CLOUD LOAD BALANCING

Abstract

Cloud computing has become integral to modern IT, with major companies like Google, Microsoft, IBM, and Amazon providing various cloud services to users. This paper explores the significance and importances of load balancing in cloud computing, a crucial aspect of maintaining efficient and responsive cloud services. It delves into different load-balancing algorithms and mechanisms, offering a comprehensive understanding of the subject matter through charts, graphs, and extensive research analysis.

The study begins by outlining the core concepts of cloud computing, which includes Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). It emphasizes the importance of cloud service providers (CSPs).

The paper then shifts its focus to load balancing, a critical component of cloud maintenance. It addresses the challenges of overloading within the cloud infrastructure and introduces Load Balancing as a Service (LBaaS) as a solution. Static load balancing algorithms like Round Robin and Weighted Round Robin are discussed in detail, highlighting their approach to distributing loads among servers. Dynamic load balancing algorithms, such as Min-Min and Max-Min, are also explored, each with its distinct advantages and limitations.

Furthermore, the paper delves into additional load balancing techniques like Logical Ring Redirection, Load Buffer Range Method, Random Early Detection Method, and Page Caching. Distributed Web Server (DWS) is presented as a reliable solution for managing increasing loads, with its architecture and components outlined.

Author

Prabhmilan Singh
Department of Computer Science and Engineering
Amity University
Noida, Uttar Pradesh, India
Prabh20021@gmail.com

In conclusion, this paper offers a comprehensive insight of load balancing in cloud servers, addressing issues of overloading and under loading servers. It discusses the rationale behind various algorithms and techniques. The study encourages further exploration of load-balancing solutions and highlights the need for algorithm complexity consideration and real-time implementation testing in future research and development efforts.

Keywords: Cloud computing, Google, Microsoft, IBM, and Amazon.

I. INTRODUCTION

In recent decades the use of Cloud Computing has increased a lot. Big IT companies like Google, Microsoft, IBM, and Amazon have been providing a lot of cloud services to the end-users. Benefits like online storage, easy access to data anytime you want, and sharing of information much easier than before were some points that attracted end-users towards it. A simple definition of cloud computing is that it's a platform where one can store, protect, access, and do a lot more with their data, with the help of the services offered by the cloud. The technology can be used by different businesses and by individuals at the same time, providing amazing benefits. Cloud computing helps in cost reduction, increasing the flexibility of the business, access to automatic updates, etc.

The figure of a cloud is not something new to us, the internet from its preliminary days has been depicted as a cloud that can give you access to data across the whole world and the concept of Cloud Computing is not very different from it. The software helps its end-users to share, store, and access data using modern networking solutions by deploying hardware and software available. Cloud computing has 3 models related to it – SaaS known as Software as a Service, IaaS known as Infrastructure as a Service, and PaaS known as Platform as a Service.

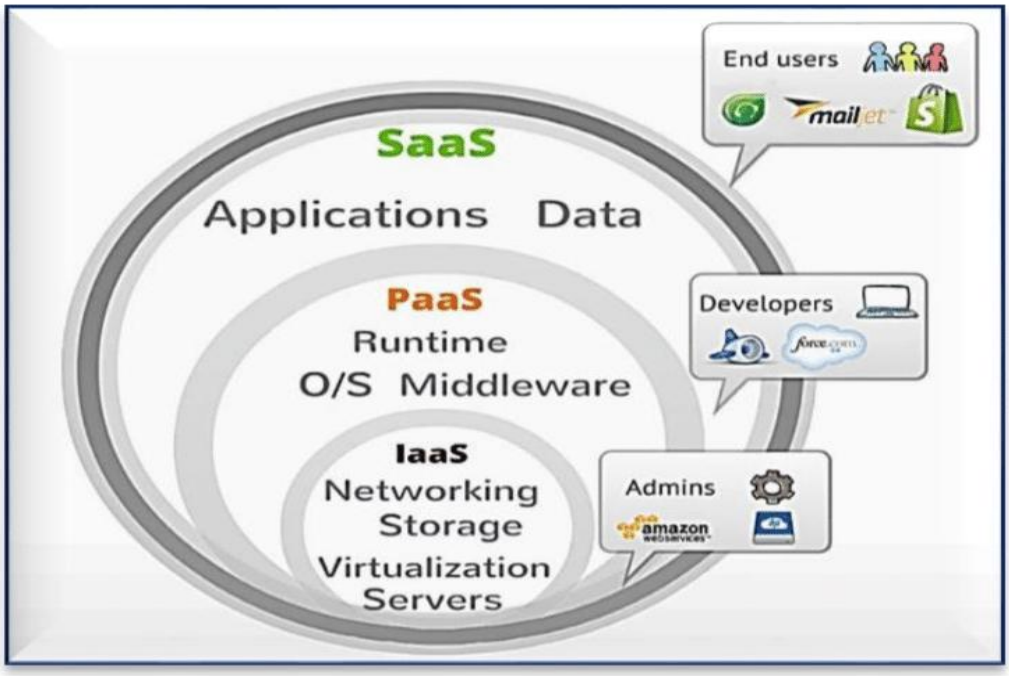

Figure 1: Different Components of Cloud Computing

1. **Software as a Service (SaaS):** In software as a service (or SaaS), a cloud service provider hosts the software to the end-users. The company offering SAAS has its own infrastructure and only provides the software to the end users. An Independent Service Provider sometimes may hire a third party in order to host its software over the internet. Sometimes, companies like Google, IBM, and Amazon may also charge a fee in the form of a monthly subscription.

2. **Platform as a Service (PaaS):** Paas is a part of cloud computing in which the end-user can run, initiate, and develop without being tensed about the cost of inflexibility, which results in building your own platform. The architecture of PaaS consists of storage, servers, and networking which enables the end to use the services by the CSP (Cloud Service Provider) without any complexity. Some examples of PaaS are – Cloud Foundry and Red Hat Open Shift

3. **Infrastructure as a Service (IaaS):** IaaS is a platform that offers you networking, computing, and storage solutions. You can store and access your data without bothering about the costs of database management systems. IaaS provides you with online storage, sometimes paid and sometimes free (depending on the Cloud Service Provider), cutting off the need to physically buy storage solutions and carrythem with them to access the data in it. The model is also helpful in providing you with real-time updates and hassle-free access to data.

Cloud computing is anon-demand paid service. Cloud Service Providers (CSP) like Apple, Microsoft, IBM, etc. charge a specific amount of fees in order to offer end users the services discussed. The fees could be an orderly monthly subscription or a one-time payment. CSPs have invested a lot of money in database management systems. They have big infrastructures where all the data uploaded by an end user is stored.

Cloud Computing is divided into 2 parts. The first is the provider of services which defines the service provided by the CSP. Models like SaaS, IaaS, and PaaS come under the same heading. The second one is based on the size and capabilities of the algorithm. NIST (National Institute of Standards and Technology) named four models under this heading, known as public, hybrid, private, and community clouds.

The execution and provision of so many services to the user, and maintenance of the cloud are also very important. Providing a hassle-free frontend and backend experience to the end-user has also to be taken care of. When many end users send their requests to the central cloud, the chances of overloading increase. Overloading can result in delays in accepting requests. To prevent this, load balancing mechanisms come into the picture. Unbalancing of servers can bring down the performance and efficiency of the service provided by the Customer Service provider. It may also result in not matching the threshold Quality of Service (QoS) in the Service Level Agreement (SLA) between the consumer and the CSP. The service of balancing the overloaded servers is known as load balancing also known as (lbaas)- Load Balancing as a Service.

In this paper, we are going to study the uses of load balancing in cloud computing. We will also study the different algorithms and mechanisms used in the process. Charts and graphs have been used in order to give you the best understanding of the concept. Many research papers have been revised in order to cover all the techniques and algorithms used in load balancing. The main aim of the paper is to provide the reader with a piece of in-depth knowledge about cloud load balancing using the pre-existing research done across the globe. Different types of algorithms used in load balancing are stated in such a manner that it should be very easy for readers to classify and understand the contrast between them. I hope that after reading the following paper you will have a better

understanding of why load balancing is important, how is it done, and the different types of techniques used in it.

II. CLOUD LOAD BALANCING

Cloud load balancing means allocating the overall load, to the central cloud or VM (Virtual machine), to different nodes using various kinds of algorithms. Whenever a bunch of requests lands over the Virtual Machine (VM), it gets exhausted or overloaded and, as a result, is not able to respond to other incoming requests. This results in decreasing the efficiency, speed, and performance of the cloud system.

In some cases, overloading the server can also cause damage to the data stored in the databases of the cloud service providers. Load balancing is an optimization technique that is used to improve parameters like system stability, response time, task execution, etc. Many researchers around the globe for a long time now have shown their concern on the topic. Past research has been revolving around topics like data management, improved response time, and improved task execution time.

Cloud load balancing also helps enterprises to manage their resources by distributing them to a variety of resources. It has also been very useful in cost reduction in relation to Document Management Systems and maximizes the availability of resources. LBaaS (Load Balancing as a Service) has also become an important model in cloud computing for businesses using the technology.

III. CLASSIFICATIONS IN CLOUD LOAD BALANCING

Cloud load balancing works on the principle of type-specific algorithms to make decisions in order to maintain the overall load. It makes decisions about how the load must be distributed among the servers to prevent any degradation in the overall performance. According to the system, the algorithm can be separated into two types: -

1. Static load balancing algorithms
2. Dynamic load balancing algorithms

In static load balancing the previous state of the system is not taken into picture while distributing the overall load. Whereas, in dynamic balancing the previous state is also considered when balancing the load. The dynamic and static algorithm are further divided into their subcategories listed below.

1. Static load balancing algorithms:

- **Round Robin:** Round-robin load balancing is just concerned with the availability of the servers to pass the overall load. A DNS server has all the distinctive IP addresses, from which, when requested an IP address, related to a domain name, is received then the addresses are returned in a rotating manner. The concept would be clearer from the following diagram:

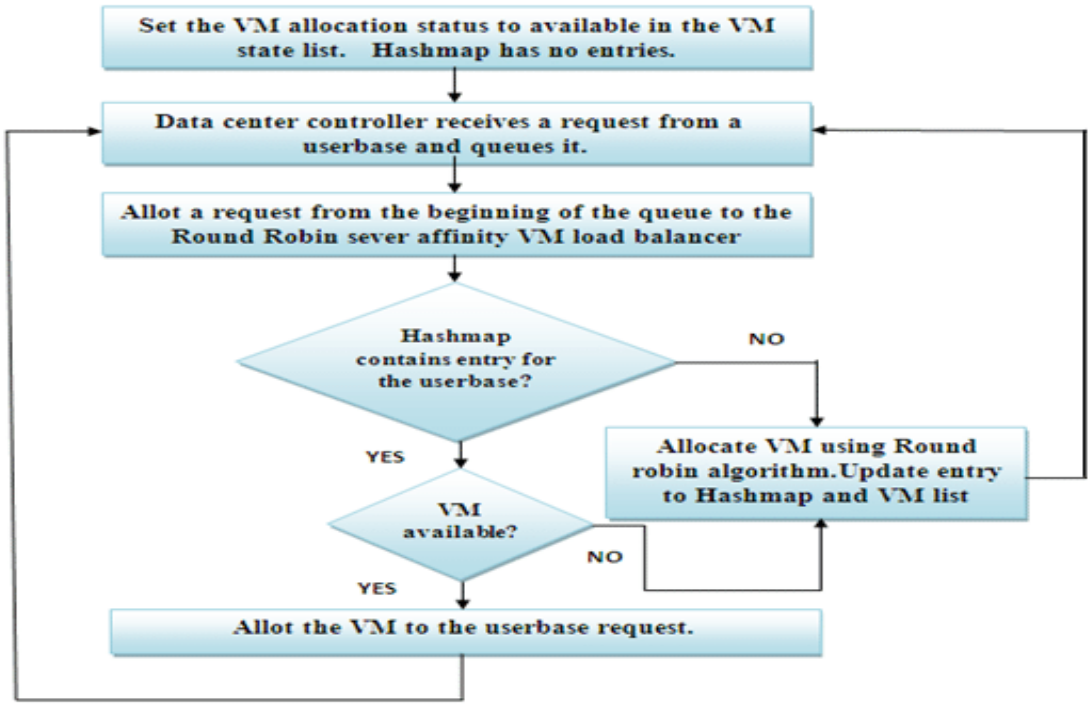

Figure 2: Round Robin Load Balancing

- **Weighted Round Robin method:** In weighted round-robin load balancing, the administrator can distribute the overall load amongst the pre-existing servers according to their capacity of holding load. The basic difference between Weighted Round Robin and Round robin balancing is that in round robin the load is distributed in an equal ratio among all the servers whereas in weighted round robin, the administrator can distribute the overall load UNEQUALLY among the servers, according to their capacity to handle the total load. The concept could be understood even better from the figure attached below: -

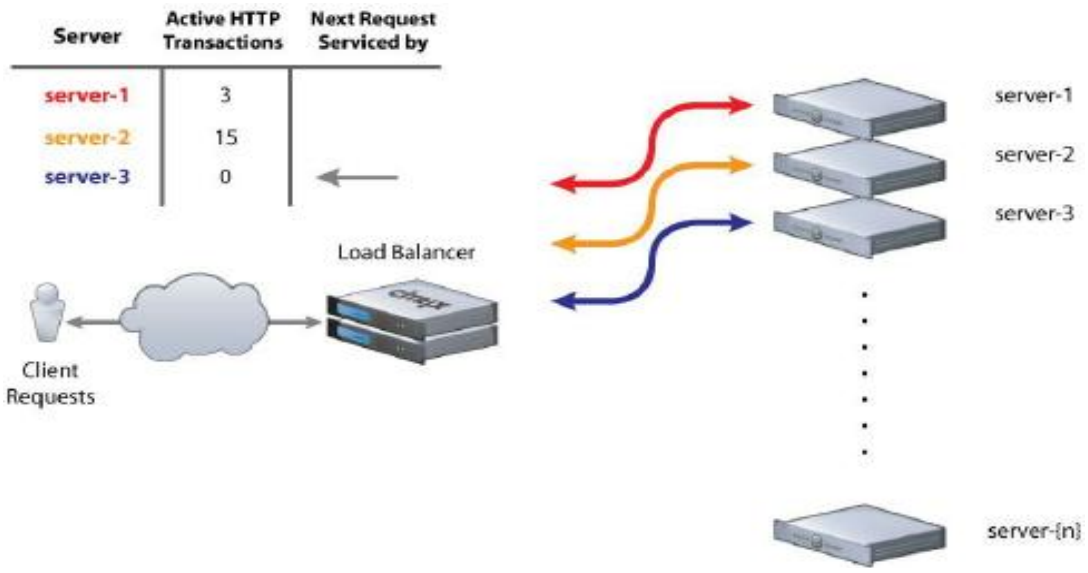

Figure 3: Weighted Round Robin Load Balancing

2. Dynamic Load Balancing:

- **Min-Min Load Balancing:** The objective of min-min load balancing is to perform the assigned tasks in the minimum time possible. The initial requests from the customer are forwarded to the server which has a minimum response, execution, and completion time. The same practice is followed until all requests are responded to.

  The limitation of min-min load balancing is that it prioritizes the smaller tasks over the larger ones. Hence, the schedule does not work out when the number of smaller tasks exceeds the number of larger ones. To overcome this picture, max-min load balancing algorithms come into the picture. Many algorithms and techniques have been put forward in the past few decades. All proposed algorithms were analyzed very carefully. Hence, an algorithm was found, which resulted in minimum execution time and stable schedule architecture. Initially, the algorithm executes a minimum-minimum load balancing algorithm and then rebuilds the schedule while keeping the minimum response time less than the previous cycle.

- **Max-Min load balancing:** Max-min load balancing works on the principle of prioritizing the larger tasks over the smaller ones. Max-min algorithm mostly works in situations with unscheduled tasks. It calculates the total time of completion of every task and then assigns it to the server with the least time of execution. The same process is repeated again until all the tasks assigned are completed.

  The max-min algorithm performs better than the min-min load balancing algorithms under circumstances when thenumber of smaller tasks is more than the number of larger tasks as min-min is unable to balance the load properly in such situations. The concept could be better acknowledged from the figure below: -

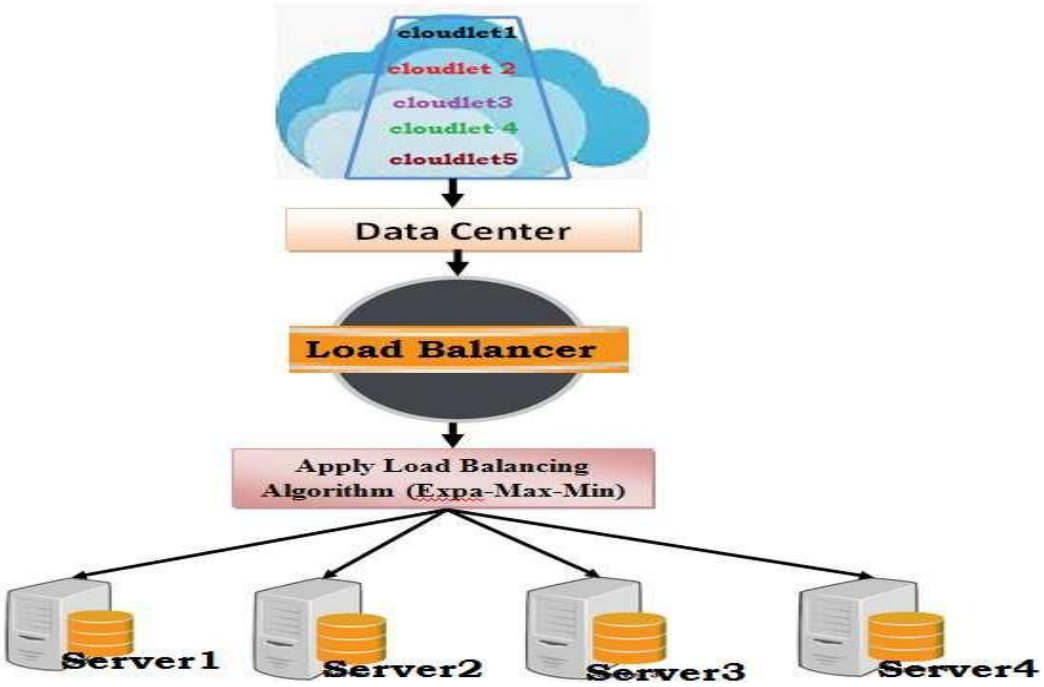

Figure 4: Max Min is applied to the Incoming Load

Hence, the two very important components of load balancing in the zone of cloud computing, dynamic load balancing and static loading balancing, play essential roles on their own parts in order to reduce the total load inthe cloud by distributing and allocating the load according to their respective methods and provide the customer with a hassle-free and smooth experience.

Some more types of cloud load balancing are: -

- **Logical ring redirection:** Web server is assembled in various logical rings in order to gain load information on the servers and solve overload or underload using specific techniques.

- **Load Buffer Range Method:** The web servers are not supposed to send their state information to the DNS again and again so that it does not crash because of overloading. So, to avoid this load buffer range which consists of the high and low threshold for each web server turns out to be very helpful.

- **Random Early Detection Method:** Whenever a server experiences overload or under load, it is not calculated in indefinite numbers but in probability. Therefore, the calculation of overload or under load, probabilistically, is done by the Random Early Detection Method (REDM). REDM or Random Early Detection Method is an algorithm that calculates the ratio of the total number of non-answered requests (due to overloading) to the total number of pending requests.

- **Page Caching:** Page caching takes advantage of the reference area and network traffic can be reduced and possibly reduce over-delays. In addition, it can also enhance the performance of uploads to a DNS-based web server system through appropriate temporary storage schemes.

3. **Distributed Web Server:** With an increasing number of end-users and servers on the web, the problem of overloading has become a topic of concern over the past decade. Many different companies, over the past decade, have been able to come up with many different techniques to overcome the problem.

There are many cloud load balancers in the market right now. Different companies use different load balancers to balance the overall load of their servers. Among all the load balancing options available DWS (Distributed Web Server) is amongst the most recognized and significant approaches to the task. It is among the most reliable solutions because of its distribution processes and load-handling capacity. The components of a distributed web server contains three components:

- The End-User
- Domain Name System (DNS).
- Web server

A distributed web-server system can be configured in multiple web servers and DNS cluster, which resolves all address resolution requests from local portals. Everyclient's time can be determined by resolving one address and multiple web server requests together. Initially, the end-user receives a web address for a group via a DNS address solution. In the meantime, the user sends many HTTP requests to the server. In

addition to resolving a URL in a web server's IP address, the DNS of a web server distributed can collect data on web servers having various elements. DNS can select a web server address depending on the collected data. In order to select the appropriate server address, DNS may apply a specific load-balancing technique between multiple servers to avoid overloading.

4. **DNS Load Balancing:** DNS load balancing is a practice of setting up a Domain Name System (DNS) while the end user's requests are distributed among a bunch of servers. The domain may be linked to a website, mail sending/receiving system, print machine server, or other services provided by the Internet. This helps instant access by providing multiple IP addresses for a single host name, which is responsible for distributing the requests/load coming from the end-users.

   DNS-based load balancing helps in improving the customer requests for specific domains. It can comprise number of strategies that can be used to distribute or manage, redirect, or manage the total load to give the client the best experience possible. DNS load balancing uses different strategies to perform load balancing. Some of them are listed below: -

   - **Setting up a backup server:** A secondary DNS is setup so that the main DNS server can distribute the incoming load from the end-user. Hence, it helps in managing the total load on the server.

   - **Round Robin DNS load balancing:** All the requests are rotated across the available non-exhausted server, so is the load distributed. Though it is primarily a load-balancing algorithm, it works well with DNS also.

   - **Dynamic load balancing with DNS:** Requests sent from the end-user are diverted towards the server with minimum load and best resources. Therefore, the speed of task completion is increased and the system functions smoothly.

IV. CONCLUSION

The paper consists of a brief study of the topic of the techniques of load balancing in cloud computing. It discusses the problem causing overloading or under loading a server. Different measures to be applied with the objective of solving the overloading and under loading of a server have been discussed. Complete reasoning and logic have been provided for all the algorithms and techniques used along with the reason to use them. Several research papers were studied which will be mentioned in the reference section of the paper. All the data reviewed was published between 2014-2023 which makes this study up to date.

The study focused on load balancing in cloud computing to update the existing technologies and inspire the readers to come forward with more load-balancing solutions. Further, the study has disclosed that algorithm complexity is largely ignored in deciding the performance of all the load-balancing algorithms, and as a result, 80% jobs do not consider performance testing. The real-time implementation of the load balancing is very small and should be discussed during future activities.

One of the noteworthy contributions of this paper is its focus on Distributed Web Server (DWS) architecture, which proves to be a reliable and efficient approach to load balancing in cloud servers. By explaining the components and operations of DWS, the paper highlights its potential to mitigate server overloads and ensure seamless end-user experiences.

The exploration of DNS-based load balancing, including strategies like backup server setup, Round Robin DNS load balancing, and dynamic load balancing with DNS, underscores the multifaceted nature of load distribution techniques available to cloud service providers.

In summary, this paper serves as are source for individuals, businesses, and researchers who are interested in the dynamic field of cloud computing. It provides a comprehensive overview of load-balancing techniques and encourages further exploration and innovation in this critical area. By citing recent research and emphasizing the need for real-time implementation and performance testing, the paper offers a solid foundation for future endeavours in enhancing cloud load-balancing solutions. Overall, it helps in the growing body of knowledge that is essential for harnessing the full potential of cloud computing in the digital age.

REFERENCES

[1] Hierarchy load balancing - Afzal, S., Kavitha, G. Load balancing in cloud computing – A hierarchical taxonomical classification. J Cloud Comp 8, 22 (2019). https://doi.org/10.1186/s13677-019-0146-7
[2] Load balancing in cloud computing: A big picture – Sambit Kumar Mishra, Bibhudatta Sahoo, Priti Paramita Parida – 15/01/2018 https://doi.org/10.1016/j.jksuci.2018.01.003
[3] A Predictive Load Balancing Algorithm in Cloud Services – Mahdee Jodayree, Mahmoud Abaza, Qing Tan – 14/10/2019 https://doi.org/10.1016/j.procs.2019.09.250
[4] EWPTNN: An Efficient Workload Prediction Model in Cloud Computing Using Two-Stage Neural Networks- K.Dinesh Kumar, E. Umamaheshwari – 27/02/2020
[5] R. B. Bohn, J. Messina, F. Liu, J. Tong, and J. Mao, "NIST Cloud Computing Reference Architecture," 2011 IEEE World Congress on Services, 2011, pp. 594-596, doi: 10.1109/SERVICES.2011.105.
[6] N. K. Chien, N. H. Son, and H. Dac Loc, "Load balancing algorithm based on estimating finish time of services in cloud computing," 2016 18th International Conference on Advanced Communication Technology (ICACT), 2016, pp. 228-233, doi: 10.1109/ICACT.2016.7423340.
[7] Figure 1: Shukur, Hanan &Zeebaree, Subhi & Zebari, Rizgar &Zeebaree, Qader & Ahmed, Omar & Salih, Azar. (2020). Cloud Computing Virtualization of Resources Allocation for Distributed Systems. Journal of Applied Science and Technology Trends. 1. 98-105. 10.38094/jastt1331.
[8] Figure 2: Mahajan, Komal &Makroo, Ansuyia& Dahiya, Deepak. (2013). Round Robin with Server Affinity: A VM Load Balancing Algorithm for Cloud Based Infrastructure. Journal of Information Processing Systems. 9. 10.3745/JIPS.2013.9.3.379.
[9] Figure 3: Rahman, Mazedur& Iqbal, Samira & Gao, Jerry. (2015). Load-Balancer-as-a-Service-in-Cloud-Computing-v7.
[10] Figure 4: Haladu, Mubarak & Samual, Joshua. (2016). Optimizing Task Scheduling and Resource allocation in Cloud Data Center, using Enhanced Min-Min Algorithm. IOSR Journal of Computer Engineering. 18. 10.9790/0661-1804061825.
[11] Konjaang, James & Ayob, Hakim & Muhammed, Abdullah. (2018). Cost Effective Expa-Max-Min Scientific Workflow Allocation and Load Balancing Strategy in Cloud Computing. Journal of Computer Science. Volume 14, 2018. 10.3844/jcssp.2018.623.638.

