Mr.Cirrus: A Map-Reduce approach for High level Cloud Computing

J. Lago¹; D. Ramet²; O. Torrño³; J. Karlsson⁴; J. Falgueras¹; N. Chelbat⁴ M. Krieger⁵ and O. Trelles²

(1) Fundación Iavante; Consejería de Salud de la Junta de Andalucia, Spain
(2) Dpt. of Computer Architecture; University of Malaga, Spain
(3) Dpt. of Languages and Computer Sciences, University of Malaga, Spain
(4) Institute for Bioinformatics, Johannes Kepler University, Linz, Austria
(5) RISC Software GmbH; Hagenberg, Austria
{ juan.lago@iavante.es; dramet@uma.es; oscart@uma.es; tjkarlsson@uma.es; juanfc@uma.es; chelbat@bioinf.jku.at; michael.krieger@risc-software.at; ortrelles@uma.es }

ABSTRACT

New data acquisition technological breakthroughs have revolution the bioinformatics data analysis era. Advances in high performance computing (HPC) strategies are necessary to effectively tackle the petabyte scale of the so called 'Big Data' problem. Parallel and, in particular, Cloud Computing is one of the most promising technologies to enable such important research. The capacity of Cloud Computing to dynamically adapt the amount of deployed resources (elasticity) to the demand allows institutions to access large computational infrastructures with an important reduction in maintenance and exploitation costs.

In this document we present Mr.Cirrus, a novel software –following the Map-Reduce paradigm [1]- tool aimed to make easy an efficient the exploitation of Cloud resources. The first aspect is covered by the adaptation of our jORCA [2] client software to simplify the registering, discovering, management, invocation and monitoring of services in the cloud. The second aspect affects directly the analysis of bio-sequences at genome level as the cloudification of such paradigmatic applications should allow us to test different tasks scheduling approaches in order to make the exploitation of Cloud resources efficient.

Mr.Cirrus has been benchmarked in several HPC environments such as Amazon EC2, IBM-SCE and MS-Azure and also in multiprocessors and cluster environments for porting legacy software in fast, economical and efficient way.

A comparative study including Hadoop [3] as reference software has been performed. Hadoop is a Map-Reduce framework that has gained popularity and nowadays there are several cloud providers offering Hadoop-based services. This is the case of Amazon Elastic MapReduce and IBM InfoSphere BigInsights cloud services. The results achieved in this study with MrCirrus have been also compared in terms of performance and speedup, in multiprocessors and cluster environments for porting legacy software in fast, economical and efficient way.

References.

Acknowledgments. This work has been partially supported by the Spanish National Institute of Bioinformatics (INB-ISCHI), the RIRAAF network (RD07/0064/0017), grant AT2009-0025 "Acciones Integradas-MICINN” and FP7-VENUS-C projects. We would also like to acknowledge IBM-SCE Austria for their support and access to the IBM SCE platform