Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Reward-based Markov chain analysis adaptive global resource management for inter-cloud computing☆

Ben-Jye Chang [a,*], Yu-Wei Lee [a], Ying-Hsin Liang [b]

[a] Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Taiwan, ROC
[b] Department of Multimedia Animation and Application, Nan Kai University Technology, Taiwan, ROC

HIGHLIGHTS

- Adopting the large-scale and small-scale traffic predictions.
- Based on the Markov chain model to analyze the service blocking and the required number of VMs for each request.
- Maximizing the net profit of the cloud provider.

ARTICLE INFO

Article history:
Received 10 April 2017
Received in revised form 29 July 2017
Accepted 20 September 2017
Available online xxxx

Keywords:
Cloud computing
Adaptive cloud resource management
Markov chain model analysis
The large-scale and small-scale traffic predictions
VM migration
Task redirection
Resource over-sale policy

ABSTRACT

The cloud IaaS provider supports diverse services for users to access big data of the real-time entertainment or the non-real-time working traffic. The IaaS provider builds data centers that include different types cloud resources/equipment, e.g., physical machines, virtual machines, networking, storages, power equipment, etc., and significantly increases cloud cost. An efficient cloud resource management is required for the cloud provider to maximize system reward while satisfying the QoS of various SLAs. This paper proposes a Reward-based adaptive global Cloud Resource Management (RCRM) that consists of three main contributions: the Large-scale and Small-scale traffic Predictions (LSP), Adaptive Cloud resource Allocation, and Maximum Net Profit. The M/M/m/m Markov chain model analyzes the service blocking and the required number of VMs for each request. For maximizing the system net profit, the cloud providers always oversell cloud resources. However, the cost of deploying data centers at different areas in the world is different. This paper adopts the VM migration-in/migration-out and task redirection to adaptively allocate cloud resources among global data centers. Numerical results demonstrate RCRM outperforms the others in dropping probability, SLA violation, violation penalty and net profit. Furthermore, the dropping probability of analysis is very close to that of simulation and justifies the correctness of the proposed Markov chain model.

1. Introduction

1.1. Introduction to cloud computing

The cloud computing is a navel computing structure that consists of three main elements: (1) the data center supported by cloud providers, (2) extremely high data rate wireless networking provided by networking providers, and (3) various-type smart devices for users. Typical cloud services include: Infrastructure-as-a-Service [1] (IaaS, e.g., Amazon EC2 [2], Google Cloud Platform [3], IBM Smart Cloud Enterprise [4] etc.), Platform as a service (PaaS), Software as a service (SaaS), etc. Clearly, the cloud provider aims to maximize system reward by the oversale of cloud resources while satisfying the quality of service (QoS) [5,6] of the Service Level Agreements (SLAs) for users. However, the available cloud resources may be insufficient for a (local) regional data center when the traffic is extremely high. For efficiently managing and allocating cloud resources in the global data centers, several escalations can be applied, including [7]: (1) Changing VM configuration, (2) Migrating applications from one VM to another, (3) Migrating one VM from one PM to another PM or creating a new VM on a PM, (4) Waking up an idle PM, (5) Outsource to other Cloud provider, etc. Fig. 1 demonstrates these five escalations of managing cloud resources on a data center.

In Fig. 1, the paper focuses on the 5th step, outsourcing to other clouds. Selecting the target cloud, the task will be redirected

☆ This research was supported in part by the Ministry of Science and Technology of Taiwan, ROC, under Grants MOST-105-2221-E-224-031-MY2, MOST-106-2511-S-252- 001, and MOST-106-3011-F-252-001.

* Corresponding author.
E-mail addresses: changb@yuntech.edu.tw (B.-J. Chang), t136@nkut.edu.tw (Y.-H. Liang).

https://doi.org/10.1016/j.future.2017.09.046
0167-739X/

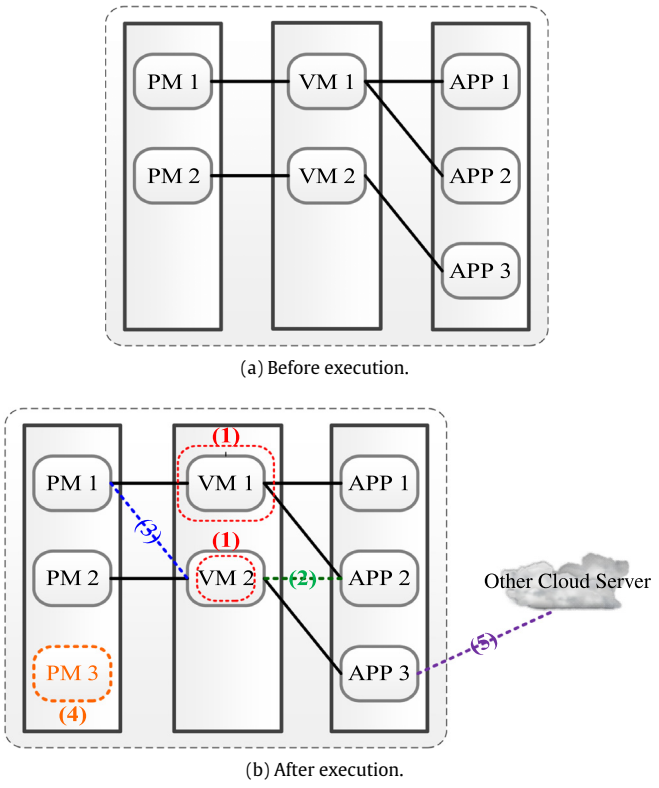

Fig. 1. Actions adopted in five escalation levels: before and after executions.

to the target cloud which has the most resources remaining [8] In the practice, there are two problems will be generated: **First**, compared to the non-real-time tasks, the real-time tasks have extremely stringent delay bound. Due to the limited delay bound, the task dropping probability is increased significantly, if the real-time tasks are redirected to another data center. **Second**, having the most resources means that there may be higher energy consumption. In the statistical information of the web site of Statista [9], the huge electricity price is different in the world [9]. Thus, the load balancing method obviously wastes power energy and increases cost, and cannot maximize the net profit for cloud providers. **Additionally**, the IaaS provider aims to maximize system profits by over-selling the cloud resources. To maximize the net profit for cloud providers by using an efficient resource management under the overselling rule becomes a critical challenge that should be addressed.

1.2. Related works of resource management in cloud computing

The related works of cloud computing can be classified into several types: (1) cloud resource management and (2) cloud performance analysis.

1.2.1. Related works of cloud resource management

In the related works of resource management [3,7,10–13], the studies aim to manage IaaS cloud that has a lot of cloud resources. In [7], the cloud resources are dynamically configured by the knowledge management (KM) approach. However, KM suffers from sudden events because of needing time to learn the knowledge In [10], Iyer et al. analyze a large-scale cloud and then eliminate any redundant Virtual CPU Instance (namely VCI-link) link for reducing system overhead and cloud processing time. In [11], Lin et al. propose a threshold-based dynamic resource allocation scheme that dynamically allocates the VMs according to the loading of the cloud application. In [12], Javadi et al. propose the hybrid cloud architecture, and adopt the failure-aware strategy to increase the user's QoS. In [13], Ardagna et al. propose the load redirection mechanism to minimize the costs of allocated resources while guaranteeing the SLA constraints. In [3], Zhu et al. propose a multi-input–multi-output feedback control model which adopts reinforcement learning to guarantee the optimal application benefit by adjusting adaptive parameters.

1.2.2. Related works of cloud performance analysis

The related studies [14–20] of the cloud performance analysis are depicted below. In [15], Beloglazov et al. use the Multisize Sliding Window workload estimation technique and dynamic consolidation of VMs to increase the resource utilization and energy efficiency. Additionally, the Markov chain model is adopted to detect the server overloading problem. In [14], Khazaei et al. use the M/G/m Markov process model to analyze the performance of the cloud server farms. In [18], Liang et al. propose the inter-domain service transfer to balance loads. Moreover, the decision making process of the service request is formulated by the semi-Markov decision process. In [16,19], they propose the M/G/m/m+r queuing system to analyze performance indicators: mean request response time, blocking probability, the probability of immediate service, etc. In [16], Khazaei et al. focus on the relationship between the number of servers and the buffer size. In [20], Khazaei et al. focus on the compound requests (i.e., a set of requests submitted by a user simultaneously) In addition, in [21], Fang et al. propose the VM Planner approach to change the VM placement and traffic flow routing by using three approximation algorithms that turns off unneeded network elements for saving power cost.

1.3. Motivation and objectives of this paper

Thus, the **motivation** of the paper is to propose a Reward-based adaptive global Cloud Resource Management (namely RCRM) of cloud computing. RCRM is based on Markov chain model analyses and it consists of several mechanisms to adaptively allocate the global cloud resources in IaaS clouds.

The main ideas of this work have three aspects. **First**, Phase 1 is to accurately predict the request traffic of different classes of traffic, and then these resources can be reserved in prior. Thus, Phase 1 can effectively reduce the service blocking probability. **Second**, based on the Markov analytical model, the Adaptive Cloud resource Allocation (ACA) is proposed to determine the required amount of cloud resources according to the predictive request traffic that will arrive at the (local) region DC and the global DCs. **Third**, when the request traffic is predicted accurately in the LSP phase and the cloud resources of the region DCs and global DCs are analyzed correctly, the goal of maximizing the carried reward and the net profit can be achieved by the proposed Maximum Net Profit (MNP) mechanism. Consequently, the net profit of a cloud system can be maximized clearly while satisfying the QoS of SLAs.

The **objectives** of this paper include: (1) adopting the large-scale and small-scale traffic predictions, (2) based on the Markov chain model to analyze the service blocking and the required number of VMs for each request and (3) maximizing the net profit of the cloud provider.

The remainder of the paper is organized as follows. The network model is defined in Section 2. Sections 3 and 4 detail the proposed reward-based adaptive global cloud resource management and the analysis of the M/M/m/m Markov chain model. Numerical results are presented in Section 5. Finally, conclusions and future works are summarized in Section 6.

Please cite this article in press as: B.-J. Chang, et al., Reward-based Markov chain analysis adaptive global resource management for inter-cloud computing, Future Generation Computer Systems (2017), https://doi.org/10.1016/j.future.2017.09.046.