Quality of Service and Load Balancing in Cloud Computing: A Review
Muna M. Taher Jawhar¹,* Hanaa Mohammed Osman²

Software department, computer and Math., College, University of Mosul ¹, Computer science department, Computer and Math College, University of Mosul ²

*Corresponding author. Email: dr.muna_taher@uomosul.edu.iq¹

Abstract
Cloud computing provides facilities. These facilities increased demand for its using as institutions and individuals moved to the cloud service. Therefore, cloud service providers must provide services to users based on the expected quality. One of the main challenges presented by the cloud computing is the Quality of Service management. QoS management is defined as allocating resources to applications to ensure service based on reliability, performance and availability. It is necessary to allocate resources based on load balancing that allows avoiding overloading or low loading in virtual machines, and this is a challenge for researchers in the field of cloud computing. This research highlights the importance of cloud computing, its types and importance, It also reviews some researches in the field of quality assurance of service in computing.

Keywords: Cloud Computing, Quality of Service, Techniques of QoS, load balancing, Scheduling, Resource Allocation, Monitoring.

I. INTRODUCTION
Cloud computing is an evolution of a computer model that provides information services in a different way from the previous model. Cloud is a new step in the chain of developing communication and cloud computing technologies by introducing a new type of virtualization service [1]. Cloud computing is a model for providing all types of services to subscribing customers with less speed and effort. Cloud computing has become the most prevalent in use, as there are an enormous number of applications available and equipped by the service provider and available on secure networks. The provision of services depends on Service Level Agreements (SLA) between the service provider and the customer [1].

An example of a cloud service provider that provides this service in the market Amazon, Google, Microsoft, IBM, etc to provide cloud services such as Governance as a Service (GaaS), Business as a Service (BaaS), Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [2]. One of the most basic issues of cloud that must be taken into account is the QoS [2]. This research aims to provide an overview of QoS for cloud computers.

Since the cloud offers a variety of resources, QoS monitoring is required to evaluate the available services in order to meet customer expectations and enforce the SLA. Several issues and challenges can arise as a result of the cloud during this phase such as load balancing, performance analysis and modeling, throughput and response time, security and privacy issues, resource management, and QoS. Actually, Load Balancing in the Cloud (LBC) is one of the most challenging problems to tackle in order to prevent virtual machine overloading or underloading during task computation. As a result, it is necessary to recognize the issue that affect LBC and to establish a load balancing strategy that is successful in cloud environment.

The objectives of this paper are:
- Getting an overview of services provided by the cloud, and quality of service structure.
- Learn the general issues and challenges faced by
the cloud and focus on load balancing in the cloud.

– Shedding light on the technologies used by researchers and opening new horizons for the future.

The paper is arranged as follow, after the introduction, cloud computing overview and the next section is QoS and their techniques. Section four contains Comparison and related survey, a comparison was made between previous work in this field by studying its strengths and weaknesses. The conclusion is in section five.

1. Cloud Computing Overview

A. Definition

There are many definitions of Cloud. The most suitable one is a model that provides resources upon request such as network, applications, services, and data storage (cloud storage). The cloud can be accessed through some special applications or even some browsers on the web. [3]

B. Characteristic

The National Institute of Standards and Technology for Cloud Computing has identified five basic characteristics of the cloud:

Self-service on demand, wide network access, pooling of resources, fast flexibility and measured service. [1][4]

C. Services model of Cloud Computing

The cloud model consists of three service models that include the Infrastructure as A Service (IaaS) layer, the Platform as A Service (PaaS) and finally the programs (Software) as A Service (SaaS) and are often visualized as layers as in the figure 1, and the layers are not required to be linked. [4]

Infrastructure as a Service (IaaS): The services provided by the cloud over the internet through high-level application interfaces that are used to abstract the low-level details of the network infrastructure such as cloud physical resources, data, security, backup, storage, processing, location and network. [5]

Platform as a Service (PaaS): The cloud allows customers to create the programming languages, libraries, tools and services. It is supported by a provider that deploys in cloud-based acquired infrastructure applications. [6]

Software as a Service (SaaS): The cloud allows customers to use installed applications that run on the cloud infrastructure through different programs or interfaces from different devices, so that the user gets access to databases and applications. Examples of applications provided are games and productivity programs such as Google Docs and Word Online. SaaS applications may be connected to cloud storage or file hosting facilities, as it is the CSE with google Docs and google Drive [7]. “Database as a Service (DaaS), Expert as a Service (EaaS), Storage as a Service (SaaS), Network as a Service (NaaS), Security as a Service (SECaaS), Communication as a Service (CaaS), Monitoring as a Service (MaaS), and Testing-as-a-service (TaaS)” are some of the other resources provided by cloud computing. [8]

D. Deployment of Cloud Computing

Private Cloud: It is a cloud infrastructure that is available for use by only one organization.

Public Cloud: Public cloud infrastructure is available for use by the general public.

Community Cloud: Its infrastructure is available for use by a specific community of cloud users.

Hybrid Cloud: Its infrastructure consists of two or more different cloud models, which remain a unique entity. The following figure illustrates cloud Deployment models [1][9].
E. Issues of Cloud Computing

Since cloud technology is a modern technology, it is therefore a renewable technology that has many aspects of the issues that must be studied, most important issues included:

Quality of Service: Service quality is one of the important problems in the issue of the cloud and due to the great demand by users to use the cloud and its expansion, it was necessary to provide services to customers upon request. There are many challenges posed by QoS, including reliability and availability provided by applications and hosted by the infrastructure. Service quality is key to cloud users who expect service providers to deliver the advertised quality as per SLA. [10]

Security: Since the majority of people and organizations start using the cloud and its applications, their data and information will be stored in the database. Therefore, this information must be protected from tampering, hacking, or even theft. Security is very important to those who use the cloud. Security is an important factor in the quality of the cloud, and for this, an integrated security policy must be implemented and not leaving any gap as a weakness, and among the security policies are authentication and licensing techniques, data encryption using the highest technologies and the protection of the cloud from attacks against it. [11]

Data storage and scalability: The cloud allows large storage capacity for consumers and organizations alike, without worrying about how data is stored or copied. [12]

2. Quality of Service (QoS)

The quality of service in the cloud can be defined as the allocation of resources to different applications or users of the cloud according to agreements besides providing some characteristics such as reliability, performance and availability.

Many researchers are looking into high-quality management methods that can take advantage of modern software and hardware resources in the cloud. [13]

2.2. Techniques of QoS

Research in the field of quality of service will facilitate the researchers in this field to obtain an overview of the techniques and methods used to solve cloud problems and obtain a somewhat better quality of service than the previous one. There is also a group of packages that are used to simulate the cloud system that the researchers used in their studies and research by preparing a simulation model with virtual cloud resources as is the case in ClousSIM. Among the applications of the QoS model are:

1. Scheduling:

Scheduling in the cloud means choosing the best resources for more efficient execution of tasks. The scheduling software used in cloud computing must be satisfactory to cloud users to meet QoS requirements. The focus of most researchers in their research was on job scheduling and workflow. [14]

2. Admission Control:

The main purpose of admission control is to provide strong performance by providing an overload protection mechanism. There are two types of it: The first type is the refusal of the service provider to new requests at the time of peak load to prevent the deterioration of the service provided to users already in the system by setting a specific limit for requests. The second type of infrastructure user is the mechanism of overloading when obtaining additional cloud resources with some delay. [15]

3. Resource Management

It is the process of assigning resources available to cloud applications. The problem occurs when the service provider wants to implement a number of virtual machines on the servers. For this, resources must be managed properly to avoid problems. The proposed solutions are for resources to be managed individually by service providers or by sharing resources to reduce management costs and obtain adequate service quality. [16]

4. Monitoring mechanisms

Monitoring mechanism is the parameters tracking technology for virtual Cloud QoS. This technology helps users to maintain the operation of cloud applications with high efficiency and monitor the performance of parameters related to the quality of service. In this way, the causes of poor service quality can be detected. [16]

5. Performance models

It can help the QoS management to predict service quality. An example of a template is the waiting list model: Queue Systems, Queue Networks, Layered Queue Networks.

Researchers also use the queue model to identify load balancing and scheduling policies to support service management activities such as rollout and provisioning, as well as to minimize user waiting time and optimize power preparation for QoS content. [16]

6. Load balancing

It is a supported feature for cloud offers, where the load balancer sends a request from the user to the infrastructure provider. The load must be stabilized and managed by a cloud provider [17].

7. Capacity allocation

The service provider determines to upload applications to an appropriate number of virtual machines that must be executed on the specified physical servers. Taking into consideration the service level agreement agreed upon with the customer [18][19].
3. Cloud Load balancing

Cloud load balancing is the task of uniformly dividing workload across virtual machines in order to maximize resource utilization. Load balancing provides a way to evenly distribute work across available resources. Its goal is to continue working even in the event of failure of a component of the cloud, to improve cloud performance and reduce response time at the lowest possible cost, as well as reduce energy consumption and carbon emissions, as well as meet the requirements of service quality to improve load balance. [20]

Cloud load balancing challenges

Many researchers have become fond of cloud computing issues in terms of both theoretical and practical aspects, and load balancing has been one of the prominent challenges in cloud computing, followed by many other problems such as virtual machine “VM migration”, virtual machine “VM security”, resource use, and user satisfaction to find solutions for the purpose of improving cloud usage. Among the problems of load balancing are Geographical Distributed Nodes, Single Point of Failure, Heterogeneous Nodes, Storage Management, Load Balancer Scalability and Algorithm Complexity [21]

4. load balancing Model

The cloud allows sharing of various resources such as server, network, and data store, which necessitates a high degree of control to control customers' requests to use such resources.

Figure 3 shows a model and workflow for the load balancing. When a cloud user requests a specific service, the user base receives these requests from users in different locations and then transfers these requests to the data center to process the requests and distribute them to the physical devices associated with the data center. After the physical device receives the tasks assigned to it, it transfers it to the load balancer that distributes the tasks to the virtual machines within this device to be executed. In the event that the virtual machine is busy performing other tasks, the task will be placed in a waiting queue until the virtual machine finishes performing its tasks, it will be transferred from the queue to the virtual machine to execute it.

The load balancer is in charge of allocating the tasks to the required virtual machine. The load balancer often prevents virtual machines from being overloaded or underloaded. If certain virtual machines are empty or underloaded and others are overloaded, system performance and quality of service will deteriorate. [22]

4.1. Strategies of load balancing

Load balancing techniques are mainly classified into static and dynamic categories, as shown in Figure 4. In static type, load balancing strategies stick to a series of laws that are independent of the system’s actual state. Static algorithms are rigid in nature and depend on prior knowledge of resources such as link time, node memory and storage capacity, processing capacity, and so on. This approach is quick and straightforward, but it fails to detect connected servers, resulting in an unequal resource distribution. The only disadvantage of this strategy is that the current state of the system is not considered during decision making. As a result, it is unsuitable for rapidly changing state systems, such as distributed systems. Some examples of static technologies are “Min-Min, Max-Min, Round Robin, Shortest Job First, and Two-stage Opportunistic Load Balancing (OLB)”. [23]

The second type of load balancing technology is dynamic, these technologies take into consideration the system’s existing condition and make a recommendation based on that. The key benefit of these innovatees is that they enable tasks to be transferred from an overloaded computer to a low-load machine. Dynamic load-balancing technologies are flexible, which improves system performance. During processing, the contract load is monitored continuously, and the contract exchanges information between them in certain periods of time to check the contract load and redistribute the workload between them. Some of the pregnancy balancing techniques are agent-based pregnancy balancing, load balancing inspired by honey bee behavior, ant colony optimization, and strangulation. Other dynamic load balancing methods
can be divided into two types: distributed and undistributed. In distributed technologies, all nodes participate in distributing the loads. Either in undistributed technologies, one node or some nodes make the decision to distribute the loads. [22]

![Classification of load balancing strategies based on system state](image)

**4.2. Load balancing metrics**

There are many measures that have been suggested by researchers in load balancing techniques for the optimal use of resources and improving performance, including: [21]

- **Performance**: The effectiveness of the system must be verified after the technology has been implemented compared to other existing load balancing techniques.

- **Response time**: the cumulative amount of time it took to finish the application submitted to the framework.

- **Productivity**: the cumulative number of activities performed on the device in a given unit of time. The higher productivity, the better system performance.

- **Scalability**: The ability of the system to achieve unified load balancing when the required number of machines increases.

- **Error tolerance**: the ability to continue performance in the event of a breakdown of any link or node.

- **Posting time**: refers to the amount of time it takes to send a request / job to an overwhelmed computer. The less time it takes to migrate, the higher the machine performance.

- **Resource utilization**: ensures that all cloud services are used properly, resulting in greater resource usage and reduced energy use in the cloud.

- **Degree of imbalance**: describes the difference between VMs.

- **Limit range**: used to represent the time spent in allocating resources to users.

**5. Review previous load balancing techniques**

We conducted a study on the load balancing techniques used by the researchers and reviewed some selected papers.

<table>
<thead>
<tr>
<th>Author name/ Year of Publication</th>
<th>Techniques and algorithm</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kokilavani. 2011/ [24]</td>
<td>Static. SJF</td>
<td>The shorter task is carried out first</td>
<td>Some task be starvation. More time need</td>
</tr>
<tr>
<td>Bala. 2016/ [27]</td>
<td>Dynamic. machine learning</td>
<td>High resource used. Low migration. Overhead</td>
<td>Not tested on a real Cloud</td>
</tr>
</tbody>
</table>
Comparison of various current type of QoS

The quality of service has been of great and continuous importance to assess, analyze and predict the quality of services provided by cloud computing. That is why it is important to know recent studies on service quality and the technologies used in it to continue working to improve service quality and provide the best services by the cloud to all users.

In this paper, we made a review of previous works that have been implemented in this field, and the aim of it is to provide an overview of the current scientific research methodology, identify directions for future research, as well as identify strengths and weaknesses as shown in the following table.

Table 2. Summary of Various Types of QoS in Cloud Computing

<table>
<thead>
<tr>
<th>Author name /Year of Publication</th>
<th>Techniques and algorithm</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hershey. 2015/[31]</td>
<td>(EMMRA) and (CC) are used in cloud.</td>
<td>Increase QoS implementation. And Prevents the denial of service attack.</td>
<td>The providers are not integrated in a real time.</td>
</tr>
<tr>
<td>Zhou Zhou. 2019/[32]</td>
<td>(MGGS) modified genetic algorithm (GA) combined with greedy strategy</td>
<td>optimize task scheduling process compared to the algorithms used in this paper</td>
<td>The proposed algorithm should be compared with other scheduling algorithms to see its quality.</td>
</tr>
<tr>
<td>Panda SK. 2018/[33]</td>
<td>Use three task partitioning scheduling algorithms, CTPS, CMHTPS and CMAXMTPS.</td>
<td>The task partitioning scheduling algorithms CTPS is an online algorithm.</td>
<td>Connection time and cost were not calculated in the proposed algorithm</td>
</tr>
<tr>
<td>Khorsand R. 2017/[34]</td>
<td>constrained workflow scheduling considering run-time circumstances</td>
<td>“Use makespan, number of messages exchanged, percentage of workflows that meet the deadline and VM usage cost”</td>
<td>Algorithm does not improve QoS. having more than one parameter</td>
</tr>
<tr>
<td>Mallikarjuna B. 2018/[35]</td>
<td>Use Bee colony optimization algorithm deals with load balancing</td>
<td>The results of the proposed algorithm are relatively good compared to FCFS and DBL algorithms.</td>
<td>does not compute cost Effectiveness.</td>
</tr>
<tr>
<td>Keshanchi B. 2017/[36]</td>
<td>GA and heuristic-based HEFT</td>
<td>The algorithm was used on a proposed cloud that was modeled</td>
<td>does not focus at reliability and energy parameters</td>
</tr>
<tr>
<td>Rafieyan. 2020/[37]</td>
<td>It combines the best-worst multi criteria decision-making method (BWM), and the VIKOR in load balance</td>
<td>Reduce the makespan, VM usage cost and waiting time</td>
<td>Makespan of the algorithm is High</td>
</tr>
<tr>
<td>Fernández-Cerero. 2018/[38]</td>
<td>Resource allocation, task scheduling for the hibernation of virtual machines.</td>
<td>Reduces the energy consumption of the cloud computing system</td>
<td>Dose not take security of each task and level.</td>
</tr>
<tr>
<td>Garg. 2018/[39]</td>
<td>Task deadline. Scheduling model.</td>
<td>decreases the average energy consumption per unit work completed by the host</td>
<td>High complexity</td>
</tr>
<tr>
<td>Danlam. 2017/[40]</td>
<td>A Dynamic MultiObjective Orthogonal Taguchi Based- Cat Swarm Optimization</td>
<td>Proposed dMOOTC was able to scaled by returning better execution time, execution cost and QoS as compared with the comparison techniques.</td>
<td>The algorithm is ineffective with significant computing workloads</td>
</tr>
<tr>
<td>Atyaf Dhari. 2017/[41]</td>
<td>Task scheduling algorithms are used to achieve the load balancing and QoS</td>
<td>The proposed LBDA is more efficient than the existing algorithms</td>
<td>Not all QoS transactions are handled</td>
</tr>
<tr>
<td>Er. Manoj. 2018/[43]</td>
<td>Resource Management</td>
<td>Experience and novelty in research and their usefulness in management have been studied</td>
<td>No study and comparison of previous research and work has been done</td>
</tr>
<tr>
<td>Linlin. 2012/[44]</td>
<td>The author suggested ProfminVM, ProfRS, and ProfPD algorithms to admission control and schedule SaaS providers</td>
<td>reducing costs and improving customer satisfaction</td>
<td>There are some errors in this estimated service time due to variable virtual machines performance in the cloud.</td>
</tr>
</tbody>
</table>
7. Conclusion

The number of internet users and modern technologies have increased, and the good performance of the requirements of networks has boosted as well. Therefore, service providers have to give the best proposals and services to support the latest technologies as well as the programs for users. Many people and institutions are turned to the cloud, which delivers an effective experience to clients at a cost by the level of service, which is given by the cloud from anywhere in the world. Therefore, service providers were required to compete in providing the services provided to the consumer with the best quality and lowest cost. Many researchers have conducted their researches on finding new explanations and innovative notions to deal with the service quality in the cloud. In this research, we conducted a study to fulfill the latest research and work in the field of cloud service quality. This research tries to highlight the quality of service and load balancing in cloud computing. Thus, each of cloud computing, the quality of service, and load balance have just explained in detail. By doing so, the research aims at showing the importance of the quality of service and load balancing in cloud computing. In the case of static systems, the changing state of the system is not taken into account. Therefore, it can be used in small businesses or limited enterprises. As for the dynamic case, it is the general case used in distributed systems.

Acknowledgement

The authors would express they're thanks to college of computer and Math., University of Mosul to support this report.

References

جودة الخدمة وموازنة الحمل في الحوسبة السحابية:
ورقة مراجعة

منى محمد طاهر
hanoosman@uomosul.edu.iq  dr.muna_taher@uomosul.edu.iq

قسم علوم الحاسبات ، كلية علوم الحاسبات والرياضيات
جامعة الموصل ، الموصل ، العراق

تاريخ استلام البحث: 2021/7/25
تاريخ قبول البحث: 2021/5/25

الخلاصة
توفر الحوسبة السحابية تسهيلات مختلفة. زادت هذه التسهيلات من الطلب على استخدامها حيث انتقلت المؤسسات والأفراد إلى الخدمة السحابية. لذلك يجب على مزودي الخدمات السحابية تقديم خدمات متغيرة بناء على الخدمة المتوقعة.

تحدد التحديات الرئيسية التي تطرًحها الحوسبة السحابية هو إدارة JBQoS بأنها تخصيص الموارد. تتضمن النتائج التحقيرية على أساس المواصفات المذكورة أدناه والوايتورف من الضروري تخصيص الموارد بناء على معايير الخدمات التي تسمح للتسويق الميزاني والمتحمس المنخفض في الأجهزة الأتراكية، وهذا يمكن تحييتها في مجال الحوسبة السحابية.

الخلاصة: الحوسبة السحابية جودة الخدمة، تطبيقات الخدمة، موازنة الحمل، النقل، تخصيص الموارد، الرقمية.