Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Distributed Systems for Real-Time Computing in Cloud Environment: A Review of Low-Latency and Time Sensitive Applications

Nisreen Luqman Abdulnabi[1]*, Subhi R. M. Zeebaree[2]
nisreen.nabi@dpu.edu.krd, subhi.rafeeq@dpu.edu.krd
[1]ITM Dept., Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq,
[2]Energy Eng. Dept., Technical College of Engineering, Duhok Polytechnic University, Duhok, Iraq,

Article Information

Submitted : 8 Mar 2024
Reviewed: 14 Mar 2024
Accepted : 8 Apr 2024

Keywords

Cloud computing, load balancing, load balancing algorithms, Low Latency, cloud analyst

Abstract

As a result of its many benefits, including cost-efficiency, speed, effectiveness, greater performance, and increased security, cloud computing has seen a boom in popularity in recent years. This trend has attracted both consumers and businesses. Being able to process and provide data or services in a quick and effective manner while adhering to low latency and time limits is the hallmark of an efficient distributed system that is designed particularly for real-time computing in cloud environments. It is essential to place a high priority on low latency and time sensitivity while developing and putting into action a distributed system for real-time computing in a cloud environment. In order to fulfil the particular requirements of the application or service, consideration must be given to a number of different aspects. In particular, the topic of load balancing will be discussed in this paper. It is possible to ensure a more effective distribution of workload and reduce latency by using load balancers, which distribute incoming traffic over many servers or instances. The throttled algorithm is believed to be the most efficient load balancing strategy for reducing service delivery delay in cloud computing. This research investigates a hybrid method known as Equally Spread Current Execution (ESCE), which is known for its combination with the throttled algorithm.

A. Introduction

The provision of computer services via the internet, like databases, servers, networking, storage, and software, is known as cloud computing (CC)[1]. Cloud computing promotes higher cost effectiveness, quicker innovation, and more flexible resource usage. One of the greatest technologies for information technology services and problem solving [2] . Virtualization technology, which is used in cloud computing, offers end users a wide range of services, including physical resources and application levels [3]. In addition, cloud computing offers other characteristics that can help scientists, such as the ability to scale or reduce the computer infrastructure based on the needs of applications and the user's budget. But now that cloud computing technology is available, we can offer a lot of dispersed infrastructures with a good setting [4]. The customer usually only pays for the cloud services used, which helps save money and runs the infrastructure more efficiently [5], as shown in the figure)1).

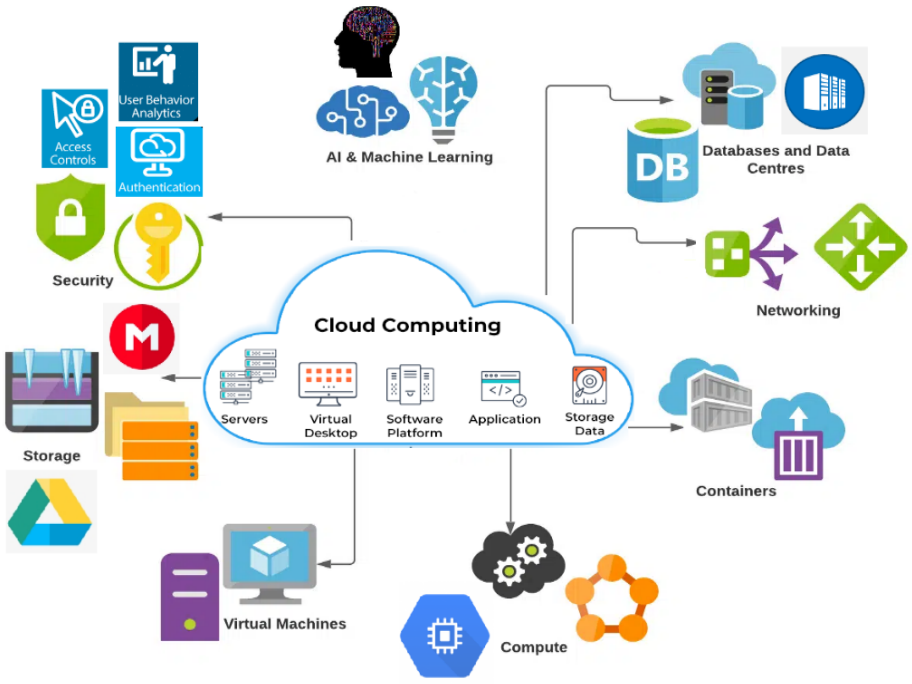


Figure 1. Cloud Computing Architecte [6]

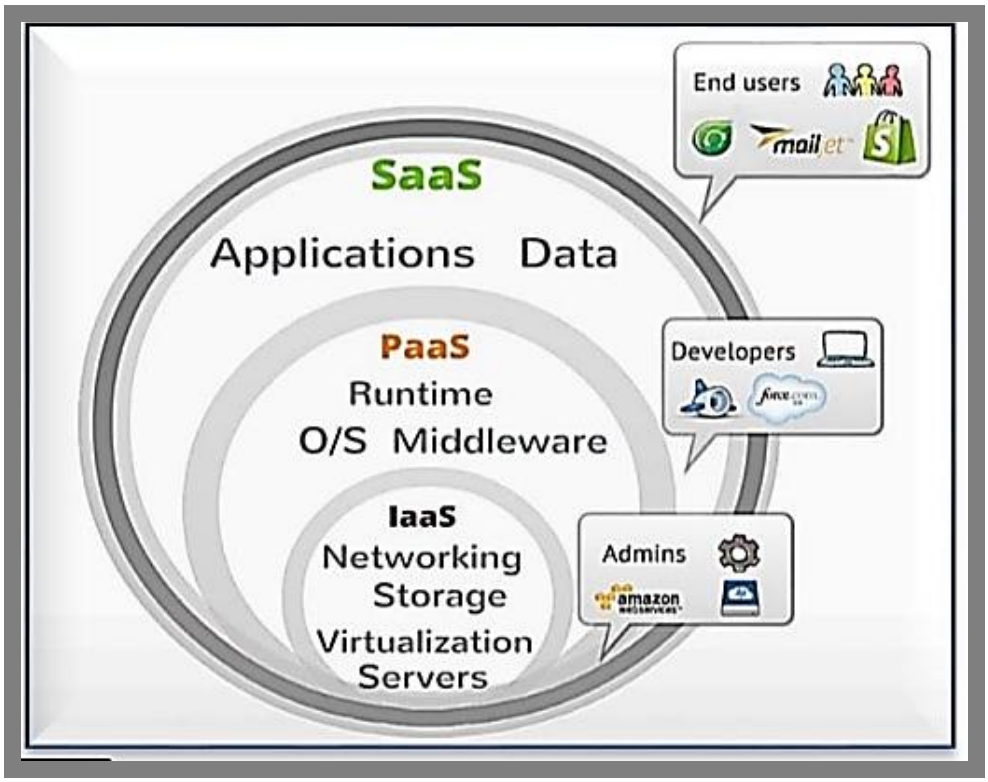


Figure 2. Cloud computing service models [12]

There are two parts to a typical cloud computing environment: the front side and the behind. The user's side is the front end, which can be accessed online, while the cloud service model is handled by the back end [7]. A wide range of dispersed and diverse resources can be used to access cloud computing services [8]. Distributed computers are the source of on-demand services [9]. Platform as a Service (PaaS) for physical resources, hardware/infrastructure, and software as a service (SaaS) are examples of services [10]. One example is Amazon Elastic Cloud, or Amazon EC2[11]. as depicted in figure (2). The fundamental benefits of cloud computing include lower costs, better speed and limitless storage [13].

1.1 Types of Cloud Computing

When seen from the perspective of an organisation, the Cloud model may be broken down into a variety of distinct categories that are distinct from one another. The degree to which the organisational units of the customers and the providers are not in close proximity to one another is the decisive element in the degree to which these categories diverge from one another[14]. The distinction between public clouds, private clouds, and hybrid clouds is shown in Figure 3, which shows that there are three distinct kinds of clouds. Every one of these categories has its own set of distinguishing qualities. Consumers make use of the user interfaces of their web browsers in order to ease the process of accessing public clouds when they wish to do this transaction. In the event that this does place, it is an indication that consumers are expected to pay for the services or resources that they use. Organisations such as Microsoft, Amazon, and Google are examples of companies that provide services via the public cloud. Additionally,

there are other businesses that provide similar services to their customers. Private clouds are clouds that are only utilised for the purpose of providing assistance for activities that are carried out inside enterprises [15]. Private clouds are clouds that are placed within businesses and are used exclusively for this reason. There is another name for private clouds, which is cloud computing. A private cloud gives you more control over the security of your data than a public cloud offers, which is a significant benefit if you are worried about the security of your data. For the purpose of storing their data, a significant number of enterprises, including medical institutions, financial organisations, and other types of businesses, make use of private clouds[16]. There is a kind of cloud computing known as a hybrid cloud, which is characterised by the distribution of services over both public and private regions. Because of its qualities, the term "hybrid cloud" is used to describe it. Certain apps that are seen as being of the highest relevance are maintained on the network of the organisation, but other services may be excluded from the network [17].

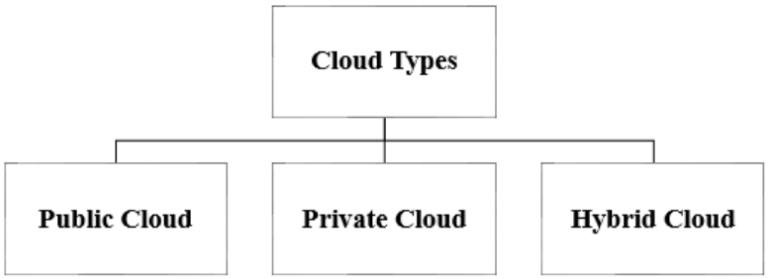


Figure 3. Types of cloud computing [18]

1.2 Components of Cloud Computing

The three fundamental elements of cloud computing are as follows:

• The Client Computers: Client computers supply communication between end users and the cloud.

• Distributed Servers: Despite being spread out over the world, the servers collaborate with one another.

• Data Centers: A collection of servers is called a data center 19].

• Cloud Architecture for Computing.

Users are able to send requests to virtual machines (VMs) via the usage of the internet, and these requests are then kept in the environment. It is the responsibility of the cloud service provider (CSP) to ensure that the quality of service (QoS) is maintained regardless of the delivery type that is used. This is achieved by making certain that user requests may be handled and completed within a certain amount of time from the time they are received. A scheduling strategy, which is also sometimes referred to as a Data Broker, is responsible for assigning user tasks to the virtual machines (VMs) that are the most appropriate for them. This is one of the roles of a scheduling strategy. Since this is the case, the burden that is distributed among the workstations and servers is distributed in a way that is fair and equal. It is feasible that the development and construction of a dynamic load balancer will result in the suitable allocation and utilisation of the resources that are conveniently available. This is something that can be accomplished.

1.3 Distributed System

A distributed system is a term that is often used to describe the infrastructure of cloud computing facilities [20]. A approach that is used rather often is this one. The delivery of services is one of the roles that cloud computing takes on, and it does so via using distributed systems. In a cloud environment, scalability, reliability, and fault tolerance are all made feasible owing to the fact that data and processing are distributed over a number of servers and data centres [21]. This makes it possible for the cloud environment to be fault tolerant. The fact that the cloud environment is scalable makes it possible to achieve scalability. When it comes to cloud computing and distributed systems, the utilisation of shared computer resources is a component that should be considered concurrently. With a distributed system, these resources are dispersed throughout a network of devices that are connected to one another. This type of system is known as a distributed system[22]. There is a kind of computing that is known as hybrid computing that is located on the cloud. The notion of cloud computing makes it feasible for several individuals or enterprises to collaboratively share resources with one another. This is accomplished via the use of virtualization [23] .

Users have the opportunity to access data, software, shared resources, and other services whenever it is convenient for them, and in line with the requirements that they have decided for themselves. This is made possible by cloud computing. When discussing the area of the internet, the term "on-demand service" is often used among users. The whole internet may be compared to a cloud, which is a metaphor that can be used to explain the internet. Through the use of cloud technology, it is possible to cut down on both operating and capital expenditures[24]. To meet the significant challenge of load balancing in cloud computing, it is essential to have a distributed method. This is because of the extensive difficulties involved [25]. In order to establish load balancing that is both efficient and inexpensive, it is difficult to do so due to the complicated architecture of the cloud and the great dispersion of its components. This makes it almost impossible to perform load balancing. Because of this, the responsibility of allocating tasks to the proper servers and customers on an individual basis is one of the most challenging aspects of the job[26]. Continuing to provide one or more services that are not being used in order to satisfy the needs that are being specified is not only not practicable but also not cost-effective. The results of each of these scenarios are bad [27].

1.4 Low latency

The length of time that elapses between a client issuing a request to a server and the server later delivering a response to the request is referred to as "latency." Latency takes into account the amount of time that has passed. When referring to this period of time, the term "latency" is often used. If you are going to utilise cloud computing, it is strongly suggested that you make use of the data centre that is situated in the location that is geographically nearest to the user. In addition, load balancers have to be constructed in order to enable the distribution of incoming requests over a large number of servers[28]. By doing so, the servers will be protected from being overloaded, and the length of time it takes for them to reply to requests will be significantly cut down. A crucial performance parameter for

cloud-based systems is reaction time. This is due to the fact that it has an impact not only on the user experience but also on the overall efficiency of the system. Due to the fact that it has an effect on both of these features, this is the case. When it comes to the length of time it takes for cloud computing to react, there are a good number of distinct aspects that might potentially have an influence. Some of the factors that are taken into consideration in this context include the kind of request that is being made, the amount of demand that is being placed on the servers, and the distance that separates the user from the servers. In general, quicker response times are preferable because they have the potential to result in a user experience that is both more fulfilling and productive. This is because faster reaction times have that potential. In addition, the likelihood of success increases with the speed with which responses are provided [29].

B. Background Theory

Many factors should be taken into account to reduce latency of distributed systems in cloud environments such as load balancing.

2.1 Load balancing

In cloud computing, load balancing refers to the practice of dividing the workload of servers in a way that is both fair and equitable. In order to prevent any one server from being overloaded or underloaded, which are both situations that might lead to a variety of possible problems, load balancing is primarily designed to prevent these situations from occurring. In addition to this, it provides a set of criteria that may be taken into consideration when determining whether or not a given virtual machine should be assigned to a specific job[30]. It is fully reliant on the capacity of the virtual machine, which is determined by the amount of demand that is made on the virtual machines, with regard to the amount of time that is required to do each task. After that, the tasks are distributed among the available resources in a way that is equitable in order to make the most effective use of the resources that are available. We utilise load balancing tactics to distribute workloads across virtual machines that are appropriate and conveniently accessible in order to reduce the amount of time it takes for operations to be finished. This helps us cut down on the amount of time it takes to accomplish jobs [31].

Load balancing is a technique used in cloud environments to distribute workloads among servers and efficiently manage the load on such devices [32]. Two benefits that load balancers provided were increased cloud resource availability and improved performance. Preventing server overload and potential failure is the primary goal of load balancing. It has been applied to provide a high throughput and short reaction times and it is frequently utilized to enhance the speed and functionality of all devices. An apparatus that distributes client requests among a group of servers is called a load balancer [11]. as depicted in Figure 4.

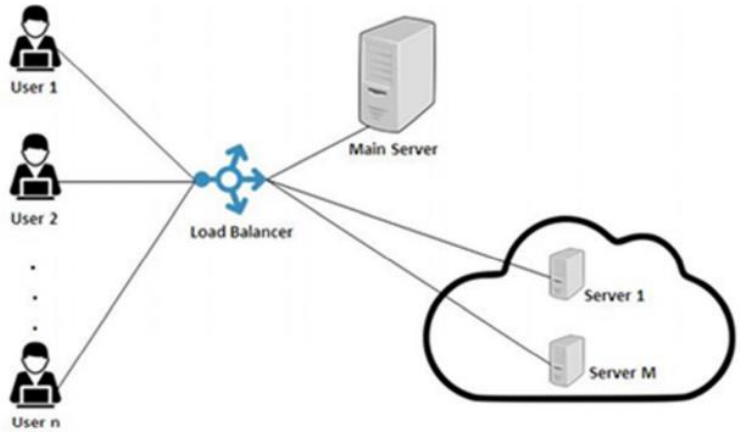


Figure 4. Load balancing mechanism [33].

a. Algorithms of Load Balancing

The main objective of load balancing is to employ the fewest resources while achieving the low latency. Additionally, the following are the most widely used load balancing algorithms in cloud analysts:

• Round Robin Algorithm

This technique is based on the round-robin method, which allots a resource to each user in turn in an equal proportion[34]. This is the most traditional and basic scheduling technique that allows for starvation-free execution and is frequently applied to multitasking. As seen in figure (5), each ready job has to execute in a cyclic queue using the round-robin approach (RR) for a predetermined period of time [35].

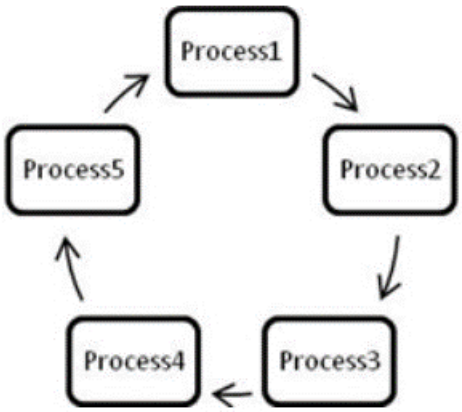


Figure 5. Round Robin Load Balancing technique [36]

• Algorithm for Equitable Spreading of Current Execution (ESCE)

In accordance with this approach, the requests that are being received are divided up and distributed among a large number of servers. The concept that the load ought to be distributed in a manner that is both fair and equitable across all of the servers that are available serves as the basis for this system. It is the responsibility of the ESCE algorithm, which is responsible for calculating the number of processes that are presently running on each server, to ensure that

incoming requests are directed to the server that has the fewest processes running at any one moment. As a consequence of this, the completion of the requirements will be ensured to be completed. As a result of the fact that this is the case, it is simple to ensure that all servers are used in an equitable manner and that none of them are overworking themselves. It is general known that ESCE has a number of problems, one of which is that it may result in overhead whenever the load balancer and the data centre controller connect in order to update the index table. This is one of the drawbacks that ESCE has. ESCE is notorious for having a number of flaws, and this is one of them [37].

• Throttled Load Balancing Algorithm

Additionally, the implementation of this technique is carried out concurrently on each and every virtual computer. Additionally, the current condition of each virtual machine is shown next to it in the same window. The current state of the machine indicates whether it is currently being used or if it is available for utilisation. In order for the client or server to be able to carry out the tasks that have been assigned to it, it is required for the client or server to first make a request to the data centre to identify a virtual machine (VM) that is appropriate for the task. In order to guarantee that the virtual machine is distributed in a manner that is both equitable and uniform across the data centre, it is essential to have a load balancer. In the process of looking through the index table, it is the responsibility of the load balancer to search through it in a sequential manner, beginning at the top and working its way down until it finds the first virtual machine that is accessible. An exhaustive analysis of the whole index table has been carried out in the past. This investigation has been carried out thoroughly. The existence of the virtual machine is taken into consideration throughout the process of picking the data centre that will handle the demand. A request will be sent to the virtual machine (VM) that can be identified by its ID in a way that is different from any other request that has ever been submitted. This request will be made by one of the data centres. In addition, the data centre notifies the load balancer of the modification to the assignment, and it also makes adjustments to the index table. Additionally, both of these acts are carried out at the same time [38].

As a result of their function as a go-between for user bases and data centres, the server broker is responsible for ensuring that services are provided. The primary objectives of utilising a service broker are to reduce the amount of time that users are required to spend making requests, to distribute the resources of the data centre in such a way that they are able to fulfil the requirements of users, and to direct user requests to the data centre that is the most effective in meeting those requirements. In accordance with the policy of the service broker [39], it is of the highest significance that the data centre that is the most appropriate for the assignment be chosen as quickly as is practically possible. The response time and the processing time of the data centre are two of the most important factors that have an influence on the efficiency of the policy that the service broker has in place. Both of these factors are vital to the efficacy of the policy. The policy that optimises response time, the policy that uses the data centre that is closest to them, and the policy that uses dynamic reconfigurable routes with load balancing

are the three service broker policies that cloud analysts utilise. These are the policies that are used by cloud analysts [40]. The primary goal of utilizing a service broker is to:

- Reduce user request latency.
- Designate which data center will respond to user inquiries.

Only the virtual machines located within the cloud system's data center are in communication with the data center controller. To reduce the latency should choose one of the server broker polices.

• Policies for Data Center Service Brokers

In order to distribute global data over a number of different data centres at the same time, the service broker policy serves as a go-between for cloud service providers and the customers of those providers. More specifically, it is intended to facilitate communication between the two parties in a more straightforward manner. The information that is supplied enables the broker to determine which data centre is the most appropriate for fulfilling the needs of the users. This is because the broker is able to determine which data centre is the most suitable. When seen from the most fundamental perspective, the regulation makes it easier to create a connection between consumers and data centres. As a consequence of this, it is now possible to provide services in response to certain requests that are generated by users. An example of a rule that is user-friendly is one that optimises response time, gives priority to proximity to data centres, and adjusts routing configurations in a manner that is adaptable [40]. Below is a list of the rules that people have expressed their preference for the most. Cloud Analyst provides users with the option to choose from three unique routing rules when it comes to the delivery of user queries to the data center [13].

b. Policy of Closest Data Center

When it comes to routing, the first policy, which is one that applies to routing, is comprised of an algorithm that is dependent on the proximity of services. This particular data centre is the one that gets the request, as the name of this particular data centre suggests, and it is the data centre that is located geographically nearest to the user. By compiling a list of the data centres that are in close proximity to one another, the objective is to reduce the length of time that the network is significantly behind schedule. A single data centre is selected at random from the list of possibilities that are located in close proximity to one another in the event that there are many data centres that are located in close proximity to one another. The usage of the datacenter technique that is geographically nearest to the user leads in greater performance inside a cloud environment that is really cloud-based. This is in comparison to the other two options that are available [41].

c. Enhance the Reaction Time Protocol

Utilising a performance-based routing strategy is the second policy, which is an extension of the policy that is based on the data centre that is geographically nearest to the user. This policy is an extension of the policy that was previously mentioned. An expansion of the policy that was discussed before, this policy is an extension of that policy. You should begin by locating the data centre that is located in the closest vicinity to you. This is the first thing you need to accomplish. In the event that the response time of the data centre that is situated in the

geographic location that is geographically closest to the user starts to decrease, the search is carried out in order to find the data centre that has the response time that is the most ideal. In the event that this is the case, then this particular data centre is regarded as the data centre with the highest speed. To determine which data centre will be selected as the final data centre, it is necessary to determine which data centre has the best speed among all of the data centres. In the event that the two data centres cannot be compared to one another, a decision will be made at random from among the possibilities that are, in addition to being the nearest, the fastest. Utilising a probability distribution that is balanced will be the means by which this objective will be fulfilled [40].

d. Load Balancing and Dynamically Reconfigurable Routing

This third policy, which is an enlargement of the same policy, builds upon the policy that chooses the data centre that is geographically closest to me. This policy is an expansion of the policy that was previously mentioned. Additionally, this policy makes use of the routing that was used in the policy that came before it. Additionally, in order to carry out the installation of dynamic routing with load balancing, it is necessary to make adjustments to the routing tables of network devices at the appropriate schedule. For the purpose of ensuring that traffic is distributed along the route in the most efficient manner feasible, this is done in order to guarantee that resources are used in the most effective manner possible [42]. The broker process is able to create connections with cloud computing systems and make use of a broad range of physical cloud computing technologies that have been built expressly for the purpose of carrying out a variety of broker activities. These technologies have been intended to facilitate the broker process. There is a chance that this will occur. Within the context of the data centre, this characteristic is often communicated via the use of a term that is known as the selection policy. Rules of the service broker, which contain special laws that make the process more easy, make it easier to construct a data centre in order to fill an upcoming request. This is because the rules include specific laws that make the process simpler. Furthermore, they provide a graphical user interface (GUI) that is both individualised and standardised, which allows customers to distribute and manage their operations across a variety of clouds. This is not the only benefit they offer [11].

• Cloud analyst

It's a graphical user interface simulator made to examine how big internet apps behave in cloud computing settings. Software developers and designers can utilize cloud analyst to determine the best approaches for allocating resources among various data centers and choosing data centers to serve specific demands [44]. It's an open-source simulation tool that was created directly on top of the cloud sim.

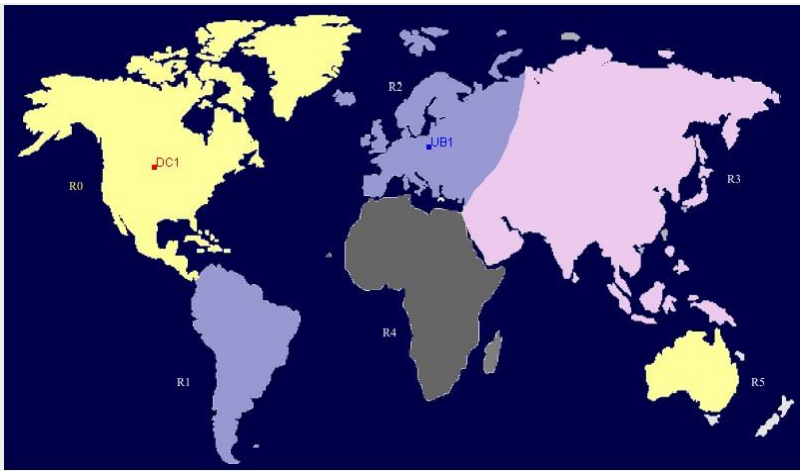


Figure 6. Regions in the Cloud Analyst [44]

• Components of Cloud Analyst Simulator

There are various components that make up Cloud Analyst, such as:

- Region

For each of the six continents, the program divides the entire planet into six sections. In order for the user bases and data of the center to be dispersed over different regions and interact with one another through them: (Asia, Europe, Africa, Australia, North and South America). As shown in figure (6).

- Data Center

This module is used to route user requests made via the Internet to virtual computers and to manage various data center operations, such as the creation and removal of virtual machines [45]. Figure (7) show the architecture of a data center.

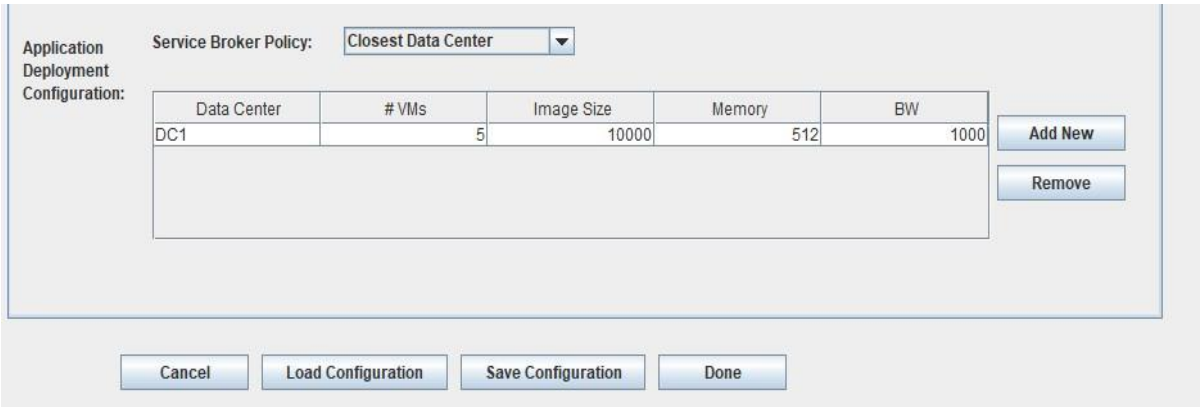


Figure 7. Data Center in Cloud Analyst [46]

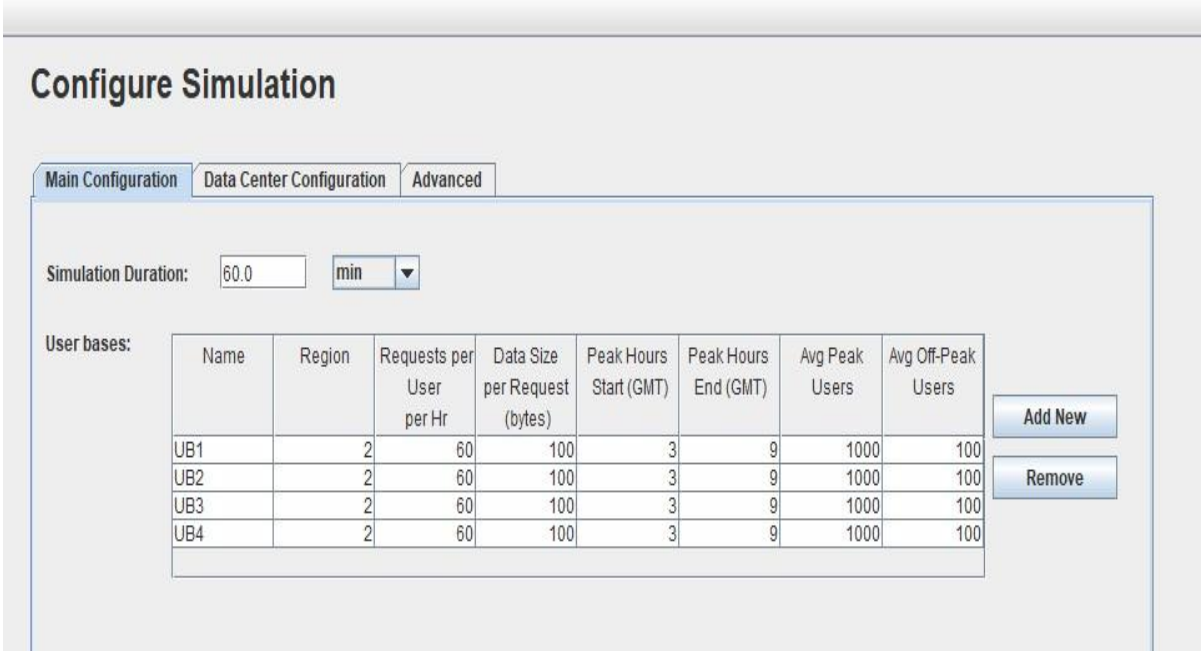


Figure 8. User Based in Cloud Analyst [31]

- User-Based

For the purpose of referring to a collection of users that are considered to be a single entity within the context of a simulation, the word "user base" is the particular term that is used. inside the context of the simulation, the most significant objective that this specific group of users has established for themselves is to enhance the amount of activity that takes place inside the simulation. The typical practice of treating a user base, which may consist of thousands of individuals, as if it were a collective entity is one that happens very often and is a behaviour that is quite prevalent. The amount of data that is broadcast in bursts is determined by the size of the user base, and these bursts occur concurrently despite the fact that they are given in bursts. This is the case despite the fact that they are supplied in discrete chunks. It is possible for a simulation model to select to represent a single user by making use of a user base; however, it is normally more beneficial to make use of a user base in order to depict a bigger number of users, which eventually results in the simulation efficiency being maximised. Users may be represented by a user base. In addition to that, this is the situation [41]. Figure (8) displays the configuration of the user base.

- Broker of Services

This component serves as a mediator to control traffic flow between user bases and data centers. It can use any one of the following three networking strategies: dynamically reconfigured routing with load, optimum response time, and closest data center.

- Load Balancer for VMs

In order to determine which virtual machine should be assigned to handle the requests (Cloudlet), the VM load balancer is crucial [47]. The policies are now

available in Cloud analyst: round-robin, throttled load balancer and load balancer with active monitoring.

- Cloud-let

A cloudlet is a group of user requests. In Cloud Analyst, you may specify how many requests should be joined together to produce a single Cloudlet. Data like the size of a request, the sequence in which it is processed, the sizes of the input and output files, the source and target usage of Internet routing, and the quantity of requests are all stored in the Cloudlet [48].

- Virtual Computer

The term "virtual machine" (VM) refers to a specific kind of actual computer hardware, such as a server or laptop. The abbreviation "VM" is often used to refer to this type of hardware. A storage disc, a central processing unit (CPU), and random access memory (RAM) are all components that are included in the devices that are used for storing data. These components are all included in the devices. In addition to this, it is equipped with a connection to the internet, which makes it possible to get internet access quickly and easily in the event that it is necessary. In addition to being often referred to as VMs, virtual machines are also frequently referred to as software-defined or virtualized computers. These computers are housed inside physical servers. The host component is the one that is accountable for ensuring that the virtual machine is maintained in a continuous way for the whole of its lifecycle. The server is able to generate several virtual machines at the same time and allocate cores to each of them in accordance with the requirements that are unique to each individual machine [48].

C. Literature Review

Here, we summarize the most important earlier research that addressed various algorithms and other methods that have been proposed for application in cloud computing environments.

Mishra et al. in (2020) was discussed the central queue algorithm this method maintains a FIFO queue with pending requests and current activities. Every new activity adds to the queue. When a request is made, the first action in the queue is eliminated. If the requested activity is not available in the queue, the request will be delayed until another one becomes available. Since this is a centralized system, there must be a lot of communication between nodes [49].

Mishra et al. (2020), was proposed A taxonomy of load balancing techniques, including static and dynamic algorithms, along with a discussion of the advantages and disadvantages of each. Dynamic load balancing algorithms use a number of policies, such as information, placement, transfer, and selection policies [49] .

Tabatabaee et al. (2021) an attempt to improve the weighted round robin (WRR) method, considered both server weight and job execution time. The advanced WRR reduces reaction time by assigning the work with the longest running time to the server with the highest weight. However, it does not take into account the availability or busy status of the virtual machines (VMs) before assigning them a job [50].

In 2021, Aloof et al. talked about a wide range service of cloud computing. The provision of such services is made possible by the use of services. One falls under the IaaS cloud type, whereas the other two stand for PaaS and SaaS clouds,

respectively. Three groups of fundamental cloud components: Client computers used by users to access the cloud, Data centers are collections of servers, whereas scattered servers are dispersed over several sites but nevertheless function together [51].

Alyouzbaki and al-Rawi (2021) introduced a novel strategy to increase data center energy efficiency: the three-threshold energy saving algorithm (TESA) virtual machine implementation method. This method relies on a straight line between processor resource usage and energy consumption. Based on workload, hosts within data centers are categorized into four groups in TESA: hosts with a light workload, hosts with an appropriate load, hosts with a moderate load, and hosts with a heavy workload. Virtual machines on a host with a light load or virtual machines on a host with a heavy load are relocated to a different host with the appropriate load by describing TESA. This study would fit well with the present work because load balancing is a common use case for it [52].

Alsaidy et al. (2022) proposed the Minimum Completion Time (MCT) technique. It is based on allocating the task to the virtual machine (VM) or resources that are available at that time, which completes the task in the shortest amount of time. This algorithm considers the criteria for distributing the load on all VM at the time of scheduling because it calculates the shortest time for an implementation from among the available resources only, which helped to achieve the load balancing principle [53] .

With reference to ant colony technology, Yong Li et al. (2022) introduced a novel technique. Ants use the strength of their pheromone to determine which path will get them to their destination. In a similar vein, every node in the network has a pheromone. The routing option for each target is displayed in each row of the pheromone list, and each column indicates the possibility that selecting a neighbor will be the next step. The ant cannot choose; in the absence of pheromones, it will be selected at random. If the pheromone is present, the node with the highest probability is chosen, and its probability is increased while the probabilities of the other nodes are dropped, updating the pheromone table [54].

According to Dhanpal (2022), the Min-Min method selects the task that has the shortest finish time among all the information available and assigns it to the machine that finishes it in the least amount of time. This algorithm assigns tasks independent of loads. Assigning minor jobs to faster machines before deciding on a virtual machine scheduling method may put large tasks at a disadvantage [55].

According to Pandit et al (2022), the First Come First Served (FCFS) method allocates resources to jobs based on the order in which they arrive, with the oldest job in the queue being completed. This method is reliant on the practice of receiving requests in a buffer when resources are used [56].

Shafiq, Dalia Abdulkareem et al. (2022) used throttled load balancing amongst virtual machines in a multi-data center to low latency. They discovered that, while using the fewest processing resources, the throttled approach provides the best overall summary reaction time and processing time in data centers [57] .

In order to achieve equitable workload distribution, Shafiq et al. (2022) created the hybrid Approach (TA & ESCE) to maintain a value that is taken as each VM's priority. In addition to having a faster reaction time, it is also more affordable [57].

Based on the round robin algorithm, T Shined et al. (2023) provide a novel load balancing technique that they adjusted by adding a dynamic time quantum that changed based on the algorithm completion round. The response time was shown to be faster as compared to the usual round-robin technique. This study's drawback is that the authors refrained from talking about processing expenses, instead focusing only on strategies to shorten reaction times. They must also contrast their results with those of other algorithms, including ESCE and Throttled, in order to assess the findings[58].

The Max-Min method was discussed by Banupriya et al. (2024). It is extremely similar to the Min-Min algorithm, with the exception that it assigns the task to the device that finishes it in the shortest amount of time by choosing the largest or longest task out of all the available tasks. However, before making a routing decision, the Min-Min technique distributes tasks independently of virtual machine workloads; in this scenario, though, small tasks could go unnoticed as the system gives priority to finishing larger tasks [59].

D. Discussion and Comparison

The purpose of this section is to summarize a useful reference for load balancing algorithms and find the advantage and disadvantage of this algorithms. In [37] was discussed the central queue algorithm this method maintains a FIFO queue the request will be delayed until another one becomes available. And in [41] proposed the Minimum Completion Time (MCT) technique this algorithm considers the criteria for distributing the load on all VM at the time of scheduling because it calculates the shortest time for an implementation from among the available resources only. Compare to [42] introduced a novel technique ant colony technology. Ants use the strength of their pheromone to determine which path will get them to their destination. Where this technique consumed more response time. The [43]

Explain the Min-Min method selects the task that has the shortest finish time among all the information available and assigns it to the machine that finishes it in the least amount of time. The reference [44] discuss the First Come First Served (FCFS) method this method is reliant on the practice of receiving requests in a buffer when resources are used. The [45] used throttled load that is provides the best overall summary reaction time and processing time in data centers. Also in [45] has been merged (TA & ESCE) the author found faster reaction time. by highlighting these algorithms in the cloud analyst simulation, as explained in most previous research, it was concluded that the throttled algorithm gives less response time and data center processing time, and when combined with another algorithm, the response time is reduced in a better way, and the two algorithms the round robin and ESCE give a convergent response time.

Table 1. Summary of literature review related to the algorithm of load balancing algorithm to low latency.

Ref.	Algorithm	Consumed Time	Advantage	Disadvantage	Tool

[38] **2021**	weighted round robin (WRR) method	Considered both server weight and job execution time.	• Better than round robin. Beneficial for nodes with varying capacity.	Not helpful if the duration of the tasks varies.	Cloudsim
[40] **2021**	the three-threshold energy saving algorithm (TESA) method	Consumed time is low.	• to increase data center energy efficiency.	The host overload detection is dependent on establishing three fixed criteria, which is unsuitable given the dynamic nature of clouds	Cloudsim
[41] **2022**	Minimum Completion Time (MCT) technique.	The response time is shortest.	• Qui ck run time • Ease of implementatio	an unbalanced demand on resources	MATLAB
[42] **2022**	ant colony technology	High reaction time	• Adapts to changing environment • has excellent fault tolerance.	There is less throughput.	Cloudsim
[43] **2022**	Min-Min method	selects the task that has the shortest finish time	• It is easy to do. • Dynamic in nature, taking into account the virtual machine's capacity, job size, and current load.	• The algorithm's main problem is that it can result in starvation. • The nature is centralized.	MATLAB
[44] **2022**	First Come First Served (FCFS)	Delay in execution the request	when scheduling tasks, according to the FIFO rule.	Completing a task requires a lot of time.	Cloudsim

[45] 2022	throttled load balancing	Better reaction time	• It's simple to implement • It works well for small, static systems • Only one scheduler is needed.	• Characterized by centralization. • Waiting times are typically lengthy.	Cloudsim
[45] 2022	hybrid Approach (TA & ESCE)	Achieve low latency.	having a faster reaction time		Cloudsim
[46] 2023	Dynamic round robin algorithm	Response time faster from natural round robin.	• Setting priorities for scheduling is not necessary • The starving effect is not a worry.	• Takes longer to complete the activity • Takes longer to switch between contexts.	Cloudsim

E. Extracted Statistics

The chart is a donut chart representing the response times of various algorithms as used in previous work. Each segment's size correlates to the response time of the respective algorithm. The color-coded legend on the right corresponds to the different algorithms like Round Robin, ESCE, throttled and hybrid (throttled& ESCE). A larger segment suggests a longer response time for that algorithm in the context it was tested, while a smaller segment suggests a quicker response time.

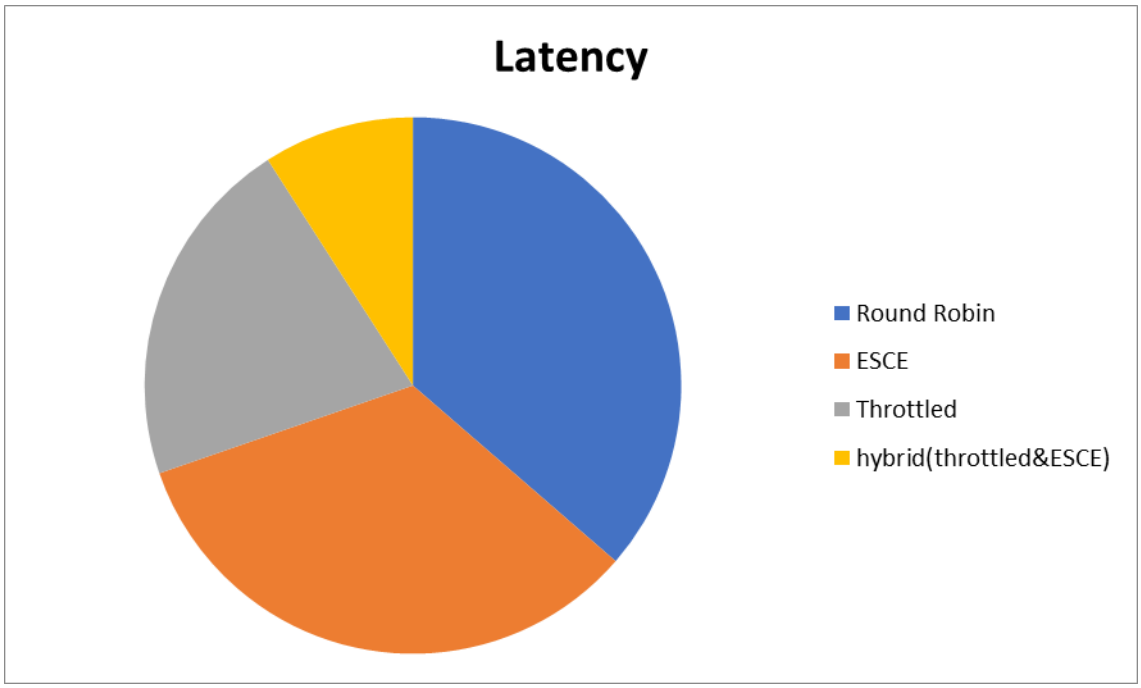


Figure 9. Statistic Chart for latency used by algorithms of load balancing.

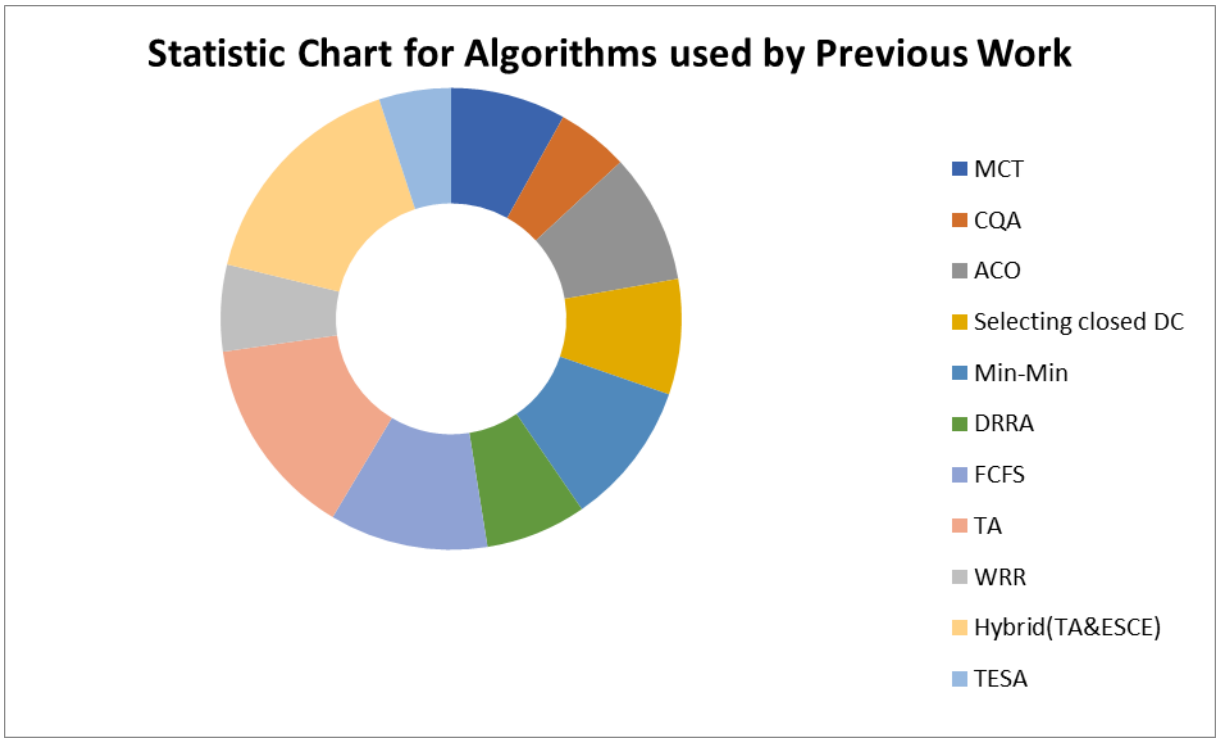


Figure 10. Statistic chart for algorithms used by previous work.

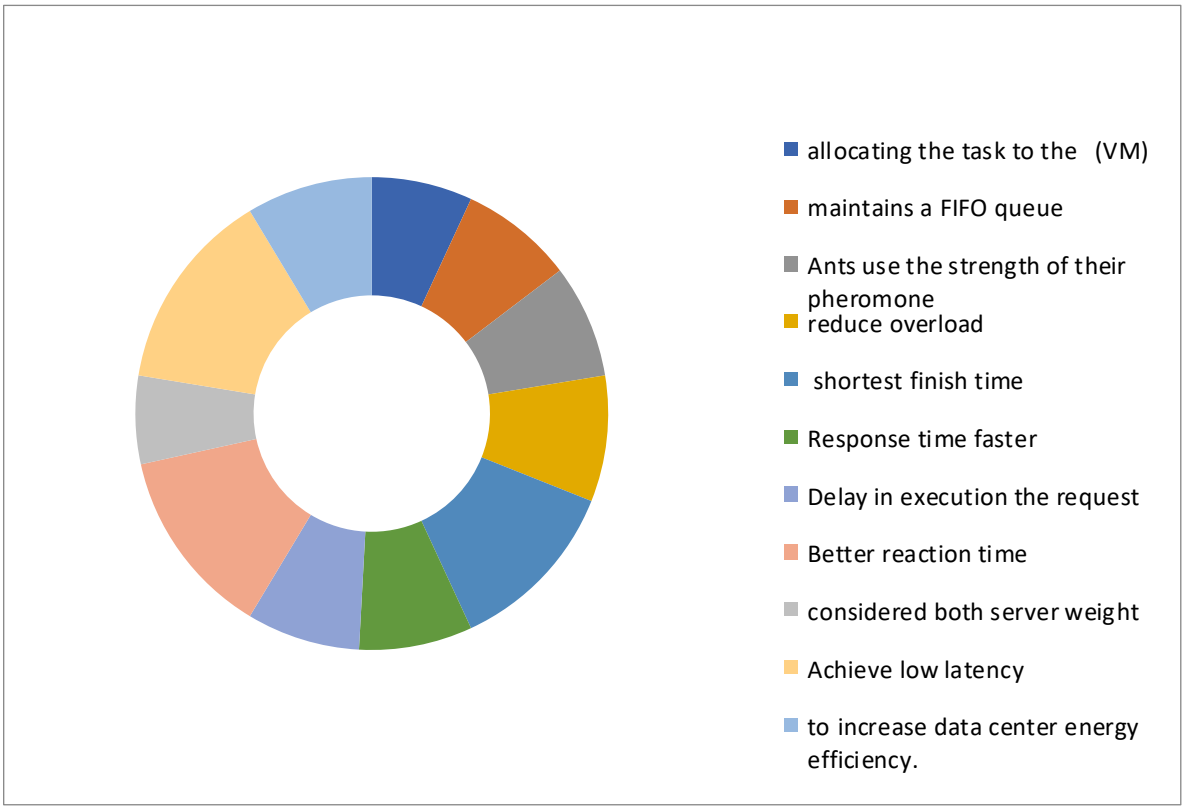
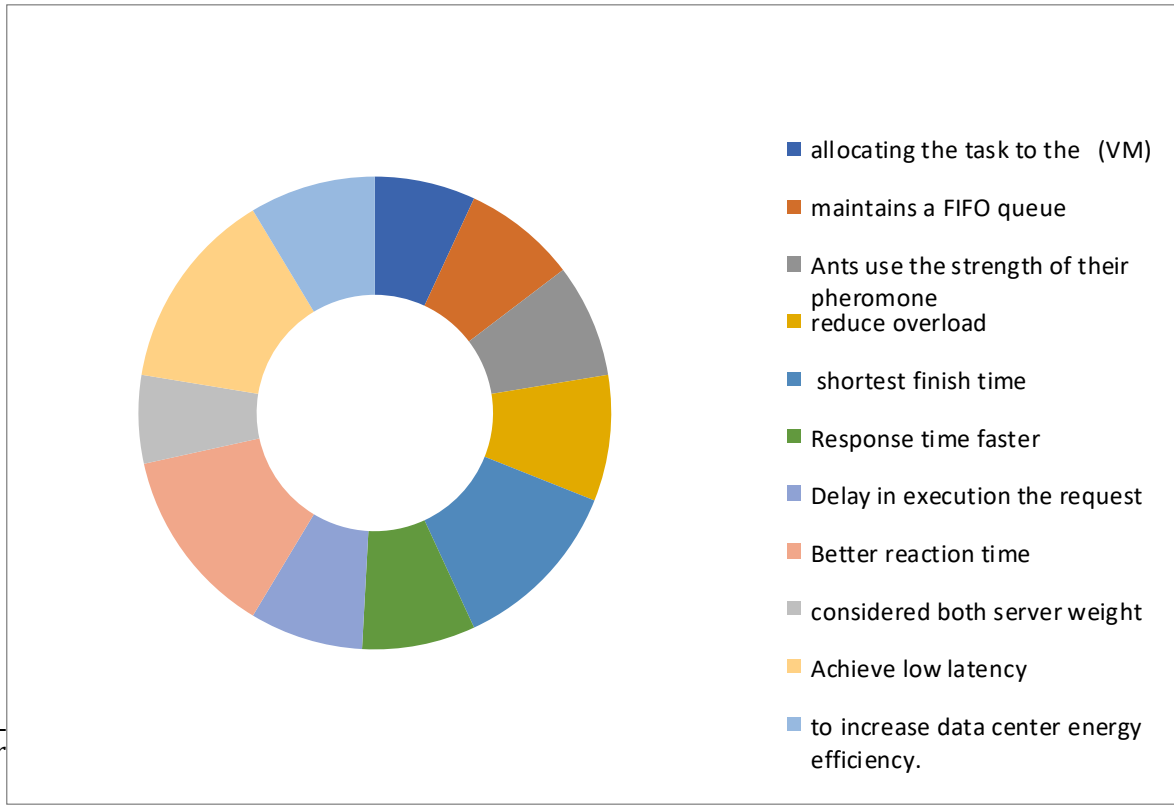


Figure 11. Statistic Chart for Result Obtains used by previous work.

F. Recommendations

The objective of this section is to provide an overview of a useful cloud computing reference. Additionally, discover the benefits and drawbacks of a review paper. In [12] have been studied the round robin. Through [13] which ESCE load balancing algorithm was studied? While in [15] throttled was discussed. By highlightin g these algorithms in the cloud analyst simulator and comparing them in terms of request response speed was concluded that low latency is achieved



using the throttled algorithm, and to more reduce latency, the throttled algorithms were combined with ESCE in a hybrid load balancing algorithm. Figure (11) Statistic Chart for latency used by algorithms of load balancing.

G. Conclusion

As a result of our analysis, which leads us to the conclusion that the response latency reaches a degree of convergence in the data centre that is geographically nearest to us, we are able to enhance the server broker policy in order to obtain speedier reaction times. This is something that we are able to do. These two approaches, which make use of networks and transmission delays, may be used to select the optimum data centre for service providing. Both of these methodologies are described before. Furthermore, the optimal performance plan takes into account the workload of the data centre by continually monitoring the performance of the data centre with the goal of achieving optimal performance. The service broker should make the optimisation of response latency their preferred approach since this often results in the lowest numbers. This is the reason why the service broker should prioritise this strategy. The hybrid load balancing technique is the one that has the lowest response latency, according to the conclusion that may be reached after thorough analysis. Furthermore, we arrived at the conclusion that the utilisation of a server broker that was designed for response time was the primary factor that contributed to the attainment of low latency. Due to the fact that it is situated in the most immediate area of the user, this web server has been selected. In circumstances in which there are a significant number of web servers situated in close proximity to one another, it will use a random selection to choose one of them. In order to determine which data centre has the quickest response time, the first step in the optimum response time approach is to calculate the flow response time for each data centre. This technique is used to determine which data centre has the shortest response time.

H. Acknowledgment

Acknowledgment to those who have provided support for the research. [Cambria 12, space single]

I. References

[1] Zeebaree, S. R., Sallow, A. B., Hussan, B. K., & Ali, S. M. (2019, April). Design and simulation of high-speed parallel/sequential simplified DES code breaking based on FPGA. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 76-81). IEEE.

[2] N. B. Ruparelia, Cloud computing. Mit Press, 2023.

[3] Hasan, D. A., Hussan, B. K., Zeebaree, S. R., Ahmed, D. M., Kareem, O. S., & Sadeeq, M. A. (2021). The impact of test case generation methods on the software performance: A review. International Journal of Science and Business, 5(6), 33-44.

[4] Zeebaree, S. R. (2020). DES encryption and decryption algorithm implementation based on FPGA. Indones. J. Electr. Eng. Comput. Sci, 18(2), 774-781.

[5] A. R. Kunduru, "THE PERILS AND DEFENSES OF ENTERPRISE CLOUD COMPUTING: A COMPREHENSIVE REVIEW," Cent. Asian J. Math. Theory Comput. Sci., vol. 4, no. 9, pp. 29–41, 2023.
[6] Malallah, H., Zeebaree, S. R., Zebari, R. R., Sadeeq, M. A., Ageed, Z. S., Ibrahim, I. M., ... & Merceedi, K. J. (2021). A comprehensive study of kernel (issues and concepts) in different operating systems. Asian Journal of Research in Computer Science, 8(3), 16-31.
[7] Zebari, I. M., Zeebaree, S. R., & Yasin, H. M. (2019, April). Real time video streaming from multi-source using client-server for video distribution. In 2019 4th Scientific International Conference Najaf (SICN) (pp. 109-114). IEEE.
[8] Mohammed, S. M., Jacksi, K., & Zeebaree, S. (2021). A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms. Indonesian Journal of Electrical Engineering and Computer Science, 22(1), 552-562.
[9] Shukur, H., Zeebaree, S., Zebari, R., Ahmed, O., Haji, L., & Abdulqader, D. (2020). Cache coherence protocols in distributed systems. Journal of Applied Science and Technology Trends, 1(3), 92-97
[10] Khalid, Z. M., & Zeebaree, S. R. (2021). Big data analysis for data visualization: A review. International Journal of Science and Business, 5(2), 64-75.
[11] F. K. Parast, C. Sindhav, S. Nikam, H. I. Yekta, K. B. Kent, and S. Hakak, "Cloud computing security: A survey of service-based models," Comput. Secur., vol. 114, p. 102580, 2022.
[12] Zeebaree, S. R., & Jacksi, K. (2015). Effects of processes forcing on CPU and total execution-time using multiprocessor shared memory system. Int. J. Comput. Eng. Res. Trends, 2(4), 275-279.
[13] Jghef, Y. S., Jasim, M. J. M., Ghanimi, H. M., Algarni, A. D., Soliman, N. F., El-Shafai, W., ... & Abbas, F. H. (2022). Bio-Inspired Dynamic Trust and Congestion-Aware Zone-Based Secured Internet of Drone Things (SIoDT). Drones, 6(11), 337.
[14] Zeebaree, S. R., Haji, L. M., Rashid, I., Zebari, R. R., Ahmed, O. M., Jacksi, K., & Shukur, H. M. (2020). Multicomputer multicore system influence on maximum multi-processes execution time. TEST Engineering & Management, 83(03), 14921-14931.
[15] I. M. Ibrahim, "Task scheduling algorithms in cloud computing: A review," Turkish J. Comput. Math. Educ., vol. 12, no. 4, pp. 1041–1053, 2021.
[16] Haji, L. M., Zeebaree, S. R., Jacksi, K., & Zeebaree, D. Q. (2018). A State of Art Survey for OS Performance Improvement. Science Journal of University of Zakho, 6(3), 118-123.
[17] R. Kollolu, "Infrastructural constraints of Cloud computing," Int. J. Manag. Technol. Eng., vol. 10, pp. 255–260, 2020.
[18] A. Kadhim Gabbar Alwaeli and K. E. Kareem Al-Hamami, "Task Scheduling Algorithms in Cloud Computing," Azerbaijan J. High Perform. Comput., vol. 5, no. 2, pp. 131–142, 2022, doi: 10.32010/26166127.2022.5.1.131.142.
[19] K. Gulia and S. K. Maakar, "A Review Paper on Cloud Computing".

[20] Zeebaree, S., & Zebari, I. (2014). Multilevel client/server peer-to-peer video broadcasting system. International Journal of Scientific & Engineering Research, 5(8), 260-265.

[21] Sadeeq, M. A., & Zeebaree, S. R. (2023). Design and implementation of an energy management system based on distributed IoT. Computers and Electrical Engineering, 109, 108775.

[22] Sami, T. M. G., Zeebaree, S. R., & Ahmed, S. H. (2023). A Comprehensive Review of Hashing Algorithm Optimization for IoT Devices. International Journal of Intelligent Systems and Applications in Engineering, 11(6s), 205-231.

[23] Ibrahim, I. M., Zeebaree, S. R., Yasin, H. M., Sadeeq, M. A., Shukur, H. M., & Alkhayyat, A. (2021, July). Hybrid Client/Server Peer to Peer Multitier Video Streaming. In 2021 International Conference on Advanced Computer Applications (ACA) (pp. 84-89). IEEE.

[24] Kako, N. A. (2021). DDLS: Distributed Deep Learning Systems: A Review. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 7395-7407.

[25] Abdulkhaleq, I. S., & Zeebaree, S. R. State of Art for Distributed Databases: Faster Data Access, processing, Growth Facilitation and Improved Communications. Science and Business. International Journal, 5(3), 126-136.

[26] Osanaiye, B. S., Ahmad, A. R., Mostafa, S. A., Mohammed, M. A., Mahdin, H., Subhi, R., Obaid, O. I. (2019). Network data analyser and support vector machine for network intrusion detection of attack type. REVISTA AUS, 26(1), 91-104.

[27] Fawzia, H., Ahmeda, D., Mostafac, S. A., Fudzeec, M. F. M., Mahmoodd, M. A., Zeebareee, S. R., & Ibrahimf, D. A. (2019). A REVIEW OF AUTOMATED DECISION SUPPORT TECHNIQUES FOR IMPROVING TILLAGE OPERATIONS. REVISTA AUS, 26, 219-240.

[28] Zeebaree, S. R. M., Cavus, N., & Zebari, D. (2016). Digital Logic Circuits Reduction: A Binary Decision Diagram Based Approach. LAP LAMBERT Academic Publishing.

[29] Zangana, H. M., & Zeebaree, S. R. (2024). Distributed Systems for Artificial Intelligence in Cloud Computing: A Review of AI-Powered Applications and Services. International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 5(1), 1-20.

[30] Ageed, Z. S., & Zeebaree, S. R. (2024). Distributed Systems Meet Cloud Computing: A Review of Convergence and Integration. International Journal of Intelligent Systems and Applications in Engineering, 12(11s), 469-490.

[31] A. F. S. Devaraj, M. Elhoseny, S. Dhanasekaran, E. L. Lydia, and K. Shankar, "Hybridization of firefly and improved multi-objective particle swarm optimization algorithm for energy efficient load balancing in cloud computing environments," J. Parallel Distrib. Comput., vol. 142, pp. 36–45, 2020.

[32] Abdullah, H. S., & Zeebaree, S. R. (2024). Distributed Algorithms for Large-Scale Computing in Cloud Environments: A Review of Parallel and Distributed Processing. International Journal of Intelligent Systems and Applications in Engineering, 12(15s), 356-365.

[33] M. A. Shahid, N. Islam, M. M. Alam, M. M. Su'Ud, and S. Musa, "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing

Environment and a Novel Fault Tolerance Approach," IEEE Access, vol. 8, no. c, pp. 130500–130526, 2020, doi: 10.1109/ACCESS.2020.3009184.

[34] Ibrahem, A. H., & Zeebaree, S. R. (2024). Tackling the Challenges of Distributed Data Management in Cloud Computing-A Review of Approaches and Solutions. International Journal of Intelligent Systems and Applications in Engineering, 12(15s), 340-355.

[35] B. Alankar, G. Sharma, H. Kaur, R. Valverde, and V. Chang, "Experimental setup for investigating the efficient load balancing algorithms on virtual cloud," Sensors (Switzerland), vol. 20, no. 24, pp. 1–26, 2020, doi: 10.3390/s20247342.

[36] A. A. A. AlKhatib, T. Sawalha, and S. AlZu'bi, "Load balancing techniques in software-defined cloud computing: an overview," in 2020 Seventh International Conference on Software Defined Systems (SDS), IEEE, 2020, pp. 240–244.

[37] A. H. Zamri, N. S. M. Pakhrudin, S. Saaidin, and M. Kassim, "Equally Spread Current Execution Load Modelling with Optimize Response Time Brokerage Policy for Cloud Computing," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 2, 2023.

[38] S. Y. Mohamed, M. H. N. Taha, H. N. Elmahdy, and H. Harb, "A proposed load balancing algorithm over cloud computing (balanced throttled)," Int. J. Recent Technol. Eng., vol. 10, no. 2, pp. 28–33, 2021.

[39] D. Yu, Z. Ma, and R. Wang, "Efficient Smart Grid Load Balancing via Fog and Cloud Computing," Math. Probl. Eng., vol. 2022, 2022, doi: 10.1155/2022/3151249.

[40] B. Nayak, B. Bisoyi, and P. K. Pattnaik, "Data center selection through service broker policy in cloud computing environment," Mater. Today Proc., vol. 80, pp. 2218–2223, 2023.

[41] A. I. El Karadawy, A. A. Mawgoud, and H. M. Rady, "An empirical analysis on load balancing and service broker techniques using cloud analyst simulator," in 2020 international conference on innovative trends in communication and computer engineering (ITCE), IEEE, 2020, pp. 27–32.

[42] A. Jyoti and M. Shrimali, "Dynamic provisioning of resources based on load balancing and service broker policy in cloud computing," Cluster Comput., vol. 23, pp. 377–395, 2020.

[43] R. M. Singari and P. K. Kankar, "Challenges and Issues of Load Balancing Algorithms in Cloud System," Adv. Prod. Ind. Eng. Proc. ICAPIE 2022, vol. 27, p. 150, 2022.

[44] P. Payaswini, "Comparative study on load balancing and service broker algorithms in Cloud computing using cloud analyst tool," Int. J. Next-Generation Comput., pp. 49–61, 2021.

[45] M. Alagarsamy, A. Sundarji, A. Arunachalapandi, and K. Kalyanasundaram, "Cost-awareant colony optimization based model for load balancing in cloud computing.," Int. Arab J. Inf. Technol., vol. 18, no. 5, pp. 719–729, 2021.

[46] A. Singh and R. Kumar, "Performance evaluation of load balancing algorithms using cloud analyst," Proc. Conflu. 2020 - 10th Int. Conf. Cloud Comput. Data Sci. Eng., pp. 156–162, 2020, doi: 10.1109/Confluence47617.2020.9058017.

[47] R. Mathur, V. Pathak, and D. Bandil, "Parkinson disease prediction using machine learning algorithm," in Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, Springer, 2019, pp. 357–363.
[48] A. Y. Ahmad and A. Y. Hammo, "A Comparative Study of the Performance of Load Balancing Algorithms Using Cloud Analyst," Webology, vol. 19, no. 1, pp. 4898–4911, 2022.
[49] S. K. Mishra, B. Sahoo, and P. P. Parida, "Load balancing in cloud computing: a big picture," J. King Saud Univ. Inf. Sci., vol. 32, no. 2, pp. 149–158, 2020.
[50] S. M. Tabatabaee, J.-Y. Le Boudec, and M. Boyer, "Interleaved weighted round-robin: A network calculus analysis," IEICE Trans. Commun., vol. 104, no. 12, pp. 1479–1493, 2021.
[51] B. Alouffi, M. Hasnain, A. Alharbi, W. Alosaimi, H. Alyami, and M. Ayaz, "A Systematic Literature Review on Cloud Computing Security: Threats and Mitigation Strategies," IEEE Access, vol. 9, pp. 57792–57807, 2021, doi: 10.1109/ACCESS.2021.3073203.
[52] Y. A. G. Alyouzbaki and M. F. Al-Rawi, "Novel load balancing approach based on ant colony optimization technique in cloud computing," Bull. Electr. Eng. Informatics, vol. 10, no. 4, pp. 2320–2326, 2021.
[53] S. A. Alsaidy, A. D. Abbood, and M. A. Sahib, "Heuristic initialization of PSO task scheduling algorithm in cloud computing," J. King Saud Univ. Inf. Sci., vol. 34, no. 6, pp. 2370–2382, 2022.
[54] Y. Li, J. Li, Y. Sun, and H. Li, "Load Balancing Based on Firefly and Ant Colony Optimization Algorithms for Parallel Computing," Biomimetics, vol. 7, no. 4, pp. 1–16, 2022, doi: 10.3390/biomimetics7040168.
[55] D. Singh, "Dynamic Resource Allotment in Cloud Computing with Predetermined Waiting Queue," vol. 10, no. 4, 2022.
[56] R. Pandit and M. Dwivedi, "Resources Load Sharing using RR, FCFS, SJF, MSJF, GP Algorithms in Cloud Computing," NEUROQUANTOLOGY, vol. 20, no. 11, pp. 7853–7872, 2022.
[57] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, "Load balancing techniques in cloud computing environment: A review," J. King Saud Univ. Inf. Sci., vol. 34, no. 7, pp. 3910–3933, 2022.
[58] A. I. M. P. E-journal, "Vidhyayana - ISSN 2454-8596," vol. 8, no. 7, pp. 707–728.
[59] M. R. Banupriya and D. Francis Xavier Christopher, "Efficient Load Balancing and Optimal Resource Allocation Using Max-Min Heuristic Approach and Enhanced Ant Colony Optimization Algorithm over Cloud Computing," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 1s, pp. 258–270, 2024.