A Way For Proposing Dynamic Load Balancing Algorithm to Resource Allocation in the Cloud DataCenter

Samaneh Mokhtar Jozani[1] , Abolfazl Toroghi Haghighat[2]

Abstract— Cloud computing is an important technology in order to comprehensive share and access to a series of resources without the need to interfere with developing services providers. The important issue about cloud computing, which is also very much considered today, is the load balancing and resource allocation, which has been considered in this essay.

Issues such as load balancing in cloud data centers, speed improvement, migration between servers, efficiency and performance, management, and agreements of service level, appropriate scalability, energy consumption reduction, and so on are among challenges of this area. The purpose of this essay is to examine and balance the load in cloud data centers in such a way that this operation causes to improve the quality of services in the cloud environment. Migrating between existing servers can affect the system's performance and efficiency, but the drop and loss of efficiency can be prevented by providing an optimal method that has been raised in this research.

Based on the performed evaluation, the method presented in this essay, compared to other methods, has been caused to reduce energy consumption and improve power of consumption, which will lead to improve the quality of the final service and the proper balancing of load, and can meet the needs of users in using the cloud and its services.

Keywords — load balancing, resource allocation, cloud computing.

1 INTRODUCTION

Cloud computing has been created from a combination of two words of computing and cloud. Cloud here means a network or network of vast networks, such as the Internet, that the typical user does not know a lot from behind the scenes and what happens there. The reason for likeness of the Internet to cloud is that the Internet, like cloud, hides its technological details from users and creates a layer of abstraction between these technical details and users.[1] For example, what a cloud computing software service provider presents is online business applications that are provided through web browsers or other software to users. Applications and information are stored on servers and provided to users upon request. Details are hidden from the user's perspective, and users do not need expertise or control over the cloud infrastructure technology that they use. In some texts, computations and processing are used instead of the word "computing". Of course, computations and processing are not complete equivalent of this word. Because according to the definition of valid dictionaries, this word means using a computer or affairs that a computer does, and computation and processing is just one of these things and works.

Cloud computing is a new concept that was emerged in late 2007. In fact, this concept is generalization on the discussion of change based on the need, which states while the needs of users are changed, the manufacturer must provide the hardware, software, and services related with that need. Today, with the rapid development of the Internet, user's need often to be met through the Internet, and this matter has formed the basis of cloud computations. [2]

The National Institute of Standards and Technology (NITS) introduces cloud computing as a model for easy access, demand-based access to large, collective computing resources that can be set in the network (such as networks, servers, storage resources, applications, and services), which can be quickly prepared by minimum effort of management or minimum interaction with the serve. [3]

Cloud computing technology uses the Internet and remote access servers to keep data and applications. This technology allows consumers and business managers to use remote programs without the need to install and access their data by Internet from anywhere. Cloud computing technology is being implemented on data centers. The data center is a basic physical unit in cloud computing, which includes large software in structure, data storage resources and hardware platform. [1]

•*MSc student, Department of Computer Engineering,*
Science and Research Branch,
Islamic Azad university, Tehran, Iran
E-mail: (romina_mokhtar@yahoo.com)
•*Assistant Professor, Department of Computer Engineering,*
Qazvin Branch, Islamic Azad university,
Qazvin, Iran(at_haghighat@yahoo.com)

Virtualization is the key part of designing data centers. By using virtualization, the complexity of lower layers of cloud structure such as hardware and operating system can be hidden from the user. The purpose of virtualization in cloud computing is to reduce the physical device needs, data center risks, and maintenance and repairing costs. [1]

The increasing dynamics and complexity of cloud computing systems has led to an increase in the need for load balancing, which is an effective approach to improve the scalability of cloud services. Load balancing means sharing workload among several multiple sources, and one of the benefits of virtualization that allows to use of maximum power of hardware and physical resources in the cloud environment. Load balancing is also a necessary mechanism to increase the agreement of service level and better use of resources. Here, the meaning of load balancing can be the processor load, the amount of memory used, the delay, or the load of network. In fact, it can be said that purpose of load balancing is to find the proper mapping of works on processors in the system, so that in each processor, an approximately equal amount of works to be executed until the overall execution time reaches its lowest amount. [1]

Load balancing can be done in two types: static and dynamic. In the static model, the current state of the system does not matter and we need to awareness and knowledge about the previous state of the system. But in the dynamic model, according to the current state of the system, it is decided to load balancing. In load balancing dynamically, user's request is sent to less-used machines. This kind of load balancing will cause to increase user satisfaction, reduce response times, increase resource utilization, and thus increase system operational capacity and power.

In a cloud data center, operational construction costs are very high. Since a data center consists of physical servers, two major problems are raised about servers in the data center, which are the server overhead and the quality of service provided to the user. [1]

A server should not remain idle or have a small workload to save on costs and energy consuming. We try to manage the load volume on the servers in order to improve data center operation power and obtain proper speed of response by providing a dynamic load balancing algorithm dynamically to meet the elasticity property and dynamic allocation of resources according to the user's need. The algorithm provided by the virtual machines migration and load balancing will cause to improve the elasticity and management of the service level agreement, and this matter will cause to increase the quality of service.

Considering that in cloud computing, the conditions and input parameters of the client have a great effect on the choice of the kind of load balancing algorithm, so, in addition to the client's conditions, the quality of services and cost in the cloud data center should also be considered in the load balancing algorithm.

2 CLOUD COMUTING SERVICES

The main cloud services include Software as Service (SaaS)[3] Platform as Service (PaaS)[4] , Infrastructure as Service (IaaS)[5] . [5], [7], [9]

2.1 Software as a Service (SaaS)

In this model, commercial application software is supplied as a service to users / clients because the client uses software components of different providers in terms of need; therefore, the main purpose is to protect the information that have been formed by these services. The services of this layer include databases, database management programs and security tools. CoreCloud and Host Analytic are examples of this service that are used to examine business efficiency and performance.

2.2 Platform as a Service (PaaS)

In this model, applications are developed on a platform or operating system, allowing the user to run their applications in the cloud.Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

2.3 Infrastructure as a Service (IaaS)

In this model, computer hardware is presented similar to servers, network technology, storage, and data center space as a service, which in order to better manage resources, including operating system and virtualization technology.

3 HIGH LEVEL CLOUD COMPUTING ARCHITECTURE

3.1 User Layer

Various kinds of users, such as clients, application programmers, and administrators, communicate with cloud software through the user layer. This layer has been consisted of two application sub-layers and a programming environment. Cloud applications are visible through the application layer for end-users. Programming environment users are the application cloud layer developers who are responsible for developing and using applications on the cloud.

3.2 Cloud System Management Layer

This layer provides the management of applications and virtual infrastructures for business solutions. This layer is responsible for providing virtual resources for services such as service level management, management of using policies and licensing management. This layer supports the scattering of applications through dynamic allocation of resources, thus demand for resources using is minimized. The key components of the cloud

[3] *Software as a Service(SaaS)*
[4] *Platform as a Service(PaaS)*
[5] *Infrastructure as a Service(IaaS)*

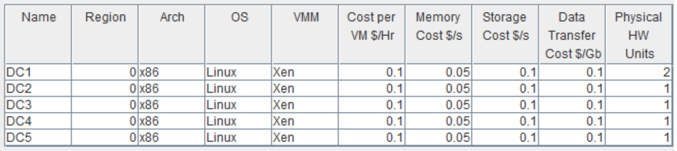
services management layer have been listed in following part.

1. SLA[6] Supervisor: When a client first sends a service request, the request is reviewed by the SLA's supervisor in order to determine the quality control requirements for QOS[7] service, acceptance or denial request of service, monitoring progress, and so on. [8] If any violation of the SLA is observed by the SLA's supervisor, it will act immediately for corrective action.
2. Provision of Resources: Virtual Machine Accessibility (VMs)[8] and required resources are followed through this mechanism and manages various requests from virtual servers, by creating multiple versions of Virtual Machine Accessibility.[4] The Resources provider is dynamically set to that processing is complete, even at peak of load. [6]
3. Arranger and Scheduler: Based on SLA supervisior information and resource provision, the arranger of tasks according to the purposes of the service provider, prioritizes or arranges them. Scheduler causes to create resource allocation by having the latest status of resource provision information, including resource accessibility and workload processing.

4 SIMULATION ENVIRONMENT

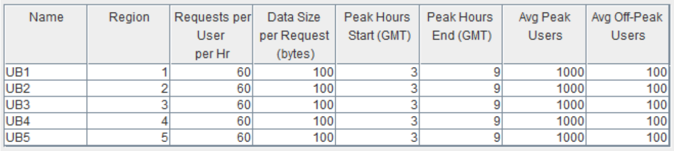
To implement the proposed algorithm from a distributor, we used three charge communicators and four computational nodes per each balancer in the Cloudsim and Java language environment. All virtual machines included in this evaluation are running on the Windows 7 operating system. The operating environment for implementation is presented in Table 4- 1

TABLE (4-1).
CONFIGURING THE DATA CENTER



Name	Region	Arch	OS	VMM	Cost per VM $/Hr	Memory Cost $/s	Storage Cost $/s	Data Transfer Cost $/Gb	Physical HW Units
DC1	0	x86	Linux	Xen	0.1	0.05	0.1	0.1	2
DC2	0	x86	Linux	Xen	0.1	0.05	0.1	0.1	1
DC3	0	x86	Linux	Xen	0.1	0.05	0.1	0.1	1
DC4	0	x86	Linux	Xen	0.1	0.05	0.1	0.1	1
DC5	0	x86	Linux	Xen	0.1	0.05	0.1	0.1	1

TABLE (4-2).
CONFIGURING USER BASES



Name	Region	Requests per User per Hr	Data Size per Request (bytes)	Peak Hours Start (GMT)	Peak Hours End (GMT)	Avg Peak Users	Avg Off-Peak Users
UB1	1	60	100	3	9	1000	100
UB2	2	60	100	3	9	1000	100
UB3	3	60	100	3	9	1000	100
UB4	4	60	100	3	9	1000	100
UB5	5	60	100	3	9	1000	100

4.1 General Specifications

In Table 4-2, all test specifications including number of users, number of requests per minute for all three methods are proposed to evaluate the proposed algorithm for comparison with the original paper algorithm.

TABLE (4-3).
TEST CONDITIONS

Category	number of users	Number of requests per minute
Test 1	500	500

In each of the experiments, various parameters are considered in the simulation step. To better understand the test and the ability to compare the results of different experiments with each other, the simulation parameters are defined in the table below. At the beginning of each experiment, a similar table is given, in which the simulation parameters of the experiment are similar to those in the table below.

1. The proposed algorithm:

Step 1: Create a DatacenterBroker and maintain a VM status indicator table and the current task assigned to each VM and determine whether it has completed processing status. At the same time creating DatacenterBroker, no VM assigned job.
Step 2: When there is a request to assign a VM, DatacenterBroker will analyze the status index table, estimating the completion time of any VM based on the formula presented below. The calculation also includes jobs available in the queue of each VM. The virtual machine will finish with the earliest time chosen for this task. If there is more than one, it is one of the first choices.

$$Capacity = \frac{\sum_{i=1}^{np} Cap(i)}{Max(\sum_{j=1}^{\delta} cores(j), np)} \quad (1)$$

Step 3: Return selected VM algorithms to DatacenterBroker.
Step 4: The DatacenterBroker post is working on the VM that is specified by the code.
Step 5: DatacenterBroker informs the algorithm about the new assignment.
Step 6: Get the algorithm to the VM status indicator table and update task.
Step 7: When the VM finishes the required processing and the DatacenterBroker is responding to the job, the update is the job that completes in the status table and reduces 1job in the index table.
Step 8: Go to Step 2

5 THE PROPOSED METHOD

Service providers must provide more satisfactory services to obtain users and, therefore, more profits. Researches have shown that the highest cost to service providers is related to the maintenance of existing infrastructures and hardware, not the creation of infrastructures. It has been also shown that the highest maintenance cost is related to energy consumption. Other problem that has been created caused by the high energy consumption is the introduction of many greenhouse gases into the environment. Improving these issues depends on the right use of hardware equipment, especially data centers. If it

[6] Service level agreement(SLA)
[7] Quality of Service(QOS)
[8] Virtual Machins(VMs)

is possible to respond to users' requests by keeping active at least the physical machines, so that the level of performance required by the contract with the service provider is not violated, then the energy consumption can be reduced, which will reduce the cost of service providers and consumers, and ultimately helps a lot to environment.

Our main purpose in this research is to increase the quality of service (QOS) in the cloud environment, which according to the following proposed algorithm that its flowchart has been shown, more and better resources can be provided in such an environment.

5.1 Figures and Tables

In Figure 1, a flowchart of a proposed algorithm has been presented in order to improve the changes in virtual machine migration, the amount of SLA violation, the average runtime, and the optimization of the standard deviation time.

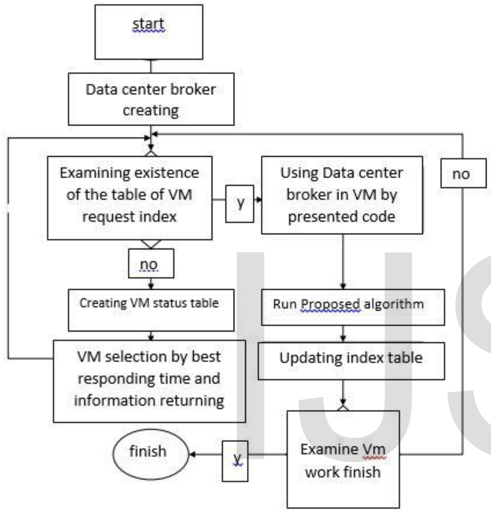

start
Data center broker creating
Examining existence of the table of VM request index
y
Using Data center broker in VM by presented code
no
no
Creating VM status table
Run Proposed algorithm
VM selection by best responding time and information returning
Updating index table
finish
y
Examine Vm work finish

Fig 1. Flowchart of the proposed algorithm

Energy consumption in cloud environments is examined from two perspectives. The first perspective of static power management, which is more related to hardware equipment, and is beyond the scope of this essay and the second perspective is the dynamic management of energy consumption which is the main subject of our discussion.

The allocation of resources can also be divided into two categories:

The first part is accepting new requests and getting virtual machines for them and putting these virtual machines on the hosts. The second part of the resource allocation process addresses to optimize previous allocations.

It should be noted that in the main essay method, we have reduced the average response time and less processing requests.

6 EVALUATION

In this section, we will evaluate and observe the improvement of energy efficiency for the declared test conditions. we used three load balancers and four computational nodes per each balancer in the Cloudsim setting and in Java language to implement the proposed algorithm from a distributor. In each test, 500 users send a total of 500 requests per minute for processing.

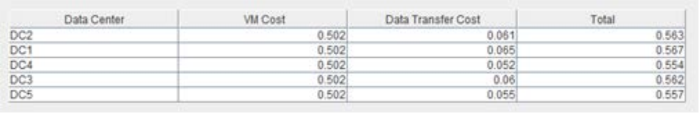
6.1 Examine and compare energy consumption

We have compared the cost of data transfer and the virtual machine cost in two ways in Table 1.

TABLE 1

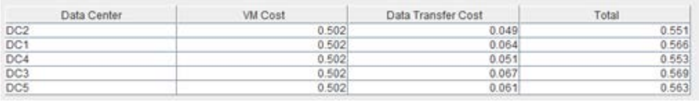
OPTIMIZING ENERGY CONSUMOTION

a) Main essay method



Data Center	VM Cost	Data Transfer Cost	Total
DC2	0.502	0.061	0.563
DC1	0.502	0.065	0.567
DC4	0.502	0.052	0.554
DC3	0.502	0.06	0.562
DC5	0.502	0.055	0.557

b) proposed algorithm



Data Center	VM Cost	Data Transfer Cost	Total
DC2	0.502	0.049	0.551
DC1	0.502	0.064	0.566
DC4	0.502	0.051	0.553
DC3	0.502	0.067	0.569
DC5	0.502	0.061	0.563

As it can be shown in Table 1, the amount of energy consumed in the main algorithm and their average of the proposed algorithm are higher.

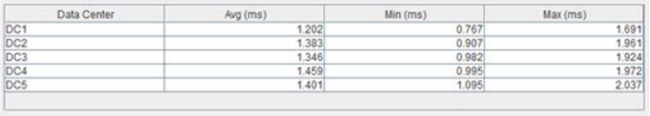
6.2 Examine and compare the power consumption rate of the system

In this section, we will have charts of power consumption rate in each cluster, then in the whole system, and the average of power consumption of the processor in the system.

TABLE 2.

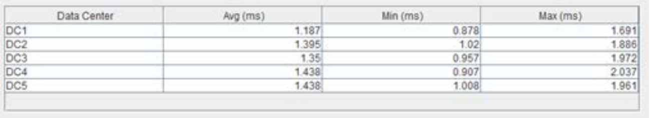
AVERAGE OF SYSTEM POWER CONSUMPTION

a) Main essay method



Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	1.202	0.767	1.691
DC2	1.383	0.907	1.961
DC3	1.346	0.982	1.924
DC4	1.459	0.995	1.972
DC5	1.401	1.095	2.037

b) proposed algorithm



Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	1.187	0.878	1.691
DC2	1.395	1.02	1.886
DC3	1.35	0.957	1.972
DC4	1.438	0.907	2.037
DC5	1.438	1.008	1.961

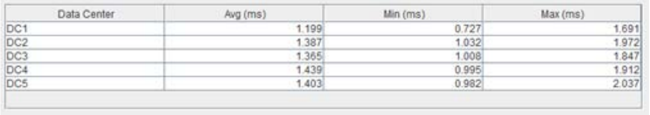
As power consumption can be seen in Table 2, the average of power consumed in the main algorithm is higher than the average of total consumed balance of the proposed method.

6.3 Examine and compare the processor consumption rate

In this section, we have examined and compared processor consumption rate in computing nodes in both algorithms. Reducing processor consumption rate in computing nodes will mean reducing final energy consumption.

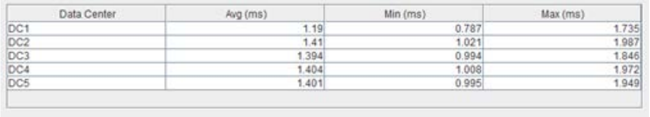
TABLE 3. EXAMINE AND COMPARE THE PROCESSOR CONSUMPTION RATE

a) Main essay metho



Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	1.199	0.727	1.691
DC2	1.387	1.032	1.972
DC3	1.365	1.008	1.847
DC4	1.439	0.995	1.912
DC5	1.403	0.982	2.037

b) proposed algorithm



Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	1.19	0.787	1.735
DC2	1.41	1.021	1.987
DC3	1.394	0.994	1.846
DC4	1.404	1.008	1.972
DC5	1.401	0.995	1.949

7 CONCLUSION

Due to the growing popularity of cloud computing, if the energy that is consumed in the cloud service provider's resources to be not controlled, then the cost of the service providing will increase in the first step and will subsequently affect the cost of paying of clients. Load balancing is one of the major challenges in cloud computing, which requires balanced local workload distribution to provide user satisfaction. Dynamic load balancing is essential to increase operational capacity and power and reduce the time of return. Although much effort has been made to obtain load balancing algorithms in the cloud environment, but given the rapid advancement of available technologies in the cloud environment and the widespread use of the cloud, a greater examining of load balancing and resource allocation is considered a must. In this essay, an algorithm was presented and simulated for dynamic load balancing in this environment. Given the performed simulation and the obtained findings in the field of load balancing in the cloud data centers, it can be seen that energy consumption has improved compared to other methods, while server operational capacity has not decreased and performance is acceptable. As a result, it must be claimed that the use of our proposed method has largely been able to fix the problems that exist in load balancing to allocate resources in the cloud environment. As a result, user satisfaction will increase in comparison with other methods. Reducing error rates and tolerance, reducing latency time, expanding, etc. are issues that need to be addressed in future researches in order to improve the quality of service.

REFERENCES

[1] S. Mustafa, B. Nazir, A. Hayat, A. Rehman Khan, S A. Madani."Resource management in cloud computing: Taxonomy, prospects, and challenges". In: computers & Electrical Engineering; Volume 47, October 20115, Pages 186-203.

[2] EPA Datacenter Report Congress. (2007). Retrived from http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Con.

[3] K. Powell ,"Quick start guide to increase data center energy efficiency",GSA Public Buildings Service, Applied Research Program and DOE Federal Energy Management Program, pp. 1,October 2008.

[4] W.Shi, and B. Hong," Towards Profitable Virtual Machine Placement in the Data Center",IEEE,,International Conference on Utility and Cloud Computing,pp.138,2011.

[5] V. Vinothina, Dr. R. Shridaran, and Dr. Padmavathi Ganpathi, "A survey on resource allocation strategies in cloud computing", International Journal of Advanced Computer Science and Applications, 3(6):97–104,2012.

[6] T. Ferreto , M. Netto, R. Calheiros,andC.De Rose," Server consolidation with migration control for virtualized data centers",Elsevier, Future Generation Computer Systems 27 -1027–1034,pp. 1028-1030,2011.

[7] A Bala, I Chana. "Prediction-based proactive load balancing approach through VM migration". In: Engineering with Computers, January 2016, pp 1-12.

[8] Saurabh Kumar Garg , Adel Nadjaran Toosi , Srinivasa K. Gopalaiyengar,Rajkumar Buyya "SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter", Journal of Network and Computer Applications , 2014.

[9] ZhixiongChen,JongP.Yoon,"International Conference on P2P, Parallel,Grid,Cloud and Internet Computing",2010 IEEE:pp 250-257.