Enabling Large-scale Scientific Workflows on Petascale Resource Using MPI Master/Worker

Mats Rynge, Gideon Juve, Karan Vahi, Gaurang Mehta, Ewa Deelman
*Information Sciences Institute, University of Southern California*

Scott Callaghan, Philip J. Maechling
*Southern California Earthquake Center, University of Southern California*
Outline

Introduction

Pegasus Workflow Management System

pegasus-mpi-cluster

Integration

SCEC Cybershake – Example application

Conclusions
Introduction

• Loosely coupled applications structured as scientific workflow, containing mix of parallel and serial tasks

• Petascale systems are optimized for parallel jobs

• Common solution: Condor glideins for the serial tasks
Glideins on NICS Kraken?

- Cray XT System Environment / ALPS / aprun
  - Login node
  - aprun node
  - Compute node
Approach

- Partition workflow into subgraphs
- Execute partition as a self-contained MPI job
Abstract Workflows - Pegasus input workflow description

- Workflow “high-level language”
- Only identifies the computation, devoid of resource descriptions, devoid of data locations

Pegasus

- Workflow “compiler” (plan/map)
- Target is DAGMan DAGs and Condor submit files
- Transforms the workflow for performance and reliability
- Automatically locates physical locations for both workflow components and data
- Provides runtime provenance
pegasus-mpi-cluster

- Master/worker paradigm
- Master manages the subgraph tasks, handing out work to the workers
- Efficient scheduling / handling of input/outputs
- Subgraph described in a DAG-similar format
- Failure management / rescue DAG
diamond.dag

<table>
<thead>
<tr>
<th>TASK</th>
<th>Command</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>/bin/echo &quot;I am task A&quot;</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>/bin/echo &quot;I am task B&quot;</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>/bin/echo &quot;I am task C&quot;</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>/bin/echo &quot;I am task D&quot;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EDGE</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>
Pegasus Mapper / pegasus-mpi-cluster integration

Original Workflow

Workflow After Label Based Clustering

C1 is a clustered job that uses PMC to execute the subgraph
Southern California Earthquake Center

CyberShake

Example Application
CyberShake

Probabilistic seismic hazard analysis workflow

• How hard will the ground shake in the future?
• Considers a set of possible large earthquakes
• 415,000 earthquakes is typical

Uses Pegasus and Condor DAGMan for workflow management

• Hierarchal workflows
• Small set of large parallel jobs
• 840,000 serial jobs, in 78 sub workflows
Probabilistic Seismic Hazard Analysis (PSHA) curve. Estimates the probability that earthquake ground motions will exceed some intensity measure.

Set of PSHA curves interpolated creates hazard map for an area.
Conclusions and Future Work

• Demonstrated efficient execution of fine-grained workflows on petascale resources by partitioning workflow into MPI master/worker jobs

• Size of partition?

• Size of MPI job?

• Handling tasks with mixed requirements?
  • pegasus-mpi-cluster now considers memory to be a consumable resource
Thank you!

Pegasus: http://pegasus.isi.edu

SCEC: http://www.scec.org