Resource Provisioning in Cloud Environment for Enhanced Quality of Service

Baldev Singh
Lyallpur Khalsa College, Jalandhar, India

Abstract- Cloud computing enables Internet based services and infrastructure to the cloud users on pay-as-usage basis. The services provided by using cloud are Infrastructure as a Service, Platform as a Service and Software as a Service. Provisioning of these services is through the use of various hardware and software resource of the cloud environment. Cloud users require better quality of service from the cloud providers. On the other hand cloud providers require optimal utilization of their cloud resources. The under-utilization and over-utilization leads to poor quality of services which crafts conflict of services between the consumers and producers of the cloud computing. Better trade-off can be crafted between cloud users and cloud providers if the resource provisioning is made in an optimal way and workload over cloud virtual environment be allocated and managed within the limits of threshold that leads to better quality of services (QOS) to the users. Resource provisioning mechanisms are classified as static and dynamic that have their positives and negatives. But the dynamic resource provisioning grants better workload management and QOS.

Keywords: Cloud computing, resource provisioning, quality of service, SLA

I. INTRODUCTION

Cloud computing is one of the latest evolution in Internet based technologies and services through which variety of resources are offered to the cloud users on pay-as-usage basis. The resources (computer servers, storage, networks, applications and services) are offered on the basis of mainly three types of cloud services Infrastructure [1, 15] as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Various characteristics of cloud computing are pay-per-use, elasticity, self-service provisioning, scalability, Multi-tenancy, broad network access for uploading and accessing cloud data, resource pooling, workload resilience and workload migration [2, 11].

Various operations performed in cloud environment [3,21] includes allocation of information and computing resources to the cloud user applications, load balancing on the basis of consumption level of resources in the cloud, power and energy management. These operations are preformed dynamically because the manual implementation is time consuming and more prone to errors. Various instances (virtual machines) [19, 22] of cloud are initiated, stopped and reinitiated automatically under automated cloud resource provisioning [4, 18] which is supported by cloud that leads to significant improvement in operational efficiency and lesser manual interventions. Cloud services are delivered after deployment of clouds.

Three types of cloud deployment models are in use that are private cloud, public cloud and hybrid cloud. To avail the cloud services from any of these private, public or hybrid cloud, resource provisioning and workload management are the main tasks that are to be performed in an optimal way which is research area of this paper.

Another important domain of research related to cloud computing is cloud security which is not part of this work. In this new world of digital transformation, cloud computing is evolving at a very fast pace and the uses of cloud will be increased many folds especially business companies and entrepreneurs will take lead to adapt cloud environment and reap the benefits of cloud.

II. CLOUD COMPUTING ENTITIES

Cloud resources can be easily spin up. Even deployment of cloud services related technologies is a matter of very short time and the scalability is also not an issue in cloud environment. With the grow and shrink capacity needs of the cloud customers, the resource requirement changes so there is elasticity to scale up or down the cloud resources. Mainly two types of entities exist in cloud computing environment. One is known as cloud (service) providers and the other are cloud users. Cloud providers are the owners of the resources that they have in the form of data centres and these resources are provided to the cloud users on rent as-per-usage basis. Cloud providers have the interest to maximize their cloud resource by way of providing to the cloud users means to say that they host as many workload as possible on each data centre resources so that resources can be utilized at the maximum.
The cloud users want to minimize their cost by running their required applications on the rented cloud. These applications are basically the workload for the cloud providers. Cloud workload is executed over the cloud resources and it should fulfill the service level agreement which is made between the cloud providers and cloud users so that Quality of Service (QoS) [12, 17] requirements are fulfilled. There are research challenges [10, 18] in context of cloud resource provisioning as QoS is the challenging task. Resources are provisioned by the cloud providers [16] on the basis of workloads. Figure-1 below shows Resource Allocation and Provisioning Cycle of cloud environment.

The workload to be provided to the cloud can be of different types like online transaction processing, storage services, e-commerce activities, websites hosting, computation tasks, program testing etc. These types of requests or applications require IaaS level resources like, server computation, storage, memory, networking and backup etc. [3, 5]. There exists the mechanism to host the load over cloud resources. First of all workload is scheduled and put that in a workload-queue. Then there exists Workload Analyser that have the information about the cloud resources for workload allocation and sharing [3, 23]. Provisioning of resources is made to cloud user requests and QOS aspect is utmost consideration at the same time. Although there exists variety of workloads and they also have different levels of defined service level agreements to be fulfilled. So the workloads are hosted separately on the basis of QOS requirements and these are also analysed on the basis of selected commitments. The workloads can be clustered on the basis of selected criteria if the workloads are large in size. There exists workload queue and resource pool so that workloads can be issued to the available resources in the pool for execution.

### III. CHALLENGES IN RESOURCE PROVISIONING IN CLOUD

Challenging in resource provisioning basically related to the optimal use of the cloud resources so that cloud users can avail enhanced quality of services in an economical way as well as the cloud providers can maximize the resource utilization and reap the benefits.

- Complexity in computation
- Multivariate uncertainty in terms of demand, availability and price of resources
- Fulfilment of Quality of service
- Efficient matchmaking of ideal resources
- Fulfilment of Service level agreements
- Multiplicity of cloud providers
IV. LITERATURE REVIEW

Resource provisioning in cloud is an important aspect in context of quality of service (QOS) to be provided to the cloud users as well as optimal resource utilization by the cloud providers so that overall system efficiency and utilization can be achieved. Various resource provision mechanisms and techniques of cloud computing are suggested by researchers from time to time out of which some are reviewed and highlighted here in this section.

Cloud provisioning or cloud resource provisioning approach propounded by [1] is based on the three key aspects of provisioning that includes Virtual Machine (VM) Provisioning, Resource Provisioning and Application provisioning. VM provisioning is made on the basis of specific hardware and software requirements and their configurations, Resource provisioning deals with scheduling of VMs over cloud servers, and Application provisioning leads to the deployment of particular and specific application software. This paper mainly focuses on VM provisioning and Application provisioning. The objective of provisioning of cloud resources in this paper is to meet the quality of service level agreement by exploiting sufficient number of computation power, memory and storage. Quality of service required for an application is taken into account on the basis of the parameters response time and rejection rate of the requests.

Authors [1] proposed an adaptive cloud provisioning approach to deal with the dynamic workload related problems like uncertain behaviour and some estimation error. They propounded three components of the proposed provisioning technique which are Application provisioner (that will work automatically and there will be no human intervention), Workload analyser, Load predictor & performance modeller to decide and allocate the workload. Simulated scenarios are setup for implementing the proposed algorithmic technique. Various complexities of the proposed need further research to overcome these complexities and to address the problems of service level agreements for trade-offs of quality of service issues.

Workload provisioning for clouds is proposed in [2] for heterogeneous workloads on grids and clouds. The techniques proposed here is based on the clustering approach to detect patterns and trends for the optimization of virtual machines provisioning. Authors take into account the energy requirements and operating costs of cloud data centres along with the challenges of cost effectiveness and maximize efficiency of the provisioning resources of cloud. They rightly opined that inappropriate provisioning of VMs leads to under-utilization of resources. Authors proposed automatic mechanism to reduce the overprovisioning of virtual resources at two levels. Short term provisioning of VMs and long term provisioning of resources is proposed here. Authors proposed dynamic and decentralised clustering based mechanism of VM provisioning. Main focus of workload provisioning suggested by the authors is by considering overprovision of resources and lacks in consideration of under-provisioning of resources.

Jing Bi et al [3] proposed dynamic approach based cloud resource provisioning. They suggested that the proposed model can be a combination of an M/M/c model and can be based on the schema of multiple M/M/1 queuing models methods. Clustering of VMs is suggested for allocation to multi-tier application for the fulfillment of required response time of cloud users. Secondly VM configuration is important consideration to achieve the goal of optimal utilization. Main parameter used in simulated experiment to test is CPU capacity only and ignored the other parameters like memory, storage disk I/O consumption.

Zhu and Agrawal [4] propounded an algorithmic approach for resource provisioning in cloud environment by using control theory. The focus of this study is on the use of resource budget [14] along with dynamically changing factors of resource provisioning. They stressed over pricing of resources i.e. better quality of service with minimum resource budget and fine grained allocation of resources. Work is performed here using synthetic cloud environment. In the virtualized environment, cluster mechanism is also adopted for emulation. In the pricing models that have used the resources mainly CPU and memory usage that makes it more compute-intensive.

P. Padala et al [5] proposed virtualized resource allocation scheduling mechanism in their work. Their approach is based on automatic and dynamic workload change strategy that applied over shared virtualized environment to meet service level agreement of the customers. The focus of the work is to reduce the costs both infrastructure and operative and to enhance resource utilization.

Deelman et al. [6] proposed resource provisioning and allocation in virtualized environment so that cost of resource (compute and storage) allocation be reduced without affecting the application performance.
significantly. The parameters of quality of service (QOS) are important to achieve which includes reliability, availability, scalability and security. Authors used the Amazon services model for experimentation of resource provisioning. This study also focuses to examine the trade-offs in resource provisioning plans and execution and for that they posed several questions related to the goal.

Chieu et al. [7] proposed provisioning and dynamically allocation of cloud resources to the cloud users. The approach demonstrated here, focus on handling of sudden load surges over the cloud with maximum resource utilization as well the better delivery of resources. This approach also discusses the scalability capabilities of cloud. The scalability indicators comprises the number of concurrent users, active connections, requests per second and average response time. This scalability concept basically demonstrate the provisioning of required additional resource of cloud environment in an optimal usage basis.

Jung et al. [8] presented their research work which encompasses an optimization system of dynamic workload provisioning in cloud environment. The ingredients they considered are power consumption of data centres, applications performance. They also highlighted that virtual machine migrations also affects power consumption and performance. Study also focuses that resources occupied by underutilized applications can be taken away to accommodate temporary demand spikes in the cloud that leads to energy-saving.

Table-1 below shows the different areas of resource provisioning mechanisms [5,10] on the basis of which researchers mostly made their studies.

<table>
<thead>
<tr>
<th>Sr.No.</th>
<th>Various Areas of Resource Provisioning in Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QOS based resource provisioning [1,9,17]</td>
</tr>
<tr>
<td>2</td>
<td>Service Level Agreement based resource provisioning [20,22]</td>
</tr>
<tr>
<td>3</td>
<td>Adaptive based resource provisioning [13]</td>
</tr>
<tr>
<td>4</td>
<td>Optimization based resource provisioning</td>
</tr>
<tr>
<td>5</td>
<td>Bargaining based resource provisioning</td>
</tr>
<tr>
<td>6</td>
<td>Cost based resource provisioning</td>
</tr>
<tr>
<td>7</td>
<td>Energy based resource provisioning</td>
</tr>
<tr>
<td>8</td>
<td>Time based resource provisioning</td>
</tr>
<tr>
<td>9</td>
<td>Rule based resource provisioning</td>
</tr>
<tr>
<td>10</td>
<td>Dynamic resource provisioning [10,16]</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

Cloud systems are emerged as new horizon to provide infrastructure as and when pay-as-usage basis. A variety of applications are deployed over cloud infrastructure for execution. The cloud services and resources are managed and provisioned by cloud service providers. This paper presents various mechanisms that are proposed by researchers and have their pros and cons. These existing cloud resource provisioning mechanism are still away to support SLA –concerned resource provisioning [22] and workload allocation. Dynamic and automatic resource provisioning is prevailing primarily due to its advantages over static and manual provisioning. Cloud computing will definitely flourish in the days to come and need more attention of better resource provisioning for enhanced quality of service that is one of the main requirement of cloud users along with cloud security requirement so that providers can meet the rapidly changing requirements of cloud computing environments.

REFERENCES


