Host Utilization Algorithm to Reduce Energy Consumption in Cloud Computing

S. Kanagasubaraja, Hema M, Vineeth M, Shreyas Rao K, Siva Anandh S

Abstract- The utilization of distributed computing server farm is developing quickly to fulfill the large increment required for systems administration, High-Performance Computing(HPC) as well as stockpiling assets for executing business and logical applications. The process of Virtual Machine (VM) solidification is inclusive of VMs getting relocated in order to make use of less physical servers. As a result, it enables the shut down or low-power mode of more number of servers which enhances the vitality utilization effectiveness, working expense and CO2 discharge. An urgent advance in VM union is have over-burden discovery, which endeavors to foresee whether a physical server is going to be oversubscribed with VMs. On the contrary to usual studies which performed utilization of CPU being the standalone indicator for host overload, a multiple correlation host overload detection algorithm was proposed in the recent study by considering a lot of factors in this regard. A higher load balance model was introduced in this text for the general public cloud, supported by the concept of cloud partitioning, in addition to a switch mechanism used to strategize differently under different scenarios. The IP address is generally shared by a true server and carbo balance. In this regard, the load balancer considers the interface developed with IP address which accepts request packets and the packets are directed to the selected servers. With an aim to improve the efficiency in public cloud environment, the algorithm employed the sport theory in the load balancing strategy.

Key words: CO2, CPU, Virtual Machine.

I. INTRODUCTION

Cloud computing is a pay-as-you-go on-demand computing service that enables its users to access applications, storage and processing power over the web. Companies can easily rent the computing infrastructure, data centers, applications or storage from a cloud service provider in order to conduct their business operations instead of incurring huge cost of owning it. The primary advantages in choosing cloud computing service are the cost incurred as investment in the equipment, followed by challenges involved in ownership and the maintenance of own IT infrastructure. In cloud computing service, the users exert control over the usage of resources. In parallel, numerous advantages are available for cloud computing service providers by providing identical services to wide range of consumers. In addition to the standard fundamental operations like storage, networking and processing power, the cloud computing services now offer tongue processing as well as AI. If a service can be delivered without the manual presence, then such services can also be delivered via cloud computing.

A virtual machine is an OS that's installed on special software called hypervisor. The hypervisor emulates CPU, memory, network, hard disk, and other resources completely enabling VMs to share the resources. You can use this method to put in several virtual machines/operating systems on one physical server and therefore the end-user who uses these virtual machines will have the identical experience that they'd have if they used dedicated hardware. In the cloud, the Virtual machine technology plays a significant role as most of the cloud hosting are done as shared hosting. So several VMs are created on the Host server, and users use these VMs to host and run their programs as separate servers. This can be true for many cases unless we use bare-metal servers within which we directly employ an actual dedicated server. Within the Cloud server, a hypervisor are going to be installed, allowing Virtual machines to be placed on top of the hypervisor. This has enabled a replacement stream of Remote desktop client represented by the term DaaS (Desktop as a Service). The DaaS computers allow users to use virtual computers hosted on a cloud as personal computers by connecting them via a client program; the web powers of these. As DaaS solutions are more convenient for remote working, many organizations turning into them. When the service provider and the client enters a mutual business commitment, it is commonly termed as Service-Level Agreement (SLA) which covers the specific deliverables with regards to quality, availability, responsibilities of both service provider as well as the service user. The commonly cited component in an SLA is the supply of service to the customer as per the agreement shared between the parties. For instance, the Internet Service Providers (ISPs) and Telco's enter into service level agreements in line with their contracts with customers. These SLAs written in simple language define the level(s) of service to be rendered. In this scenario, the SLA general covers technical definitions for MTBF and MTTR with the former being unit of time between failures while the latter being unit of time to repair or unit of time to recovery; accountability between the parties for reporting faults and paying fees incurred and responsibility for different data rates throughput or likewise measurable details.

II. RELATED WORKS

Lian et al [1] conducted a study in which the researchers developed a vibrant as well as combined resource scheduling algorithm that enables both dynamic as well as integrated resource scheduling and integrates the memory, CPU as well as the network bandwidth for physical machines.

Revised Manuscript Received on May 25, 2020.
* Correspondence Author

Dr. S. Kanagasubaraja*, Easwari Engineering College, Chennai, India.
Hema M, Easwari Engineering College, Chennai, India.
Vineeth M, Easwari Engineering College, Chennai, India.
Shreyas Rao K, Easwari Engineering College, Chennai, India.
Siva Anandh S, Easwari Engineering College, Chennai, India.

Though it calculates the combined magnitude of the full imbalance cloud data center level with average imbalance, this algorithm do not take energy efficiency into account for cloud data centers. Tang et al [21] developed a GA-implemented hybrid algorithm which accounts energy consumption as a prominent element in both physical machines as well as in communication network. This algorithm is considered to be the best when compared with heuristic algorithms, one such search algorithm whereas HGA is the optimum universal search algorithm. The computation time gets higher downside the hierarchy of heuristic algorithms. In the study [3] conducted by Castro et al, the CPU as well as RAM energy were used since it enhances VM assignment by following an influence model. In this model, the energy consumed by CPU as well as RAM was considered leaving aside the energy consumed by memory and communication network in the data center setup. Li et al [4] developed a study which leveraged multi resource double threshold method as well as modified PSO. This study designed a double threshold model along with multi resource utilization in order to induce the VMs that planned to shift. In general, the swarm optimizations do not look at the local optima, a common issue faced by heuristic algorithms. While at the same time, its disadvantage is its non-consideration of memory and communication network. In the study conducted by Fareahnakian et al [5], prediction aware VM consolidation was used. This algorithm is unique since the UP-VMC counts both CPU utilization as well as memory SLA violations and is found to be a combination of memory and CPU utilization. The emergence of SLA violations is common due to memory mix and CPU utilization. However, one should assure that the network resource utilization and traffic are considered for the purpose of optimizing VM placement. The multivariate analysis host overloaded detection was used by Adele et al [6] that notably reduced the amount of energy consumption while at the same time, there was no compromise in adhering to SLA i.e., achieved high level of SLA adherence. However, it makes use of the workload that remains not-so-significant for the bandwidth.

III. PROPOSED SYSTEM

System design:

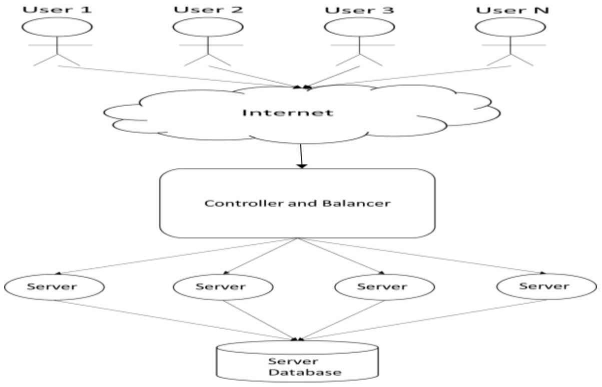


Fig 3.1

The figure 3.1 explains how the components in the architecture are distributed throughout the system

Multi-dimensional regression host utilization for host overload and underload detection algorithms(MDRHU-AS):

For the subsequent project, the host overload and also the underload are found with the assistance of a multivariate analysis based algorithms. In order to seek out the overload and also the underload we use three basic factors and that they are memory, bandwidth and also the CPU utilization. Also, they need two major components with which the entire algorithm depends on and one amongst them is that the data that is available for the independent factors of executing the VMs. the opposite component could be an important component when the working of the algorithm for the host underload and overload is taken into account is to use the three different metrics and mix them specified a novel derivative which holds the worth of the general host utilization. We hence used different models out of which one is Geometric Relation (GR) in order to execute host utilization formulation. Euclidean Distance (ED) in addition to Absolute Summation (AS) is the two alternate models which are used to propose the formula for host utilization.

MDRHU-AS: The MDRHU-AS algorithm is the alternative algorithm that is similar to MDRHU-ED, which estimates host utilization using:

AbsoluteSummation:

Utilization value= $\frac{(UC+RAM+BW)}{NCA}$

NCA= $|d(UC)| + |d(UM)| + |d(UN)|$

where NCA is the normalization constant for absolute summation. It becomes critical to make a note that the substitute approaches proposed herewith i.e., (MDRHU-AS) is utilized to determine the host utilization with the help of profiled data in relation with memory, CPU as well as BW utilizations respectively. In line with figure 1, multiple regression algorithm is used to perform the next step i.e., to develop a general model which tracks down the independent variables such as CPU, memory and BW so that it aligns with the respective dependent variable i.e., host utilization. Being an add-on element of simple linear regression, the multiple regression has a primary objective i.e., to forecast the dependent variable's value. i.e., utilization of the host based on the values of different independent variables (such as BW, memory as well as CPU). The multiple regression algorithm results in the forecasting of future host utilization. As soon as the predicted host utilization is attained via the regressor, the former's value is assessed. The host is deemed to be overloaded when the predicted host utilization outperforms the threshold as per the previous study recommendations. As a next step, a VM is chosen so that it can be shifted instead of the overloaded host. With the purpose of analyzing the correlation that exists between multidimensional approach and the utilization of single factor parameters, the researcher calculated the VM utilization through GR, for every individual factor as well as the host utilization, as shown in the figure 2

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

against the two proposed multidimensional models (MDRHU-ED and MDRHU-AS) at different time slots. It can be understood from the observation that GR host utilization averages the individual utilization behavior to some extent. AS was able to track the overall utilization behavior with better values using highly notified differences in the curves.

METHODOLOGY:

In order to avoid the challenges associated with performing the repeatable industry-scale investigations in direct and real infra, the simulations are selected for academic research for testing the proposed algorithms' efficiency. In this research work, the layered shift toolkit was utilized as a simulation framework. This might be attributed to the following reasons. The layered shift allows the positioning of VM at dual levels in the order to host level and VM level whereas it also executes space-shared provisioning techniques and time-shared provisioning techniques while the former allot specific CPU cores among the VMs. These techniques exhibit the same behavior to the first return FCFS scheduling law. In time-shared technique, the capability of 1 core is differently distributed among the VMs. The performance of these techniques are in line with Round-Robin (RR) scheduling technique. The layered shift allows the virtualized frameworks to be modelled and it sustains on both demand resource provisioning and resource management. So, the layered shift tool kit was selected as the simulation platform in the current study. The researchers extended the layered shift toolkit with few energy-aware simulations that are unavailable in the original core framework. In case of multi-dimensional regression, normal Least sq. (OLS) multiple correlation operate is utilized for the statistic calculation, then host utilization prediction. Its price note that LR is already enforced within the Layered shift machine. However, we've got an inclination to feature academic degree implementation of HLRHOD and MRHOD. In line with this, the proposed algorithms such as MDRHU-ED and MDRHU-AS were also used in the Layered shift.

The threshold need to know whether the host is full or incomplete and this is one of the important parameters to be modifiedfor host overload and underload detection methods as it is considered as the safety parameter in layered shift. The protection parameter, on the other hand, provides a sharp definition about the system coordinates' VMs on the physical servers. In case of too rigid protection parameters, there are only less opportunities available for energy savings to occur. On the other hand, when the protection parameter is found too relaxing, there is a possibility for occurrence of high degree SLA violations. So, the researchers inclined to perform the academic degree investigation (via simulation) choosing for the protection parameter price in order to deliver the product's academic degree as an acceptable exchange between SLA violation and energy saving as denoted by the final performance metric. The results attained from the subsequent sections are considered as mistreatment of the experimentally-adjusted safety parameter. It is to be noted that after detecting bunch overload and underload, the next step is to make a choice on specific VMs to be shifted from the total host to different hosts. To finalize, the Minimum Migration Time (MMT) rule was used for VM choice whereas the modified Best match Decreasing (BFD) rule was proposed for the migration of VMs. The MMT as well as BFD are already implemented in the layered shift, an important point to consider.

POWERMODEL:

The cooling devices, disk storage, memory, power providers and computer hardware are few components that consume the power from the cloud data centers. It is challenging to propose an actual analytical model to calculate the power consumption due to the evolution of qualified and capable models of multicore CPUs down the time. As a result, the real information on power consumption, retrieved from the SPECpower benchmark results, was used against the associate degree analytical power consumption model. At every programming interval that is fixed at 300 seconds, the host overload as well as underload is assessed. The host sorts were HP Pro-Liant ML110 G4 (Intel Xeon 3040/2 cores/1860 MHz/4GB), and HP ProLiant ML110G5 (Intel Xeon 3075/2cores/2660 MHz/4 GB), and their power consumption features area unit.

PERFORMANCE METRICS:

With an aim to determine the outcome of the proposed algorithms, the researcher used the metrics listed below.

TOTAL ENERGY CONSUMPTION:

This value denotes the total energy consumed by the physical resources present in the data center.

NUMBER OF VM MIGRATION:

This reflects the time required to shift the VMs from overload and underloaded hosts to other intermediate hosts and underloaded hosts to intermediate hosts.

SLA VIOLATIONS:

The host overload and underloading and performance degradation which causes the performance degradation is captured

ENERGY AND SLA VIOLATIONS:

It is a combination of energy and the performance degradation in SLA violations

PROPOSED SYSTEM:

Cloud Load balancing is the process of distributing workloads and computing resources across one or more servers. Load balancing schemes depending on whether the system dynamics are important can be either static or dynamic. Static schemes do not use the system information and are less complex while dynamic schemes will bring additional costs for the system but can change as the system status changes. A dynamic scheme is used here for its flexibility. The model has a main controller and balancers to gather and analyze the information. Thus, the dynamic control has little influence on the other working nodes. The system status then provides a basis for choosing the right load balancing strategy. The workload is segregated among two or more server, network interface or other computing resources, enabling better resource utilization and system response time. Thus, for a high traffic website, effective use of cloud load balancing can ensure business continuity.

In addition we also use help of threshold values to improve host utilization. For this purpose we set two dynamic threshold values known as upper threshold and lower threshold. The upper threshold value is used for determining the overload state of the server and the lower threshold values is used to determine the under load state of the server. According to the states the host is redirected to the servers in a manner which provides better host utilization and even energy consumption. The following diagrams represents the work flow of determining the states.

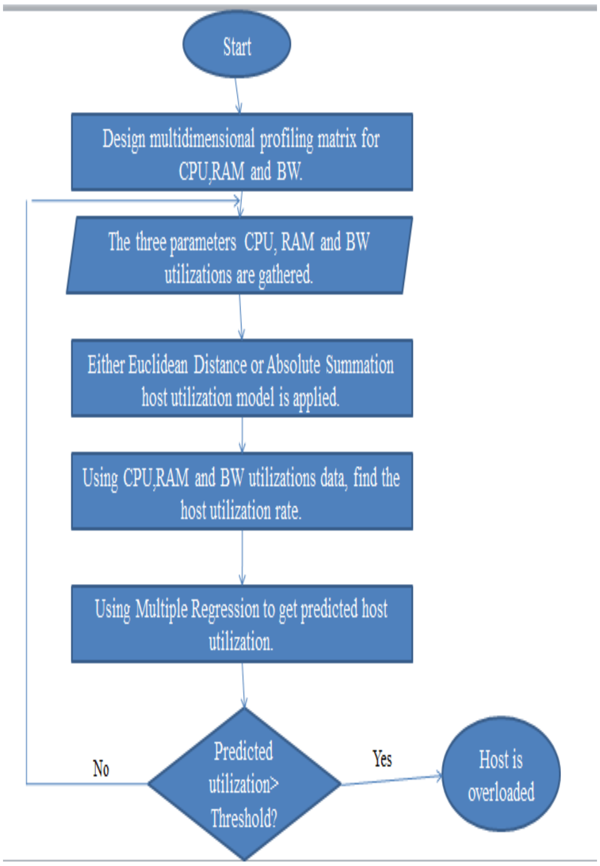

Start
Design multidimensional profiling matrix for CPU,RAM and BW.
The three parameters CPU, RAM and BW utilizations are gathered.
Either Euclidean Distance or Absolute Summation host utilization model is applied.
Using CPU,RAM and BW utilizations data, find the host utilization rate.
Using Multiple Regression to get predicted host utilization.
Predicted utilization> Threshold?
No
Yes
Host is overloaded

Fig. 3.2 Multi-Dimensional Regression Host Utilization for Overload Detection (MDRHU)

The flowchart in fig.3.2 proposes Multi- Dimensional Regression Host Utilization (MDRHU) algorithm for the overload condition is presented above. The flowchart gives a detailed step by step procedure during overload scenario. Where the steps are as follows:

Initially a multidimensional profiling matrix is designed for the parameters like CPU, RAM and BW.

After the designing of the matrix the utilization values of the above mentioned parameters CPU, BW and RAM are obtained separately for performing the further calculations.

After getting the utilization values of the parameters, either Euclidean distance or Absolute Summation host utilization model is applied.

After that the host utilization rate is calculated using the CPU, BW and RAM data that obtained earlier.

After that using the multiple regression algorithm to predict the host utilization rate.With the obtained host utilization rate check whether the rate is greater than the threshold value. If it is greater than the host is overloaded else above steps is repeated again to make it efficient.

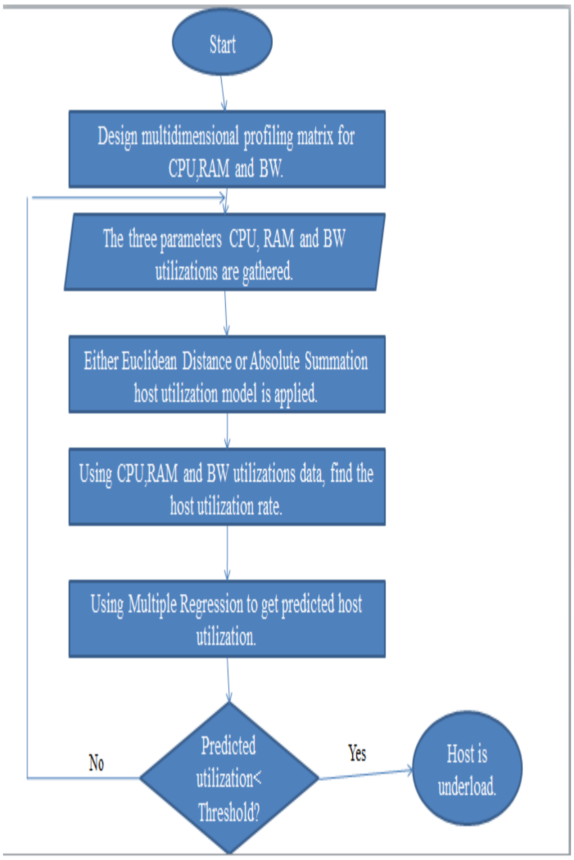

Start
Design multidimensional profiling matrix for CPU,RAM and BW.
The three parameters CPU, RAM and BW utilizations are gathered.
Either Euclidean Distance or Absolute Summation host utilization model is applied.
Using CPU,RAM and BW utilizations data, find the host utilization rate.
Using Multiple Regression to get predicted host utilization.
Predicted utilization< Threshold?
No
Yes
Host is underload.

Fig. 3.3Multi-Dimensional Regression Host Utilization for under load Detection (MDRHU)

The flowchart in Fig.3.3 proposes the Multi- Dimensional Regression Host Utilization (MDRHU) algorithm for the under load condition is presented above. The flowchart gives a detailed step by step procedure during under load scenario. Where the steps are as follows:

Initially a multidimensional profiling matrix is designed for the parameters like CPU, RAM and BW.

After the designing of the matrix the utilization values of the above mentioned parameters CPU, BW and RAM are obtained separately for performing the further calculations.

After getting the utilization values of the parameters, either Euclidean distance or Absolute Summation host utilization model is applied. After that the host utilization rate is calculated using the CPU, BW and RAM data that obtained earlier.

After that using the multiple regression algorithm to predict the host utilization rate.With the obtained host utilization rate check whether the rate is lesser than the threshold value. If it is lesser then the host is under load else above steps is repeated again to make it efficient.

IV. MODULE

Module 1: Main controller and balancers:

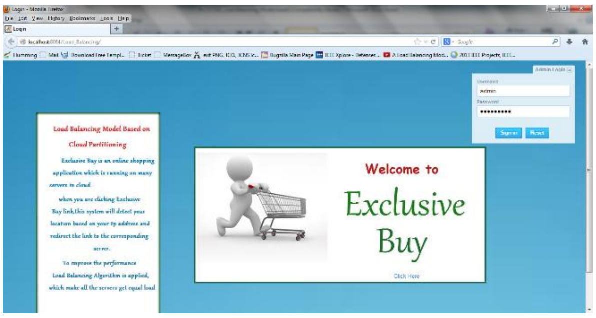
The main controller and the balancer is the main solution for the balancing of the load.

Retrieval Number: E9547069520/2020©BEIESP
DOI: 10.35940/ijeat.E9547.069520
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

The assigning of the jobs are done to the cloud partition which are suitable and then it gets in contact with the balancers in each partition so that the status information is refreshed. Higher rate of processing are yield from smaller data sets as the information for each partition is dealt by the main controller . The node status information are collected by the balancers from each partition and then the jobs are distributed by selecting the right partition.

Input:



Load Balancing Model Based on Cloud Partitioning

Exclusive Buy is an online shopping application which is running on many servers in cloud

when you are clicking Exclusive Buy link,this system will detect your location based on your ip address and redirect the link to the corresponding server.

To improve the performance Load Balancing Algorithm is applied, which make all the servers get equal load

Fig 4.1

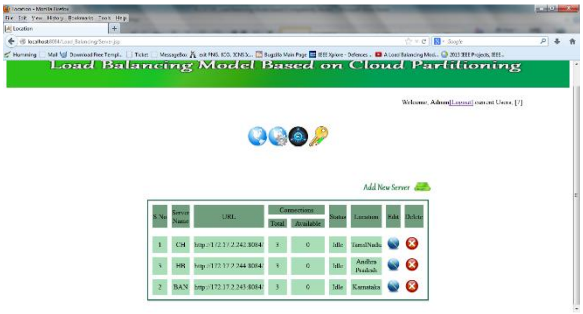
The fig 4.1 explains about the server admin page where the admin logs in to the website via a username and password to sign in to the admin page where the user can monitors the users and the different hosts that are accessing to the webpage using an external network.

Output:

Load Balancing Model Based on Cloud Partitioning
Welcome, Admin[Logout] current Users, [?]

Fig 4.2

The fig 4.2 explains about the web page after the admin has logged in to the page where in the admin can perform four main functions that are Add server, add location, Monitoring the users and change of password.



Load Balancing Model Based on Cloud Partitioning
Welcome, Admin[Logout] current Users, [?]
Add New Server

S.No	Server Name	URL	Connections		Status	Location	Edit	Delete
			Total	Available				
1	CH	http://172.17.2.242:8084/	3	0	Idle	TamilNadu		
3	HR	http://172.17.2.244:8084/	3	0	Idle	Andhra Pradesh		
2	BAN	http://172.17.2.243:8084/	3	0	Idle	Karnataka		

Fig 4.3

The fig 4.3 explains about the web page after the admin has logged in to the page where in the adding of the new servers for the different networks is done which contains the server name , url, total connections and availability, the status which is automatically which sets depending on the number of users in a server. This can be both edited and deleted.

Module 2: Assigning jobs to the cloud partition:

First, the right partition has to be chosen, when there is an arrival of a job at a public cloud. The cloud partition status can be divided into four types:

Idle: When the percentage of idle nodes exceeds balancer A, change to idle status.

Normal: When the percentage of the normal nodes exceeds balancer B, change to normal load status.

Overload: When the percentage of the overloaded nodes exceeds balancer C, change to overloaded status.

Under load: When the percentage of the under loaded nodes exceeds balancer D, changes to under load status.

Output:

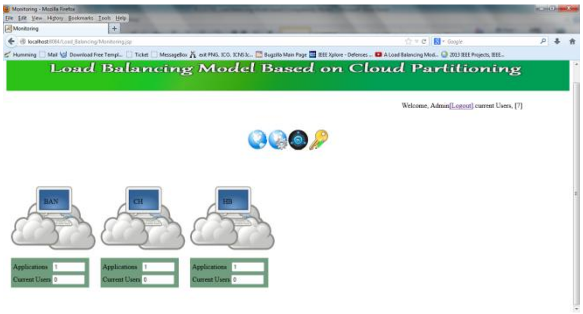


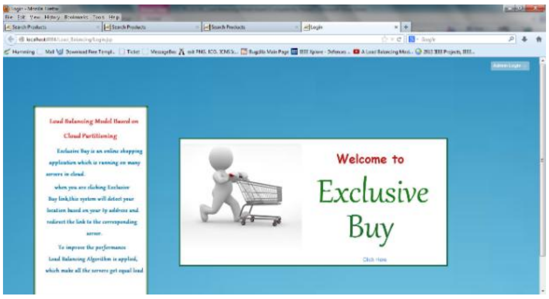
Fig 4.4

The fig 4.4 explains about the page where the different servers can be monitored where in the image for each of the server is shown with the number of applications and the number of current users that are present in a server. This keeps changing dynamically when the number of the users increase and decrease accordingly. The status information of the servers which are holding the users is recorded at regular intervals.

Module 3: Load balance strategy for the idle status:

A lot of simple load balance algorithm methods such as the Dynamic Round Robin, the Weight Round Robin and the Random algorithm. A simple algorithm here is the round robin algorithm and hence that is used. The Round Robin algorithm passes the requests that are new to the next upcoming server in the queue. There is no status information available because the status information is not recorded by the algorithm. In the regular Round Robin algorithm, there is an equal opportunity given to each node for it to be chosen. But, in a cloud that is public, the the performance of each node and configuration will be not similar; therefore, some nodes may be overloaded in this method. Hence, an alternate Round Robin algorithm is used, which called “Round Robin based on the evaluation of the load degree”. This is a very simple one. Before the Round Robin step, based on the degree of load the ordering of the nodes in the load balancing table from the lowest degree to the highest degree are made.

Input:



Load Balancing Model Based on Cloud Partitioning

Exclusive Buy is an online shopping application which is running on many servers in cloud.

when you are clicking Exclusive Buy link,this system will detect your location based on your Ip address and redirect the link to the corresponding server.

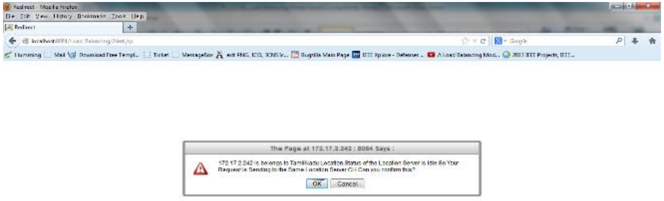
To improve the performance Load Balancing Algorithm is applied, which make all the servers get equal load

Welcome to Exclusive Buy

Fig 4.5

The fig 4.5 explains about the of the main website which acts as a front end for the project where the users via different servers. The users are given access to the server through this page where the website is accessed

Output:



The Page at 172.17.2.242 : 8084 Says :

172.17.2.242 is belongs to TamilNadu Location Status of the Location Server is Idle So Your Request is Sending to the Same Location Server CH Can you confirm this?

Fig 4.6

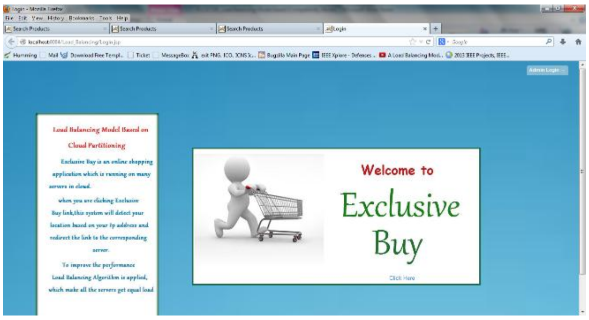
In fig 4.6, the process is such that initially when a new user tries to login to the website, the user is connected to the server with the least number of users. This is done using the algorithm in a round robin fashion. Hence, the users which arrive at the website are connected to the servers in this fashion and are allocated to the servers accordingly with the message that is given above

Module 4: Load balancing strategy for the Overload status:

When the partition of the cloud is normal, the arrival of jobs are much faster than in the idle state and the complexity of the situation is more, so an alternative strategy is applied for the balancing of the load. Every user wants their jobs to be completed in the lowest time possible, so a method is required for the public cloud where a reasonable time is taken for completing the jobs of each user. In this a static load balancing strategy based on multiple regression for distributed systems. And this work provides us with a new review of the load balance problem in the cloud environment. The jobs that are assigned to the servers that have crossed the level of threshold will be further be assigned to another server which is idle or which has the number of users being way lower than the level of threshold. Hence, each of the servers present is balanced efficiently.

Similarly for a condition when there is an under load in the server (i.e A server 1 carrying low amount of jobs where in there is another server 2 which are already having the lower threshold value) then the jobs in the server 1 are assigned to the server 2 so that a balance is maintained.

Input:



Load Balancing Model Based on Cloud Partitioning

Exclusive Buy is an online shopping application which is running on many servers in cloud.

when you are clicking Exclusive Buy link,this system will detect your location based on your Ip address and redirect the link to the corresponding server.

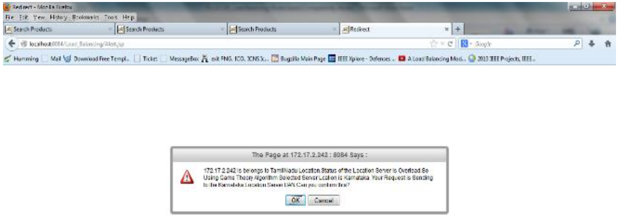
To improve the performance Load Balancing Algorithm is applied, which make all the servers get equal load

Welcome to Exclusive Buy

Fig 4.7

The fig 4.7 explains about the main website which acts as a front end for the project where the users via different servers. The users are given access to the server through this page where the website is accessed

Output:



The Page at 172.17.2.242 : 8084 Says :

172.17.2.242 is belongs to TamilNadu Location Status of the Location Server is Overload So Using Game Theory Algorithm Selected Server Location is Karnataka. Your Request is Sending to the Karnataka Location Server LAN Can you confirm this?

Fig 4.8

In fig 4.8, the process is almost same until the user arrives to the website. Here, the process is such that initially when a new user tries to login to the website, the user is tried to get connected to a server but when the algorithm uses a round robin fashion it first tries to get connected to a server which has reached a threshold level, it gives a warning message for the same and later it skips to a server which is way lower than the threshold level.

V. EXPERMENTAL RESULTS

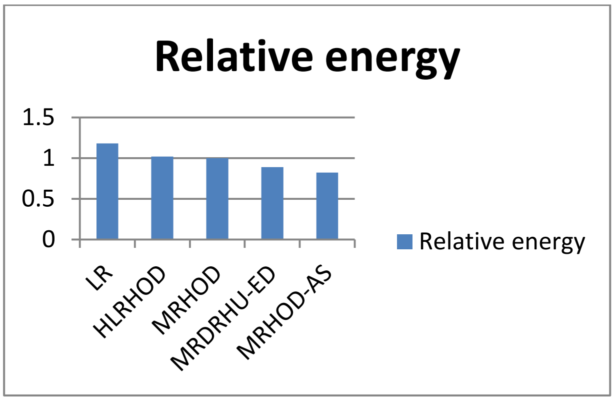


Fig 5.1

The figure 5.2 explains about the comparison of relative energy usage by the different algorithms

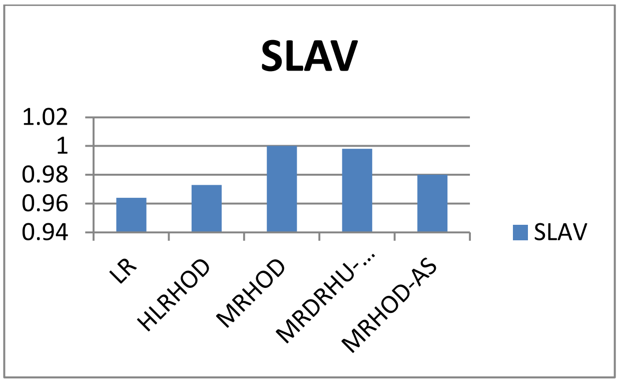


Fig 5.2

The figure 5.3 explains about the comparison Service level agreement violations by using the different algorithms

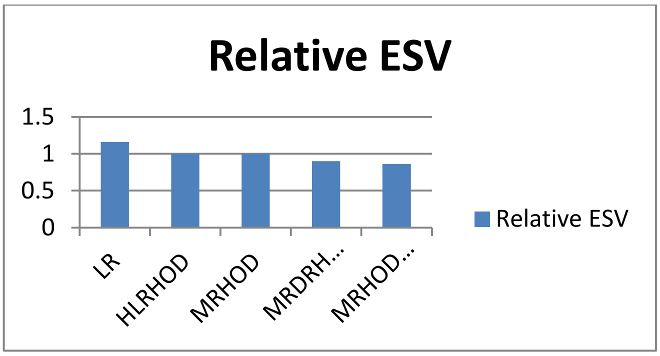


Fig 5.3

The figure 5.4 explains about the comparison of the relative energy and the Service level agreement violation by using the different algorithms

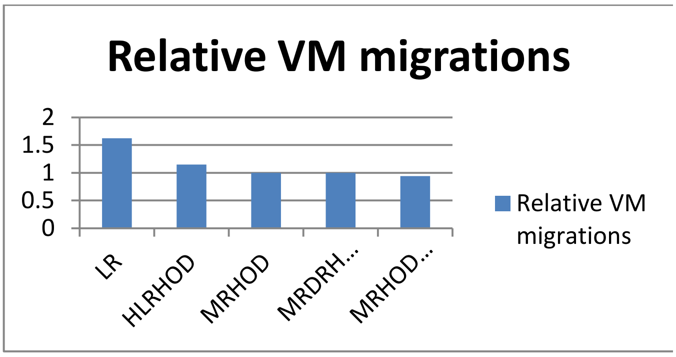


Fig 5.4

The figure 5.4 explains about the comparison of relative VM migrations taking place in different algorithms

The graph shows that the consideration of the BW utilization in workloads causes an improvement in all metrics for multiple factor algorithms. Using MRHUD-AS leads to an better energy consumption as compared to the other algorithms like MRHUD-ED, MRHOD, HLRHOD and LR which is the main goal of the project. SLAV metrics in MRHOD is reduced drastically when compared to the other algorithms, but MRHOD violated service level agreements more than 50% as compared to newly proposed algorithm MRHUD-AS. Hence, the ESV metric is greatly improved for MRHUD-AS as to using the other algorithms. The number of VM migrations is also reduced significantly by the proposed algorithms. Hence the ESM metric shoes a significant improvement for the proposed algorithm.

VI. CONCLUSION

In contrast to the majority of the past work, which utilize the CPU usage as the sole pointer for have over-burden, this project thought about different components: CPU, memory and system BW use. This is persuaded by the approach how HPC applications are con-stressed by CPU in addition to memory and BW prerequisites. In this way, this project introduced a family of novel multi-dimensional relapse have over-burden recognition calculations, which join CPU, memory and system BW usage by means of Euclidean Distance (ED) and Absolute Summation (AS), separately. The expected outcome of this project is two-crease. To start with, the exhibited calculations depend on multi-dimensional relapse, prompting improved outcomes as fare as vitality utilization and administration level understanding infringement. Second, the proposed calculations were tried utilizing genuine world HPC remaining tasks at hand. In this research, an improved version of load balance model was introduced for public cloud using the concept of cloud partitioning. In this model, a switch mechanism was also present in order to select a wide of mechanisms for dynamic scenarios. Game theory was applied by the algorithm to the load balancing strategy only with the purpose of improving the efficacy of the public cloud environment.

REFERENCE

1. Tian W, Zhao Y, Zhong Y, Xu M, Jing C (2011) A dynamic and integrated load-balancing scheduling algorithm for cloud datacenters. In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems. IEEE. pp 311–315.
2. Tang M, Pan S (2015) A hybrid genetic algorithm for the energy-efficient virtual machine placement problem in data centers. Neural Process Lett 41:211–221.
3. Castroa PP, Bareretoa V, Corrêaa SL, Granville LZ, Caredoso KV (2016) A joint cpu-ram energy efficient and sla compliant approach for cloud datacenters. ComputNetw 94:1–13.
4. Li H, Zhu G, Cui C, Tang H, Dou Y, He C (2016) Computing 98:303–317.
5. Fareahnakian F, Pahikkala T, Liljeberg P, Plosila J, Hieu NT, Tenhunen H (2016) Energy-awaree vm consolidation in cloud data centers using utilization prediction model. IEEE Trans Cloud Comput.
6. Abdelsamea A, El-Moursy AA, Hemayed EE, Eldeeb H (2017) Virtual machine consolidation enhancement using hybrid regression algorithms. Egypt Inform J 18(3):161–170.energy-efficient consolidation of virtual machines in cloud data centers.MGC, Bangalore, India. Copyright 2010 ACM 978-1-4503-0453-5/10/11. T.R. V
7. Kaushar H, Ricchariya P, Motwani A (2014) Comparison of sla basedenergy efficient dynamic virtual machine consolidation algorithms. Int JComput Appl 102:0975–8887.
8. Zhu F, Li H, Lu J (2012) A service level agreement framework of cloud computing based on the cloud bank model. Comput Sci Autom Eng (CSAE), IEEE 1:255–259.
9. Vigliotti FLDMPA, Batista MD (2014) A green network-aware vms placementmechanism. In: Proceedings of the IEEE Globecom. IEEE, Austin.
10. Sharma O, Saini H (2016) Vm consolidation for cloud data center using median based threshold approach. Twelfth Int Multi-Conference Inf Process-2016 (IMCIP-2016), Procedia Comput Sci 89:27–33.

11. Zhou Z, Hu Z, Song T, Yu J (2015) J Cent South Univ 22:94–98 Beloglazov A, Buyya R (2012) Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. Concurr Comput:Pract Experience (CCPE) 24:1397–1420.
12. Monil MAH, Rahman RM (2016) VM consolidation approach based on heuristics, fuzzy logic, and migration control. J Cloud Comput 5:8.
13. Moges FF, Abebe SL (2019) J Cloud Comput 8(1):2. https://doi.org/10. 1186/s13677-019-0126-y.
14. Wang H, Tianfield H (2018) Energy-aware dynamic virtual machine consolidation for cloud datacenters. IEEE Access 6:15259–15273. https://doi.org/10.1109/ACCESS.2018.2813541
15. Yousefipour A, Rahmani AM, Jahanshahi M, Energy and cost-aware virtual machine consolidation in cloud computing. Softw: Pract Experience 48(10):1758–1774. https://doi.org/10.1002/spe.2585,https://onlinelibrary. wiley.com/doi/abs/10.1002/spe.2585.
16. Anandharajan, Bhargavan D, Bhagyaveni AM (2013) Vm consolidation techniques in cloud data center. J Theor Appl Inf Technol 53:267–273
17. Tamiz M, Jones D, Romero C (1998) Eur J Oper Res 111:569–581. Wooldridge JM (2015) Introductory econometrics: A modern approach. Nelson Education