Contribution ID: **62** Type: **Presentation**

Cloud infrastructure for the on demand provisioning of Worker Nodes

Thursday, 20 September 2012 14:00 (30 minutes)

Wider impact of this work

This work, in a way, introduces elasticity to the Grid. Using if not the same but also similar implementations administrators of Grid resources can modify on demand the number computing resources offered based on certain thresholds (i.e. the total number of submitted jobs and the underlying number of physical job slots). We feel that the user experience of the Grid will be enhanced through such implementations, especially for users that rely heavily on parametric job types or large parallel jobs (or both).

Printable Summary

We showcase the development and usage of a Quattor based Openstack reference cloud which is used for the on demand provisioning of additional Worker Nodes under our Grid based infrastructure. The need of being able to add computing resources on-the-fly has gradually emerged over the years as a way to leverage the large number of jobs that may occasionally be routed towards a Grid site. After considering several PaaS based alternatives that could be used to harnest such a need we decided to deploy and operate an Openstack based cloud.

Description of the work

Using an Openstack reference cloud we introduce the concept of elasticity to our Grid site by adding and removing on demand computing resources in the form of virtualized Worker Nodes (vWNs). The vWNs are added under the central batch job queueing system whenever a large number of jobs is directed towards the site (i.e. whenever the number of queued jobs exceeds the total number of physical job slots offered). Such situations are not uncommon in our experience as several users rely heavily on the usage of parametric jobs. Large batches of parametric jobs may be directed by the WMS to a single site as at the time of submission the WMS does not calculate dynamically the impact of the total number of jobs on the Grid site but rather treats them independently using the information supplied at some earlier given point in time by the information system. By bringing elasticity to the Grid we manage to leverage such abrupt demands on computing resources by deploying vWNs to handle the large number of jobs. Distinguishing among serial and parallel jobs we do so by provisioning two types of instrances one for single CPU jobs and one for multi-CPU jobs consisting in the later case of 8 CPUs per instance. Once the number of queued jobs descreases to a number lower than the number of available job slots (physical not virtualized) no additional vWNs are created and the system is left to 'cool off'.

Primary authors: KANELLOPOULOS, Christos (GRNET); TRIANTAFYLLIDIS, Christos (GRNET); KOROSOGLOU, Paschalis (GRNET)

Presenter: TRIANTAFYLLIDIS, Christos (GRNET)

Session Classification: Virtualised Resources

Track Classification: Virtualised Resources: challenges and opportunities (Michel Drescher: track leader)