Computers and Electrical Engineering 47 (2015) 138–160

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

# Q-aware: Quality of service based cloud resource provisioning ☆

Sukhpal Singh *, Inderveer Chana

*Computer Science and Engineering Department, Thapar University, Patiala, Punjab 147004, India*

ARTICLE INFO

*Article history:*
Received 15 May 2014
Received in revised form 4 February 2015
Accepted 4 February 2015
Available online 25 February 2015

*Keywords:*
Cloud metrics
Workload patterns
Resource provisioning
Cloud computing
Cloud workloads
Quality of service

ABSTRACT

Provisioning of appropriate resources to cloud workloads depends on the Quality of Service (QoS) requirements of cloud workloads. Based on application requirements of cloud users, discovery and allocation of best workload – resource pair is an optimization problem. Acceptable QoS cannot be provided to the cloud users until provisioning of resources is offered as a crucial ability. QoS parameters based resource provisioning technique is therefore required for efficient provisioning of resources. In this paper, QoS metric based resource provisioning technique has been proposed. The proposed technique caters to provisioned resource distribution and scheduling of resources. The main aim of this research work is to analyze the workloads, categorize them on the basis of common patterns and then provision the cloud workloads before actual scheduling. The experimental results demonstrate that QoS metric based resource provisioning technique is efficient in reducing execution time and execution cost of cloud workloads along with other QoS parameters.

## 1. Introduction

Cloud Computing offers pay per use based services such as Infrastructure as a Service, Platform as a Service and Software as a Service through different Cloud providers [1]. Cloud provider provides the data and computing resources dynamically to the Cloud users based on their requirements over the internet. This is a challenging task to maintain the required Quality of Service (QoS) level of service to fulfill the expectations of Cloud consumers [2]. The complexity of management of resources in Clouds is increasing day by day, so an efficient technique is required for management of resources. To solve this problem, there should be a focus on provisioning before actual scheduling of resources. Resource provisioning is the practice of implementing policies and procedures that improve the efficiency of computing resources in such a way so as to reduce the execution time and cost while considering other QoS parameters like availability, network bandwidth, serviceability and customer confidence level. Further resource scheduling executes workloads to the provisioned heterogeneous resources based on scheduling decisions [3].

Presently, provisioning of resources can be done in two ways: On-demand (providing resources quickly to urgent workloads) and long term reservation (reserving resources for later usage). In on-demand criteria, executing too many workloads on a single resource will cause problem of interference which leads to performance degradation and over provisioning. In long term reservation, many resources are in idle state which leads to under provisioning. Under provisioning and over

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Raja K Murugesan.

* Corresponding author.
*E-mail addresses:* ssgill@thapar.edu (S. Singh), inderveer@thapar.edu (I. Chana).

http://dx.doi.org/10.1016/j.compeleceng.2015.02.003
0045-7906/© 2015 Elsevier Ltd. All rights reserved.

provisioning of resources leads to wastage of time and resources that increases cost [4]. To handle this problem, there is a need of resource provisioning techniques for better management of resources to: (i) identify the Cloud workloads, (ii) analyze the Cloud workloads to find their QoS requirements, (iii) classification of Cloud workloads and (iv) provisioning of resources to the workloads without violation of Service Level Agreement (SLA). The scheduling of resources is generally done on best effort approach, which is not efficient in terms of cost and execution time [5]. For example: we cannot execute the second workload until the first workload completes which leads to the problem of starvation. To solve this problem, we need provisioning of resources before actual resource scheduling, in which second workload will be executed immediately after the completion of first workload. Thus, the waiting time of workload can be reduced to a large extent.

The motivation of this research work stems from the challenges in managing and an efficient utilization of the resources. In real life situations, there are many constraints including (i) satisfying the QoS (ii) reducing the resource's cost, (iii) minimizing the workload's execution time by meeting the desired deadline described by cloud consumer and (iv) improves customer satisfaction. To provision the resources along with the QoS constraints is the main objective of this research work. This research paper proposes QoS metric based resource provisioning technique in which resources are provisioned by clustering of workloads after assigning weights to quality attributes of each workload. By using the proposed technique, real resource provisioning and resource scheduling can be predicted. Thus the queuing time, over and under-utilization of resource can be avoided or be assuaged.

The paper is structured as follows: In Section 2, related work of cloud workloads and resource provisioning has been presented. Workload identification and analysis has been presented in Section 3. In Section 4, Resource provisioning technique has been proposed. Experimental setup and results has been presented in Section 5. In Section 6, conclusions and the future scope have been presented.

## 2. Related work

Provisioning of resources for Cloud workloads is an important part of resource management in cloud. The research work done in the area of cloud workloads and resource provisioning is described in this section.

### *2.1. Cloud workloads*

Cloud workload is an abstraction of work of that instance or set of instances that are going to be executed. For example: Running a web service is a valid workload [4]. Mian et al. [6] proposed cost based framework to find the execution cost of workload executing on different computing resources which considers and balances Service Level Agreement (SLA) penalties and cost of resource to predict the performance of framework without considering execution time. Smith et al. [7] described impact of various Cloud workloads on energy consumption of resources to predict and monitor the energy consumption of different resources accurately but not considered execution time and cost. Ciciani et al. [8] proposed an autonomic resource provisioning technique to predict the forecasting resource requirement and chances of SLA violations but not considering QoS requirements like execution time, cost etc.

Breternitz et al. [9] presented Synthetic Workload Application Toolkit which creates, deploys, provisions and executes the Cloud workloads automatically without considering SLA. Delimitrou et al. [10] proposed a workload based approach iBench to find the impact of resource scheduling and resource provisioning to efficiently design an application with maximum resource utilization but not considered execution time and cost. Zhang et al. [11] found various research issues of dynamic management of Cloud workload in heterogeneous environments of Cloud to fulfill demand of cloud user and reduce resource consumption and response time. Son et al. [12] proposed SLA based resource allocation framework to balance the load on available resources without violation of SLA but not considered execution time and cost. LaCurts [13] proposed an approach "Cicada" used to predict the Cloud workload to reduce the execution time and SLA violations only for homogeneous workloads. Chang et al. [14] proposed neural network model based resource allocation mechanism to predict the Cloud workload for efficient resource provisioning without considering heterogeneous workloads.

Here existing work of workloads in the context of Cloud has been presented. They have not considered heterogeneous cloud workloads and QoS requirements like cost, execution time simultaneously [3,4]. To develop an efficient resource provisioning technique, there is need to identify the heterogeneous Cloud workloads and their QoS requirements.

### *2.2. Resource provisioning*

The resource provisioning techniques in distributed systems frequently have the objectives of distributing the workload on resources and increasing their resource consumption but reducing the time of workload execution. Existing resource provisioning techniques have been presented in this section.

#### *2.2.1. Resource provisioning techniques – without quality of service*

Quiroz et al. [15] proposed a distributed approach to determine resource scheduling of workload on innovativeness clouds. To handle with erroneous request for resource that causes to over utilization of resources delivered by cloud workload request, their approach has revealed a design-based technique for approximating the workload execution time specified