==========

Title: Data Analysis on Large Heterogeneous Infrastructures for Science

Eugen Feller - DALHIS

The worldwide scientific community is generating large datasets at increasing rates causing data analysis to emerge as one of the primary modes of science. A scientific data analysis environment needs to address three key challenges:

a) programmability: easily user composable and reusable programming environments for analysis algorithms and pipeline execution,
b) agility: software that can adapt quickly to changing demands and resources, and,
c) scalability: take advantage of all available resource environments including desktops, clusters, grids, clouds and HPC environments.

In the first part of the talk, we will present a brief overview of the research agenda of the DALHIS associate team, a collaboration between the ACS department at LBNL and the Inria Myriads project-team, that aims at creating a software ecosystem to facilitate seamless data analysis across desktops; HPC and cloud environments.

The Hadoop MapReduce framework has recently evolved to an efficient parallel programming framework for processing large amounts data. In the second part of the talk, we will focus on our recent study of data locality in MapReduce and of the trade-offs brought by executing MapReduce applications in virtualized data centers. In order to leverage data locality Hadoop MapReduce is commonly deployed with collocated data and compute management layers. However, with recent advances in networks and the need to enable elastic MapReduce in the cloud, it is believed that data locality is most likely to become less important in the future. Moreover, while data centers consume tremendous amounts of energy, only a few works have investigated the power profiles of MapReduce applications, a first step towards devising energy saving mechanisms. We will present the results of a performance and energy efficiency evaluation of Hadoop MapReduce with anti-collocated compute and data management layers in two different environments: clusters and clouds.

==========