Journal of Network and Computer Applications 65 (2016) 144–154

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Virtual resource prediction in cloud environment: A Bayesian approach

Gopal Kirshna Shyam [a,*], Sunilkumar S. Manvi [b]

[a] Department of Computer Science and Engineering, Reva Institute of Technology and Management, Bengaluru 560064, India
[b] School of Computing & Information Technology, Reva University, Bengaluru 560064, India

ARTICLE INFO

Article history:
Received 9 June 2015
Received in revised form
20 February 2016
Accepted 2 March 2016
Available online 4 March 2016

Keywords:
Bayesian network
Resource prediction
Cloud computing

ABSTRACT

With increase in requirement for dynamic execution of user's application in cloud, resource prediction techniques are gaining a lot of importance as the foundation for online capacity planning and virtualized resource management in data centers. There is a wide scope for the development of accurate resource requirement prediction methods to ensure that the virtualized resources do not suffer from over or under-utilization. We propose a Bayesian model to determine short and long-term virtual resource requirement of the CPU/memory intensive applications on the basis of workload patterns at several data centers in the cloud during several time intervals. However, the model is applied to predict resource(s) of all applications in general.

The parameters considered for prediction in the model are day of week, time-interval of application access, workload, benchmarks, and availability of virtual machines etc. The model is simulated by using the SamIam Bayesian network simulator and workload traces of Amazon EC2 and Google CE data centers in dynamic scenarios. The performance is evaluated by considering benchmarks of CPU intensive applications (web based). The proposed model is able to predict virtual resources in a cloud environment with better accuracy as compared to other models.

1. Introduction

Cloud computing is a new trend in Internet computing where resources like storage, computation power, network, applications etc. are delivered as a service. This service is categorized as Infrastructure-as-a-service (IaaS), Platform-as-a-service (PaaS), and Software-as-a-service (SaaS). The IaaS delivery model provides storage, hardware and networks as the basic subcomponents of services. It offers on-demand creation of virtual machines (VMs) for various users and applications, which enables dynamic management of VMs for maximizing the resource utilization in the data centers (DCs).

In the present day scenario, information technology (IT) companies demand usage of non-stop running servers to increase productivity. This leads to increased hardware purchase costs and also contributes to environmental pollution due to increased CO_2 emissions as mentioned in Meisner et al. (2013). The cost to operate underutilized servers is significantly higher as mentioned in Kukeraja and Singh (2012). Furthermore, a server requiring additional resources during peak load demands provisioning of better hardware and increased number of dynamic servers.

As the size of cloud scales up, cloud computing service providers must handle massive requests. Thus, there is a challenge to scale up the cloud performance. In spite of glorious future of cloud computing, many critical problems are yet to be investigated and explored for efficient and robust realization as presented in Jiang et al. (2013b) and Shyam and Manvi (2015). On-demand provisioning of CPUs in cloud environment, as discussed in Jiang et al. (2013a, 2013b), is a challenge for many DCs. This is because of the lack of instant cloud resource provisioning.

Modern CPU and memory intensive applications in cloud environment (that require virtual CPUs (vCPUs) and virtual memory (vRAMs)) have fluctuating resource demands which leads to complex behavior in resource usages, as their intensity and composition change over time. Further, optimization of VMs performance in cloud scenario is required with massive increase in data/information/applications on day to day basis. For example, the discount sales day(s) offered by e-commerce gaints such as Amazon, Flipkart, Snapdeal etc. sometime fails to deliver as per the requirement of all the customers due to lack of resource management planning. Hence, VMs prediction is an immediate challenge to be tackled as given in Manvi and Shyam (2014). Some examples of CPU intensive applications are file compression, video editing, 3D modelling, audio, video encoding and decoding etc. Examples of memory intensive applications are computational fluid dynamics, weather resource forecasting etc.

[*] Corresponding author.
E-mail addresses: gopalkrishna@revainstitution.org (G.K. Shyam), sunil.manvi@revainstitution.org (S.S. Manvi).

http://dx.doi.org/10.1016/j.jnca.2016.03.002
1084-8045/© 2016 Elsevier Ltd. All rights reserved.

Resource prediction will help to achieve the service-level agreements (SLA) signed with the customers (or users). Effective prediction of resource can facilitate load balancing and proactive job scheduling across DCs which results in improved resource utilization, lowering DC costs, and improved job performance. Although different models have been proposed for resource prediction in Jiang et al. (2013a, 2013b), these models are either based on history of jobs execution or correlated resources. Existing prediction work in high performance computing (HPC) systems has focused mainly on using moving averages, auto-regression, and noise filters as mentioned in Akioka and Muraoka (2012), Carrington et al. (2012), Dabrowski and Hunt (2009), Dinda and Hallaron (2010), and Tirado et al. (2011). These prediction methods have been evaluated with traces of load in Grids or HPC systems. When applied to bursty cloud workloads, they have limited accuracy. Moreover, these works do not attempt to predict the short and long term future workload pattern.

Research on developing tools and techniques has been investigated and implemented in computational grids, such as network weather service as discussed in Wolski et al. (2013) and Birje et al. (2014), which monitors the network and computing resource QoS (Quality of Service) and periodically forecasts the QoS of application workload. Other monitoring tools, such as the one in Zhang et al. (2013) and Zanikolas and Sakellariou (2013), were popular in grid and cluster computing. Some other tools go by the name Hawkeye, Ganglia, MDS-I, and MDS-II. These tools are only concerned with monitoring the QoS parameters for the hardware resources (CPU, storage, and network), while application-specific QoS parameters and service level agreements (SLA) requirements are not taken into account. But most of these tools do not consider the issue of auto scaling and de-scaling primitives supported by virtualized cloud resources. Clearly, such tools are not suitable to tackle the challenges of cloud computing environments and for hosted application types.

Current challenges with respect to resource prediction are: (i) prediction of the virtualized resources dynamically in order to handle variable workloads, (ii) management of network complexity, (iii) accurate representation of training data, and (iv) management of resources in IaaS data centers. Existing schemes tackle these challenges by random selection, multiple training, cross validation etc. which are not optimal.

To address these challenges, we need a model which can identify dependencies considering different parameters in virtualized cloud environment, which contributes to nearly accurate predictions, and satisfies QoS parameters without SLA violation.

Hence, motivated by the above challenges, we construct a Bayesian model by identifying different variables that are discrete and optimally select training data which contributes in predicting the virtual resources. Bayesian networks (BNs) provide the modeling needed for prediction of resources in cloud. The advantages of using BNs include (i) predicting target variable in the face of uncertainty, (ii) representing a casual relationship by an arc between two nodes and the conditional probabilities of one node conditioned on its parents, and (iii) providing a valid output when any subset of the modeled variables is present. Such models would be helpful for system maintenance and application schedule in multiple DCs. We restrict the discussions to VMs, which can be configured to obtain required number of vCPUs and vRAMs, as both of these are important, expensive and most sought resource in cloud environment.

There are four major contributions in this work. (i) Identification of dependencies among variables in virtualized cloud scenario based on analysis of workloads at different DCs of a cloud. (ii) Novel method for optimally selecting training data. (iii) Prediction of mean VMs (for CPU or memory intensive applications) and the mean response time for user's application by capturing the patterns of workload at DCs, and (iv) using the statistics obtained to manage the capacity planning of DCs.

Rest of the paper is structured as follows. Section 2 presents related works, Section 3 describes the proposed model, Sections 4 and 5 discuss the simulation and results respectively, and Section 6 concludes the work.

2. Related works

Some of the resource prediction models are discussed in this section. We classify them into three types: (i) analytical, (ii) computational intelligence, and (iii) simulation models. Table 1 presents works on resource prediction in cloud computing. Analytical models are based on Markovian chains, regression models and Queing theory. A stochastic Markovian model has been proposed to measure the scalability and tractability of infrastructure resources in cloud. Researchers have worked on determining the capacity supply model for virtualized servers which can be used for predicting a set of resource utilization metrics. Researchers have also devised a model for end-to-end network service delivery systems in network virtualization environments. The technique for end-to-end performance of network service delivery in the virtualization-based cloud computing has been analyzed.

Computational intelligence models are based on fuzzy logic, neural networks, genetic algorithms and multi-agent systems. Using computational intelligence models, the effect of shared resources in virtualized environments has been studied. Simulation tools used for resource prediction are Overdriver, Memory-buddies, and Vmctune.

The existing resource prediction models in cloud consider excessive allocation of resources in order to handle peak demands. High demand periods in the data centers of cloud are characterized by high occupancy levels. It poses significant risks to the user's application as communication issues play a major role. Further, it is observed that very few works are available in workload prediction using the BN model, but none of them address prediction of virtual resource requirement of applications. This has motivated us to investigate and explore the use of BNs for addressing this issue.

3. Proposed work

The proposed resource prediction model is designed based on data collected from topmost cloud service provider Amazon EC2 and Google CE DCs, as reported in Park and Pai (2014) and Reiss et al. (2012) respectively. The collected data includes CPU and memory utilization for different applications in various DCs. After an analysis of data, we observed the following. (i) DCs of Amazon EC2 and Google CE vary in their resource usage during certain time-intervals, as well as suffer from over-utilization and under-utilization problems at times. (ii) Resources are utilized during weekdays efficiently, but sometimes resources' availability is not met during peak loads; whereas, resource usage is below normal during weekends. Hence, we want to predict the virtual resource requirements of the application categories at various DCs, ranging from small to large time intervals to enhance the quality of customer services. In this section, we discuss the scenario considered for the proposed work, the steps to construct BNs model, and Bayesian inference for query solving.

3.1. Scenario description

We consider a scenario where an IT organization(s) (say, customers) uses an infrastructure of a cloud service provider (CSP) to