Chapter 8
Scheduling Data Intensive Scientific Workflows in Cloud Environment Using Nature Inspired Algorithms

Shikha Mehta
Jaypee Institute of Information Technology, India

Parmeet Kaur
Jaypee Institute of Information Technology, India

ABSTRACT

Workflows are a commonly used model to describe applications consisting of computational tasks with data or control flow dependencies. They are used in domains of bioinformatics, astronomy, physics, etc., for data-driven scientific applications. Execution of data-intensive workflow applications in a reasonable amount of time demands a high-performance computing environment. Cloud computing is a way of purchasing computing resources on demand through virtualization technologies. It provides the infrastructure to build and run workflow applications, which is called 'Infrastructure as a Service.' However, it is necessary to schedule workflows on cloud in a way that reduces the cost of leasing resources. Scheduling tasks on resources is a NP hard problem and using meta-heuristic algorithms is an obvious choice for the same. This chapter presents application of nature-inspired algorithms: particle swarm optimization, shuffled frog leaping algorithm and grey wolf optimization algorithm to the workflow scheduling problem on the cloud. Simulation results prove the efficacy of the suggested algorithms.

INTRODUCTION

This chapter presents a study of scheduling data-intensive scientific workflows in IaaS clouds using nature inspired algorithms. Workflows are a commonly used computational model to perform scientific simulations (Juve et al., 2013). These models are mostly used to visualize and manage the computations as well as activities happening in scientific processes. They are employed to illustrate the applications

DOI: 10.4018/978-1-5225-5852-1.ch008

involving a series of computational tasks with data- or control-flow reliance among themselves. They are used to represent complex scientific problems prevailing in diverse fields such as Bioinformatics, Physics, weather data analysis and modelling, structural chemistry etc (Juve et al., 2013). However, scientific applications of this nature are, in general, data-driven, and use files to communicate data between tasks. The data and computing requirements of such these applications are ever-growing which are further featured by complex structures and entail heterogeneous services. Executing such data intensive workflow applications pose numerous challenges such as quality of service (QoS), scalability, data storage, computing resources along with heterogeneous and distributed data management (Juve & Deelman, 2011). Therefore, these demand a high-end computing environment in order to complete the task in a considerable duration. As the scientific data is growing at a pace faster than ever, it is no longer going to be feasible to transfer data from data centres to desktops for analysis. In contrast, processing will time and again take place on high-performance systems with local storage of data.

Cloud computing is basically a on demand method to purchase computing as well as storage resources via virtualization technologies (Buyya et al., 2009). Such services are available most prominently as per three models, namely, Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). Internet is the backbone for all the cloud computing models. Software as a Service (SaaS) model of cloud computing allows users to access providers applications on a client's system without being bothered about the administration of the services which is being done by the vendor itself. SaaS is commonly used for providing cloud applications such as Web-based e-mail, social networking websites; online document editors etc. The second category of cloud model is the Platform as a Service (PaaS) that provides frameworks for use in development or customization of applications. This makes possible fast and cost-efficient application development, testing, and deployment. The last category of popular cloud model is Infrastructure as a Service (IaaS) that offers a heterogeneous collection of resources from which users can lease resources according to their requirements.

This chapter lays focus on task scheduling and resource provisioning particularly in Infrastructure as a Service (IaaS) clouds. Due to the availability of unlimited and diverse types of resources, IaaS cloud models are more appropriate for applications involving scientific workflows. IaaS provides a large shared pool of heterogeneous resources or virtual machines (VMs) to execute computationally expensive workflow applications. Cloud computing is a computing paradigm that is apt to deal with most of the challenges listed above. It provides a technique of acquiring compute and storage resources according to a user's requirement through virtualization technologies. Virtualization technology of clouds enables easy deployment, management and execution of workflow applications in clouds. This is the result of the benefits presented by virtualization such as migration of code and data, fault tolerance, process isolation and customization of services for users. This has made it possible for cloud computing platforms to allocate virtual machines dynamically as Internet services (for e.g. Amazon EC2/S3).

However, cloud services come at a pay-per-use basis and hence, it is necessary to schedule workflow applications in a way that reduces the total cost of leasing resources for execution, besides other possible criteria. Scheduling tasks on resources is a NP hard problem (Ullman, 1975; Lin &Lu, 2011). Therefore, using meta-heuristic algorithms is an obvious choice for the same. This chapter will present the application of nature-inspired algorithms, namely particle swarm optimization, shuffled frog leaping algorithm and bat algorithm to the scheduling problem for workflows on cloud (Zhan et al., 2015). The next section of the chapter will introduce the scheduling problem. Subsequently, the nature inspired algorithms and their application to the current problem will be explained in brief. Next, the metrics that can be used to evaluate computational performance of workflows in cloud environment will be discussed. This will be