A SURVEY OF LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING

Urja N
Email: [1]urja.uday@gmail.com

Abstract:
Cloud computing is a model for enabling ubiquitous network access to a shared pool of configurable computing resources. Cloud load balancing is a type of load balancing that is performed in cloud computing which is the process of distributing workloads across multiple computing resources. Cloud computing brings the advantages of 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 of 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 concern. There are various scheduling algorithms that maintain load balancing through 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 comparative study of the same.
Keywords: Cloud computing, load balancing, scalability

i. Introduction:

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources like networks, servers, applications, services etc that can be rapidly provisioned and released with minimal management effort or service provider information. Cloud provides a cost effective 'pay-as-you-go' model, in which the end users pay only for the resources which they consume. Cost is one of the main reasons for the success of the cloud.

A. Characteristics of the cloud:

1. On demand service: Cloud computing provides 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 anywhere on the network including laptops, tablets and smart phones.
3. Resource pooling: Both the storage and computing resources are pooled to achieve multi-tenancy.
4. Rapid elasticity: The cloud services can be rapidly 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. Architecture of the Cloud system

The architecture of the cloud refers to the components. These components consists of a back end platform and a front end platform. The front end may contain thin clients or thick clients or mobile devices. The back end refers to the servers and storage.

✓ Clients:

A cloud client consists of computer hardware and/or software that relies on cloud computing for application delivery, or that is specifically designed for delivery of cloud services and that,

in either case, is essentially useless without it. [3]

◆ Thin clients: These typically are used for display and od not do any kind of computation. They do not have any internal memory as the servers do all the computation and other work.

◆ Thick clients: These typically use different browsers to connect to the cloud and internet.

◆ Mobile client: The end devices are mobile devices like smart phones and tablets. Mobile cloud computing is a branch of cloud computing where at east some of the devices are mobile [2]

✓ Datacenter

◆ Datacenter is collection of servers hosting different applications. An end user will have to connect to the data centre to make use of the cloud applications

◆ A datacenter, geographically may be located at any distance from the cloud.

✓ Distributed servers

◆ A distributed server is a server that repeatedly checks the services of their hosts.

◆ Distributed servers are the part of a cloud which are hosted on the internet to provide services for various applications.

◆ The user gets a feel that the application is being run on the user's machine when one accesses the cloud.

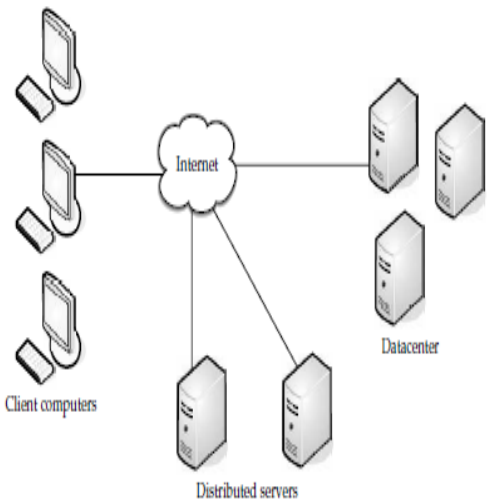


Fig 1. Components of the cloud [1]

iii. Cloud models

3.1 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 virtual machine instances and virtual storage.Amazon Web Services (AWS), Microsoft Azure, Google Compute Engine (GCE), Joyent. Are all example for IaaS.

✓ Platform as a service (PaaS):

PaaS provides the users the capability to develop and deploy application in the cloud using development tools, application programming interfaces (APIs), software libraries and services provided by the cloud provider. The cloud service manages the underlying cloud infrastructure including servers, networks, operating systems and storage. Apprenda is an example for enterprise platform as a service model.

✓ Storage as a Service (SaaS):

SaaS provied the users a place for storage. These servers may be distributed all over the world

3.2 Deployment models

Public cloud
Private cloud
Hybrid cloud
Community cloud

Iv. Virtualization:

Virtualization refers to something that is virtual and not real, but provides all facilities like that of the real one. Virtualization is a software implementation that separates physical infrastructures to create various dedicated resources. Virtualization software makes it possible to run multiple operating systems and multiple applications on the same server at the same time.

2 types of virtualization are found in case of clouds :

✓ Full virtualization:

In case of full virtualisation a complete installation of one machine is done on the another machine. It will result in a virtual machine which which will have all the softwares that are present in the actual server.

✓ Para virtualiztion:

In paravitualisation, the hardware allows multiple operating systems to run on single machine by efficient use of system resources such as memory and processor. e.g. VMware software. Here all the services are not fully available, rather the services are provided partially.

V. Load balancing:

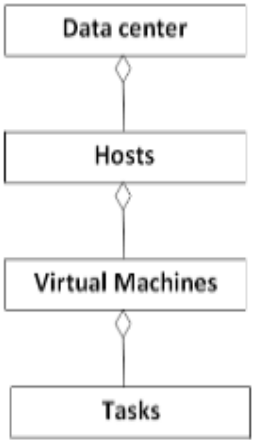
Load balancing is the process of redistributing and reassigning the larger processing load to the smaller processing nodes, in order to improve the performance of the system. IT basically must have a mechanism to process user requests and make the application run faster. Most of the cloud vendors are provide an automatic load balancing service, which allows clients to increase the number of CPUs 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

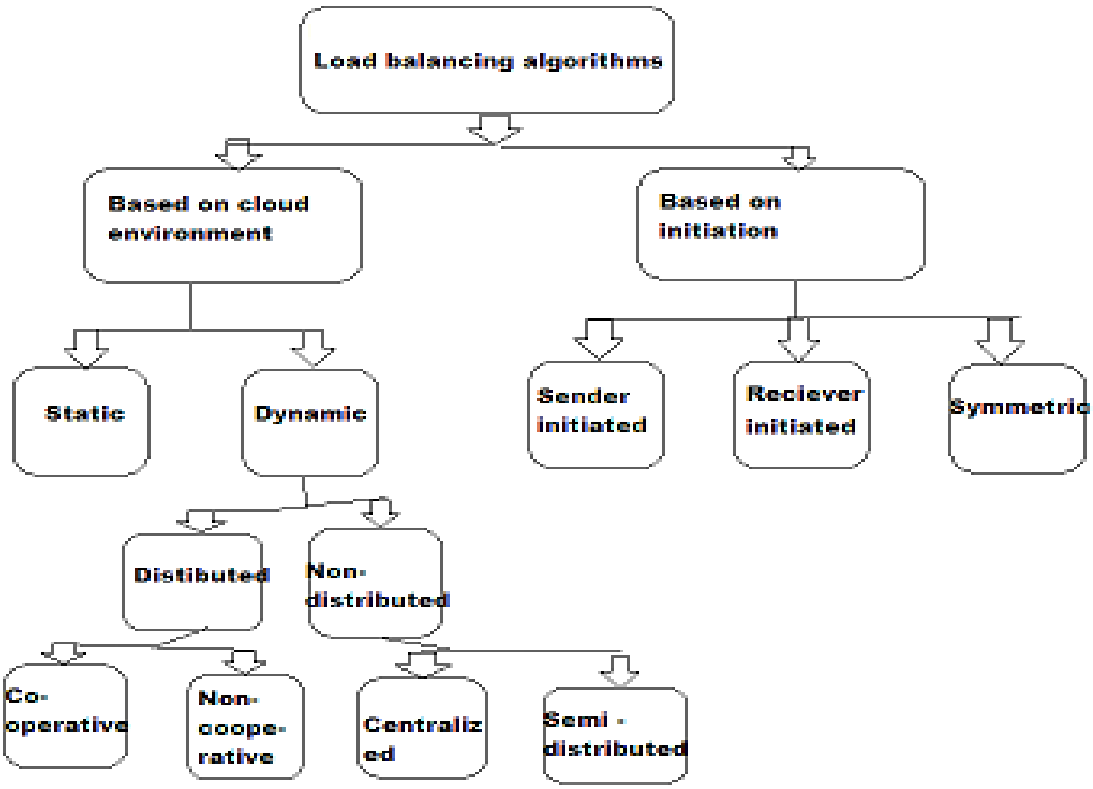
Load balancing helps in the allocation of computing resources to achieve proper resource utilization. High resource utilization with proper load balancing helps in minimizing resource consumption. It helps in implementing scalability and avoiding Bottlenecks. Load balancing techniques help networks and resources by providing a maximum throughput with minimum response time. Load balancing is dividing the traffic between all servers, so data can be sent and received without any delay with load balancing. [8][9][10]

For effectively measuring the efficiency of Load Balancing algorithms a simulation environment are required. CloudSim [7] is a productive tool that can be used for modeling of Cloud. During the lifecycle of a Cloud, CloudSim allows VMs to be managed by hosts which in turn are managed by 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.



1) Static approach: This is an approach that is mostly defined in the design or implementation of the system. In this approach, the load balancing algorithms divide the traffic equally between all servers.
2)Dynamic approach: This is the approach that considers the current state of the system during load balancing decisions. This approach is more suitable for widely distributed systems such as cloud computing .
Dynamic load balancing approaches have two 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 have their own benefits. [11][12] [13] [14]

1. 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 remote processors. Though the load distributions between processors are equal, the time taken for processing these jobs are 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.

2. Task Scheduling: This algorithm mainly consists of two levels of task scheduling
Mechanisms . Both of these are based on load balancing to meet dynamic requirements of

users . It achieves high 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 .

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

4. Randomized: This algorithm is static in nature. ere 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.

5. 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 the 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.

6. 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 is 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.

7. Honeybee Foraging Behavior: It is a nature inspired Algorithm for self-organization. Honeybee achieves global load balancing through local server actions. The performance of the system is enhanced with increased system diversity. The main problem is that throughput is not increased with an increase in sy380 stem size. When the diverse population of service types is required then this algorithm is best suited.

8. 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.

9. Honeybee foraging behavior: Whenever certain Virtual Machines are overloaded then no more tasks should be send to overloaded virtual machine if under loaded virtual machines are made available.For optimized solution and better response time the load has to be balanced among overloaded and under loaded virtual machines. The honey bee behavior based load balancing (HBB-LB) targets to achieve well balanced load across virtual machine.

10. Lock-free multiprocessing solution : It proposed a lock-free multiprocessing load balancing solution that avoids the use of shared memory in contrast to other multiprocessing load balancing solutions which use shared memory and lock to maintain a user session. It is achieved by modifying kernel. This solution helps in improving the overall performance of load balancer in a multicore environment by running multiple load-balancing processes in one load balancer.

11. Ant Colony Optimization: :- Ant algorithms is a multiagent 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 behaviour is directed more to the survival of the colonies .They not think for individual.

12. Shortest Response Time First: In this each process is assigned a priority which is allowed to run. In this equal priority processes are

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.

13. Randomized: Randomized algorithm is of type static in nature. In this algorithm a process can be handled by a particular node n with a probability p. The process allocation order is maintained for each processor independent of allocation from remote processor.

This algorithm works well in case of processes are of equal loaded. However, problem arises when loads are of different computational complexities. Randomized algorithm does not maintain deterministic approach. It works well when Round Robin algorithm generates overhead for process queue.

Vi. Conclusion:

One of the major issues in cloud computing is load balancing. It helps in the efficient utilization of resources and hence in 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 consist many factors like scalability, better resource utilization, high performance, better response time.

References:

[1] Michael Armbrust, Armando Fox, Gunho Lee, Ion Stoica(2009) “Above the Clouds :A Berkeley View of Cloud Computing” University of California at Berkeley Technical Report No. UCB/EECS- 2009-28

[2]Sowmya Ray, Ajanta De Sarkar ”Excecution analysis of load balancing algorithms in cloud computing” Birla Institute of Technology, Mesra -Kolkata IJCCSA - Vol2 No 5

[3]https://en.wikipedia.org/wiki/Category:Cloud_clients - Category Cloud clients - Definition of cloud clients

[4] 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.