Problem statement

To be cost effective, many applications hosted on cloud environments have to use the automatic scaling capabilities [1]. The currently existing scaling facilities base on executing actions triggered by exceeding thresholds of particular low-level metrics (e.g., CPU consumption). Such approach is very convenient to make the resource consumption most effective but may fail to optimize business objectives described in a Service Level Agreement.

We propose to employ high-level, customized metrics only relevant to a particular system, as decision-making criteria for auto-scaling. Triggers for dynamic resource provisioning are based on the concepts directly bound to the application under consideration. Such an approach is easier to maintain for the user and focuses on business objectives. To implement the approach we exploit the SAMM system [2,3].

Approach

Integrating SAMM into Cloud

SAMM uses the rules to configure an event processing engine

Cloud infrastructure

Input pattern of numerical integration requests

Test environment using Future Grid

A sample numerical integration application was used to carry out experiments. The architecture of the whole environment with the SAMM involved is presented on the figure below.

Experiments and results

To evaluate the new approach we compare the results of automatic scaling of sample application based on the CPU-usages and the average time spent by computation requests while waiting for processing metrics.

Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>Dynamic workload</th>
<th>Static workload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average instances number</td>
<td>4.52</td>
<td>2.28</td>
</tr>
<tr>
<td>Average wait time (ms)</td>
<td>100.877.06</td>
<td>187.564.69</td>
</tr>
<tr>
<td>Average queue length</td>
<td>904.06</td>
<td>275.46</td>
</tr>
</tbody>
</table>

Conclusions and future work

- Automatic scaling can greatly lower computation costs, depending on the workload strategy adopted
- In our experiments with a numerical integration service the average wait time metric turned out to have a positive influence on the system from business perspective
- In some cases, conventional CPU usage may also be a reasonable criterion for resource management
- On-going work includes the development of a web interface for SAMM and support for other stacks, like OpenNebula or Open Stack.

Acknowledgments

The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.02.03.00-00-007/08-08 as part of the PL-Grid Project (www.plgrid.pl) and ACC,Cyfrent AGH grant 520-08. The experiments were carried out on the Future-Grid Infrastructure [4].

References: