Combination of Scheduling and Dynamic Data Replication for Cloud Computing Workflows

Kouidri Siham, Department of Computer Science, Faculty of Exact and Applied Sciences, University Oran1 Ahmed Ben Bella, Oran, Algeria
Yagoubi Belabbas, Department of Computer Science, Faculty of Exact and Applied Sciences, University Oran1 Ahmed Ben Bella, Oran, Algeria

ABSTRACT

Cloud computing is a powerful and high-capacity system, because it can satisfy various demands and share resources for users. It also to benefits from a capacity for treatment and unlimited storage. However, it is burdensome for the providers of internet services that the user demands are increasing as computer capacity is growing stronger and stronger. Therefore, the techniques of workflow scheduling and data replication are used to decrease the costs of the data intensive application. Unfortunately, these two approaches, which are very complementary, are used separately. In this article, a combination of workflow scheduling based on the clustering of data and dynamic data replication strategies has been introduced together. A Cloud simulator, Cloudsim, is used to evaluate the performance of the proposed algorithm. Simulation results show the effectiveness of the proposed algorithm in comparison with well-known algorithms such as random data placement and the Build Time algorithm.

KEYWORDS
Cloud Computing, CloudSim, Clustering, Data Placement, Data Replication, Scientific Workflow, Tasks Scheduling, Virtual Machine

1. INTRODUCTION

Cloud computing is heavily based on a more traditional technology: Grid Computing, which has been researched for more than 20 years. Cloud computing focuses on the sharing of information and computation in a large network of nodes, which are quite likely to be owned by different vendors/companies (Goyal & Agrawal, 2013). It is believed that cloud computing has been one of the sources for success in several major companies such as Google and Amazon. It delivers infrastructure, platform, and software that are made available as subscription-based services in a pay-as-you-go model to consumers as shown in Figure 1. Further, the provided abstract, virtual resources, such as networks, servers, storage, applications and data, can be delivered as a service rather than a product. These services are referred to as Infrastructure as a Service (IaaS), this model encompasses the storage

DOI: 10.4018/IJIRR.2019100103

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
as a service, computation resource as a service, and communication resource as a service. Example for this kind of service is: Amazon-S3 for storage, Amazon-EC2 for computation resources, and Amazon-SQS for communication resources. Platform as a Service (PaaS) the PaaS model provides the user to deploy user-built applications on top of the cloud infrastructure, that are built using the programming languages and software tools supported by the provider (e.g., Java, python, .Net), and Software as a Service (SaaS) this model provides the software applications as a service (Jang, Kim, & Lee, 2012).

In general, Cloud computing systems require the management of a massive number of data sets to make the data available to the many clients accessing the data. Data is typically replicated in these large-scale applications to improve the data availability and job response time. Because the behavior of the thousands of cloud users is very dynamic, it is a challenging problem to determine where and when to make data replications to increase the system availability, in which millions of files will be generated from these scientific experiments and thousands of clients world-wide will access these data. This huge volume of data sets requires new strategies to determine how to make the data more available (Goyal & Agrawal, 2013). Replication of data is typically viewed as the solution to improve file access time and reliability. Replication is a frequently used technique in the cloud, such as GFS (Google file system) (Ghemawat, Gobioff & Leung, 2003), HDFS (Hadoop Distributed File System) (Shvachko, Hairong, Radia & Chansler, 2010). They provide reliable storage and access to large scale data by parallel applications. The number of data replicas is usually less than 3, in order to meet the high availability, high fault tolerance and high efficiency requirement, it is necessary to dynamically adjust the popular data files, the number of data replicas and the sites to place the new replicas according to the current cloud storage (Vignesh, Kumar & Jaisankar, 2013).

In scientific workflow applications data plays an important role. The jobs submitted by the users in these applications require huge input data sets distributed geographically and transferring these large-sized data takes tremendous amount of time. Scheduling and Replication are two well-known techniques to boost the performance of cloud computing. In the literature, most of these techniques are focusing on job scheduling or data replication separately.

In this work we propose to integrating task scheduling and data replication into one framework for the sole goal of minimizing the total workflow execution time and data displacement number of the cloud computing.

The rest of this paper is structured as follows: Section 2 explains some related works. Section 3 presents the adopted system model. Section 4 introduces our proposed approach. Section 5 evaluates the performance of simulation experiments using CloudSim. Conclusion and future works are presented in Section 6.
Personalized Content-Based Image Retrieval
www.igi-global.com/chapter/personalized-content-based-image-retrieval/28074?camid=4v1a