International Journal of Research in Engineering and Innovation (IJREI)
journal home page: http://www.ijrei.com
ISSN (Online): 2456-6934

ORIGINAL ARTICLE

Simulation of load balancing algorithms using cloud analyst for distinct regions IT resources

Aditiya[1], Rashim Rana[2]

[1]*M. Tech student at Himalayan Institute of Engineering and Technology's Department of Computer Science and Technology, India*
[2] *Assistant Professor at Himalayan Institute of Engineering and Technology's Department of Computer Science and Technology, India.*

Article Information

Received: 02 February 2022
Revised: 05 March 2022
Accepted: 10 March 2022
Available online: 25 March 2022

Keywords:

Cloud Computing
Load Balancing Algorithms
Response time minimization
Cloud Analyst
Cloud Security

Abstract

Cloud computing is currently on the rise as new technology and as a new business model, according to experts. Data archiving, servers (both real-world and virtual), telecommunications, machine learning tools for data analysis, and a cloud services provider's (CSP) data center are all examples of cloud computing resources that are available on demand over the internet. Cloud computing allows client devices to access data from faraway servers over the internet, which is referred to as "cloud computing." Rather than purchasing, operating, and maintaining physical data centers and servers, we can take advantage of technology services provided by third parties. A pay-per-use pricing structure is in place. It is spreading over the world simply because of its simple and service-oriented business approach. It delivers the greatest computing service to the user and makes the execution of large-scale activities much easier than before. Through the use of a cloud analyst tool, this study provides an overview of the evaluation of the response time and processing time of several existing algorithms that are used in cloud computing.

1. Introduction

Cloud computing is a worthwhile innovation in the era of the high-level software engineering field. Cloud computing is overwhelming software program that has the ability to exchange IT Software Industry and software making program more attractive. The concept of the Cloud has revolutionized the IT industry. According to the Gartner's report [1] on August 2, 2021, one of the senior research directors Mr. Henrique Cecci said that "The pandemic serves as a catalyst for the adoption of cloud services and digital innovation due to its harmful impact on economy and on various organizations. There are four trends that are expanding the area of cloud capabilities and offerings such as Cloud Ubiquity, Regional Cloud Ecosystem, Carbon-Intelligent Cloud and Sustainability. Customers don't have to pay attention to the internal mechanism of the cloud and can use the computing resources over the internet. It is an on demand service that can also be called as pay-as-you-go services. It means that you have to pay only for those resources that have been used by you. Users can access the resources like data storage, application servers, and development tools and data centers (DCs) from the remote areas via internet. These IT resources are managed and hosted by the service providers called as Cloud Service Providers (CSP). These CSPs charge the customer monthly subscription fee according to the usage. It facilitates the customers to use the cloud services at lower cost, more scalable architecture and improved agility. Cloud Computing chips away at the premise of sharing of assets to accomplish intelligence. Instead of saving data over the proprietary hard drive, cloud computing facilitates the customers to store the data on remote database. Customer can access the data and software till the devices has access to the

Corresponding author: Aditiya
Email Address: aadigurung@gmail.com
https://doi.org/10.36037/IJREI.2022.6206

web. It is termed as cloud computing because the information being accessed is remotely present in virtual space or cloud. The CSPs facilitates the customers to store the data on remote servers and access remotely from anywhere in the world whenever required through internet. Various researches over cloud computing stated that you don't have to carry the huge computer clusters and applications with you and internet connectivity becomes a cloud for you.

As per the new investigates done by National Institute of Standards and Technology (NIST) in excess of 92% of the enormous associations utilize more than one cloud. Their research shows a data where the prediction says that 55 percent of the enterprise workloads will rely on public clouds by the end of 2021. It shows the expansion of cloud computing over the IT world vigorously. Wide network access, measured service, distributed resource pooling, Rapid Elasticity, and Self-service that is available on demand are among some of the essential characteristics provided by cloud computing [2]. The use of cloud computing technology is becoming a popular approach to outsource the external IT resources at cheap cost. There are four types of cloud models that are represented according to the environment. It depends upon the users which type of cloud is best suited for their needs. These four cloud computing models are: Private cloud (It is provisioned only for a single organization and is also termed as single tenant environment), Cloud computing in the public domain (It is owned by third party owners that are called cloud service providers and the resources are shared by multiple customers and will be charged on a pay-as-you-use basis), Hybrid cloud (It comprises of two or more different private or public clouds that are compatible with application and data portability e.g. BMC, NetApp, AWS VPC), Community cloud (It is termed as community cloud because it is a multi-tenant platform and it allows multiple company user groups to collaborate on same project that shares the same interest e.g. Healthcare, Education etc.).

The cloud computing provides three very lucrative conventional services like:

- Platform-as-a-Service (PaaS): It involves the provisioning of platform hardware, infrastructure, and development tools on-demand, the term refers to the process of developing, running, and managing applications at a low cost and with minimal management. The upkeep of the platform on-site will become much simpler and less complex as time passes. Google App Engine, IBM Cloud Platform, and others are examples of Platform as a Service.
- Software-as-a-Service (SaaS): Using a web browser or an API, you can access the cloud-hosted application software on your desktop, laptop, or mobile device. The customers that use SaaS system usually has to pay monthly subscription fee or annual subscription fee. Few benefits offered for using SaaS are Automatic upgrades and Protection from data loss. In most of the commercial organization software today it acts as primary delivery model. Few examples of SaaS are AWS, Saleforce, Microsoft Office 365, Dropbox and Adobe Creative Cloud etc.
- Infrastructure-as-a-service (IaaS): All of this offers pay-per-use access to basic IT resources like virtual and physical servers, storage, and networking over the internet. It helps in reducing the large capital expenditure on computing resources. It enables the user for the provision of scaling and shrinking of resources on the basis of needs. IaaS is the only model that provides the lowest level of control of computing resources to the user in the cloud. Some popular IaaS providers include are AWS EC2, Linode, Google Compute Engine, and Azure Virtual Machines [3].

The three service models of SaaS, PaaS, and IaaS are responsible for data security in the cloud. Data remains in two states in the Cloud host, namely, at rest and in transit. In Rest state, the data remains stable which means that data remains static in cloud. In rest state, sometimes it becomes too difficult for an organization to protect data at rest if they are not having their own private cloud because they don't have any physical control over data. While in Transit state, the data transfer takes place which means data will move in and move out of the cloud. The data in transit can be in encrypted or in unencrypted form. The data can be a file, document or a database stored in cloud. A user who is located at some remote location can access or upload data whenever required in transit state. At transit state the data becomes very sensitive due to its exposure during transfer [4]. Due to these two states, the situation of data theft arises which will affect the confidentiality and integrity of data. Now a day's various security software providers provide the security for the cloud services. According to Gartner, McAfee which is one of the security software providers claim that 52% companies experiences better security than on premises. By the end of year 2020 IaaS cloud service will experience 60% lesser security incident than on traditional data centres.

A survey on simply server less emphasize over the importance of server less computing. In that survey, the author defines that over the last decades how server less computing is affectionating all the researchers towards it. Its influence is growing due to its benefits of reducing costs, improving scalability, decreasing latency and elimination of server-side management. It is a computing model that provides a relief on managing the backend management tasks such as scaling, scheduling, provisioning and patching. After deploying this model all the developers can work freely and focus their time on coding and business logic of specific applications. It works on the basis of per request which means that it supports scale up and scale down of the architecture of cloud storage on the basis of request done by customer. Customer has to pay only for the time during which the application is running. No need to pay for the idle capacity. Function-as-a-service (FaaS) is a service which is a subset of server less. Sometimes FaaS is assumed as a server less computing. It allows the developers to execute a portion of code and response to specific event. For the execution of code, the cloud service providers will manage the computing resources such as physical hardware, virtual machine operating system and storage space for particular interval of time. Billing for the use of resources will start when the execution starts and stops when the execution stops or finished [5].

2. Load Balancing

When a workload is split or distributed across multiple nodes at the same time, this is referred to as load balancing or load distribution. It is the process which ensures that no computing machine will overload or under-load (idle state). The data can be sent and received by dividing the workload traffic on distinct servers to achieve less response delay. Its goal is to increase the overall performance of the cloud infrastructure response by speeding up the various constraints such as execution time, response time, system stability, and reducing the delay time, among others. Load balancing algorithms are used to achieve the highest possible throughput and the shortest possible response time by employing resources in the most efficient manner [7]. It ensures that all the computing resources will get utilize properly and completely. It is an optimization method that focuses on the allocation of tasks, the scheduling of tasks, the allocation of resources, and the management of resources. In the given article [6], it is stated that the process of load balancing works on two level load balancers. In a level 1 load balancer, the workload of a physical machine is distributed across its respective virtual machine; however, in a level 2 load balancer, the workload is balanced by spreading the workload across multiple virtual machines running on a variety of hardware. There are two types of load balancing methodologies: dynamic and static.

2.1 Static load balancing

The static load balancing algorithm is used by cloud service providers to distribute workload across homogeneous assets. Progressive provisioning is done here in which the suppliers established suitable properties before the administration began in accordance with the agreements made. Non versatile properties are used due to the steady state. The cloud requires prior information of the hub's range, insights of power, storage memory, performance and client requirements that are not variable [11].

2.2 Dynamic load balancing

The cloud is responsible for ensuring the heterogeneous property in the dynamic load balancing theorem. Dynamic provisioning is done with adaptable assets in which the specialist cooperative splits more assets as expected from the customer and when they are not used they are expelled. The cloud does not require any prior learning and the prerequisites of the customers are versatile. It is difficult to reproduce the dynamic state yet very adaptable with the distributed computing condition [9].

3. Load balancing algorithms

The ability to keep up with activities in a cloud computing environment is critical for staying on top of things. It reduces the response time in requests in order to avoid overburdening the framework; additionally, it increases throughput while obtaining optimal asset utilization [15]. The fundamental of presenting calculations in load adjusting is to stay away from over-burdening and inactivity of nodes in a cloud framework. Accordingly, calculations will guarantee that every one of the nodes is allocated to a similar measure of responsibility in a cloud framework. An increment in distributed computing stages like Windows Azure Platform, Amazon S3, and so on, and use in Artificial Intelligence [16] will improve the advancement of many web look with particular elements, from cloud service providers.
Load balancing algorithms are critical because they ensure that clients receive consistent administrations without interruption. Static load balancing occurs in a static environment when the algorithm's display ignores the current state of the system. Client requirements do not alter during runtime in this way. Adjusting the display of the algorithms in dynamic load is highly dependent on the system's state. Because assets are adaptive in nature, algorithms successfully execute load balancing in a dynamic context.

3.1 Round robin algorithm

It operates on the principle of the round robin method, as the name suggests. After receiving a request from a client, the data center controller regulator directs the round robin load balancer to allocate another virtual machine (VM) for the purpose of processing the request. The weighted round robin algorithm is depicted in Fig. 1 as a basic representation of its structure.

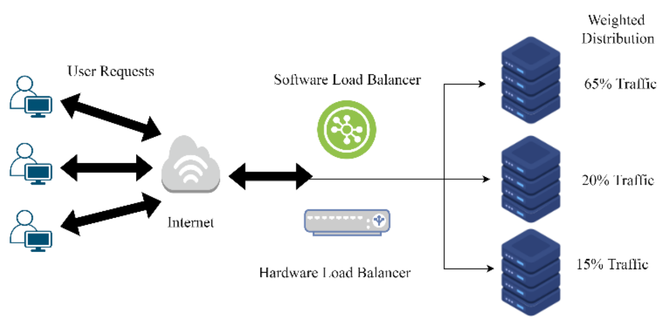

Figure 1: Basic structure of weighted round robin algorithm

It is selected at random by the Round Robin Load Balancer (RRLB), which sends the information gathered along with the VM ID to the data center controller for processing. This is how the demands that arise as a result of a roundabout request are resolved [10]. Round robin balancing algorithms are used to distribute the data workload evenly among the various servers participating in the system. The traffic generated by these route servers will be easily despised if this approach is followed, further complicating the situation. A round-robin algorithm is the name given to this algorithm, which distributes traffic in an even manner. With the help of weighted round robin, a number of issues can be addressed and resolved. When the weighted round robin algorithm was introduced, it was designed to address the critical challenges associated with the round robin algorithm. In this algorithm, each server is assigned a load or weight, and the server with the highest load receives more connectivity than the other servers. The server will receive balanced traffic if all of the nodes are equal [8]. In weighted round robin balancing method, the network administrator plays an important role of assigning a fixed numerical weight to each and every server existing in the pool. Server that is assigned with high value of weight will receive more requests from the users and vice versa. Providing weight to the different servers seems like assigning the server priority.

3.2 Active Monitoring Load Balancing Algorithm

This algorithm is also referred to as equally spread current execution (ESCE) load adjusting computation, which is an abbreviation for Active Monitoring Load Balancing Algorithm. It makes use of an innovative active monitoring load balancer (AMLB) to distribute the stack in a consistent manner across the various nodes and servers. The means remember for this algorithm is portrayed in fig. 2. The AMLB is responsible for maintaining a file table of servers and requests in the data center's record table, which is accessible to all users. As a result, when the framework receives another request, it anticipates that the data center index table will identify servers that are the least stacked or latent in the data center. It is a unique burden adjusting calculation that distinguishes the least burden or inert virtual machine and allocate them load [12]. At the point when the client's solicitation for provisioning of virtual machine (VM) is gotten by the data center controller. Regulator advances the solicitation to AMLB to assign a VM and parses the record table to discover which VM is least loaded or inactive. Parsing will go on until the entire record table gets parsed and the most un-stacked VM is found. It immediately informs the data centers controller of the presence of the VM by returning the VM id. If more than one VM with the lowest weight is found, the VM with the lowest weight is chosen using the First Come First Serve (FCFS) method. Meanwhile, the AMLB sends the VM's ID to the data center controller, which is stored in the database. Furthermore, the data center controller communicates with the virtual machine that has been identified. By that point, the data center controller has encouraged the AMLB to implement the new VM distribution strategy in the data center. By that point, AMLB has updated the designation table by increasing the count for that VM by one. When a virtual machine (VM) has completed its reasonable task of meeting the predefined demand, it propels a response to the information server regulator. On getting a response, it illuminates AMLB about the de-assignment regarding the virtual machine. Subsequent to getting the reaction, the data center controller refreshes the assignment table by lessening the portion number 1 for that VM. In the wake of getting the reaction, the data center controller refreshes the allocation table by lessening the distribution number 1 for that VM.

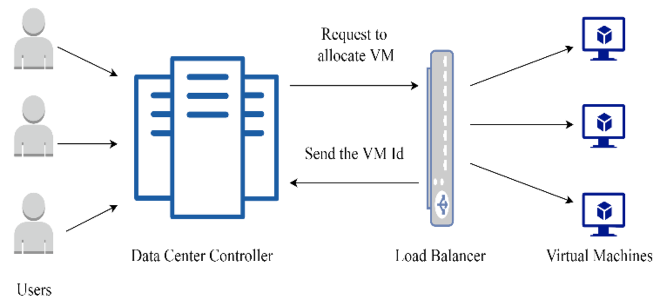

Figure 2: Active Monitoring Load Balancing

3.3 Throttled Algorithm

This algorithm makes use of a throttle load balancer (TLB) to keep track of the amount of load and heap on each virtual machine in turn. In this case, each virtual machine is assigned to complete only one task at a time. Other task can be assigned only if the present work has been done successfully. The various steps associated with throttle algorithm are as per the following: The TLB maintains the index table which consists of information about all the VMs and their state. State of VM represents whether VM is available or busy. The client requests data center controller to allocate the appropriate VM to perform particular job. The data controller then demands TLB to allocate the VM.

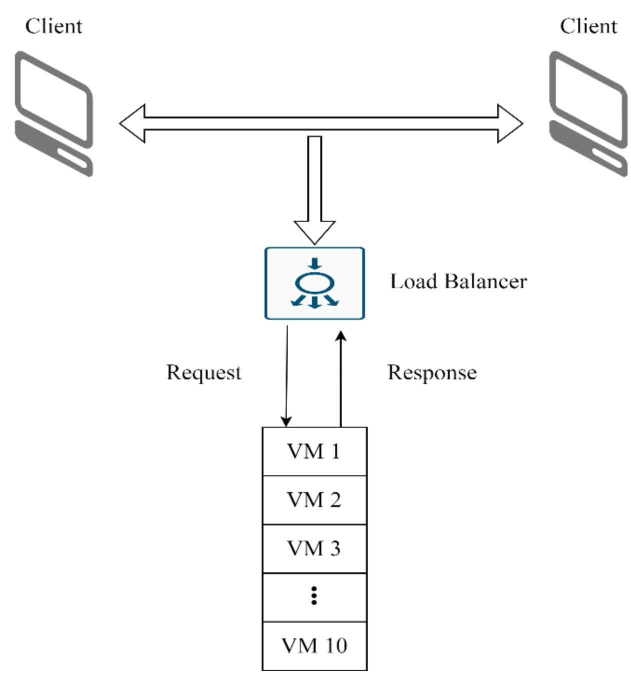

Figure 3: Throttled scheduling process

TLB is responsible for scanning index table thoroughly from start to the very end until it finds the respective VM. After finding out VM, the TLB returns the VM with its respective id. The index table is then revised by TLB and incrementing of allocation is done for that VM by 1. While for another situation, in the event that TLB observes no VM in the accessible state then it essentially brings invalid back. For this situation, the data center will tell everyone of client's solicitations to remain in line until VM is accessible. At the point when the execution of the all-around apportioned VM process wraps up. It provides the data center controller with information about the VM's territory. When the data center controller receives a response, it sends an affirmation to TLB regarding the de-portioning of virtual machines [13]. The TLB then, at that point, diminishes the portion count by 1 for VM and refreshes the data in the allocation table. The throttle scheduling process can be shown in the fig. 3.

4. Methodology

There are two factors on which the performance of load balancing methods is analyzed. These factors are load and load performance. Performance under load is measured in terms of the average response time achieved by the user. Load is defined as the number of requests received by the client and the amount of CPU time used. The algorithm for load balancing is based on a variety of factors such as length of cloudlet tasks, response time, arrival time of task, VM configuration and time to complete task. Response time is the addition of cost of client's request and process execution time. It is possible to calculate response time using the equation presented in research [14], which can be stated as follows:

Response Time, $R(t) = F_{fin}(t) - A_{arr}(t) + T_{delay}(t)$ (1)

where,

$F_{fin}(t)$ = time to complete the task.
$A_{arr}(t)$ = arrival time of client's request
$T_{delay}(t)$ = transfer time of the task/transmission delay.

This formula can be used to calculate $T_{delay}(t)$:

$T_{delay}(t) = T_{latency} + T_{transfer}$

This is defined as the time taken for data to be transferred from a source location to a destination location for a single request, with $T_{latency}$ denoting the network latency and $T_{transfer}$ denoting the time it takes to transfer data from a source location to a destination location for a single request (S_{req}). After applying Poisson distribution, we get $T_{latency}$ from latency matrix given in internet characteristics.

$T_{transfer} = S_{req}/B_{puser}$

where, $B_{puser} = B_{total}/N$

When looking at the Internet's characteristics, B_{total} is the total available data transfer bandwidth, and N is the number of active clients requesting data during the current transmission. The web attributes also keep track of the client's demand N, which is a one-way trip between two different locations.
In this algorithm, the adjusting is finished by Data center broker. It can influence the execution time in the neighborhood climate of the data center. Consequently, the postponed boundary during correspondence can be discarded.

Expected task completion time can be calculated using various equations in:

This is the formula defined in for scheduling policies involving Spaceshare-Spaceshare or Timeshare-Spaceshare, respectively:

$$eft(p) = est + \frac{rl}{capacity * cores} \quad (2)$$

where,

eft(p): Cloudlet P's expected completion time in hours.
est: Arrival time of Cloudlet P.
rl: Total no. of instructions to be executed on a processor of Cloudlet P.
capacity: Average processing power of a core for Cloudlet P.
cores(p): Cloudlet anticipates a certain number of centers

Capacity calculation defined in is given as:

$$capacity = \sum_{i=1}^{np} \frac{cap(i)}{np} \quad (3)$$

For Spaceshare-Timeshare or Timeshare-Timeshare scheduling policy, the expected task completion is determined by the formula in equation (4) & (5):

$$eft(p) = ct + \frac{rl}{capacity * cores(p)} \quad (4)$$

where,
cap: Processing power of the core.
ct: Current simulation time

5. Simulation methods used to deploy load balancing algorithms

5.1 CloudSim

CloudSim is an open-source structure, which is utilized to re-enact cloud computing frameworks and administrations. It is created by the "CLOUDS Lab" association and is composed completely in Java. It is utilized for demonstrating and mimicking a distributed computing climate as a method for assessing a theory before programming improvement to duplicate tests and results. For instance, if you somehow happened to send an application or a site on the cloud and needed to test the administrations and burden that your item can deal with and furthermore tune its presentation to beat bottlenecks prior to gambling with organization, then, at that point, such assessments could be performed by basically coding a recreation of that climate with the assistance of different adaptable and versatile classes given by the CloudSim bundle, liberated from cost. Both exploration-based cloud improvement and industry-based cloud advancement have profited from the CloudSim venture's far and wide application. When taken from a calculated standpoint, CloudSim offers class definitions that address data centers, physical hosts, virtual machines, cloud administrations, cloud clients, inward centers networks, and the energy consumption of physical hosts, data centers components, and virtual machines from one perspective and virtual machines from the other. However, CloudSim allows for the dynamic insertion of simulation elements as well as the provision of message-passing applications and data centers network topologies on the other hand [17] [18]. The basic architecture of CloudSim is shown in fig. 4.

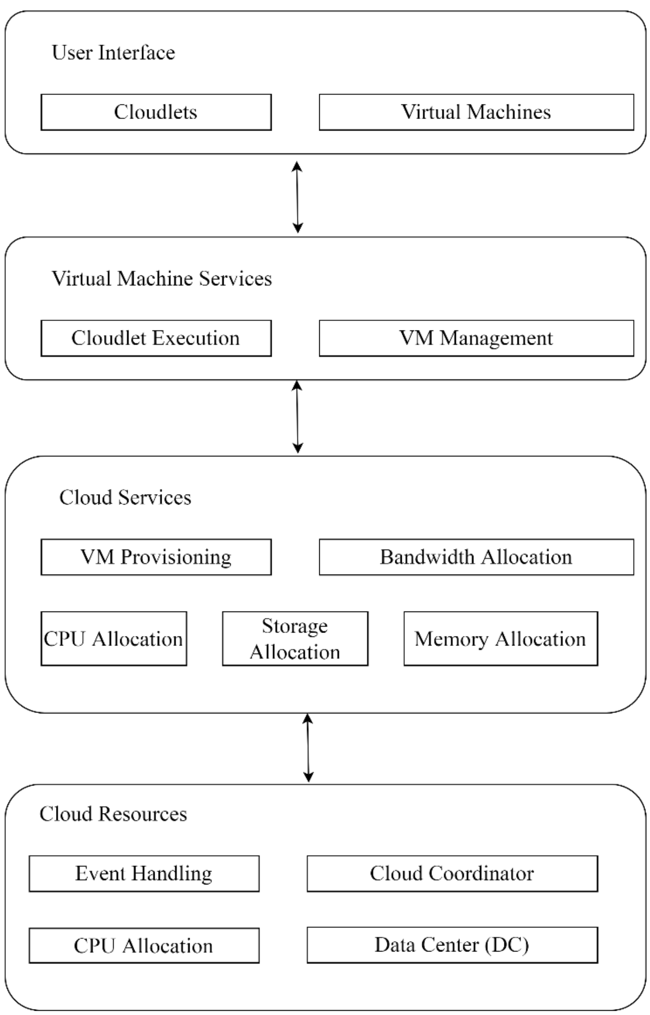

Figure 4: Architecture of CloudSim

5.2 Cloud Analyst

Cloud Analyst is a reproduction apparatus planned in light of CloudSim. It gives a graphical UI that will uphold geologically appropriated huge scope cloud applications. The general reason for Cloud expert is to concentrate on the conduct of such applications under various organization arrangements. One of the fundamental goals of Cloud Analyst is to isolate the reproduction trial and error practice from a programming exercise. Cloud Analyst is based upon CloudSim toolbox by expanding its functionalities. Notwithstanding the CloudSim parts, the central parts are a Graphical User Interface (GUI) pack, as displayed in fig. 5.

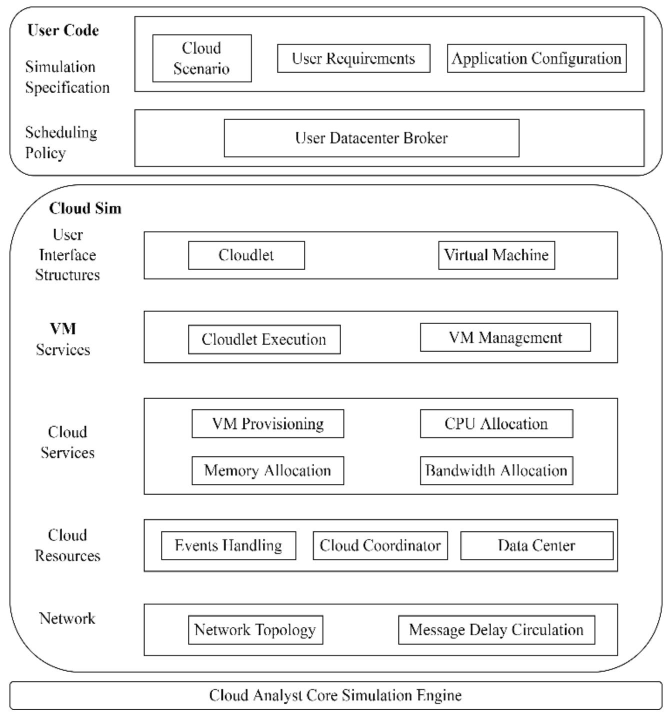

Figure 5: Architecture of Cloud Analyst

There are various tasks performed by Cloud Analyst. It comprises of reaction time, handling time, cost, and so forth by performing different recreation activities the cloud supplier can zero in on the best methodology for dispensing the assets, picking the data center, advancing expense on the basis of request made. A summary of the various activities carried out by the cloud analyst tool is depicted in Fig. 6.

6. Components of Cloud Analyst

6.1 Geography & Region

The world is split into six distinct regions or districts, and each of these regions has a user base (UB) and a data center (DC) that can be found in at least one of the other regions.

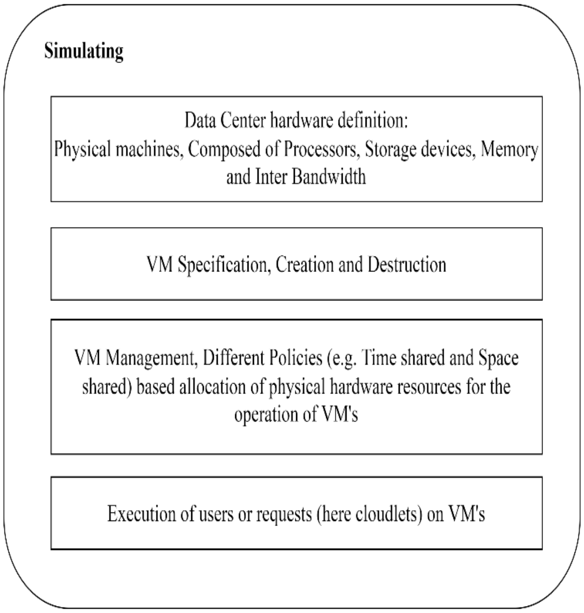

Figure 6: Cloud Analyst Tasks

6.2 Internet

A true representation of the word web. Reasonable data transmission and move delays are presented and can be configured for each of the world's six regions.

6.3 Internet Cloudlet

Gathering of solicitations from clients. Requests can be packaged as a single Internet Cloudlet in the advance tab, with information, for example, the size of the solicitation execution request and the quantity of info and result archives remembered for the bundle.

6.4 Cloud App Service Broker

When using Cloud Analyst, the Service Broker is in charge of managing traffic between data centers and userbases, and it has the authority to determine which data center is most advantageous for a particular request from a client base. Using one of the three routing arrangements provided, the service broker can choose either the closest data center or the while dynamically reconstructing itself in accordance with the load routing strategy, the one that has the quickest reaction time. In terms of network idleness, the closest data center or optimize response time scheduling policy routes traffic from the source client base to the nearest data center. When a node is overloaded, dynamic reconfiguration with load routing strategy distributes loads across different nodes. When the load on a node increases to the point where it causes data center degradation even below a predetermined threshold value, this technique distributes the load evenly among the available data centers using a distributed computing model.

6.5 Data center controller

As a key component of the Cloud Analyst area concept, it deserves special mention. It is the front face that the Cloud Analyst uses to gain access to the core of CloudSim's tool stash capabilities and abilities.

6.6 VM Load Balancer

When it comes to balancing the workload of a server, a Virtual Load Balancer gives you greater options because it distributes traffic across numerous network servers. The goal of virtual load balancing is to simulate software-driven infrastructure by utilizing virtualization technology. It emulates the functionality of a physical load balancing appliance by running the software on a virtual machine.

6.7 GUI

GUI is short for Graphical User Interface. It is responsible for providing an interaction interface between cloud analyst software and user. Using GUI, one can set parameters for userbases, virtual machine configuration, saving or loading configuration as well as execute and drop the simulation.

6.8 Simulation Configuration

For the cloud analyst tool to perform load balancing policy analysis, it is critical to configure the various components of the tool.
Table 1, 2, and 3 depict how we configured the cloud analyst tool for the Application Deployment, Data Center configurations, and User Base using cloud analyst tool. The world is divided into six distinct regions or districts in cloud analyst. In configuration of simulator we used simulation time setting for 60.0 minutes and Closest Data Center policy as "Service Broker Policy".

Table 1: User Bases Configuration

User bases					
Name	UB1	UB2	UB3	UB4	UB5
Region	0	1	2	3	4
Requests per User per Hr	8	15	12	7	9
Data Size per Request(bytes)	1000	1000	1000	1000	1000
Peak Hours Start (GMT)	5	12	17	16	1
Peak Hours End (GMT)	11	15	22	24	7
Avg Peak Users	150000	1000	50000	300000	400000
Avg Off-Peak Users	15000	100	1500	30000	100

Table 2: Application Deployment Configuration

Application Deployment			
Data Center	DC1	DC2	DC3
#VM	125	100	75
Image Size	1000	1000	1000
Memory	2048	2048	512
BW	1000	1000	1000

Table 3: Data Center Configuration

Name	UB1	UB2	UB3
Region	0	1	2
Arch	x86	x86	x86
OS	Linux	Linux	Linux
VMM	Xen	Xen	Xen
Cost per VM $/Hr	0.1	0.1	0.1
Memory Cost $/s	0.05	0.05	0.05
Storage Cost $/s	0.1	0.1	0.1
Data transfer Cost $/Gb	0.1	0.1	0.1
Physical HW Units	2	1	1

7. Results and Analysis

Following the simulation, the cloud analyst was capable of achieving the results depicted in the accompanying illustrations. Figs. 7-9 shows the arrangement for each heap adjusting strategy based on the outcome determined for measurements such as response time, cost, and loading time in satisfying the request.

7.1 Response time

All of our calculations are carried out with the help of the cloud analyst tool, and the findings are given in tables 4, 5, and 6, correspondingly, for overall regional response time and data center request processing time of the load balancing algorithms Round robin (RR), Equally

Round Robin Algo:

Table 4: Overall Response Time of RR

Overall Response Time Summary			
	Average (ms)	Minimum (ms)	Maximum (ms)
Overall Response Time	242.81	38.74	465.30
Data center processing time	16.40	0.47	57.16

Response Time by Region					
User base	UB1	UB2	UB3	UB4	UB5
Avg (ms)	67.486	52.584	52.621	357.692	304.508
Min (ms)	46.344	42.696	38.738	265.696	254.021
Max (ms)	89.055	63.155	63.859	465.302	362.224

Spread Current Execution Algorithm (ESCE), and throttle respectively. Based on the simulation results, we discovered that the RR and ESCE algorithm have nearly identical overall response times in practice. While in the case of Throttled policy it creates a large difference between it and the other two policies. We've also compiled a list of the response times of various user bases from different regions in relation to different data centers.

ESCE Algo:

Table 5: Overall Response time of ESCE

Overall Response Time Summary			
	Average (ms)	Minimum (ms)	Maximum (ms)
Overall Response Time	241.43	38.74	455.22
Data Center Processing Time	14.60	0.47	39.78

Response Time by Region					
Userbase	UB1	UB2	UB3	UB4	UB5
Avg (ms)	66.924	52.59	52.559	355.746	303.845
Min (ms)	46.344	42.696	38.738	265.696	254.021
Max (ms)	85.759	63.155	63.859	455.222	358.461

Throttle Algo

Table 6: Overall Response Time of Throttle

Overall Response Time Summary			
	Average (ms)	Minimum (ms)	Maximum (ms)
Overall Response Time	236.94	38.74	453.82
Data Center Processing Time	8.32	0.47	25.01

Response Time by Region					
Userbase	UB1	UB2	UB3	UB4	UB5
Avg (ms)	64.22	52.603	52.541	349.868	304.079
Min (ms)	46.344	42.696	38.738	265.696	254.021
Max (ms)	85.496	63.155	63.859	453.817	362.9

7.2 Request Servicing Time of Data Centers (DCs)

Data centers' request servicing times for each loading policy are presented in Tables 7, 8, and 9 correspondingly, for each data center.

Round Robin Algorithm:

Table 7: Overall DC Request Servicing Time of Round Robin

Data Center	DC1	DC2	DC3
Avg (ms)	9.526	2.83	20.135
Min (ms)	1.546	0.645	0.465
Max (ms)	18.467	3.382	57.162

ESCE Algorithm:

Table 8: Overall DC Request Serving Time of ECSE

Data Center	DC1	DC2	DC3
Avg (ms)	8.882	2.817	17.695
Min (ms)	1.542	0.645	0.645
Max (ms)	14.871	3.382	39.776

Throttle Algorithm

Table 9: Overall DC Request Servicing Time of Throttle

Data Center	DC1	DC2	DC3
Avg (ms)	5.425	2.821	9.944
Min (ms)	1.545	0.645	0.465
Max (ms)	10.455	3.382	25.013

7.3 Virtual Machine Migration

In virtual machine migration, the main goal is to distribute a load of an overloaded data center among different data centers. It is just like the process of unloading the heavy machine by shifting its load to various different machines. We consider a whole machine to be a file or series of files using virtualization, move virtual computers across real machines thanks to this feature.

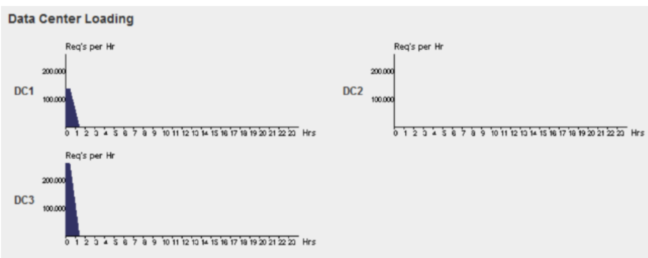

Figure 7: DC Loading Time

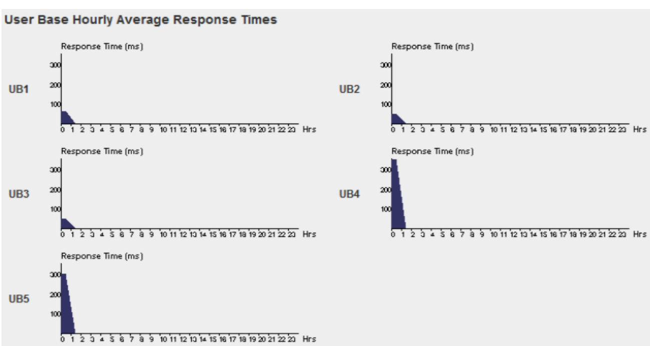

Figure 8: UB Hourly Average Response Time

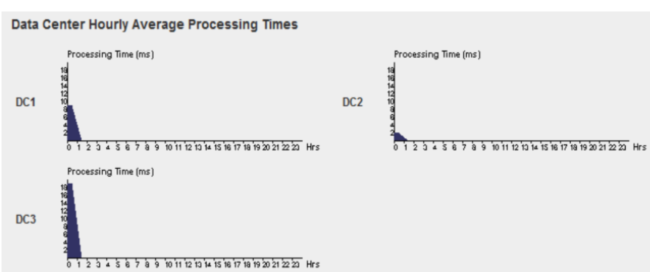

Figure 9: DC Hourly Average Processing Times

7.4 Energy Management

The world has seen an expansion in energy guidelines as of late demonstrating the more noteworthy need for energy management instruments. In the worldwide market, particularly for service organizations, energy advisors and energy administrators, it has become unbelievably vital to have the option to serve and oversee numerous clients from anyplace. Energy management becomes need of any association in any industry. Hence, explicitly cloud-based energy management software has turned into the predominant instrument.

7.5 Rising of Small Data Centers

The main reason behind the rise of small data centers is as they are more affordable. Small suppliers convey the cloud computing services prompting geo-variety figuring. Continuous load adjusting will become a global concern in order to ensure adequate reaction time while maintaining an asset's optimal course.

7.6 Data Storage Management

With the passage of time, keeping track of a big volume of data becomes a formidable task. Maintaining the continually rising volume of data generated by an organization in data centers is proven problematic for cloud computing service providers. To attain the optimum response time for accessing data and optimum storage in cloud will be a great challenge in future.

8. Conclusions

The response time and data moving cost are tests that each architect must pass in order to foster items that can improve business performance in the cloud-based area. Few systems require effective planning and burden-adjusting asset distribution procedures, resulting in increased functional costs and lower consumer satisfaction. In a cloud-based environment, a more substantial test in reaction time reduction is normally seen for each and every architect of the IT area in order to develop the things that can raise the effectiveness of company execution and client fulfillment. These factors were taken into account when simulating three distinct load-adjustment methods for executing client demands in a cloud climate: round robin, Equally Spreaded Current Execution Load (ESCE) and throttled. For each computation, a model is developed that predicts the average response time as well as the data center processing time. During examining the response time of data centers for particular regions we have found that more nearer the userbases and data centers, more efficient will be the performance of cloud and vice versa.

References

[1] Cecci, H. (2021, August 2). Gartner Says Four Trends Are Shaping the Future of Public Cloud. Gartner; www.gartner.com. https://www.gartner.com/en/newsroom/press-releases/2021-08-02-gartner-says-four-trends-are-shaping-the-future-of-public-cloud

[2] NIST's Definition of Cloud Computing. (2021, December 2). Https://Reciprocity.Com/Blog/Nists-Definition-of-Cloud-Computing/.

[3] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility, Future Generation Computer Systems", Volume 25, Number 6, Pages: 599-616, ISSN: 0167-739X, Elsevier Science, Amsterdam, The Netherlands, June 2009.

[4] Albugmi, Ahmed A - Al Assaf, Madina AU - Walters, Robert AU - Wills, Gary PY - 2016/08/01 SP - T1 - Data Security in Cloud Computing DOI - 10.1109/FGCT.2016.7605062 ER –

[5] Hassan, H.B., Barakat, S.A. & Sarhan, Q.I. Survey on serverless computing. J Cloud Comp 10, 39 (2021).

[6] Afzal, S., Kavitha, G. Load balancing in cloud computing – A hierarchical taxonomical classification. J Cloud Comp 8, 22 (2019).

[7] R. Shimonski. Windows 2000 & Windows Server 2003 Clustering and Load Balancing. Emeryville. McGraw-Hill Professional Publishing, CA, USA (2003), p 2, 2003.

[8] R. X. T. and X. F. Z.. A Load Balancing Strategy Based on the Combination of Static and Dynamic, in Database Technology and Applications (DBTA), 2010 2nd International Workshop (2010), pp. 1-4.

[9] Monika Choudhary, Sateesh Kumar Peddoju "A Dynamic Optimization Algorithm for Task Scheduling in Cloud Environment" IJERA ISSN: 2248-9622 Vol. 2, Issue 3, May-Jun 2012, pp.2564-2568.

[10] Saroj Hiranwal and Dr. K.C. Roy, "Adaptive Round Robin Scheduling Using Shortest Burst Approach Based On Smart Time Slice" International Journal Of Computer Science And Communication July-December 2011 ,Vol. 2, No. 2 , Pp. 319-323.

[11] Chaczko Z, Mahadevan V, Aslanzadeh S, Mcdermid C. Availability and load balancing in cloud computing. In International Confer-ence on Computer and Software Modeling, Singapore 2011 Sep 16 (Vol. 14)

[12] Rai, S., Sagar, N., & Sahu, R. (2017). An Efficient Distributed Dynamic Load Balancing Method based on Hybrid Approach in Cloud Computing. International Journal Of Computer Applications, 169(9), 16-21. doi: 10.5120/ijca2017914876

[13] Soumya Ranjan Jena and Zulfikhar Ahmad, " Response Time Minimization of Different Load Balancing Algorithms in Cloud Computing Environment", International Journal of Computer Applications, Volume 69,No.17,Pages:22-23 May 2013.

[14] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov ,C´ esar A. F. De Rose and Rajkumar Buyya (2010), "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms", Software: Practice and Experience (SPE), Volume 41 Number 1, pp.23-50.

[15] Y. Zhu and Y. Hu, "Efficient, Proximity-Aware Load Balancing for DHTBased P2P Systems," IEEE Trans. Parallel and Distributed Systems, vol. 16(4), pp. 349-361, Apr. 2005.

[16] Tadapaneni, N. R. (2017). Artificial Intelligence In Software Engineering. Available at SSRN: 3591807 or doi: 10.2139/ssrn.3591807

[17] R. Buyya, A. Beloglazov, J. Abawajy, Energy-efficient management of data centers resources for Cloud Computing: a vision, architectural elements, and open challenges. arXiv preprint arXiv:1006.0308 (2010)

[18] R. Buyya, R. Ranjan, R.N. Calheiros. Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities. High Performance Computing & Simulation, 2009. HPCS'09. International Conference on. IEEE, (2009)

Cite this article as: Aditiya, Rashim Rana, Simulation of load balancing algorithms using cloud analyst for distinct regions IT resources, International journal of research in engineering and innovation (IJREI), vol 6, issue 2 (2022), 132-141. *https://doi.org/10.36037/IJREI.2022.6206*.