Dynamic Data Replication Based on Tasks scheduling for Cloud Computing Environment

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

ABSTRACT
Cloud computing provides IT resources (e.g., CPU, memory, network, storage, etc.) based on virtualization concepts and a pay-as-you-go principle. It comprises an accumulation of inter-related plus virtualized calculating resources which are managed by one or more amalgamated calculating resources. With the development of a computerized scientific workflow, the amount of data is increasing exponentially. Workflow scheduling and data replication have been considered the major challenges in cloud computing. Nevertheless, many researchers focus on scheduling or data replication separately. In this article, a combination of workflow scheduling based on the clustering of data and dynamic data replication strategies, has been introduced together and evaluates several performance metrics using a CloudSim simulator. The aim of this proposed algorithm is to minimize the completion time and transfer time. The performance of this proposed algorithm has been evaluated using the CloudSim toolkit.

KEYWORDS
Cloud Computing, Clustering, Dynamic Replication, Scheduling, Scientific Workflow, Virtual Machines

1. INTRODUCTION
Due to the development of virtualization and Internet technologies, Cloud Computing has emerged as a new computing platform. Cloud computing provides high performance computing resources and mass storage resources that are used in large scale scientific applications such as high energy physics, bioinformatics, climate modeling (Jang, Kim, Kim, & Lee, 2012).

Cloud computing provides one or more consolidated IT resources based on SLA between service providers and service consumers (Goyal & Agrawal, 2013). The data requirements for these scientific applications have been growing at an unprecedented rate in both volume and scale with huge input data sets. 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, many techniques are there, which trying to reduce the job execution time for high performance and good throughput. From these techniques job scheduling and data replication algorithms. When these techniques are implemented in cloud Environment, give different results, and many researchers day by day are

DOI: 10.4018/IJSITA.2017100104
creating new techniques for that. In contrast, most of these techniques are focusing on job scheduling or data replication separately.

A good scheduling strategy will allow shortest access to the required data, therefore reduce the data access time. Vice versa, replication strategy that allows place data in a wisely manner will offer a faster access to files require by cloud jobs, hence increase the job executions performance. This technique involves creating multiple copies of data in multiple sources to reduce access time and bandwidth consumption. It also guarantees data reliability and system load balancing (Mansouri, 2014).

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

The rest of this paper is structured as follows: Section 2 explains some related works. Section 3 introduces our proposed approach. Section 4 evaluates the performance of simulation experiments using CloudSim. Conclusion and future works are presented in section5.

2. RELATED WORK

To cover related literature, this section is divided into two groups. The first group discusses schedule strategies, the second group of related work discusses The replication strategies. Scheduling of tasks is considered a critical issue in the Cloud computing environment, A comparative study of task scheduling algorithms on the Cloud computing environment has been done in (Vignesh, Kumar & Jaisankar, 2013):

- **Round Robin:** It is the simplest algorithm that uses the concept of time quantum or slices. Here the time is divided into multiple slices and each node is given a particular time quantum or time interval and in this quantum, the node will perform its operations.
- **Preemptive Priority:** Priority of jobs is an important issue in scheduling because some jobs should be serviced earlier than other those jobs cannot stay for a long time in a system. A suitable job scheduling algorithm must be considered priority of jobs.
- **The Shortest Job First (SJF):** An SJF algorithm is simply a priority algorithm where the priority is the inverse of the next CPU burst. That is, the longer the CPU burst, the lower the priority and vice versa.

Yuan and all (Yuan, Yang, Liu, & Chen, 2010) in their paper propose a clustering strategy matrix k-means based on data placement for scientific applications in the cloud, this paper contains two strategy (Build–Time stage and Runtime stage), that group the existing datasets in k data centers during the workflow in the first stage and dynamically clusters newly generated datasets to the most appropriate data centers in the second stage.

The replication strategies can be classified according to whether the strategy is static or dynamic. The static replication strategy can be applied as it creates and manages the replicas manually (Tang, Lee, Tang, & Yeo, 2006) (i.e., a replica remains in the system until it is deleted by users or its lifetime has expired (Venugopal, Buyya, & Ramamohanarao, 2006)). In other words, static replication will not change with the changes in user behavior.

In (Wang, Yan, & Wang, 2010), the Google File System (GFS) is a scalable distributed File system designed for large data intensive applications. Files in GFS are divided into fixed size bloc, the system GFS use a static distributed cloud data replication algorithm. In the Hadoop distributed file system (Chang & Chang, 2008), an application can specify the number of replicas for each file,
Related Content

Using Semantics to Discover Web Services Based on Partial Data: An Update of Previous Research
[www.igi-global.com/article/using-semantics-to-discover-web-services-based-on-partial-data/103866?camid=4v1a](www.igi-global.com/article/using-semantics-to-discover-web-services-based-on-partial-data/103866?camid=4v1a)

Innovation, Information Systems Strategic Alignment and Business Value
[www.igi-global.com/article/innovation-information-systems-strategic-alignment/43612?camid=4v1a](www.igi-global.com/article/innovation-information-systems-strategic-alignment/43612?camid=4v1a)
The Future of Supply Chain Management: Shifting from Logistics Driven to a Customer Driven Model
www.igi-global.com/chapter/future-supply-chain-management/28705?camid=4v1a

Decision Making by Emergency Room Physicians and Residents: Implications for the Design of Clinical Decision Support Systems
www.igi-global.com/chapter/decision-making-emergency-room-physicians/36770?camid=4v1a