Chapter 11

Fairness-Aware Task Allocation for Heterogeneous Multi-Cloud Systems

Sanjaya Kumar Panda  
Indian School of Mines Dhanbad, India & VSS University of Technology Burla, India

Roshni Pradhan  
VSS University of Technology Burla, India

Benazir Neha  
VSS University of Technology Burla, India

Sujaya Kumar Sathua  
VSS University of Technology Burla, India

ABSTRACT

Cloud computing is rapidly growing for its on-demand services over the Internet. The customers can use these services by placing the requirements in the form of leases. In IaaS cloud, the customer submits the leases in one of the form, namely advance reservation (AR) and best effort (BE). The AR lease has higher priority over the BE lease. Hence, it can preempt the BE lease. It results in starvation among the BE leases and is unfair to the BE leases. In this chapter, the authors present fairness-aware task allocation (FATA) algorithm for heterogeneous multi-cloud systems, which aims to provide fairness among AR and BE leases. We have performed rigorous experiments on some benchmark and synthetic datasets. The performance is measured in terms of two metrics, namely makespan and average cloud utilization. The experimental result shows the superiority of the proposed algorithm over the existing algorithm.

INTRODUCTION

The popularity of Cloud Computing is rapidly growing because of its advancement in virtualization technology (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009; Li et al., 2012; Huang et al., 2013). This technology facilitates the cloud service providers (CSPs) to create the virtual machines (VMs) instantly (Desmarais, 2013, pp. 18-24). Moreover, the VMs are deployed in a data center to access a service (Sotomayor, Montero, Llorente, & Foster, 2008). Thus, the CSP provides on-demand service over the internet. These services are delivered to customers on pay-per-use basis. But these services are
limited to the capacity of a data center. In order to handle the customer’s demands, some of the services are transferred to another data center (Forell, Milojicic, & Talwar, 2011). In Infrastructure-as-a-Service (IaaS) cloud, the services are demanded by customers in the form of leases. The lease may be submitted in one of the form, namely Advance Reservation (AR) or Best Effort (BE) (Sotomayor, Keahey, & Foster, 2006; Sotomayor, Montero, Llorente, & Foster, 2009; Li et al., 2012; Nathani, Chaudhary, & Somani, 2012).

AR lease is represented by a 3-tuple as follows.

\(<S, E, N>\)

where

\(S = \text{Start time of a lease,} \)
\(E = \text{Execution time of a lease (Note that, End time (ET) = S + E), and} \)
\(N = \text{No preemption. (Note that, AR lease is non-preemptive in nature.)} \)

In contrary, BE lease is also represented by a 3-tuple as follows.

\(<E, ST, P>\)

where

\(E = \text{Execution time of a lease,} \)
\(ST = \text{Starvation time of a lease (i.e., how long a BE lease can wait), and} \)
\(P = \text{Preemption. (Note that, BE lease is preemptive in nature (Sotomayor, 2008)).} \)

AR lease has higher priority over BE lease. Hence, AR lease can preempt BE lease on its arrival. Moreover, if the AR lease is remaining or arriving in the cloud systems, the BE lease cannot be executed. Thus, it leads to starvation among BE leases. Therefore, a significant problem is to schedule the AR and BE leases such that starvation among the leases is reduced up to some extent. This problem is referred as task (or lease) allocation problem in heterogeneous multi-cloud environment, which is not studied in the recent literatures. The primary objective of this problem is to find the execution order of the leases so that the starvation is reduced.

In this chapter, we address the above problem for heterogeneous multi-cloud environment and propose a novel algorithm Fairness-Aware Task Allocation (FATA). The algorithm allows few BE leases to schedule before AR leases in a regular interval. However, the interval depends on the arrival rate of AR and BE leases respectively. For example at time instance \(t = 0\), eighty AR leases and twenty BE leases are arrived. If there is no fairness, it schedules eighty AR leases followed by twenty BE leases which leads to starvation among BE leases. From here onwards we will refer the no fairness (i.e., stated above) as existing algorithm. We also refer lease, workload and task interchangeably in this chapter. In order to overcome the starvation problem, we introduce FATA algorithm. We perform extensive experiments on the proposed algorithm using synthetic and benchmark data sets. The experimental results clearly