Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.mosuljournals.com

Quality of Service and Load Balancing in Cloud Computing: A Review

Muna M. Taher Jawhar[1],* Hanaa Mohammed Osman[2]

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.iq[1]*

Article information

Article history:
Received : 25/5/2021
Accepted : 25/7/2021
Available online :

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.

Correspondence:
Author : Muna M. Taher Jawhar
Email: dr.muna_taher@uomosul.edu.iq

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]

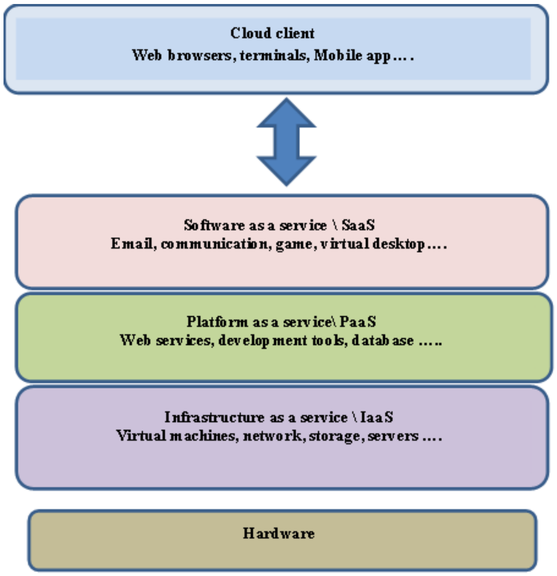
Cloud client
Web browsers, terminals, Mobile app….
Software as a service \ SaaS
Email, communication, game, virtual desktop….
Platform as a service\ PaaS
Web services, development tools, database …..
Infrastructure as a service \ IaaS
Virtual machines, network, storage, servers ….
Hardware

Figure 1: Cloud Service Models

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].

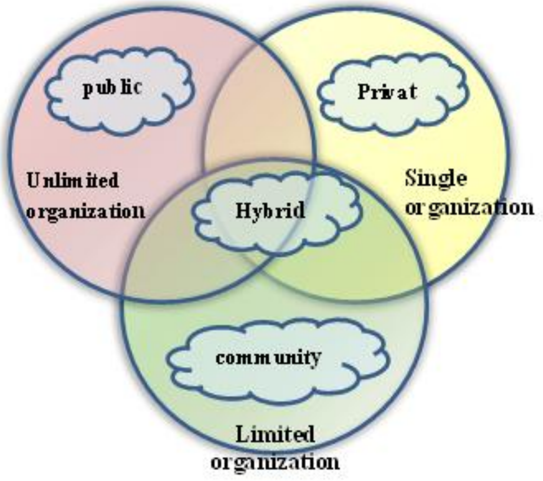

Figure 2: cloud deployment models

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]

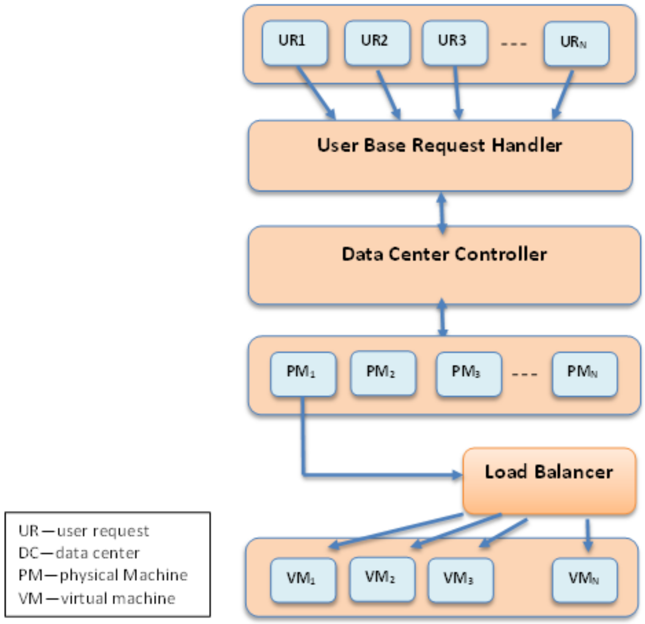

Figure 3: load balancing model

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 innovateeons 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]

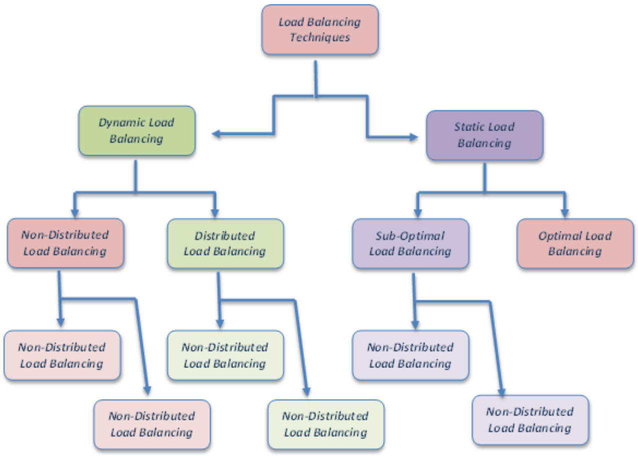
Load Balancing Techniques
Dynamic Load Balancing
Static Load Balancing
Non-Distributed Load Balancing
Distributed Load Balancing
Sub-Optimal Load Balancing
Optimal Load Balancing
Non-Distributed Load Balancing
Non-Distributed Load Balancing
Non-Distributed Load Balancing
Non-Distributed Load Balancing
Non-Distributed Load Balancing
Non-Distributed Load Balancing

Figure 4. classification of load balancing strategies based on system state

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 unite 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 1. review previous load balancing techniques

Author name/ Year of Publication	***Techniques and algorithm***	***Advantages***	***Disadvantages***
Kokilavani. 2011/ [24]	Static. SJF	The shorter task is carried out first	Some task be starvation. More time need
Pasha. 2014/[25]	Static. round-robin	useful for web servers.	Not useful for cloud environment
Cho. 2015/[26]	Dynamic. ACO with PSO	Increased resource used. Low computation time.	High cost. Homogeneous server support.
Bala. 2016/[27]	Dynamic. machine learning	High resource used. Low migration. Overhead	Not tested on a real Cloud
Babu. 2016/[28]	Dynamic. Bee colony	Response time is low. High resource. Low migration task.	Low scalability. Complexity.
Devi. 2016/[29]	Dynamic. round-robin.	Low response Time.	Homogeneous environment execution
Chen. 2017/[30]	Dynamic. Novel load-balancing.	Improved makespan and QoS	High response time. Low parameters.

6. 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

Author name /Year of Publication	*Techniques and algorithm*	*Advantages*	*Disadvantages*
Hershey. 2015/[31]	(EMMRA) and (CC) are used in cloud.	Increase QoS implementation. And Prevents the denial of service attack.	The providers are not integrated in a real time.
Zhou Zhou. 2019/[32]	(MGGS) modified genetic algorithm (GA) combined with greedy strategy	optimize task scheduling process compared to the algorithms used in this paper	The proposed algorithm should be compared with other scheduling algorithms to see its quality.
Panda SK. 2018/[33]	Use three task partitioning scheduling algorithms, CTPS, CMMTPS and CMAXMTPS.	The task partitioning scheduling algorithms CTPS is an online algorithm.	Connection time and cost were not calculated in the proposed algorithm
Khorsand R. 2017/[34]	constrained workflow scheduling considering run-time circumstances	"Use makespan, number of messages exchanged, percentage of workflows that meet the deadline and VM usage cost"	Algorithm does not improve QoS . having more than one parameter
Mallikarjuna B. 2018/[35]	Use Bee colony optimization algorithm deals with load balancing	The results of the proposed algorithm are relatively good compared to FCFS and DLB algorithms.	does not compute cost Effectiveness.
Keshanchi B. 2017/[36]	GA and heuristic-based HEFT	The algorithm was used on a proposed cloud that was modeled	does not focus at reliability and energy parameters
Rafieyan. 2020/[37]	It combines the best-worst multi criteria decision-making method (BWM), and the VIKOR in load balance	Reduce the makespan, VM usage cost and waiting time	Makespan of the algorithm is High
Fernández-Cerero. 2018/[38]	Resource allocation, task scheduling for the hibernation of virtual machines.	Reduces the energy consumption of the cloud computing system	Dose not take security of each task and level.
Garg. 2018/[39]	Task deadline. Scheduling model.	decreases the average energy consumption per unit work completed by the host	High complexity
Danlami. 2017/[40]	A Dynamic MultiObjective Orthogonal Taguchi Based- Cat Swarm Optimization	Proposed dMOOTC was able to scaled by returning better execution time, execution cost and QoS as compared with the comparison techniques.	The algorithm is ineffective with significant computing workloads
Atyaf Dhari. 2017/[41]	Task scheduling algorithms are used to achieve the load balancing and QoS	The proposed LBDA is more efficient than the existing algorithms	Not all QoS transactions are handled
Subrana. 2015/[42]	Making decision rule for middle point of data.	The proposed method applicability in cloud computing.	The restricted and complexity of the calculation of the algorithm.
Er. Manoj. 2018/[43]	Resource Management	Experience and novelty in research and their usefulness in management have been studied	No study and comparison of previous research and work has been done
Linlin. 2012/[44]	The author suggested ProfminVM, ProfRS, and ProfPD algorithms to admission control and schedule SaaS providers	reducing costs and improving customer satisfaction	There are some errors in this estimated service time due to variable virtual machines performance in the cloud.

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 fulfil 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 loud 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

[1] Mariem Jelassi, Cherif Ghazel, Leila Azzouz Saïdane, A survey on quality of service in cloud computing. Conference Paper · September 2018.

[2] Helen Anderson Akpan, B. RebeccaJeya Vadhanam, A Survey on Quality of Service in Cloud Computing, International Journal of Computer Trends and Technology (IJCTT) – volume 27 Number 1, October 2015.

[3] Shiva Prakash, Quality of Service in Cloud Computing: A Survey, International Journal of Advance Research in Science and Engineering(IJARSE), Vol. 06 Issue No. 09, September 2017, pp 296-301.

[4] Peter Mell, Timothy Grance. The NIST Definition of Cloud Computing (Technical report). National Institute of Standards and Technology: U.S. Department of Commerce.2011. doi:10.6028/NIST.SP.800-145. Special publication 800-145.

[5] Mohammad Hamdaqa, Tassos Livogiannis and Ladan Tahvildari, A Reference Model For Developing Cloud Applications , CLOSER 2011 - International Conference on Cloud Computing and Services Science.

[6] Gartner; Massimo Pezzini; Paolo Malinverno; Eric Thoo. Gartner Reference Model for Integration PaaS. Retrieved 16 January 2013.

[7] Davis Geeta and Shiva Prakash, Role of Virtualization Techniques in Cloud Computing Environment, will be published Springer book series –Advances in Intelligent Systems and Computing(AISC), 2017.

[8] Roopali Punj and Rakesh Kumar. Technological aspects of WBANs for health monitoring: A comprehensive review. Wireless Networks (2018), 1–33.

[9] Yang, C.; Huang, Q.; Li, Z.; Liu, K.; Hu, F. Big Data and cloud computing: innovation opportunities and challenges. 2017, International Journal of Digital Earth.

[10] Danilo Ardagna, Giuliano Casale, Michele Ciavotta, Juan F Pérez and Weikun Wang, Quality-of-service in cloud computing: modeling techniques and their applications, Journal of Internet Services and Applications 2014, 5:11.

[11] Hashem H. Ramadan and Dr. Divya Kashyap, Quality of Service (QoS) in Cloud Computing, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 8 (3), 2017, 318-320.

[12] Kzxk B. P. Rimal, A taxonomy and Survey of cloud Computing systems, Fifth International Joint Conference on INC , IMS and IDC, 2009.

[13] Juhi Singh, Shalini Agarwal and Jayant Mishra, A Review: Towards Quality of Service in Cloud Computing, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2015): 78.96 | Impact Factor : 6.391.

[14] Snehal A.Narale, P.K.Butey, Cloud Computing Techniques to meet QoS, IJACEN December 2015.

[15] D. Ardagna, Quality-of-service in cloud computing:modeling techniques and their applications, Journal of Internet Services and Applications, 2014.

[16] Abdelzahir Abdelmaboud, Quality of service approaches in cloud computing: Asystematic mapping study, The Journal of Systems and and Software, no. 101, 2015.

[17] Isaac Odun-Ayo, Member, IAENG, Olasupo Ajayi, and Adesola Falade, Cloud Computing and Quality of Service: Issues and Developments, Proceedings of the International MultiConference of Engineers and Computer Scientists 2018 Vol I IMECS 2018, March 14-16, 2018, Hong Kong.

[18] Massimo Coppola, Emanuele Carlini, Daniele D'Agostino, Jörn Altmann and José Ángel Bañares, Economics of Grids, Clouds, Systems, and Services, 15th International Conference, GECON 2018, Pisa, Italy, September 18–20, 2018, Proceedings, Part of the Lecture Notes in Computer Science book series (LNCS, volume 11113), 2018.

[19] Deepshikha and Dr. Shiva Prakash, A Survey On Qos In Cloud Computing Environment, Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019) IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4.

[20] Aditya Bhardwaj and Challa RamaKrishna. 2018. Efficient multistage bandwidth allocation technique for virtual machine migration in cloud computing. Journal of Intelligent & Fuzzy Systems 36 (2018), 1–14. DOI:https://doi.org/10.3233/JIFS-169819

[21] Pawan Kumar and Rakesh Kumar, Issues and Challenges of Load Balancing Techniques in Cloud Computing: A Survey. ACM Computing Surveys, Vol. 51, No. 6, Article 120. Publication date: February 2019.

[22] Shang liang Chen, Yun yao Chen, and Suang hong Kuo. 2017. CLB: A novel load balancing architecture and algorithm for cloud services. Computers and Electrical Engineering 58 (2017), 154–160. DOI:https://doi.org/10.1016/j.compeleceng.2016.01.029

[23] Aarti Singh, Dimple Juneja, and Manisha Malhotra. 2015. Autonomous agent based load balancing algorithm in cloud computing. Procedia Computer Science 45 (2015), 832–841.

[24] T. Kokilavani and D. I. George Amalarethinam. 2011. Load balanced min-min algorithm for static meta-task scheduling in grid computing. International Journal of Computer Applications 20, 2 (2011), 43–49.

[25] Nusrat Pasha, Amit Agarwal, and Ravi Rastogi. 2014. Round robin approach for VM load balancing algorithm in cloud computing

environment. International Journal of Advanced Research in Computer Science and Software Engineering 4, 5 (2014), 34–39.

[26] Keng-Mao Cho, Pang-Wei Tsai, Chun-Wei Tsai, and Chu-Sing Yang. 2015. A hybrid meta-heuristic algorithm for VM scheduling with load balancing in cloud computing. Neural Computing and Applications 26, 6 (2015), 1297–1309. DOI:https://doi.org/10.1007/s00521-014-1804-9

[27] Anju Bala and Inderveer Chana. 2016. Prediction-based proactive load balancing approach through VM migration. Engineering with Computers 32, 4 (2016), 1–12. DOI:https://doi.org/10.1007/s00366-016-0434-5

[28] Remesh Babu and Philip Samuel. 2016. Enhanced bee colony algorithm for efficient load balancing and scheduling in cloud. Innovation in Bio-Inspired Computing and Application 4 (2016), 135–142. DOI: https://doi.org/10.1007/978-3-319-28031-8_6

[29] Chitra Devi and Rhymend Uthariaraj. 2016. Load balancing in cloud computing environment using improved weighted round robin algorithm for nonpreemptive dependent tasks. The Scientific World Journal 2016 (2016), 1–14. DOI:https:// doi.org/10.1155/2016/3896065

[30] Shang liang Chen, Yun yao Chen, and Suang hong Kuo. 2017. CLB: A novel load balancing architecture and algorithm for cloud services. Computers and Electrical Engineering 58 (2017), 154–160. DOI:https://doi.org/10.1016/j.compeleceng. 2016.01.029

[31] P. C. Hershey, S. Rao, C. B. Silio and A. Narayan, System of systems for Quality-of-Service observation and response in cloud computing environment, IEEE Systems Journal, Volume: 9, Issue:1, pp. 1-5, 2015.

[32] Zhou Zhou, Fangmin Li, Huaxi Zhu, Houliang Xie, Jemal H. Abawajy and Morshed U. Chowdhury, An improved genetic algorithm using greedy strategy toward task scheduling optimization in cloud environments, Neural Computing and Applications, Springer-Verlag London Ltd., part of Springer Nature 2019.

[33] Panda SK, Pande SK, Das S. Task partitioning scheduling algorithms for heterogeneous multi-cloud environment. Arabian Journal for Science and Engineering. 2018 Feb 1;43(2):913-933.

[34] Khorsand R, Safi-Esfahani F, Nematbakhsh N, Mohsenzade M. ATSDS: adaptive two-stage deadline-constrained workflow scheduling considering run-time circumstances in cloud computing environments. The Journal of Supercomputing. 2017 Jun 1;73(6):2430-2455.

[35] Mallikarjuna B, Krishna PV. A nature inspired bee colony optimization model for improving load balancing in cloud computing. International Journal of Innovative Technology and Exploring Engineering (IJITEE). December, 2018;8(2S2):51–54.

[36] Keshanchi B, Souri A, Navimipour NJ. An improved genetic algorithm for task scheduling in the cloud environments using the priority queues: formal verification, simulation, and statistical testing. Journal of Systems and Software. 2017 Feb 1; 124:1-21.

[37] Rafieyan E, Khorsand R, Ramezanpour M. An adaptive scheduling approach based on integrated best-worst and VIKOR for cloud computing. Computers & Industrial Engineering. 2020 Jan; 8:106272–106293.

[38] Fernández-Cerero D, Jakóbik A, Grzonka D, Kołodziej J, Fernández-Montes A. Security supportive energy-aware scheduling and energy policies for cloud environments. Journal of Parallel and Distributed Computing. 2018 Sep 1; 119:191-202.

[39] Garg N, Goraya MS. Task deadline-aware energy-efficient scheduling model for a virtualized cloud. Arabian Journal for Science and Engineering. 2018 Feb 1;43(2):829-841.

[40] Danlami Gabi, Abdul Samad Ismail, Anazida Zainal and Zalmiyah Zakaria, Quality of Service (QoS) Task Scheduling Algorithm with Taguchi Orthogonal Approach for Cloud Computing Environment, International Conference of Reliable Information and Communication Technology, IRICT 2017: Recent Trends in Information and Communication Technology pp 641-649, , DOI 10.1007/978-3-319-59427-9_66.

[41] Atyaf Dhari and Khaldun I. Arif, An Efficient Load Balancing Scheme for Cloud Computing, Indian Journal of Science and Technology, Vol 10(11), DOI: 10.17485/ijst/2017/v10i11/110107, March 2017.

[42] C. Subarna, Optimal Data Center Scheduling for Quality of Service Management in Sensor-cloud, IEEE Transactions on Cloud Computing, 2015. IEEE Transactions on Cloud Computing PP(99):1-1, DOI:10.1109/TCC.2015.2487973.

[43] Er. Manoj Kumar, Cloud Computing in Resource Management, International Journal of Engineering and Management Research, ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962, Volume-8, Issue-6, December 2018, DOI: doi.org/10.31033/ijemr.8.6.8.

Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya, SLA-based admission control for a Software-as-a-Service provider in Cloud computing environments, Journal of Computer and System Sciences, Volume 78, Issue 5, September 2012, Pages 1280-1290.

جودة الخدمة وموازنة الحمل في الحوسبة السحابية:

ورقة مراجعة

منى محمد طاهر **هناء محمد عصمان**

dr.muna_taher@uomosul.edu.iq hanaosman@uomosul.edu.iq

قسم علوم الحاسبات ، كلية علوم الحاسوب والرياضيات

جامعة الموصل، الموصل ، العراق

تاريخ استلام البحث2021/5/25 تاريخ قبول البحث 2021/7/25

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 الخدمات للمستخدمين بناءً على الجودة المتوقعة. أحد التحديات الرئيسية التي تطرحها الحوسبة السحابية هو إدارة جودة الخدمة. تعرف إدارة **QoS** بأنها تخصيص الموارد للتطبيقات لضمان الخدمة على أساس الموثوقية والأداء والتوافر. من الضروري تخصيص الموارد بناءً على موازنة الحمل التي تسمح بتجنب التحميل الزائد أو التحميل المنخفض في الأجهزة الافتراضية ، وهذا يمثل تحديًا للباحثين في مجال الحوسبة السحابية. يسلط هذا البحث الضوء على أهمية الحوسبة السحابية وأنواعها وأهميتها ، كما يستعرض بعض الأبحاث في مجال ضمان جودة الخدمة في الحوسبة.

الكلمة المفتاحية : الحوسبة السحابية ، جودة الخدمة ، تقنيات جودة الخدمة ، موازنة الحمل ، الجدولة ، تخصيص الموارد ، المراقبة.