Scientific Workflows on Clouds with Heterogeneous and Preemptible Instances

Fabio TORDINI [a,1], Marco ALDINUCCI [a], Paolo VIVIANI [a], Ivan MERELLI [b], Pietro LIÒ [c]

[a] *Computer Science Dept, University of Torino, Torino, Italy*
[b] *IBT – CNR, Segrate (MI), Italy*
[c] *Computer Laboratory, University of Cambridge, Cambridge, UK*

Abstract. The cloud environment is increasingly appealing for the HPC community, which has always dealt with complex scientific applications. However, there is still some skepticism about moving from traditional physical infrastructures to virtual HPC clusters. This mistrusting probably originates from some well known factors, including the effective economy of using cloud services, data and software availability, and the longstanding matter of data stewardship. In this work we discuss the design of a framework (based on Mesos) aimed at achieving a cost-effective and efficient usage of *heterogeneous* Processing Elements (PEs) for workflow execution, which supports *hybrid cloud bursting* over preemptible cloud Virtual Machines.

Keywords. HPC, Scientific Workflows, Cloud Engineering, Resource Provisioning

1. Introduction

In the HPC landscape, workflows play a very important role for scientific computing and application coordination, because they provide means to formalize and organize complex scientific processes by supporting the modeling, execution and monitoring of the data analysis process. In this paper we mainly focus on Bioinformatics workflows, commonly referred as *pipelines*. They typically exploit a pure *Data flow* behavior [6].

For scientists to switch to the cloud for their analysis and computations, an important metric to consider is the cost for resource usage, taking into account application's workload and performance requirements. Typically, the cost per unit of time of Infrastructure-as-a-Service (IaaS) grows in a linear-affine fashion with Virtual Machines (VMs) reliability, core count, memory size and storage (in increasing order of weight); VMs executing tasks using only a fraction of their cores are entirely payed. Bioinformatics pipelines are typically long-running and performance-demanding applications. Nowadays, their execution on cloud might be a quite expensive option.

Notwithstanding, the cloud makes it possible to execute a scientific workflow with a close to zero investment in infrastructures, according to the pay-per-use model. Also, the cloud execution model offers a much more dynamic execution model with respect to tra-

[1] Corresponding Author: tordini@di.unito.it