Optimal Double Renting Scheme Based Dynamic Virtual Machine Provisioning and Allocation for Improving QoS and Profit Maximization

M. Rajadarshini and Dr. R. Alageswaran

1PG student - Department of CSE, K.L.N.College of Engineering, Pottapalayam, Sivagangai 630612
2Professor- Department of CSE, K.L.N.College of Engineering, Pottapalayam, Sivagangai 630 612

Received 27 May 2016; Accepted 28 June 2016; Available 12 July 2016

Address For Correspondence: M. Rajadarshini, PG student - Department of CSE, K.L.N.College of Engineering, Pottapalayam, Sivagangai 630612

Copyright © 2016 by authors and American-Eurasian Network for Scientific Information (AENSI Publication). This work is licensed under the Creative Commons Attribution International License (CC BY).

ABSTRACT

One of the major challenging problems for cloud providers is designing efficient mechanisms for Virtual Machine provisioning and allocation. Such mechanisms enable the cloud providers to effectively utilize their available resources and thereby make them achieve higher profits. In this paper, the strategy of Double-rent queuing model is proposed for service providers. This includes Short-term renting and Long-term renting schemes as major elements, which plays a vital role in improving Quality-of-Service. The allocations of resources are based on Double renting scheme which uses the concept of heuristics. Proposed solution provides achieving results and outcomes for both the cloud providers and the consumers.

KEYWORDS: Cloud computing, virtual machine provisioning, dynamic resource allocation, guaranteed service quality, profit maximization, double-rent queuing model, service-level agreement.

INTRODUCTION

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. It relies on sharing computing resources rather than having local servers to handle applications. Therefore it is a model which enables on-demand access to a shared pool of configurable computing resources. Self service provisioning, scalability and pay-as-per-use are the three main benefits for cloud computing.

Services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). IaaS provides the users to migrate their workload to the Virtual Machines. Users have an allocated storage capacity and start/stop access and can configure the VM and storage as desired. PaaS is used for general software development and it provides a virtual platform or a simulated environment for web services, thereby reducing the cost of installation. SaaS is a distribution model, where software applications are delivered over the internet. They are often referred to as the web-services.

In this project, Infrastructure-as-a-Service is considered. The existing VM instance makes appropriate decisions on effective allocation of existing resources and thereby improves the QOS.
Quality of service to the requested users based on the method in which the requests are submitted. This process follows the double rent queuing technique for resource categorization. Cloud storage is the virtual storage space, where data is stored online and can be accessed by multiple clients from different geographic locations. The virtual machine combines multiple servers to reduce the work load and reduce the execution time. Virtualization technology creates several Virtual Machines on a physical server. Therefore the virtualization concepts, reduces amount of hardware in use and improves the utilization of resources.

A major challenging problem for cloud providers is difficultly in designing efficient mechanisms for Virtual Machine (VM) provisioning and allocation [1].

- A cloud centre can have a large number of facility (server) nodes, typically of the order of hundreds or thousands, traditional queuing analysis rarely considers systems of this size.
- The coefficient of variation of task service time may be high.
- The coefficient of variation of task response time may be high.
- Due to the dynamic nature of cloud environment, diversity of user’s requests and time dependency of load, cloud centers must provide expected quality of service at widely varying loads.
- The cloud provider provisions VMs based on the requests of the users and determines their requests. Efficient mechanisms must be considered to treat different users in a fair manner.

Related Work:

In [4], Urgaonkar et al focus on effective resource management mechanisms for sharing resources in commodity clusters. To address this issue, we present the design of Sharc a system that enables resource sharing among applications in such clusters. Sharc depends on single node resource management mechanisms such as reservations or shares, and extends the benefits of such mechanisms to clustered environments. Techniques for managing two important resources CPU and network interface bandwidth—on a cluster-wide basis. Our techniques allow Sharc to support reservation of CPU and network interface bandwidth for distributed applications, dynamically allocate resources based on past usage, and provide performance isolation to applications. Sharc extends the benefits of single node resource management mechanisms to clustered environments. Sharc typically requires no changes to the operating system so long as the operating supports resource management mechanisms such as reservations or shares, Sharc can be built on top of commodity hardware and commodity operating systems.

Beloglazov et al proposed an energy efficient resource management system for virtualized Cloud data centres that reduces operational costs and provides required Quality of Service (QoS) [3]. Rapid growth of the demand for computational power by scientific, business and web-applications has led to the creation of large-scale data centres consuming enormous amounts of electrical power. Energy savings are achieved by continuous consolidation of VMs according to current utilization of resources, virtual network topologies established between VMs and thermal state of computing nodes.

Menasce et al proposed the concept of QoS Broker (QB) in [5]. QB selects a service provider, from among participating service providers, which maximizes a global utility for a client under a cost constraint. Additionally, our architecture supports a flexible and loosely coupled integration scheme and serves as a mediator to route requests and responses between service providers and consumers. The QoS broker uses analytic queuing models to predict the QoS values of the various services that could be selected under varying workload conditions.

GuiyiWei et al proposed the concept of Game Theory to solve sophisticated parallel computing problem by requesting the usage of resources across a cloud-based network, and cost of each computational service depends on the amount of computation in [6]. A practical approximated solution with the following two steps is proposed. First, each participant solves its optimal problem independently, without consideration of the multiplexing of resource assignments. A Binary Integer Programming method is proposed to solve the independent optimization. Second, an evolutionary mechanism is designed, which changes multiplexed strategies of the initial optimal solutions of different participants with minimizing their efficiency losses. The algorithms in the evolutionary mechanism take both optimization and fairness into account.

Junwei Cao et al proposed the approach of treating a multiserver system as an M/M/m queuing model, such that the optimization problem can be formulated and solved analytically in [7]. Two server speed and power consumption models are considered, namely, the idle-speed model and the constant-speed model. The probability density function of the waiting time of a newly arrived service request is derived. The expected service charge to a service request is calculated. The expected net business gain in one unit of time is obtained.

Proposed Solution:

The resources in the service system are normally rented for a long duration. The incoming service requests will have to be queued up for a long duration in order to process the service. The limited resources, available in
the system may be engaged in processing a single service for quite a long term. Hence, as requests are not processed as per the SLA, the QOS drops dramatically. This leads to loss of revenue.

Therefore, the resources are scheduled based on double renting model. This approach provides higher Quality-of-Service as well as increased revenue. The scheme combines the concept of Long Term and Short Term renting techniques. It serves in reducing the resource wastage to a great extent.

The service consumption may be requested on the two methods. It may be either on the Reservation approach or the On-Demand technique.

The decision on Long term or Short term renting is made based on Heuristics data. When the user is registered for the very first time, a record is added for the corresponding user in the heuristic database with the credit value. From thereon, when the same user requests for the second time, the resource allocation will be decided based on the updated heuristic record.

Fig 1 represents the Architecture diagram of the three-tier-system interface.

![Architecture Diagram](image1)

**Fig. 1:** Architecture Diagram

The three typical parties mentioned in the structure include the Infrastructure providers, the service providers and the customers. A customer submits a service request to a service provider which delivers services either on demand or on reservation. The customer receives the desired result from the service provider with certain service-level agreement, and pays for the service based on the amount of the service and the service quality.

**Methodology:**

The below mentioned methodology is involved in resource allocation and service-request processing.

**Double Renting Scheme:**

- Long-term or Short-term renting scheme is decided based on the heuristic record. The service consumption is henceforth taken up with two approaches.
  - Service Consumption using Reservation
  - Service Consumption using On Demand

  The Reservation approach handles resources which are rented for a long term. The On-Demand technique handles resources which are taken up for a shorter duration.

  Fig.2 depicts the process of double renting scheme.

![Queue Processing](image2)

**Fig. 2: Queue Processing**

The incoming service requests are appended in the queue. The service providers will handle the request based on the heuristic approach. If the resources are decided to be allocated for a long-term, they are routed to the desired server. Similarly, the short term resource allocation is handled. Hence, the resources are allocated with a great reduction in wastage.
Algorithm:
The below algorithm describes the approach of heuristic usage in deciding allocation of resources.
The resources are allocated based on the service requested by the users. The requests can be either based on
Reservation or On-Demand.
The outputs captured in the algorithm have two major components.
a. Valuation Rate – The profit obtained on processing the request within the constraints mentioned in
   SLA.
b. Credit – It is directly proportional to the resource usage on the given time frame.

Input:
Set of servers is denoted as,$\text{S} = \{S_1, S_2, \ldots, S_n\}$
Set of resources is denoted as,$\text{R} = \{1, 2, 3, \ldots, R\}$
Set of VM Instances are represented as,$\text{VM} = \{1, 2, 3, \ldots, M\}$
Set of Users is denoted by,$\text{U} = \{1, 2, 3, \ldots\}$
Set of User Requests is denoted by,$\text{J} = \{1, 2, 3, \ldots J\}$
$\text{Wr} = \{1, 2, 3, \ldots, M\}$
$\text{Wr}$ is the Working Rate of the servers
Where $M$ is the Maximum capacity that
the server can be loaded with.
The working rate of the servers are compared.
Consider,$W_i = \text{Wr of Server1}$
$W_j = \text{Wr of Server2}$
Amount of resources of type $R(r)$, for
each Virtual Machine Instance
$\text{VM}(m)$ can be together denoted as
“$W_{rm}$”
$X_i$ denotes the allocation value. It takes two constants which can be either 0(not allocated) or 1 (allocated).
$X_i = \{0, 1\}$
$V$ is the Valuation rate,
$Cr$ is the maximum Capacity limit for
each resource ($r$).
$Lr$ is the denotation of Long Term
Resource Allocation.
$Sr$ is the denotation of Short Term
Resource Allocation.
For each user($i$) in Subset $U$
/
if(user($i$) creates login for the first time)
/
Considering each Instance($m$) in
Subset VM,
Sum of ($J_{im} \cdot W_{mr} \cdot X_i$) $\leq$ $Cr$
Record the time taken for service
processing
/*@ Account based on the usage */
Credit is recorded
[user($i$),Credit] entry is logged into
consolidation report
}
else
/
/*@ The consolidation report
captures Lr/Sr based on heuristics
report */
result = read Heuristic Report
If (result == Lr || Sr)
{
    Considering each Instance(m) in
    Subset VM,
    Sum of (Jim . Wmr . Xi) <= Cr
    /
    /
    /
}

Output:
V*, Credit

Evaluation Metrics:

The metrics are evaluated based on two parameters.
The Valuation Rate denotes the profit obtained on processing the request within the constraints mentioned in SLA. This reflects the processing done based on single term, long term or short term. The higher the value, the higher is the Quality-of-Service provided to the users.
The Credit value is directly proportional to the resource usage on the given time frame. This information is accounted in the heuristic record.

Results:
The processes involved in resource allocation are described with the results captured.

Step-1: Cloud configuration:
Unique IP is assigned for the Cloud Server.
The cloud server can be used for client access,
Fig-3 provides the snapshot of the cloud server configured.

![Cloud Configuration](image)

Fig. 3: Cloud Configuration

Step-2: Formation of Common Cloud Server:
Common cloud server is the centralized server. It is the Decision Maker of Multi cloud Environment. The centralized server helps in heuristic collection and maintenance.

Step-3: User authentication:

Fig 4 represents the authentication process.
The user registration process is done by the administrator. Here, the users fill in their required information during the registration process. After registration every user will get an identity for accessing the cloud space.
Fig. 4: User Authentication

Step-4: Service Discovery using Double Renting Scheme:

Providers are discovered by comparing the specifications listed in the Request for service (RFS) with service descriptions. The discovery is constrained by the Heuristics record and an organization can release the RFS to a limited preapproved set of providers. The service is allocated to the user either on a shorter time period or on a longer duration based on their cloud space usage.

The Double Renting Schemes are categorized as
- Short term or On Demand.
- Long term or Reservation Schemes.

Fig 5 represents the heuristics captured.

Fig. 5: Heuristic Database

Step-5: Resource Allocation:

Common cloud server acts as the Resource Allocator. Dynamic resource allocation within the cloud environment is henceforth achieved. The resources are categorized based on the double renting scheme. Heuristic approach of resource categorization is considered.

Fig 6 provides the Reservation technique, which accounts for Long Term Renting Scheme.

Fig. 6: Reservation Method
Fig 7 provides the On-Demand technique, which accounts for Short Term Renting Scheme.

**Fig. 7:** On-Demand Method

**Analysis of Results:**

The valuation rate and Credit information are captured. Graphs are generated accordingly.

Fig 8 provides the snapshot of the service which is processed for the user, requested via Reservation Scheme.

**Fig. 8:** Service Usage for Reservation Request

Fig 9 provides the snapshot of the service which is processed for the user, requested via On-Demand Scheme.

**Fig. 9:** Service Usage for On-Demand Request
Fig 10 explains the performance metrics with Time used (based on the renting scheme in minutes) vs Resource usage (calculated in %). The diagram proves that the approach greatly reduces the resource wastage.

![Graph showing performance metrics]

Fig. 10: Credit Rate

Fig 11 depicts the measurement of accounting involved in the process. The x-axis denotes the value (in Rupees) and the y-axis depicts the amount of resource used (in GB). The graph metrics proves that the approach results in greater profits.

![Graph showing accounting metrics]

Fig. 11: Valuation Rate

Conclusion:
The Cloud Providers effectively utilize the available resources to a maximum extent, thereby greatly reducing the resource wastage. This approach hence proves to guarantee a high Quality-of-Service and thereby maximizes the profit rate using a novel double-renting scheme based on heuristic approach. Hence, the method proposed in this paper serves as a best choice to the service providers creating a WIN-WIN strategy for both the cloud-providers and the consumers.

REFERENCES