Analysis of Different Load Balancing Techniques in Cloud

Venkateshwarlu Velde[1], Dr. B Rama[2],

[1] Department of Computer Science, Kakatiya University, Warangal, Telangana, India
veldevenkat@gmail.com
[2] Department of Computer Science, Kakatiya University, Warangal, Telangana, India
rama.abbidi@gmail.com

Abstract—***Cloud computing is a utility to deliver services and resources to the users through high speed internet. It has a number of types and hybrid cloud is one of them. Load balancing in cloud computing means the actions that are performed for distributing the workloads across multiple computing resources.***

Cloud computing brings the advantages such as availability and scalability with the pay-as-you- go model, which has a lot to do with the success of cloud. This model has the advantages such as scaling an application dynamically, support heavy traffic, routing traffic to the closest virtual machine and the likes. With recent emerging technology, load balancing is a primary concern.

There are various scheduling algorithms that achieve load balancing by many efficient job scheduling and resource allocation techniques. The aim of this paper is to discuss briefly some of the cloud concepts, the existing load balancing techniques and present a detailed study and comparison of the same.

Index Terms—Cloud computing, load balancing, scalability

I. INTRODUCTION

Cloud computing is a new technology .It providing online resources and online storage to the user's .It provide all the data at a lower cost. In cloud computing users can access resources all the time through internet. They need to pay only for those resources as much they use .In Cloud computing cloud provider outsourced all the resources to their client. Cloud provides a cost effective 'pay-as-you-go' model. Cost is one of the main reasons for the success of the cloud.

A. Characteristics of the cloud:

1. On demand service: Cloud computing provide its users with resources and services as per the demand of the user. This does not involve interacting with the cloud service provider
2. Broad Network Access: The cloud resources can be accessed from anywhere using internet in laptops, tablets and smart phones.
3. Resource pooling: Both the storage and computing resources are available and pooled to achieve multi-tenancy.
4. Rapid elasticity: The cloud services can be fastly scaled up or down based on demand.
   a) Horizontal scaling: It refers to launching and removing server resources as per the demand.
   b) Vertical scaling: It refers to changing the computing capacity of the already assigned server resources.
5. Measured service: Cloud computing incorporates the pay as you go model. The specific resources that are used are charged based on a previously specified metric.

II. CLOUD MODELS

Service models :

Infrastructure as a Service (IaaS): Iaas provides the users the capability to provision computing and storage resources. These instances are provided to the users as virtual storage and virtual machine instances.

Users can start, stop, configure and manage the VM instances and virtual storage.Amazon Web Services (AWS), Microsoft Azure, Google Compute Engine (GCE), Joyent are examples for IaaS.

Platform as a service (PaaS): PaaS provides the users the capability to build and use application in cloud environment using development tools, application programming interfaces (APIs), software libraries and services provided by the cloud provider.

The cloud service manages the underlying cloud infrastructure which includes servers, networks, OS and storage. Apprenda is an example for enterprise platform as a service model.

Storage as a Service (SaaS): SaaS provide the users a place for storage. These servers may be distributed all over the Planet

III. LOAD BALANCING

Load balancing is the procedure of redistributing and reassigning the larger processing load to the smaller processing nodes, in order to improve the performance of the system. IT basically must have a mechanism to process user requests and make the application run faster.

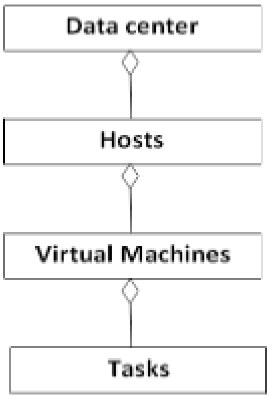
Most of the cloud service providers (CSP) provide an automatic load balancing service, which allows clients to increase the number of Processors or memories for their resources to scale with increased demands. This completely is optional and mostly depends on the clients business needs. Load balancing takes care of two important things primarily to make sure of the availability of Cloud resources and secondarily to enhance the performance [6].

This will ensure,

- Resources are easily available on demand.
- Resources are efficiently utilized under condition of high/low load.
- Reduced energy consumption in case of low load, when the usage of the CPU cycles and memory falls below a certain threshold.
- Reduction in the resource usage cost Load balancing helps in the allocation of computing resources to achieve proper resource utilization.

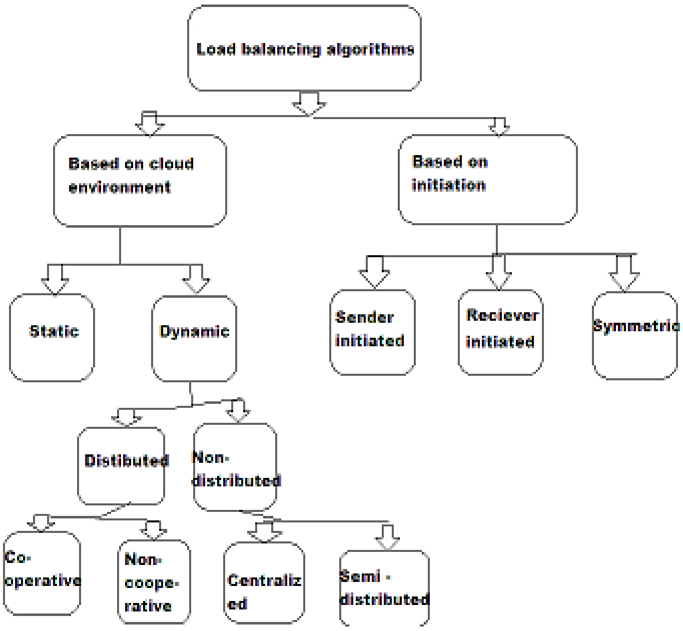
High resource utilization with efficient load balancing helps in minimization of resource consumption. Load balancing techniques facilitate networks and resources result a most outturn with minimum time interval. Load balancing is assigning the traffic among all network nodes, therefore data may be sent and received with zero delay with load balancing. [8][9][10]

For effectively scaling the working efficiency of Load Balancing algorithms a appropriate environment is required. CloudSim [7] is a productive tool that can be useful for modeling of Cloud. During the development and lifecycle of a Cloud, CloudSim allows VM's to be managed by hosts which in turn are managed by servers or datacenters.



Class diagram of the cloud [2]

Classification of load balancing algorithms:
The figure given below gives a high level classification of load balancing algorithms.



A. *Static approach:*

This is an approach that is mostly defined in the design or implementation of system. In this approach, the load balancing algorithms distribute the traffic equally among all servers.

B. *Dynamic approach:*

This is the approach that considers only the current state of the system during load balancing decisions. This approach is effective for widely distributed systems such as cloud computing. Dynamic load balancing approaches have 2 types. They are

- Distributed approach
- Non-distributed (centralized) approach.

a) Centralized approach: - In this approach, a single node is responsible for managing and distribution within the whole system.
b) Distributed approach: - In this approach, each node independently builds its own load vector. Vector collecting the load information of other nodes. The decisions are made locally using the load vectors [11].
There are many load balancing algorithms which help in obtaining better throughput and improve the response time in cloud environment. Each of them has their own benefits. [11][12] [13] [14]

C. Round Robin:

In this algorithm, the processes are divided between all processors. Each process is assigned to the processor in a particular order called round robin order.
The process allocation order is maintained locally and is independent of the allocations from the remote processors. Though the load distributions between processors are equal, the time taken for processing these jobs is not same. So at any point of time some nodes may be heavily loaded and others remain idle. This algorithm is mostly used in web servers where http requests are of similar nature and are distributed equally.

D. Task Scheduling:

This algorithm mainly consists of two levels of task scheduling Mechanisms. Both of these are based on load balancing to achieve dynamic requirements of users. It achieves maximum resource utilization. This algorithm achieves load balancing by first mapping tasks to virtual machines and then all virtual machines to host resources .It is improving the task response time .It also provide better resource utilization .

E. Opportunistic Load Balancing

This is an attempt to keep each node busy, therefore does not consider the present workload of each computer. This manages the load but it does not consider the expectation execution time of task, therefore the whole completion time becomes poor.

F. Randomized:

This algorithm is static in nature. There a process can be handled by a particular node n with a probability p. When all the processes are of equal load the algorithm works well. This algorithm is not maintaining deterministic approach, thus problems arise when the loads are of unequal computational complexities.

G. Min-Min Algorithm:

It starts with a set of all unassigned tasks .In this minimum completion time for all tasks is found. Then after that among calculates minimum times the minimum value is selected. Then task with minimum time schedule on machine. After that the execution time for all other tasks is updated on that machine then again the same procedure is followed until all the tasks are assigned on the resources. The main problem of this algorithm is has a starvation.

H. Max-Min Algorithm:

Max-Min algorithm is almost same as the min-min algorithm. The main difference is that in this algorithm first minimum execution times are found out.
Then the maximum value is selected which is the maximum time among all the tasks on any resources. After that maximum time finding, the task is assigned on the particular selected machine. Then the execution time for all tasks is updated on that machine, this is done by adding the execution time of the assigned task to the execution times of other tasks on that machine. Then all assigned task is removed from the list that executed by the system.

I. Compare and Balance:

This algorithm is uses to reach an equilibrium condition and manage unbalanced systems load. In this algorithm on the basis of probability (no. of virtual machine running on the current host and whole cloud system), current host randomly select a host and compare their load. If load of current host is more than the selected host, it transfers extra load to that particular node. Then each host of the system performs the same procedure. This load balancing algorithm is also designed and implemented to reduce virtual machines migration time. Shared storage memory is used to reduce virtual machines migration time.

J. Ant Colony Optimization

An ant algorithm is a multi agent approach to difficult combinatorial optimization problems. Example of this approach is travelling salesman problem (TSP) and the quadratic assignment problem (QAP) . These algorithms were inspired by the observation of real ant colonies. Ant's behavior is directed more to the survival of the colonies .They not think for individual.

K. Shortest Response Time First

In this each process is assigned a priority which is allowed to run. In this equal priority processes scheduled in FCFS order. The (SJF) algorithm is a special case of general priority Scheduling algorithm. In SJF algorithm is priority is the inverse of the next CPU burst. It means, if longer the CPU burst then lower the priority. The SJF policy selects the job with the shortest (expected) processing time first. In this algorithm shorter jobs are executed before long jobs. In SJF, it is very important to know or estimate the processing time of each job which is major problem of SJF.

IV. CONCLUSION

One of the major issues in cloud computing is load balancing. It helps in the efficient utilization of resources and hence it enhances the performance of the system. A few existing algorithms can maintain load balancing and provide better strategies through efficient scheduling and resource allocation techniques. This paper presents a concept of Cloud Computing along with load balancing. There are many above mentioned algorithms in cloud computing which consists of many factors like scalability, better resource utilization, high performance, better response time.

REFERENCES

[1] A. Khiyaita, H. EI Bakkli, M. Zbakh, Dafir EI Kettani, " Load Balancing Cloud Computing: State Of Art", 2010, IEEE.

[2] M. D. Dikaiakos, G. Pallis, D. Katsa, P. Mehra, and A. Vakali, "Cloud Computing: Distributed Internet Computing for IT and Scientific Research", IEEE Journal of Internet Computing, Vol. 13, No. 5, Septe9mber/October 2009, pages 10-13. https://en.wikipedia.org/wiki/Category:Cloud _clients - Category Cloud clients - Definition of cloud clients

[3] Anthony T.Velte, Toby J.Velte, Robert Elsenpeter, Cloud Computing A Practical Approach, TATA McGRAW-HILL Edition 2010.

[5] Chaudhari, Anand and Kapadia, Anushka, " Load Balancing Algorithm for Azure Virtualization with Specialized VM", 2013,algorithms,vol 1,pages 2, Chaudhari

[6] Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh, Christopher Mcdermid (2011)"Availabity and Load Balancing in Cloud Computing" International Conference on Computer and Software Modeling IPCSIT vol.14 IACSIT Press,Singapore 2011

[7] Huang, A. Software Architect, Citrix Systems Apache Cloud-Stack Architecture.

[8] Nayandeep Sran,Navdeep Kaur , "Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing ",vol 2,jan 2013

[9] Bala, Anju and Chana, Inderveer, "A survey of various workflow scheduling algorithms in cloud environment", 2nd National Conference on Information and Communication Technology (NCICT), 2011

[10] Chaczko, Zenon and Mahadevan, Venkatesh and Aslanzadeh, Shahrzad and Mcdermid, Christopher, "Availability and load balancing in cloud computing", International Conference on Computer and Software Modeling, Singapore, chaczko2011availability

[11] Rajwinder Kaur, Pawan Luthra, "Load Balancing in Cloud computing", ACEEE

[12] S.-C.Wang, K.-Q. Yan, S.-S.Wang, C.-W. Chen, "A three-phases scheduling in a hierarchical cloud computing network", in: Communications and Mobile Computing (CMC), 2011 Third International Conference on,IEEE, 2011,pp. 114–117.

[13] O. Elzeki, M. Reshad, M. Elsoud, "Improved max-min algorithm in cloud computing, International Journal of Computer Applications"vol 50 (12) (2012) pages 22–27

[14] Zhong Xu, Rong Huang,(2009)"Performance Study of Load Balanacing Algorithms in Distributed Web Server Systems", CS213 Parallel and Distributed Processing Project Report.