Optimal Resource utilization using Dynamic Load Balancing algorithm in Cloud Computing

Namrata H. Gohel[1], Mr. Krunal Panchal[2]

[1]*P.G. student, Department of computer engineering, L.J.I.E.T., Ahmedabad, Gujarat, India.*
[2]*Assistance professor, Department of computer engineering, L.J.I.E.T., Ahmedabad, Gujarat, India.*

ABSTRACT

The cloud computing is internet based computing, whereby shared resources, software, information are provided to computers and other devices on demand, pay-per-use. Due to popularity of cloud, the users of cloud are increasing day by day and that has become important issue for cloud service provider in terms of load balancing. Load balancing used to distribute a larger processing load to smaller processing nodes for enhancing the overall performance of the system. Load balancing help in fair allocation of resources to achieve a high user satisfaction and proper resource utilization. This paper presents the basic cloud fundamentals and load balancing concepts. In this paper , propose dynamic load balancing algorithm which will take the load on the server as well as type of resource in consideration for efficient load balancing and resource scheduling.

Keyword : - *cloud, load balancing, scheduling, resource utilization.*

1. Introduction

Cloud computing is internet based computing whereby shared resources, software, and information are provided to computers and other devices on demand, pay-per-use.

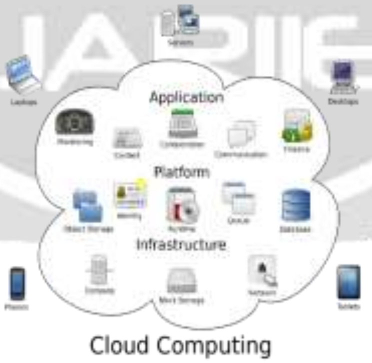

Fig-1: Cloud Computing [3]

As defined by NIST[1] "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (network, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." Cloud service provider, cloud users and cloud developers are the three participants in cloud. Physical/virtual server, storage, datacenters, networking devices, hypervisors and middleware is important component of cloud platform.

1.1 Cloud Deployment Model

Based on deployment there are four types of cloud:

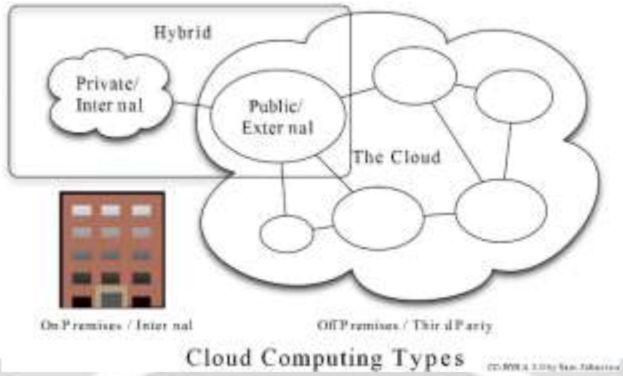

Fig -2: cloud deployment model

Public cloud – The cloud infrastructure is provisioned for open use by the general public or a large industry group and is owned, managed, and operated by a business, academic or government organization.

Private cloud – The cloud infrastructure is provisioned for exclusive use by a single organization and is owned, managed, and operated by the organization, a third party or some combination of them.

Community cloud – The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organization that have shared concern.

Hybrid cloud – Hybrid cloud is a combination of two or more clouds that remains unique entities but is bound together by standardized technology that enables data and application portability.

1.2 Cloud Service Model

There are main three service model of cloud:

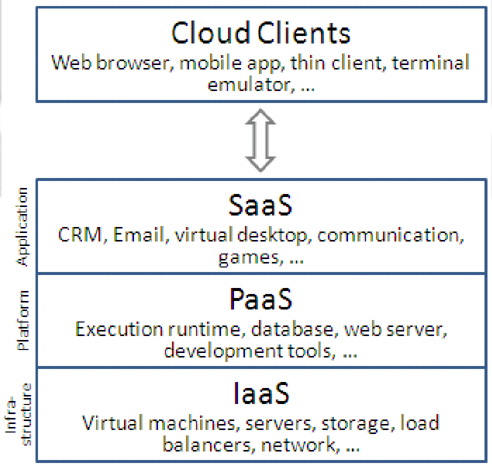
Cloud Clients
Web browser, mobile app, thin client, terminal emulator, ...
SaaS
CRM, Email, virtual desktop, communication, games, ...
PaaS
Execution runtime, database, web server, development tools, ...
IaaS
Virtual machines, servers, storage, load balancers, network, ...
Application
Platform
Infra-structure

Fig- 3: cloud service model

Software as a Service (SaaS): SaaS provided all the application to the customers which are provided by the cloud service provider.

Platform as a Service (PaaS): PaaS provides all the resources to the customers that are required for building application.

Software as a Service (IaaS): IaaS provides resources as a service which includes processing power, database, storage, computing capabilities that are offered on demand and paid on base of usage.

2. Load Balancing

Load balancing is used to distribute a larger processing load to smaller processing nodes for enhancing the overall performance of system [5]. Load balancing help in fair allocation of computing resources to achieve a high user satisfaction and proper resource utilization. Load balancing or resource scheduling between storage nodes is an important aspect in cloud computing.

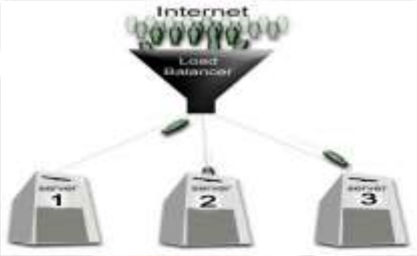

Fig- 4: Load Balancer in Cloud Computing

The goal of load balancing-

a) Enhance the system performance
b) Have a backup plan ready if any failure occurs in the system
c) Uphold the system stability
d) Allow further modifications in the system
e) Distribute the load effectively and obtain cost effectiveness

3. Literature Review

3.1 The Load Balancing Algorithm based on dynamic migration of virtual machines

This load balancing algorithm is based on dynamic migration of virtual machines, which mainly include the load monitoring module, the overload detection module and the load scheduling module Load balancing with Optimal cost scheduling algorithm. The load monitoring module is used for the periodic collection of virtual machines and the resources of the server load information. The overload detection module monitors the overloaded nodes in the system according to the data provided by the load monitoring module and database. The load scheduling module completes the virtual machine to migrate to the destination node. In this algorithm uses the load indicator vector to describe the various resources on the node. There are main three resources in this algorithm: CPU, memory and bandwidth. The trigger strategy will create a virtual machine migration to balance the load. When an indicator of node exceeds the threshold, instead of immediately triggering the migration, it predicts its future n load value according to its historical load recorded. When at least k values in the prediction are bigger than the threshold, the migration begins. In this algorithm fractal method is used to predict the CPU load, memory and bandwidth load [10].

3.2 Load balancing with Optimal cost scheduling algorithm

In this algorithm the workload is distributed evenly across all the hosts in the cloud to avoid a situation where some nodes are heavily loaded while the others have hardly any work. It is one of the Resource Scheduling Algorithm that optimizes the cost and schedules the resources based on the cost. In the proposed algorithm resources are grouped as packages in each VM. When the user requests for the resource the VM consisting of that package is executed. This technique brings down the execution cost of the service provider. It works for cost optimization at the cloud service provider, while rescheduling already accommodates requests to make space for a newly arrived request [11].

3.3 Load Balancing algorithm based on Ant Colony Optimization

This technique of load balancing is based on Ant Colony Optimization (ACO) concept. ACO is inspired from the ant colonies that work together in foraging behaviour. The ants work together in search of new sources of food and simultaneously use the existing food sources to shift the food back to the nest. The ants leave a pheromone trail upon moving from one node to another. By following the pheromone trails, the ant subsequently came to the food sources. The intensity of the pheromone can vary on various factors like the quality of food sources, distance of the food, etc. A Data Center server is known as node in this system.
The ant will use two types of pheromone for its movement [12]. They are:

- Foraging Pheromone (FP)

While moving from underloaded node to overloaded node, Ant will update FP.

- Trailing Pheromone (TP)

While moving from overloaded node to underloaded node, Ant will update TP.

The ACO Algorithm gives optimal resource utilization. The performance of the system is enhanced with high availability of resources, thereby increasing the throughput [12].

3.4 An SLA-aware Load Balancing Scheme

The two-level decentralized load balancer architecture is divided into two levels: global load balancer and local load balancer. Each global load balancer is connected to an SLA-aware local load balancer that forms a virtual zone [13].

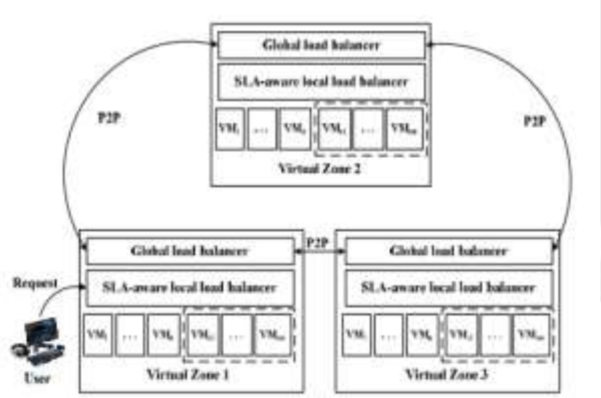

Fig -5: two-level decentralized load balancer (tldlb) architecture [13]

A local load balancer has two main tasks. The first task is monitoring the load of VMs which are in the same virtual zone. The local load balancer will provide the information (CPU, memory, network bandwidth, and disk I/O utilizations) to the global load balancer. The second task is choosing an appropriate VM using a neural network-based load balancing algorithm and then redirects the request to the VM. Global load balancers are connected to one another via P2P connections. The global load balancers exchange the load information of each virtual zone using the load information from each local load balancer. If there is no VM available in the spare VM pool to serve an overloaded virtual zone to meet the SLA requirement, the corresponding global load balancer will direct the requests to another light-loaded virtual zone to service the requests [13].

3.5 Effective Scheduling Based on Reliability

In this algorithm User requests for the resources in form of the cloudlets. Cloudlets are submitted to the Broker. There is a reliability database of the virtual machine. When a virtual machine fails, the reliability of that virtual machine is decreased and the database is updated. The cloudlets are sorted according to their priority; the highest priority cloudlet gets highest reliable virtual machine. The processing of the tasks submitted in form of the cloudlet take place at the allocated virtual machines. The results are provided back to the users [14].

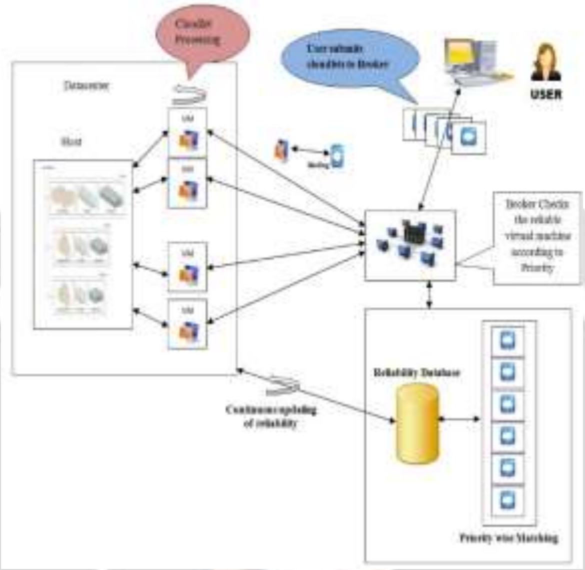
Cloudlet Processing
User submits cloudlets to Broker
USER
Datacenter
Host
VM
VM
VM
VM
Binding
Broker Checks the reliable virtual machine according to Priority
Reliability Database
Continuous updating of reliability
Priority wise Matching

Fig -6: Effective Scheduling Based on Reliability [14]

4. Proposed System

We propose dynamic load balancing algorithm which will take the load on the server as well as type of resource in consideration for efficient load balancing and resource scheduling. We will use some existing prediction model for the busty traffic. In proposed algorithm we have two modules: 1. Prediction & load calculation module 2. Request allocation module. We can define load on server as a vector <cpu,memory> Where, cpu =CPU utilization of that server, memory = memory utilization of that server

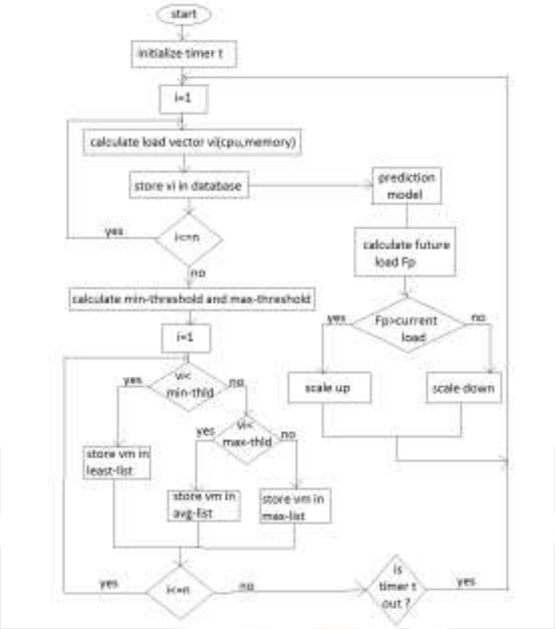

Fig-8: Flow chart of prediction & load calculation module

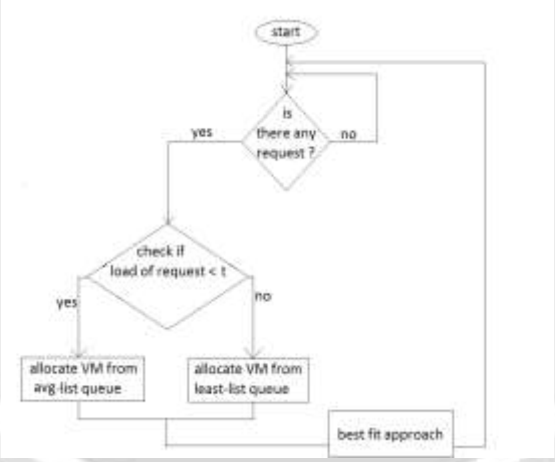
start
yes
is there any request ?
no
check if load of request < t
yes
no
allocate VM from avg-list queue
allocate VM from least-list queue
best fit approach

Fig-9: Flow chart of request allocation module

5. Simulation and Result

Tools required to implement the Optimal Resource Utilization using Dynamic Load Balancing algorithm in Cloud Computing are as below:
In this paper implement the proposed algorithm using J2EE and using the most central and well-known service of AWS known as Amazon Elastic Compute Cloud, also known as "EC2”.
Instances are launched and configured using EC2 service of Amazon Web Services.
We have used AWS Toolkit and AWS SDK.

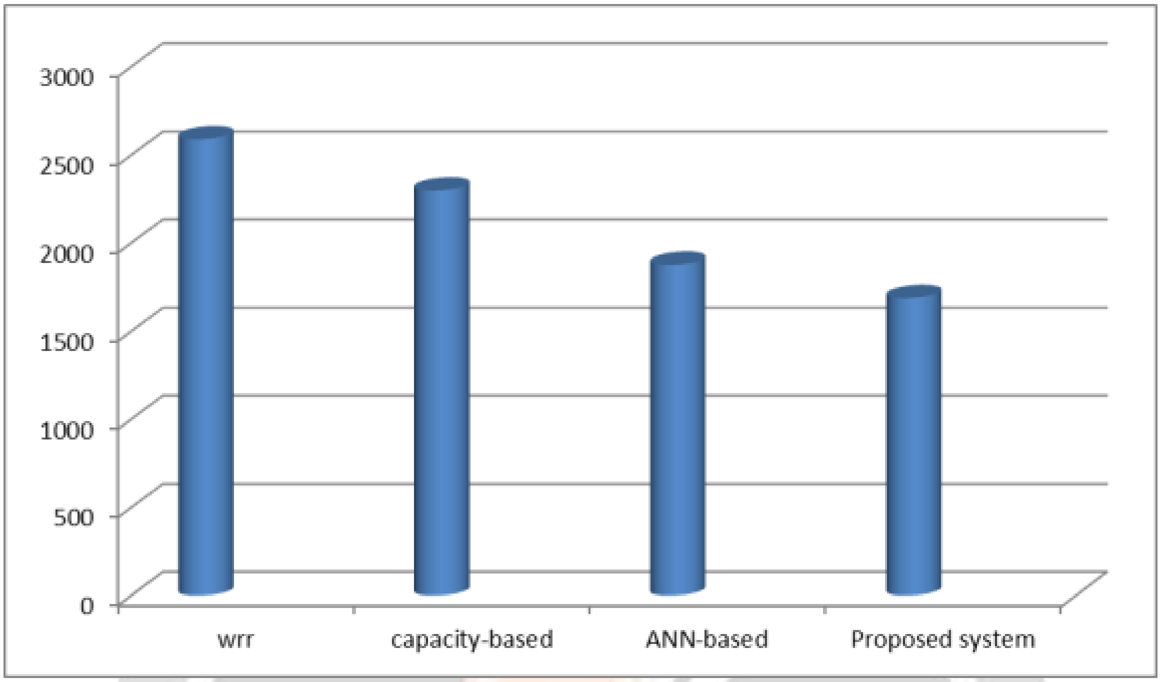

Chart-1: Average Response time for the scheduling algorithm

6. Conclusion

Load balancing is used to distribute a larger processing load to smaller processing nodes for enhancing the overall performance of system. Load balancing is a techniques that helped networks and resources by providing a Maximum throughput with minimum response time. In cloud computing environment load balancing is required distribute the dynamic local workload evenly between all the nodes. In this paper we survey different techniques of load balancing in cloud computing environment. The algorithm should not only balance the load but also help in efficient utilization of resources, increase overall throughput and decrease the response time. In this paper we proposed Dynamic load balancing algorithm. By using this algorithm we can utilize resources optimally.

7. REFERENCES

[1] Lee Badger,TimGrance, Robert Patt-Corner,JeftVoas,"Cloud Computing Synopsis and Recommendations", National Institute of Standards and Technology, Special Publication 800-146, , May 2012.

[2] Shaw,S.B.; SinghA.K.,"A survey on scheduling and load balancing techniques in cloud computing environment"Computer and Communication Technology (ICCCT), 2014 International Conferenceon 2014,Pages:8-795,
DOI: 10.1109/ICCCT.2014.7001474

[3] https://en.wikipedia.org/wiki/Cloud_computing

[4] Hisao Kameda, EL-Zoghdy Said Fathyy and Inhwan Ryuz Jie Lix, "A Performance Comparison of Dynamic vs Static Load Balancing Policies in a Mainframe, Personal Computer Network Model",Proceedings Of The 39th IEEE Conference on Decision &Control,2000.

[5] B. P. Rima, E. Choi, and I. Lumb, "A Taxonomy and Survey of Cloud Computing Systems", Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC, Seoul, Korea, August 2009, pages 44-5l

[6] Nidhi Jain Kansal, Inderveer Chana, "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI, Vol. 9, Issue 1, January 2012.

[7] Ghutke, B.; Shrawankar, U., "Pros and Cons of Load Balancing Algorithms for Cloud Computing" Information Systems and Computer Networks (ISCON), 2014 International Conference on, 2014, Pages: 123 - 127, DOI: 10.1109/ICISCON.2014.6965231

[8] Tejinder Sharma, Vijay Kumar Banga, "Efficient and Enhanced Algorithm in Cloud Computing," International Journal of Soft Computing and Engineering (IJSCE), Vol. 3, Issue 1, pp. 2231-2307, 2013.

[9] Sandeep Sharma, Sarabjeet Singh, Meenaksshi Sharma, "Performance Analysis of Load Balancing Algorithms", World Academy of Science, Engineering and Technology, 2008

[10] Haozheng Ren, Yihua Lan, Chao Yin "The Load Balancing Algorithm in Cloud Computing Environment" 2012 2nd International Conference on Computer Science and Network Technology, 978-1-4673-2964-4/12.

[11] Mrs. Nagamani H. Shahapure, Dr. Jayarekha P "Load Balancing with Optimal Cost Scheduling Algorithm" 2014 INTERNATIONAL CONFERENCE ON COMPUTATION OF POWER, ENERGY, INFORMATION AND COMMUNICATION (ICCPEIC), 978-1-4799-3826-1/14

[12] Ekta Gupta, Vidya Deshpande "A Technique Based on Ant Colony Optimization for Load Balancing in Cloud Data Center" 2014 International Conference on Information Technology, 978-1-4799-8084-0/14

[13] Chung-Cheng Li and Kuochen Wang "An SLA-aware Load Balancing Scheme for Cloud Datacenters" ICOIN 2014, 978-1-4799-3689-2/14

[14] Tingting Wang, ZhaobinLiu, Yi Chen, Yujie Xu, Xiaoming Dai "A Novel Attempt towards Effective Scheduling Based on Reliability in Cloud Environment" ICTCS '14, November 14, ACM 978-1-4503-3216-3/14/11.

[15] Yingchi Mao, Daoning Ren, Xi Chen "Adaptive Load Balancing Algorithm Based on Prediction Model in Cloud Computing" ICCC'13, December 1–2, 2013, ACM 978-1-4503-2119-8/10/06.