Load Balancing Mechanisms in Amazon Web Services using Meta Heuristic Rules

Ramya Rajamanickam, T. Hemalatha,S. Puspalatha, M. Buvana

Abstract—Cloud computing is defined as the resource that can be delivered or accessed by the local host from the remote server via the internet. Cloud providers typically use a "pay-as-you-go" model. The evolution of cloud computing has led to the evolution of modern environment due to abundance and advancement of computing and communication infrastructure. During user request, and system response generation, an amount load will be assigned in the cloud computing, where it may be over or under load. Due to heavy load, power consumption and energy management problems are created, and it makes system failure and lead data loss. Though, an efficient load balancing method is compulsory to overcome all mentioned problems. The objective of this work is to develop a metaheuristic load balancing algorithm to migrate multi-server for load balancing and machine learning techniques is used to increase the cloud resource utilization and minimize the make-span time of the task. Using an unsupervised machine learning technique, it is possible to predict the correct response time and waiting time of the servers by getting the prior knowledge about the virtual machines and its clusters. And this work involves to calculate the accuracy rate of the Round-Robin load balancing algorithm and then compared it with a proposed algorithm. By this work, the response time and waiting time can be minimized and also it increases the resource utilization and minimize the make- span time of the task.

Keywords- Cloud Computing, Load Balancing, Amazon Web Service Cloud platform, Meta Heuristics Approach, Ant Colony optimization Algorithm.

I. INTRODUCTION

One of the booming and applicable environments using widely in various real time application for service providing is cloud computing. It is a multitenant environment, in which a greater number of users request for a greater number of same or different services at the same time. Due to unimaginable request and response applied in the cloud at the same time, there will be a need for load balancing. Load balancing is applied only among the servers to fulfill the users' demand in terms of service provision. Cloud computing is also called as pay n use environment where people can pay and use the services according to their requirement. To satisfy the customer, manage resource allocation and service provision, virtualization method is used by cloud computing. In accordance to the virtualization method, the cloud computing offers virtual resources to various users at low budget. It helps to mid-level companies can use own infrastructure without heavy investment. Payment for the services depends on the usage of the resources and the duration. Based on the request, loading and unloading of request and responses are applied. Different systems get failure due to loading and unloading process in terms of power consumption, execution time, machine fault, and etc. In order to over these issues load balancing method is used in cloud computing. It also includes scheduling, resource investigation and allocation in cloud. Different types of loads in the cloud computing are processors, storage and data transmission, etc.

A mechanism used to identify the heavy load location in the cloud computing due to loading and unloading is called as load balancing. Once it detects the location of the heavy load, then the load balancing clears the load in terms of scheduling or splitting the tasks into subtasks. Various parameters involved in the load balancing process is optimized for managing the loads in the network. Different research works proposed different methods or techniques to do different kinds of load balancing in cloud computing. Since, the cloud service demand is increasing rapidly day by day, it is essential to apply an effective method for load balancing without compromising the performance of the cloud computing. Load balancing also balances the resource allocation, resource utilization, increasing performance and energy management with the help of optimizing the work load among more number of users requesting more number of different or same services.

Load balancing plays an important role in cloud computing. Due to the difference in the cloud computing capacity, the instance may either over or underutilized. Without load balancing, there may be some effects to the instance/ servers with respects to the workload distribution. Demerits may be a high response time and high waiting time of the server/client. Now- a -days load balancing for cloud resource utilization is an important research topic which automatically balances the load in order to allocate the loads equally to all instances. There is a need to develop a load balancer algorithm which will effectively use to balance the load in the cloud computing servers. We have designed a system based on Amazon cloud platform to prevent no even a single node is overloaded and to perform auto-scaling. so that we can achieve less response and waiting time. And to achieve high resource utilization and minimize the make-span of a given task. The main aim of the system is to

Revised Version Manuscript Received on 20 September, 2019.

RamyaRajamanickam, PG Scholar, CSE Department, PSNA CET, Dindigul, Tamilnadu , India.
(Email: ramyarajamanickam2596@gmail.com)
T. Hemalatha, Professor , CSE Department, PSNA CET, Dindigul, Tamilnadu , India.
S. Puspalatha, Professor, CSE Department, PSNA CET, Dindigul, Tamilnadu , India.
M. Buvana, Professor, CSE Department, PSNA CET, Dindigul, Tamilnadu , India.

balance the load for cloud-based multi- servers using Meta heuristics approach.

In addition to this, the proposed model is designed to incorporate with autonomic feature like auto-scaling module so that a new instance can be added to the pool of the running instance. Also, our work uses the machine learning technique to balance the load and then compares the work with the existing round-robin algorithm, in order to prove that our algorithm works efficiently than the other.The aim of the work described in the study is to balance the load for cloud-based multi-server and also to analyze the performance of the proposed algorithm with the Round-Robin algorithm in AWS using meta heuristics approach called Ant ColonyOptimization.

II. LITERATURE SURVEY

To balance the load for multi-servers, a game theoretic approach[1] is used. So that the response time of the task is reduced by using the iterative proximal algorithm and non-cooperative game theory. To distributed the load efficiently, a dynamic load balancing is done using a machine learning technique[2].thus, this machine learning techniques gives more fine tuned analytical data for load scheduling mechanisms. The scheduling or allocation of the task[3] in the cloud environment is a NP hard problem. This work concludes that the static load balancing algorithm are suitable for homogenous environment and the dynamic load balancing algorithm suits both for homogenous and heterogeneous environment. The work [4]mainly addresses the problem of resource allocation and load balancing along with different environment in cloud computing. For efficient resource utilization and better return on investment for cloud service providers and consumer[5], different types of static and dynamic load balancing algorithms are proposed.

III. PROPOSED SYSTEM& RESULTS

Our objective is to develop an algorithm that uses the meta heuristics approach. Ant Colony Optimization (ACO) is used to balance the load, which gets the prior knowledge about the virtual machine and its clusters so that the make-span time can be minimized and then reduces the waiting time of the task. Besides AWS uses the Round Robin algorithm for balancing the load. Hence, this will be compared with the proposedalgorithm.

A. Features of Proposed System

The advantages of the proposed work are as follows,

1. Make-span time isminimized.
2. Response time and waiting time is reduced.
3. High resourceutilization.
4. Migration time of the load isreduced.
5. Increasedperformance.

IV. IMPLEMENTATION

The proposed system is experimented using Platform as a Service available in Amazon Web Services and as follows,

A. Dataset from AWSEC2

The first part mainly focused to fetch the system parameters of the AWS cloud platform. The system parameters like CPU usage, memory usage, CPU speed, RAM size, the response time of the system, number of cores in the systems are collected from 1GB RAM with different time intervals. After fetching the parameters, the data's are smoothened or scaled up a SPSS modeler to clean the data and then all the parameters are converted into the excel sheet. Figure 1 depicts the working model of AWS cloud platform.

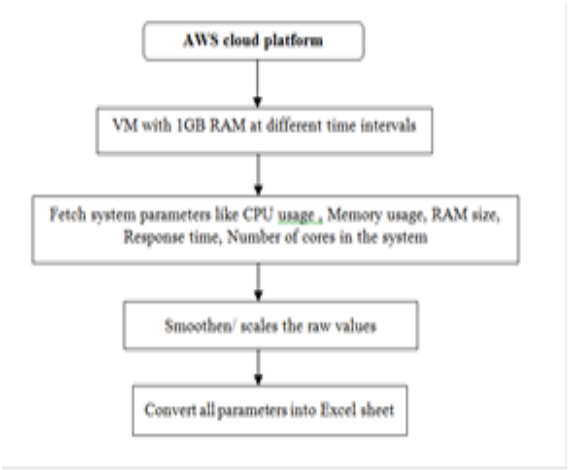

Figure 1- Working Model of AWS

B. High Level Design of PaaS

In the proposed algorithm, instead of using a game theoretic approach our work uses meta-heuristics approach. The new load balancing algorithmwhich comprises of load scheduling and load optimization technique to balance the load. And also the load estimator estimates the load of the virtual machine by calculating the load factor of the virtual machine from the database. A database stores all the information regarding the request by means of meta VMM. The purpose of meta VMM is to collect and retrieves all the information from the local instances where it has multiple virtual machines running in multiple physical machines. While moving to the part of machine learning, an unsupervised learning technique called Ant Colony Optimization (ACO) is used to optimizethe allocation of the load since the allocation is untrained. And then clustering is performed to balance or migrate the load from the virtual instances.

As a result, our work achieves less response time, less waiting time, and also minimizes the make-span time of the task. This work is used to compare the load balancing algorithm like Round-Robin with Ant Colony Optimization load balancing cloud scheduler in order to compare the efficiency of that algorithm with our new algorithm. A Web Services is created to handle all the HTTP request of the user which is easy to interoperate with both user and cloud computing environment. Figure 2 explains the high level design of PaaS.

The proposed system consists of two parts and they are,

a). Request Handler

This module is designed to handle the request. The request handler is the end point through which the cloud users integrateswith the cloud. This module is responsible for

interacting with the web server and handle the request from the cloud users. Any request can be routed to any number of instances. Instances can handle multiple requests.

b). Load Balancer

Load computes CPU using , network utilization and memory capacity load of each instance. According to that, the load balancer distributes the workload to each instance. The reliability and concurrency of cloud application are increased by using the load balancer. Rules with the support of ACO technique, the load balancer allocates the load to the VM on the basis of the following condition:

a. When the load is low and priority ishigh, the load balancer allocates the task to the VM which makes minimum make-span time.

b. When the load is high and priority is low, the load balancer allocates the task to the new Virtual Machines that reduce the make-span time of the cloudservice.

This module includes the following steps which needed to be balanced ,

1. Use Ant Colony Optimization output as an input. Then the load balancer based on this information, it applies the following conditions to allocate resources and perform the workloaddistribution.

a. When the load is low and priority is high, the load balancer allocates the task to the VM which makes minimum make-span time.

b. When the load is high and priority is low, the load balancer allocates the task to the new Virtual Machines that reduce the make-span time of the cloudservice.

2. Based on the conditions, if any condition is met, then resources are allocated accordingly and the status of resources is changed.

3. While allocating the instances for PaaS, it is necessary to meet the Service Level Agreement(SLA) made between the provides andconsumers.

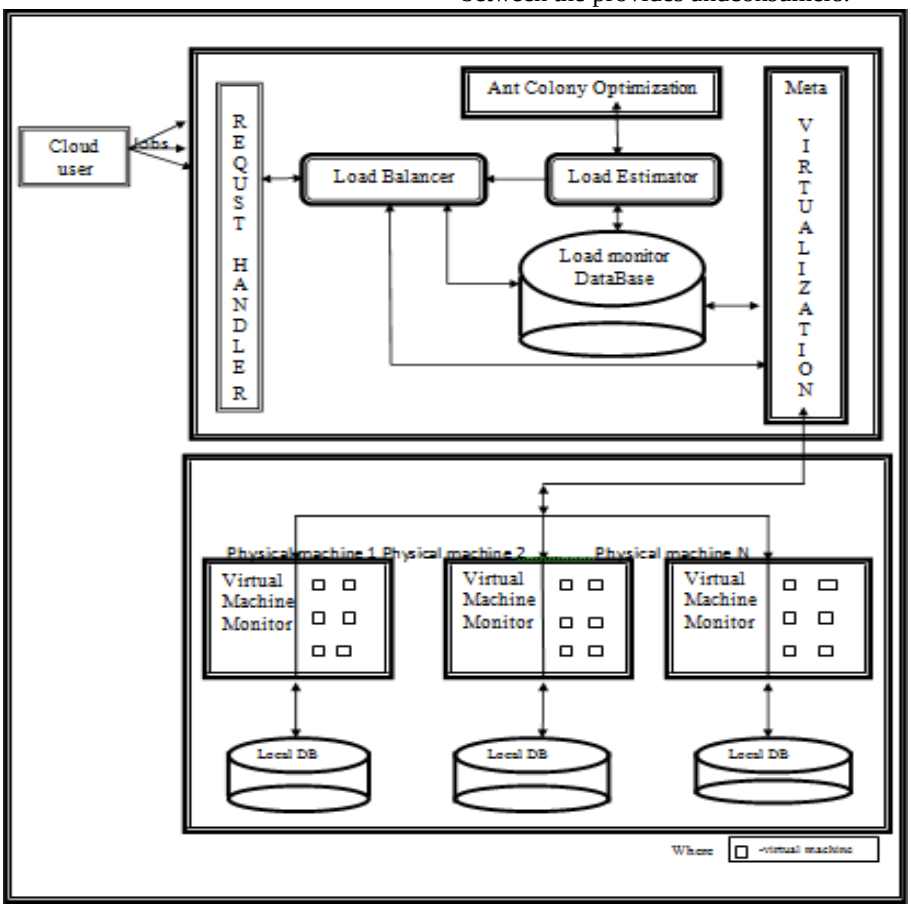

Figure 2- High Level Designof PaaS

c). Meta Virtual Machine Monitor

A Virtual Machine Monitor (VMM) is a virtualization technique that allows multiple AWS EC2 instance (guest OS) to run in a single host system all at the same time. Generally, metadata is the data about the data. Metadata describes the information about the particular datasets, objects, and resources. The set of files together of VM can be said as a meta VMM.

d). Load Estimator

The following commands are used to retrieve the system parameters from the AWS. In AWS cloud, various types of instance are provisioned on demand with various configuration which are suited for different type of applications. Few of the different types of EC2 instances which are available with the different combinations of CPU, Memory, disk and Networking in the form of Platform as a Service are in the following table I The instances are broadly categorized in to Accelerated Computing, Storage Optimized, memory optimized, compute optimized and General Purpose instance[17]. For ourexperimentation only T2 and T3 Instances are used which are General Purpose

Instance.The instances are broadly categorized in to Accelerated Computing, Storage Optimized, memory optimized, compute optimized and General Purpose instance. For our experimentation only T2 and T3 Instances are used which are General Purpose Instance.

Sl. No.	Instance Type	Specification	Details
1.	T2	3.3 GHz & 3.0 GHz Intel Scalable Processor	High Frequency Intel Xeon Processor with VCPUs of 1/2/4/8 with memory of ranging from 0.5 to 32 GB with Elastic Block store (EBS)
2.	T3	2.5 GHz Intel Scalable Processor	Max. 8 VCPU with 32 GiB (Gigabyte) Memory with EBS
3.	M4	2.5 GHz Intel Xeon	Max. 96 VCPU with384 GiB Memory with(EBS)
4.	M5	2.5 GHz Intel Xeon with Advanced Vector Instruction set	Max. 96 VCPU with 384 GiB Memory with (EBS)

Table I- Different Type of Instance Available in AWS

In this experimentation several instances of type T2 and T3 with varying memory and network capacities are used to collect the response time and make span time for various set of tasks. The tasks are chosen to simple web services to web applications with database connectivity and MySQL. Then scaling is performed to smooth the values collected at different intervals for a period of time and several durations. The Ant Colony Optimization algorithm is executed to classify collected data set along with the instance type and other features to segregate the instances so that such instances may be used for further allocation of task. Thus in AWS cloud platform, the system parameters are fetched in 1GB RAM with different time intervals in the T2 micro EC2 instances. Figure 3 and figure 4 describes the system parameters fetching. Figure 5 describes the datasets of AWS.Then the data's are smoothened and scaled by the SPSS modeler.The output is given as the input to the ACO algorithm. The algorithm clusters the significant and non-significant attributes of the system parameters and updates the information into the load monitor database.The database classifies the least load as the highest priority and high load as the lowest Priority with the use of load estimator. When the cloud user submit the request, the load balancer optimizes and allocate the load to the VM in the Oracle VM VirtualBox.

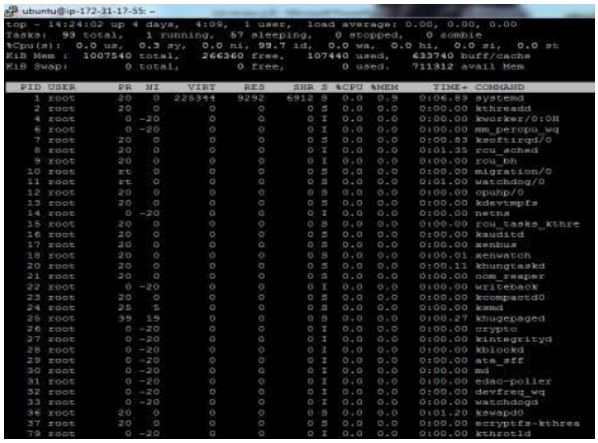

Figure 3- Fetch the System Parameters

Figure 4 – System Parameters of VM

	A	B	C	D	E	F	G
1	VIRT	RES	SRH	%CPU	%MEM	PR	NI
2	44508	4044	3456	2	0.4	20	0
3	44540	4040	3444	0.3	0.4	20	0
4	44524	4116	3500	0.3	0.4	20	0
5	44556	4068	3512	6.7	0.4	20	0
6	44556	4004	3444	6.7	0.4	20	0
7	44524	4004	3400	0.3	0.4	20	0
8	44524	3932	3328	0.3	0.4	20	0
9	44508	4032	3428	0.2	0.3	20	0
10	44524	4132	3520	0.4	0.5	20	0
11	44508	4108	3520	0.3	0.4	20	0
12	44508	3940	3388	0.7	0.8	20	0
13	44511	4008	3416	0.3	0.4	20	0
14	44508	3928	3340	1.2	1.4	20	0
15	44321	3984	3396	0.6	0.7	20	0
16	44329	4108	3520	1.6	1.9	20	0
17	44508	4108	3520	1.3	1.4	20	0
18	44207	4104	3512	0.5	0.7	20	0
19	44321	4206	3643	0.8	0.9	20	0
20	44451	3972	3384	2.1	2.3	20	0
21	44508	3940	3352	3.1	3.2	20	0
22	44524	4060	3456	1.5	1.6	20	0
23	44509	4016	3428	0.5	0.7	20	0
24	44508	4020	3432	0.3	0.4	20	0

Figure 5- AWS Datasets

V.CONCLUSION

Load Balancing approach is rooted in all areas of research, because there is some of the challenges which leads to the cloud computing performance overhead, producing more response and waiting time. So this work has an insight of less response time and minimizes the make-span time of the task in cloud services by using balancing the load using a meta heuristics approach.Experimentation is done using AWS PaaS instances of type T2 & T3 with varying VCPU, Memory and Network capacities. In the future, this work will be extended to test the proposed load balancing algorithm which will be compared with the existing round robin algorithm used in AWS Cloud.

VI. ACKNOWLEDGMENT

The authors would like to thank the Management of PSNA College of Engineering and Technology, Dindigul for the support to complete this project inside the campus by providing Lab facility.

REFERENCES

1 AlirezaSadeghiMilani, NimaJafariNavimipour (2016),' Load balancing mechanisms and techniques in the cloud environments: Systematic literature review and future trends' Journal of Network and Computer Applications, 71 (86–98), 1084- 8045.
2 Asha V, Bharath Kumar, Girish V (2018), 'Load Balancing in Cloud Computing' International Journal of Recent Trends in Engineering & Research (IJRTER), 4(101), 2455-1457.
3 AvnishThakur, Major Singh Goraya(2017), 'A Taxonomic Survey on Load Balancing In Cloud' Journal of Network and Computer Applications, 98,1084-8045.
4 AzharulKarim S .M, John J Prevost (2017),' A Machine Learning-based Approach to Mobile Cloud Offloading' Computing Conference, 978-1-5090-5443
5 BakulPanchal, SmaranikaParida (2018) 'AnEfficient Dynamic Load Balancing Algorithm Using Machine Learning Technique in Cloud Environment' International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), 4,2394-4099.
6 Bhavani B Hand H S Guruprasad (2015), 'Resource Provisioning Algorithms in Cloud Computing : A Survey' , International Journal of Research in Computer and Communication Technology(IJRCCT), 3, 278- 5841..
7 Chubo Li, Kenli Li (2018) 'A Game Approach to Multi-servers Load Balancing with Load – Dependent Server Availability Consideration' IEEE Transaction on Cloud Computing, 2168-7161(c).
8 HarshadTrivedi, vedang shah (2015), 'A Distributed Dynamic and Customized Load Balancing Algorithm for Virtual Instances', Nirma University International Conference on Engineering(NUiCONE), 978-1-4799-9991-0.
9 ManaliTrivedi, HinalSomani (2016), 'A Survey on Resource Provisioning Using Machine Learning in Cloud Computing' International Journal of Engineering Development and Research(IJEDR), 4, 2321-9939.
10 Michael Borkowski, Stefan Schulte, ChristophHochreiner (2016), 'Predicting Cloud Resource Utilization' ACM International Conference on Utility and Cloud Computing, 10(1145), 978-1-4503-4616-0.
11 MinxianXu, WenhongTian, RajkumarBuyya (2017) 'A survey on load balancing algorithms for virtual machines placement in cloud computing' Wiley, 29, 4123.
12 Narayan Joshi, KetanKotecha (2018) 'Implementation of Novel Load Balancing Technique in Cloud Computing Environment' International Conference on Computer Communication and Informatics (ICCCI), 10(1109), 978-1-5386-2238-4
13 Naveen Kumar Gandhi, Ayushi Gupta (2017) 'Survey on Machine Learning based scheduling in Cloud Computing' International Conference on Intelligent System, Meta-heuristics and Swarms Intelligence (ISMSI) , 10(1145), 978-1-4503-4798-3.
14 Padmavathi M, Shaik. MahaboobBasha (2017), 'Dynamic And Elasticity ACO Load Balancing Algorithm for Cloud Computing' International Conference on Intelligent Computing and Control Systems (ICICCS),10(1109), 978-1-5386-2745-7.
15 Raza Abbas Haidri, Katti C.P, Saxena (2014), 'A Load Balancing Strategy for Cloud Computing Environment' International Conference on Signal Propagation and Computer Technology(ICSPCT), 10(1109), 978-1-4799-3140-8.
16 Sambit Kumar Mishra, BibhudattaSahoo, PritiParamitaParida (2018) 'Load Balancing in Cloud Computing: A big Picture' Journal of King Saud University - Computer and Information Sciences, 1(03), JKSUCI 395.