Contribution ID: **110**   Type: **Presentations**

Bringing cloud technology to distributed data infrastructures

Tuesday, 9 April 2013 14:20 (20 minutes)

Summary

The talk presents work ongoing in the research and development part of EUDAT to integrate next generation storage concepts into distributed data stores. We show three projects: using iRods, the main storage component of the EUDAT infrastructure, to integrate S3 storage; another approach for doing this using DPM; and using iRods storage with HDFS and hadoop to query the storage, compute and make the results available in the native iRods name space.

Impact

The attendees will get an overview about the storage research activities related to cloud technologies in EU-DAT. We present the use cases for the different projects and display the issues one can expect when attempting to integrate next generation storage concepts. The audience will get an introduction into iRods, especially the rule engine (which is used for the hadoop processing), and different approaches to integrate S3-aware cloud storage.

Interested members of the audience can approach the speakers for test installations of the different components.

Description

In EUDAT, a european data infrastructure project started in 2011, european researchers work together build a collaborative data infrastructure based on a federation of research institutes and their local resources. Since more and more research institutes make use of and host cloud technologies, the research and development part of EUDAT investigates possibilities to integrate this infrastructure.

In the work currently done, we have shown that iRods, the main storage technology used in EUDAT, can be used integrate S3-based storages such as OpenStack Swift. Similar exploratory work has been done for the widely used grid storage DPM, which shows where further iRods/OpenStack integration might go. Both projects have different goals and thus use different implementations. In the iRods solution, the iRods server/cluster works as a cache, so it is beneficial for local processing or staging to HPC, in the DPM implementation the data never passes the DPM, aiming at scenarios where the processing also happens in the cloud.

A multitude of scientific disciplines also discover the MapReduce principle as a means to process their data. We show that HDFS storages can be integrated in both iRods and DPM and that iRods can even use MapReduce to transparently generate data using the hadoop MapReduce implementation.

Primary authors: Dr RYBICKI, Jedrzej (Jülich Supercomputing Centre); Mr BRZEŹNIAK, Maciej (Poznan Supercomputing and Networking Center); Mr HELLMICH, Martin (CERN)

Presenter: Mr HELLMICH, Martin (CERN)

Session Classification: Cloud Platforms

Track Classification: Cloud Platforms (Track Lead: M Drescher and M Turilli)