Future Generation Computer Systems 78 (2018) 379–391

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A resource provisioning framework for bioinformatics applications in multi-cloud environments

Izzet F. Senturk [a], P. Balakrishnan [b], Anas Abu-Doleh [a,*], Kamer Kaya [a,c], Qutaibah Malluhi [b], Ümit V. Çatalyürek [a]

[a] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210, United States
[b] KINDI Center for Computing Research, Qatar University, Doha, Qatar
[c] Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey

HIGHLIGHTS

- A cloud broker framework is proposed for bioinformatics applications.
- The framework simplifies extending local resources to a multi-cloud environment.
- Simultaneous use of multiple computing resources from cloud and local clusters is enabled.
- Workflow improvement mechanism enhances submitted abstract workflows by exploiting parallelism.
- Scheduling algorithm decreases the workflow execution time for a given budget.

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form
16 May 2016
Accepted 12 June 2016
Available online 4 July 2016

Keywords:
Cloud computing
Cloud broker
Interoperability
Multi-cloud
Bioinformatics

ABSTRACT

The significant advancement in Next Generation Sequencing (NGS) have enabled the generation of several gigabytes of raw data in a single sequencing run. This amount of raw data introduces new scalability challenges in processing, storing and analyzing it, which cannot be solved using a single workstation, the only resource available for the majority of biological scientists, in a reasonable amount of time. These scalability challenges can be complemented by provisioning computational and storage resources using Cloud Computing in a cost-effective manner. There are multiple cloud providers offering cloud resources as a utility within various business models, service levels and functionalities. However, the lack of standards in cloud computing leads to interoperability issues among the providers rendering the selected one unalterable. Furthermore, even a single provider offers multiple configurations to choose from. Therefore, it is essential to develop a decision making system that facilitates the selection of the suitable cloud provider and configuration together with the capability to switch among multiple providers in an efficient and transparent manner. In this paper, we propose BioCloud as a single point of entry to a multi-cloud environment for non-computer savvy bio-researchers. We discuss the architecture and components of BioCloud and present the scheduling algorithm employed in BioCloud. Experiments with different use-cases and scenarios reveal that BioCloud can decrease the workflow execution time for a given budget while encapsulating the complexity of resource management in multiple cloud providers.

1. Introduction

The last decade have witnessed rapid advances in the field of genomics thanks to the evolution of the genome sequencing technologies which lead to accelerated generation of digital biological information in unprecedented amounts. The emergence of high-throughput NGS has revolutionized genomics research by providing an astounding cost reduction making the whole genome sequencing possible for as low as $1000 [1] and hence making the technology pervasive. The availability of NGS on a wider scale with its decreased cost and high-throughput have paved the way for more complex NGS data at a rate outpacing the advances in computation and storage capacities [2].

[*] Corresponding author.
E-mail addresses: ifs5@cornell.edu (I.F. Senturk), abudoleh.1@osu.edu (A. Abu-Doleh), kaya@sabanciuniv.edu (K. Kaya), qmalluhi@qu.edu.qa (Q. Malluhi), umit@bmi.osu.edu (Ü.V. Çatalyürek).

http://dx.doi.org/10.1016/j.future.2016.06.008
0167-739X/© 2016 Elsevier B.V. All rights reserved.

Minimizing the impact of the increased data complexity requires scalable solutions for storing and analyzing massive NGS data. The scalability issues of NGS drives the efforts to cloud computing, which is converging as a frontier to address this class of problems by enabling large scale computing resources on demand, tailored to specific requirements in a pay-per-use manner [2]. Cloud computing renders maintaining large clusters unnecessary while handling peak-time loads and addressing issues such as availability, load balancing and fault tolerance.

Cloud providers tend to offer resources through their custom APIs which restrict the development of the tools with respect to the vendor specific API. In the long term, customers are restricted to the vendor and cannot migrate from one cloud provider to another seamlessly. The proposed BioCloud[1] employs a Multi-Cloud [6] model and acts as a *broker* across the resources of multiple cloud providers. Considering the vast number of hardware profiles available for selection in cloud providers, the complexity of determining the hardware profiles to be used and their quantity can be overwhelming not to mention the complexity of resource provisioning and configuration. Note that there are 38 current generation "instance types" in EC2 [7] and 19 "flavors" in Rackspace [8] available to choose. Some of these hardware profiles are optimized for memory, CPU, storage, etc. BioCloud analyzes workflow steps and evaluates hardware profiles in available resources while considering user requirements such as deadline, budget, etc. in order to determine the type and number of resources to be used for each of the workflow steps individually. BioCloud ensures the availability of resources for workflow execution by provisioning resources in such a way that the resources are neither wasted nor additional delay occurred due to the waiting time for resource initialization (i.e., booting the resource, dynamic cluster configuration on the cloud, etc.). This requires a scheduling algorithm which considers several resource options (hardware profiles in cloud providers, possibility of configuring a cluster in the cloud and determining the number of compute nodes to be used) and the possibility of exploiting parallelism which is the main focus of this paper. Scheduling may not yield the best solution unless an accurate estimation for the running times cannot be attained. BioCloud exploits its profiler to keep the execution times of the tools on the given resources considering the size of the input and output files. This enables a means to estimate the execution time of the workflow steps. Scheduler employs resource manager to ensure availability of the resources before the workflow steps are dispatched for execution. Scheduler also evaluates the workflow steps for parallelism and modifies workflows to enable parallelism if possible. Scheduler cooperates with the workflow manager and takes care of the required manipulations on the data and tool settings.

BioCloud encapsulates all the complexity of resource management and provides a single entry point to create custom workflows and run them in a simple and efficient manner through its user-friendly web-interface. The public virtual machine (VM) image we provide [5] can be employed to start a BioCloud instance. We assume BioCloud users have an existing account in at least one of the cloud providers. In order to start using BioCloud, users create a BioCloud account through the web-interface, and complete their profiles by providing the available resources to be used. The resources can be cloud account(s), local clusters, servers, and datasets. Then a BioCloud instance is started on the cloud using the provided cloud credentials. Once the instance is initialized, workflow manager interface (Galaxy [9]) is presented to the user which runs on one of the resources provided by the user. The workflows created by the user are executed over the computational resources defined earlier. If multiple computational resources are available, jobs in a multi-step workflow can be run on different resources based on the scheduling algorithm and the user requirements. BioCloud strives to exploit parallelism to reduce the overall workflow execution time by running parallel steps using different computing resources or dividing a single step into multiple parallel steps by partitioning the input data and computation, whenever possible.

BioCloud offers a loosely coupled architecture through its service oriented architecture. BioCloud Portal web-service is employed to expose some of the functionalities of the system so that some of the workflow decisions (i.e., when to dispatch a workflow step and where to run this step) are delegated to the BioCloud Portal web-service. This enables modularity where scheduling logic is separated from the core workflow system. This provides the flexibility of updating the scheduling algorithm and other features of the system (i.e., improving abstract workflows submitted by the user and presenting the new workflow for execution) without requiring a software update on the user side.

The rest of the paper is organized as follows. Section 2 compares the features of the proposed work with notable studies from the literature. Section 3 details the proposed BioCloud architecture. The proposed scheduling algorithm is discussed in Section 4. Section 5 demonstrates the features of the proposed system by evaluating BioCloud using two real-life use-cases. Finally, the concluding remarks are given in Section 6.

2. Related work

The vast amount of data generated by NGS platforms poses a challenge to store, access and manipulate data in an efficient manner within a reasonable amount of time. A single workstation is often not sufficient to complete the analysis in a reasonable amount of time and organizations need to own and maintain specific type of hardware to handle operations in such scale. To remedy the problem, some efforts have focused on parallelizing existing tools using various distributed memory parallelism schemes, such as MPI (Message Passing Interface) [10,11] or MapReduce [12,13]. However, both approaches require dealing with complex software frameworks and hence require experienced developers for efficient parallelization, and also experienced users to use developed applications.

In the rest of this section, we discuss various frameworks commonly used in bioinformatics based on the underlying infrastructure they support.

2.1. Web-based frameworks

Many frameworks, such as Grendel [14], provides a web service based architecture to access high performance computing (HPC) resources. Web services can be invoked remotely so that the functionality of the deployed tools can be exposed on the network without interoperability concerns. However, computational resources are limited with the maintained HPC resources. MG-RAST [15] is an open source platform specialized in metagenome analysis. Users can analyze their data through the offered analysis pipeline. The jobs and the data made public by the user are stored in the system indefinitely which makes MG-RAST a repository for metagenomic data.

Anvaya [16] provides a platform to conduct automated genome analysis by interfacing with several bioinformatics tools and databases. BioExtract server [17] enables the researchers to apply the analytical tools over the data extracted from several

[1] The term BioCloud is also used by the Beijing Institute of Genomics [3] to denote their bioinformatics cloud system. Recently, the term BioCloud has been used as a general term to indicate the cloud-based bioinformatics applications [4]. In this paper, the term BioCloud denotes our multi-cloud bioinformatics framework [5].