

Evaluating Scalability of a Cloud Based Platform for Biological Networks Analysis

Original

Evaluating Scalability of a Cloud Based Platform for Biological Networks Analysis / Bertone, Fabrizio; Caragnano, Giuseppe; Ruiu, Pietro; Terzo, Olivier; Vasciaveo, Alessandro; Benso, Alfredo. - STAMPA. - (2015), pp. 464-468. (4th International Workshop on Hybrid Cloud Computing Infrastructure for E-science Application (HCCIEA) Brazil July 2015) [10.1109/CISIS.2015.68].

Availability:

This version is available at: 11583/2620706 since: 2016-04-26T19:55:16Z

Publisher:

IEEE

Published

DOI:10.1109/CISIS.2015.68

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

Evaluating scalability of a cloud based platform for biological networks analysis

Fabrizio Bertone, Giuseppe Caragnano, Olivier Terzo
Istituto Superiore Mario Boella
Torino, Italy
{familyname}@ismb.it

Alessandro Vasciaveo, Alfredo Benso
Politecnico di Torino
Torino, Italy
{name.familyname}@ismb.it

Abstract — Research centers performing biomedical research collide with the problem of the treatment of large amounts of data. Several scientific fields in biomedical adopt technologies that can analyze samples in a more accurate way thanks to the high granularity which the current equipment provide the results. The cloud computing technology allows to create scalable and flexible infrastructures and data management services. In recent years the number of solutions that can be included within the phenomenon of cloud computing has increased. There are many cases of distributed solutions with high storage and processing capacity and the possibility of serving a large number of users. In this paper authors describe a biological networks modeling tool, implemented with the aid of MapReduce algorithms that works on a cluster in a cloud computing infrastructure.

Keywords — *cloud computing; resource orchestrator; Map Reduce; parallel algorithms*

I. INTRODUCTION

There are many examples of initiatives where computing infrastructures support scientific programs such as the Office of Cyberinfrastructure the US National Science Foundation [1] and the Swiss initiative SwissBioGrid[2]. The amount of data, the requirements in terms of computing power, the need to provide infrastructure for collaborative information sharing and the ability to achieve economies of scale in order to use them on-demand, are all factors that converge towards the possibility of using Cloud Computing solutions within e-Science.

The connections within the scientific communities and application domains are increasing on international scale, introducing a risk of fragmentation of computing infrastructure. This phenomenon raises the question of studying models and service infrastructure that lead to a more flexible and optimized management of computing resources and storage in order to be able to serve different user communities through integrated platforms. Regardless of the application, and geographical contexts can identify some common needs.

First of all, a greater integration of technologies to ensure better management of information with an emphasis on the availability of data (storage) and the possibility to have highly reconfigurable computing resources.

Next, a greater access to distributed computing platforms for the development and implementation of new algorithms or for their adjustment to a very distributed. Another aspect to take in consideration is the availability of a true e-Science cloud highly scalable and flexible in terms of scaling up that scaling down of contributing to the creation of high quality engineered products. The aim of the study is to evaluate the behavior of a Gene Regulatory Network (GNR) model simulation in a cloud computing environment.

II. PLATFORM

A. Overview

Biomedical research dealing with biological networks analysis can be greatly improved with the support of specific computational platforms having adequate characteristics. Some aspects are particularly important in order to increase research proficiency. Great computational power, substantial storage capacity, scalability and resiliency are essential features. Within various available infrastructure models, this study focus on a custom cloud built upon the open source framework Hadoop 2 [3].

B. Architecture

In our test environment, the platform is completely composed by VMs (virtual machines) forming a private cloud. The cloud management is supported by OpenNebula [4]. The computational and storage environment is provided by a standard Hadoop cluster, enhanced with biomedical tools.

In order to simplify the cluster management and periodical upgrading of the included software, a unique customized image is used to instantiate each VM. The role and available resources -mainly vCPUs and RAM- of each VM can then be differentiate using different templates and configurations. Three main types of VMs are used inside the cluster, plus other single-purpose machines with marginal non-essential roles. The first and most numerous type is the Hadoop worker.

This class of VMs is used to provide all the computing and storage capabilities of the cluster. A second type of VM, called YARN Master, is used to establish and manage computational resources within the cluster. A third type of VM, called HDFS Master, manages storage functionalities.

C. Function of components

Platform users -researchers-, initially configure simulation parameters using a custom *GUI*, which act as an interface and abstraction layer to the cloud platform. An custom *agent*, placed inside the YARN master VM, periodically check if new simulation request were submitted, keeps and manages a requests queue, eventually ordering them by priority, and tracks jobs in execution. When a new job should be submitted to Hadoop, the agent send a formatted and parameterized request to the YARN Master. The YARN Master, considering available resources inside the cluster -i.e. free workers-, begins and coordinates the distributed simulation process. When all the simulations are successfully completed, a second processing stage begins. The MapReduce algorithm, once more executed by the workers, is used to aggregate all simulations in a single output and eliminate redundancy introduced within parallel simulations.

Resulting data is then stored in the HDFS distributed file system and made available to the user. During all process stages, the agent probes the Masters in order to track status and update the user interface.

III. SCALABILITY PLATFORM TESTING

Scalability parameters can be assessed in order to evaluate and model different configurations efficiency. A model could then be used in order to optimize the cluster configuration and exploit available resources at best. A first test is held executing a fixed number of simulations inside clusters of increasing sizes, while using VMs with minimal resources. Next, the same tests are held using clusters composed by VMs of increasing resources.

IV. REFERENCES

- [1] <http://orci.research.umich.edu/>
- [2] http://people.isb-sib.ch/Heinz.Stockinger/publications/podvinec_SwissBioGrid.pdf
- [3] <https://hadoop.apache.org/>
- [4] <http://opennebula.org/>