Contribution ID: **83** Type: **not specified**

Dynamic Deployment and Execution of Hadoop Applications on EGI FedCloud Resources

Wednesday, 6 April 2016 17:00 (20 minutes)

With the rapid increase of data volumes in scientific computations, the importance of utilising parallel and distributed computing paradigms in data processing is becoming more and more important. Hadoop is an open source implementation of the MapReduce framework supporting processing large datasets in parallel and on multiple nodes in a reliable and fault-tolerant manner. Scientific workflow systems and science gateways are high level environments to facilitate the development, orchestration and execution of complex experiments from a user-friendly graphical user interface. Integrating MapReduce/Hadoop with such workflow systems and science gateways enables scientists to conduct complex data intensive experiments utilising the power of the MapReduce paradigm from the convenience provided by science gateway frameworks.

This presentation will illustrate how easily Hadoop clusters can be deployed on EGI FedCloud resources, Hadoop applications can be executed on these clusters, and finally resources can be released after execution. Users of the EGI FedCloud WS-PGRADE gateway can import and parameterise pre-prepared workflows for the above tasks published in a public workflow repository. Users can set the type/flavour and number of desired nodes in the Hadoop cluster, select the target EGI FedCloud site, and define the Hadoop executable and the desired data source and destination. All three functionality (create Hadoop cluster, execute Hadoop job, destroy Hadoop cluster) can be executed as a standalone job or can be combined into more complex workflows automating different user scenarios.

Presenters: BLANCO, Carlos (University of Cantabria); KACSUK, Peter (MTA SZTAKI); KISS, Tamas (University of Westminster, London, UK)

Session Classification: Federated cloud PaaS and SaaS