WorkflowGenerator

Workflow Generator

To facilitate evaluation of workflow algorithms and systems on a range of workflow sizes, we have developed a set of synthetic workflow generators. These generators use the information gathered from actual executions of scientific workflows on the Grid as well as our understanding of the processes behind these workflows to generate realistic, synthetic workflows resembling those used by real world scientific applications.

The code used to generate all of the synthetic workflows below, and many others, is available from the GitHub repository. The java workflow generator sometimes generates negative task runtimes, so watch out for that.

Simulator

WorkflowSim can be used to simulate the workflows generated by the Workflow Generator.

Traces

Traces and execution logs from real workflows are available here: here, here, and here. Data sets like these were used to parameterize the Workflow Generator.

Synthetic Workflows

Pegasus Workflows

These workflows come from a paper by Bharathi, et al. [1]. There is another paper with more information about the workflows by Juve, et al. [2].

A large collection of DAXes similar to the ones listed below is available here. Note that it is about 375 MB.

<table>
<thead>
<tr>
<th>Workflow Type</th>
<th>Example</th>
<th>DAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Montage</td>
<td><img src="image" alt="Montage Workflow Diagram" /></td>
<td>25 Node DAX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50 Node DAX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 Node DAX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000 Node DAX</td>
</tr>
</tbody>
</table>
**CyberShake**
The CyberShake workflow is used by the Southern California Earthquake Center to characterize earthquake hazards in a region.

**Epigenomics**
The epigenomics workflow created by the USC Epigenome Center and the Pegasus Team is used to automate various operations in genome sequence processing.

**LIGO Inspiral Analysis**
LIGO's Inspiral Analysis workflow is used to generate and analyze gravitational waveforms from data collected during the coalescing of compact binary systems.

**SIPHT**
The SIPHT workflow, from the bioinformatics project at Harvard, is used to automate the search for untranslated RNAs (sRNAs) for bacterial replicons in the NCBI database.

**Ramakrishnan and Gannon Workflows**
These workflows come from a report by Ramakrishnan and Gannon [3].

<table>
<thead>
<tr>
<th>Workflow Type</th>
<th>Figure in Report</th>
<th>Example</th>
<th>DAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEAD Mesoscale Meteorology</td>
<td>Figure 1</td>
<td>leadmm.xml</td>
<td>leadmm.xml</td>
</tr>
<tr>
<td>LEAD ARPS Data Analysis System</td>
<td>Figure 2</td>
<td>leadadas.xml</td>
<td>leadadas.xml</td>
</tr>
<tr>
<td>LEAD Data Mining Workflow</td>
<td>Figure 3</td>
<td>leaddm.xml</td>
<td>leaddm.xml</td>
</tr>
<tr>
<td>Storm Surge SCOOP Workflow</td>
<td>Figure 4</td>
<td>scoop_small.xml</td>
<td>scoop_small.xml</td>
</tr>
<tr>
<td></td>
<td></td>
<td>scoop_medium.xml</td>
<td>scoop_medium.xml</td>
</tr>
<tr>
<td></td>
<td></td>
<td>scoop_large.xml</td>
<td>scoop_large.xml</td>
</tr>
<tr>
<td>Floodplain Mapping</td>
<td>Figure 5</td>
<td>floodplain.xml</td>
<td>floodplain.xml</td>
</tr>
<tr>
<td>Glimmer</td>
<td>Figure 6</td>
<td>glimmer.xml</td>
<td>glimmer.xml</td>
</tr>
</tbody>
</table>
Pan-STARRS Load  
Figure 13  
psload_small.xml  
psload_medium.xml  
psload_large.xml

Pan-STARRS Merge  
Figure 14  
psmerge_small.xml  
psmerge_medium.xml  
psmerge_large.xml

McStas  
Figure 15  
mcstas.xml


