BioCloud: Resource Provisioning Framework for Bioinformatics Applications in a Multi-Cloud Environment

Izzet F. Senturk[1], Bala Krishnan[2], Kamer Kaya[1,3], Qutaibah Malluhi[2], Umit V. Catalyurek[1]
[1]Dept. Biomedical Informatics, The Ohio State University, Columbus, OH, 43230
[2]KINDI Center for Computing Research, Qatar University, Doha, Qatar
[3]Computer Science & Engineering, Sabanci University, Istanbul, Turkey

1- Introduction

Next Generation Sequencing (NGS) has paved the way to extract genetic information from biological systems in a massively parallel fashion. It also made the technology pervasive by decreasing the sequencing cost significantly. The availability of NGS in a wider scale due to decreased sequencing costs has shifted the challenge from obtaining sequencing data to analyzing it which requires addressing scalability issues. Especially for small-scale labs, the lack of dedicated hardware that is required to complete the analysis within a reasonable amount of time drives the efforts to cloud computing.

2- Motivation

Even though the Cloud helps the researchers to avoid the high ownership and maintenance costs of such dedicated hardware, the complexity of configuring the analysis services that can dynamically scale on-demand can be a barrier for non-computer savvy bio-researchers. BioCloud provides a ready-to-use and efficient analysis environment to the bio-researchers so that they can focus solely on the analysis, not on the complexities of the computational environment.

BioCloud can handle peak loads by provisioning additional resources and configuring them in such a way that the resource utilization can be maximized. The core competency of BioCloud is the ability of managing resources from multiple cloud providers and local clusters so that these resources are utilized simultaneously in a seamless and efficient manner. There are projects that provide Cloud support for NGS applications, such as Globus Genomics [1], but they are designed to work with single Cloud provider.

Cloud providers tend to develop custom APIs which complicates the interoperability of the tools to access their services. These custom APIs also make it difficult to migrate from one cloud provider to another or to use multiple providers simultaneously. BioCloud saves the researchers the trouble and enables a hybrid multi-cloud model without any interoperability complications.

3- The BioCloud Framework

BioCloud provides a single entry point to conduct bioinfomatics analysis on a multi-cloud environment. BioCloud leverages DeltaCloud [2] for resource provisioning and Galaxy [3] for abstract workflow development and execution. Through its user-friendly web-interface, researchers can define their resources and specify their time and budget constraints. Based on this information, BioCloud designates a schedule for the submitted workflow and determines the target resources for execution. It also ensures the resource availability before executing a particular workflow step. With the features it provides, BioCloud can schedule different workflow steps on different cloud environments while scaling the resources on-demand.

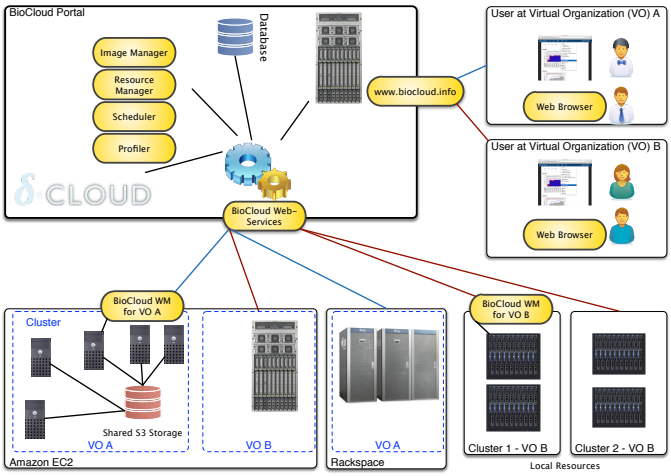

Fig. 1: Layered architecture of BioCloud

BioCloud (Fig. 1) consists of two main components: BioCloud Portal and BioCloud Workflow Managers (WM). The portal is BioCloud's single access point of entry for all users (virtual organizations) where they can manage their resources. On the other hand, WMs are the management components of the system specific to a particular virtual organization (VO).

BioCloud offers a loosely coupled architecture in which some of the decisions (e.g., when to dispatch a workflow step and where to run this step) are delegated to BioCloud Portal so that the scheduling logic is separated from the core workflow system. This provides the flexibility of updating the scheduling algorithm without requiring a

software update on the client side.

BIOCLOUD strives to exploit parallelism and reduce the workflow execution time by running parallel steps on different resources. When possible, it also divides a single step into multiple parallel steps by partitioning the data and the associated computation to further improve the response time. Once the workflow is submitted for execution, BIOCLOUD partitions the data to match the computation with available resources.

Galaxy also provides partitioning capability for the tools. However, it cannot fully optimize the workflow execution where two (or more) steps can run on the partitioned data. In such cases, Galaxy redundantly merges and partitions the data between the consecutive steps. As illustrated in Fig. 2, the partitioning scheme we employed in our scheduler postpones the merging step until it is really required.

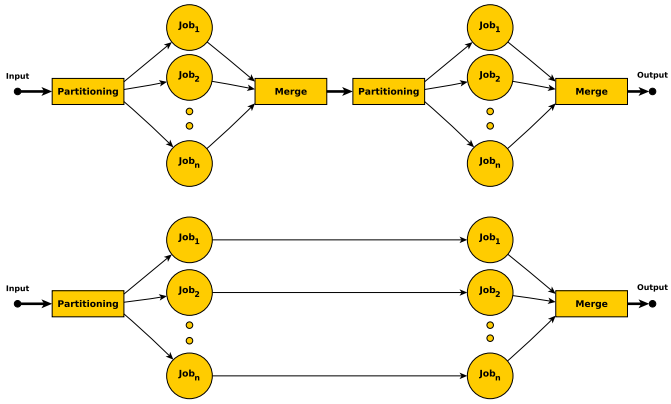

Fig. 2: Partitioning scheme in BIOCLOUD's scheduler avoids redundant merging and partitioning.

During scheduling, BIOCLOUD evaluates whether the partitioning can be applied to individual workflow steps in order to optimize the execution while increasing resource utilization. Regarding the submitted workflow as an "abstract workflow", an optimized workflow is generated that utilizes the available resources, and hence enable parallelism. For example, for the user-designed abstract EXOMESEQ workflow [4] (Fig. 3), BIOCLOUD offers an improved version as given in Fig. 4.

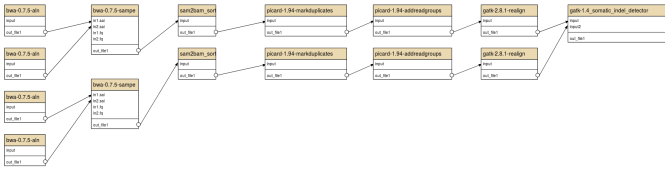

Fig. 3: Abstract EXOMESEQ workflow for BIOCLOUD.

4- Preliminary Results

We run the EXOMESEQ workflow sequentially on a single workstation and parallel on multiple cloud resources. The results are provided in Fig. 5. The total execution time is reduced by half by utilizing different computing environments to exploit parallelism. Thanks to BIOCLOUD's data

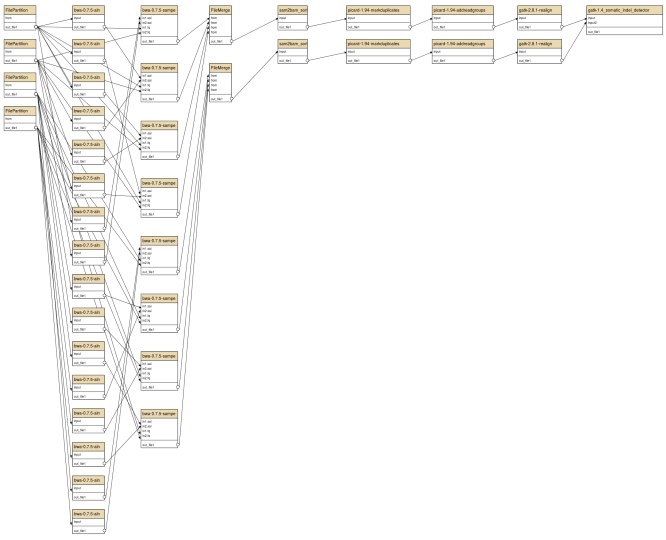

Fig. 4: Generated workflow

partitioning, a higher level of scalability is attained especially for the first two steps on the multi-cloud environment. Depending on the step dependencies (task or data), the data flow, and the nature of the tools used, total execution time can be reduced even further through simultaneous execution of the steps on different computing resources.

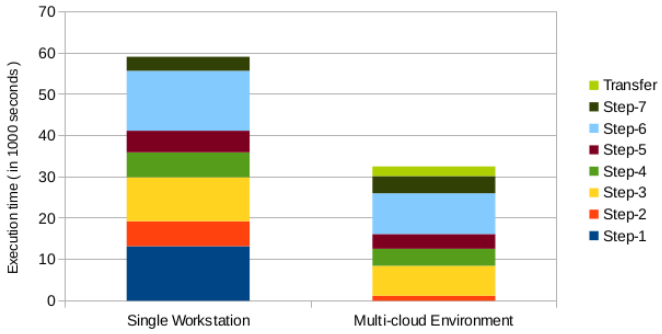

Fig. 5: Execution times of the EXOMESEQ workflow on a single workstation and multi-cloud environment

5- Research Needs

We need configurable cloud environments in order to develop and test our scheduling techniques in a controlled environment and enhance BIOCLOUD via new features and improvements. These resources will be also useful for testing multi-site scheduling techniques.

REFERENCES

[1] "Globus Genomics." [Online]. Available: https://www.globus.org/

[2] "Deltacloud." [Online]. Available: http://deltacloud.apache.org/

[3] J. Goecks *et al.*, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, pp. R86+, Aug. 2010.

[4] J. A. Woyach *et al.*, "Resistance mechanisms for the bruton's tyrosine kinase inhibitor ibrutinib," *New England Journal of Medicine*, vol. 370, no. 24, pp. 2286–2294, 2014.