» OpenMPI
  » Wide-open, direct connections
» MPICH-g2
  » Range of open ports, direct connections
» PACX-MPI
  » Relay proxies

» QCG-OMPI
  » Direct connection (if open ports available)
    » Port range
  » Traversing TCP
  » Proxy
Performances: communications

Throughput $\rightarrow$

Latency $\downarrow$

$\rightarrow$ Bandwidth
  $\rightarrow$ Shared if using relaying method
  $\rightarrow$ Same otherwise

$\rightarrow$ Latency
  $\rightarrow$ Higher if using relaying method
  $\rightarrow$ Same otherwise

<table>
<thead>
<tr>
<th>times in seconds</th>
<th>MPICH-G2</th>
<th>OpenMPI</th>
<th>QCG-OMPI direct</th>
<th>QCG-OMPI proxy</th>
<th>QCG-OMPI traversing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-Cluster</td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0001</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Orsay-Rennes</td>
<td>0.0104</td>
<td>0.0103</td>
<td>0.0103</td>
<td>0.0106</td>
<td>0.0103</td>
</tr>
<tr>
<td>Orsay-Bordeaux</td>
<td>0.0079</td>
<td>0.0078</td>
<td>0.0078</td>
<td>0.0084</td>
<td>0.0078</td>
</tr>
<tr>
<td>Bordeaux-Rennes</td>
<td>0.0081</td>
<td>0.0080</td>
<td>0.0080</td>
<td>0.0087</td>
<td>0.0080</td>
</tr>
</tbody>
</table>
MPI for Grids

- Grid-enabled MPI middleware
  - Inter-cluster connectivity
  - Efficient collective operations
- Grid-enabled MPI application
  - Hierarchical master-worker
  - Virtual topology
Topology-Aware Global Communications

- Fit with the physical topology
- Minimize inter-cluster communications
  - Naive algorithm: $O(\log N)$ messages
  - Goal: $O(1)$ messages
- Hierarchical approach
MPI_Bcast on a grid
MPI_Reduce on a grid

QCG_Reduce

message size (B)

Standard

Grid

time (ms)
MPI for Grids

- Grid-enabled MPI middleware
  - Inter-cluster connectivity
  - Efficient collective operations
- Grid-enabled MPI application
  - Hierarchical master-worker
  - Virtual topology
Ray2mesh

» Ray2mesh
  » Geophysics application
  » Collective operations
  » Master-worker computation

» Drawbacks
  » Centralized master
    » Unique data queue, bottleneck
  » Not scalable
  » Collective communications
Hierarchical Ray2mesh

» Adapted collective communications
» Hierarchical master-worker
JobProfile for Ray2mesh

<grmsJob appId="ray2mesh-job">
  <resourceTemplates>
    Define resources and network requirements
  </resourceTemplates>

  <task taskId="ray2mesh-task1">
    <resourcesDescription>
      <topology>
        <group>
          <group groupId="central_master">
            <processes>
              Describe each group in terms of number of processes and resource requirements
            </processes>
          </group>
          <processesConnection>
            Define network requirements in groups and between groups
          </processesConnection>
        </group>
      </topology>
    </resourcesDescription>
    <execution type="open_mpi">
      Give path to the application, provide command-line arguments, path to input data files and standard input/outputs
    </execution>
  </task>
</grmsJob>
Topology allocation
Experimental Setup

» Grid’5000
  » Experimental grid
  » Nation-wide testbed
  » 13 clusters in 9 cities
  » 5000 cores
  » Opteron, Xeon, Itanium, G5
  » Ethernet, Myri10g and 2000, IB10G
  » Completely reconfigurable

» We used 3 sites:
  » Orsay, Rennes, Bordeaux
  » Central master in Orsay
  » (N-1)/3 nodes per cluster
Performances

Execution time of Ray2mesh on a Grid

- Vanilla
- Grid-optimized collectives
- Topology-aware
- Topology-aware, workers and master only

Execution time (s)

# of procs
Conclusion

» MPI on grids
  » Need for an adapted communication middleware
    » QCG-OMPI

» MPI applications can give good performances on grids
  » Adapted collective operations
  » Grid-friendly application
  » Participation of the scheduler

» Investigate other patterns
  » Hierarchical domain decomposition
  » Ring of ring