Journal of Network and Computer Applications 82 (2017) 93–113

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Review

Survey on prediction models of applications for resources provisioning in cloud

Maryam Amiri*, Leyli Mohammad-Khanli

Faculty of Electrical and Computer Engineering, University of Tabriz, 29 Bahman Blvd, Tabriz, East Azerbaijan, Iran

ARTICLE INFO

Keywords:
Cloud Computing
Prediction
Application
Workload
Resources Provisioning

ABSTRACT

According to the dynamic nature of cloud and the rapid growth of the resources demand in it, the resource provisioning is one of the challenging problems in the cloud environment. The resources should be allocated dynamically according to the demand changes of the application. Over-provisioning increases energy wasting and costs. On the other hand, under-provisioning causes Service Level Agreements (SLA) violation and Quality of Service (QoS) dropping. Therefore the allocated resources should be close to the current demand of applications as much as possible. Furthermore, the speed of response to the workload changes to achieve the desired performance level is a critical issue for cloud elasticity. For this purpose, the future demand of applications should be determined. Thus, the prediction of the application in different aspects (workload, performance) is an essential step before the resource provisioning. According to the prediction results, the sufficient resources are allocated to the applications in the appropriate time in a way that QoS is ensured and SLA violation is avoided. This paper reviews the state of the art application prediction methods in different aspects. Through a meticulous literature review of the state of the art application prediction schemes, a taxonomy for the application prediction models is presented that investigates main characteristics and challenges of the different models. Finally, open research issues and future trends of the application prediction are discussed.

1. Introduction

Cloud computing is a computing paradigm that provides services such as infrastructure, platform and software based on a pay-as-you-go model (Coutinho et al., 2015; Kulkarni and Agrawal, 2014). Elasticity is one of the prominent features of cloud computing (Petcu and Vzquez-Poletti, 2012). It is the degree of the system adaptability to the workload changes by provisioning and deprovisioning the resources automatically in a way that the allocated resources match the current demand (Herbst et al., 2013). So the elastic application allocates or releases the resources according to its requirements. To comply with the obligations, cloud needs to allocate a suitable amount of resources according to the current demand of applications. Under-provisioning causes Service Level Agreements (SLA) violation, Quality of Service (QoS) dropping and the customer dissatisfaction. This may lead to the loss of customers and a decrease in revenue. On the other hand, Over-provisioning wastes energy and resources and it even increases costs like network, cooling and maintenance. So the resources management is a complicated process in cloud and an efficient resource management technique is required (Singh and Chana, 2016c). As Fig. 1 shows, the efficient resources management plan impacts on three different aspects of cloud. It fulfils SLA and satisfies cloud customers. It guarantees the cloud obligations to its users. So customers will adhere to cloud in the future. It also prevents the resources wasting. So the energy consumption and the operational cost decrease. The reduction of energy consumption leads to decrease carbon emissions, which could facilitate green cloud computing. Both of the cost reduction and the revenue increase improve the profit of cloud providers (Kumar and Buyya, 2012). Therefore, the efficient resources management allocates the minimum amount of required resources for SLA fulfillment (Manvi and Krishna Shyam, 2014) and leaves the surplus resources free to deploy more Virtual Machines (VMs) (Garg et al., 2014). For this purpose, the resources allocated to each application should be close to the application demand in a way that SLA is satisfied and resources wasting is minimized.

Furthermore, the speed of response to the workload changes to achieve the desired performance level is a critical issue for elasticity (Coutinho et al., 2015). Although the important advantage of elasticity is to match the amount of resources allocated to the application with the amount of resources it requires, the time that resources take to be

* Corresponding author.
E-mail addresses: maryam.amiri@tabrizu.ac.ir (M. Amiri), l-khanli@tabrizu.ac.ir (L. Mohammad-Khanli).

http://dx.doi.org/10.1016/j.jnca.2017.01.016
Received 9 October 2016; Received in revised form 16 December 2016; Accepted 15 January 2017
Available online 18 January 2017
1084-8045/

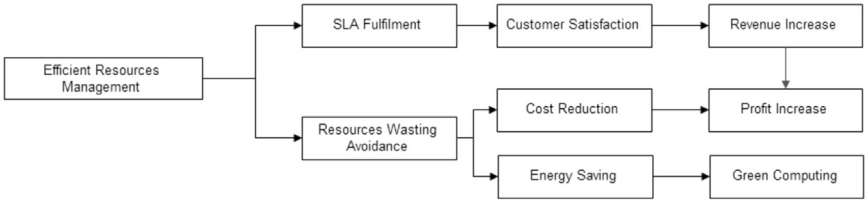

Fig. 1. The Influence of Efficient Resources Management on Different Aspects of Cloud: Revenue, Green Computing and Profit.

ready to use is a potential problem (Galante and Bona, 2012). Cloud elasticity and dynamic resources allocation are based on the virtualization techniques (Hwang et al., 2016). The VM provisioning technologies take several minutes (Jiang et al., 2013). This delay is intolerable for the tasks that need the resources scaling during the computation. It might lead to SLA violation, QoS dropping and finally a reputation loss of cloud. To reduce the delay, there are three approaches. The first approach, VM provisioning technologies, assists to ready new VMs in seconds for the requests (Jiang et al., 2013). The state of the art VM provisioning technologies, such as streaming VM technology (Labonte et al., 2004) and VM cloning (Lagar-Cavilla et al., 2009) cannot decrease time wasting of VM creation (Jiang et al., 2013). The second approach is about to ask all customers to provide a plan of the future resources demand. It is not possible according to the cloud obligations and the lack of customers' knowledge (Jiang et al., 2013). Due to VM technologies and the limitations of the customers' knowledge, the future demand prediction is the only practical and effective solution for the fast resources provisioning. A proactive prediction method predicts the future demand fluctuations in a way that the resource manager has enough time to provide the appropriate resources before occurring the workload burstiness.

If the sudden increase of the future demand is predicted, the resource manager scales up the infrastructure and prepares VMs according to the predicted future demand before the surge of demand occurs. In the same way, according to the demand reduction, the allocated resources are released. The released resources can be used to create new VMs or to allocate them to VMs that need more resources. Indeed, allocated resources are quickly matched with the demand and the rapid elasticity (Mell and Grance, 2011) is accomplished. Thus, SLA is satisfied, energy wasting is avoided and on demand provisioning is fulfilled for systems implemented by using cloud services.

However, providing cloud services that guarantee dynamic QoS requirements of users and avoid SLA violation is a big challenge. Currently, the services are provisioned and scheduled according to the resources availability, without the guarantee of the expected performance. Singh and Chana (2015a). Therefore, the future demand prediction is an indispensable step for the rapid elasticity implementation and the effective resource provisioning in the dynamic cloud environment.

Although many literatures such as Galante and Bona (2012), Huang et al. (2014), Manvi and Krishna Shyam (2014), Weingartner et al. (2015), Aceto et al. (2013), Singh and Chana (2016a), Singh and Chana (2015a), Singh and Chana (2016b), Huebscher and McCann (2008), Coutinho et al. (2015) survey cloud computing in different aspects, there is a lack of a detailed investigation of the application prediction in cloud. This paper presents a survey on the prediction of the application in different aspects such as the performance and the workload. The contributions of this paper are as follows:

- To the best of our knowledge, this paper is the first survey on the prediction of cloud applications. It presents a comprehensive review of the newest and the most prominent prediction models.
- A general taxonomy for proposed models, techniques and frameworks of the application prediction is presented. Literatures are grouped based on their proposed methods and explained briefly.
- Open research issues, challenges and the future trends of the application prediction in cloud are presented.

This paper is structured as follows: Section 2 describes different aspects of the application prediction such as main challenges, characteristics, needs and evaluation metrics for the prediction in cloud. Section 3 presents the modeling approaches for the application prediction. Section 4 investigates the proposed prediction methods and describes their advantages and disadvantages. In Section 5, different techniques are compared and challenges and directions are explained. Finally, the paper is concluded in Section 6.

2. Application prediction

The application prediction is to forecast the future behaviour of the application in different dimensions such as the workload and the performance. So the workload and performance prediction are branches of the application prediction. The application prediction is an essential step for the efficient resources management in cloud. According to the future demand of the application, the efficient resources provisioning should detect the minimum amount of resources to fulfill QoS parameters such as CPU utilization, response time, availability, reliability and security (Singh and Chana, 2016a). Table 1 shows the QoS requirements of different applications (Singh and Chana, 2016c, 2015b). We recommend that readers interested to the resources management refer to Singh and Chana (2016a), Singh and Chana (2016b). In this section, the application prediction is considered in different aspects. At first, the different dimensions of

Table 1
Cloud Applications and their QoS requirements (Singh and Chana, 2016c).

Applications	QoS requirements
Web sites	Reliable storage, high network bandwidth, high availability
Technological computing	Computing capacity, reliable storage
Endeavour software	Security, high availability, customer confidence level, correctness
Performance testing	Execution time, energy consumption and execution cost
Online transaction processing	Security, high availability, internet accessibility, usability
Central financial services	Security, high availability, changeability, integrity
Storage and backup services	Reliability, persistence
Productivity applications	Network bandwidth, latency, data backup, security
Software/project development and testing	User self-service rate, flexibility, creative group of infrastructure services, testing time
Graphics oriented	Network bandwidth, latency, data backup, visibility
Critical internet applications	High availability, serviceability, usability
Mobile computing services	High availability, reliability, portability