Mody University International Journal of Computing and Engineering Research
Vol. 1 Issue 2, 2017, pp.71-75
ISSN: 2456-9607 (Print) 2456-8333(Online)

A Comprehensive Analysis of Load Balancing Algorithms in Cloud Computing

Shalini Joshi, Uma Kumari

Computer Science & Engineering,
Mody University of Science and Technology, India

shalini123.joshi@gmail.com, umakumari.cet@modyuniversity.ac.in

Received 15 July 2017, Published 05 Nov. 2017

Abstract: Cloud computing (CC) term came into existence after decades of research using existing technologies like parallel computing, grid computing, peer to peer technology, distributing computing, and virtualization etc. Now days, the most approved applications are internet services with large number of users. Therefore as the size of cloud scales up, cloud computing service providers make it necessary to handle massive requests. Load balancing is one of the main challenges in cloud computing which distributes the dynamic workload across multiple nodes to ensure that no single resource is either overburdened or underused (idle).This paper presents cloud computing concepts, architecture and load balancing algorithm.

Keywords: Cloud computing, architecture, load balancing.

1. INTRODUCTION

Cloud computing, as a current active commercial offering computing that started to become apparent in late 2007[1].Cloud computing is a technological paradigm that providing large infrastructure, storage, scalability, resources pooling, virtualization and wide range of services over the internet[2].The main objective of cloud computing is that it's all users are permitted to take benefits from all different type of technologies without having a deep knowledge about the technologies at any instance of time. Cloud computing is the fastest developed technology in the IT industry and a new delivery model for the services on "pay-per-usage" basis [3]. By using this model cloud technologies that providing services to the users on demand throughout the internet. The users and their demands are increasing day by day, so there are various technical challenges that requires to be addressed like server consolidation, virtual machine migration, fault tolerance, scalability and high availability but main centralized issue is the load balancing, it is a method that distributes the workload among various available nodes of a distributed system to improve both resource utilization and job response time [4]**.** To make the use of resources most efficiently in cloud system, there are various load balancing algorithm that have been introduced.

2. CLOUD COMPUTING

Cloud computing has made outstanding changes in the environment of both parallel and distributed system.

According to NIST Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management, effort or service provider interaction[5]. So cloud computing is an internet based services that are provided to users on demand at any instance of time over a network with the scale and reliability of a data center [1]. Any cloud computing system must consist of three main components that are client, data center or distributed servers [2]. Each component plays a vital role in cloud computing and the purpose of these components can be presented as shown in fig. 1.

- *Client*
  End users that access the clouds to manage their information Is related to cloud. There are three types of clients-Mo*bile,Thin and Thick*
- *Datacenter*
  It is a collection of servers that host different types of applications. An end user connects to the data center to supports different applications.

- *Distributed Servers*
  Distributed servers are part of a cloud that available over the internet and also actively check the services of their hosts.The aim of a distributed web server is to store information and serve client requests [6].

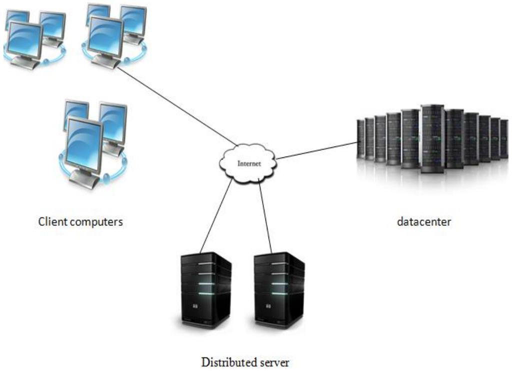


Fig. 1 Components of Cloud

3. TYPES OF CLOUD COMPUTING

Cloud computing can be divided into different parts that's based on two approaches[2]. These approaches are capability and accessibility.

A. *Based on the type of capability*

Cloud system gives three different services

i. Software as a Service (SaaS): It permits the users to use a complete application on someone else's system without purchasing and maintaining overhead [7]. An example is web based email and google documents.

ii. Platform as a Service (PaaS): It is also called service model of cloud computing in which user can develop applications using web based tools and libraries from cloud service provider[7]. An example of PaaS is Force.com and Google App Engine[8].

iii. Infrastructure as a Services (IaaS): The main focus of IaaS is to provide highly scalable resources over the internet. It provides users in virtualized manner and also adjusted on demand such as infrastructure, servers, hardware, software and storage [7]. A straightforward example of IaaS is Amazon EC2 [8].

B. *Based on the type of accessibility*

According to this type cloud is divided into three parts and that are as following;

i. Public cloud: Public cloud can be accessed by anyone in the world from anywhere over the internet [4]. Public cloud resources are based on "pay-per-usage". Amazon's and Google's cloud are example of this type of cloud.

ii. Private cloud: It is deployed inside an organization. Means it can be accessed only by the employees of that organization [3]. Organization maintains hardware, software infrastructure and also has control on accessing its resources [4].

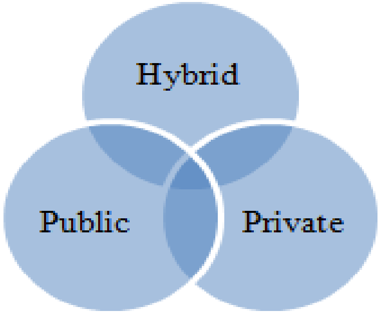


Fig. 2 Types of cloud

iii. Hybrid cloud: It is a combination of both public cloud and private cloud [4].It permits businesses to manage and control some resources internally within organization and some externally [9].

4. CLOUD COMPUTING ARCHITECTURE

The architecture of a cloud computing system is mainly developed as a set of layers. A cloud computing architecture is as shown in fig. 3

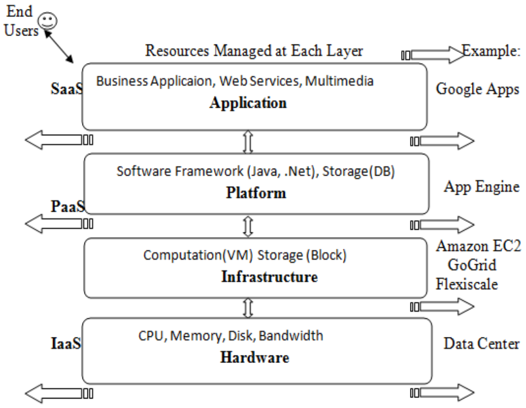
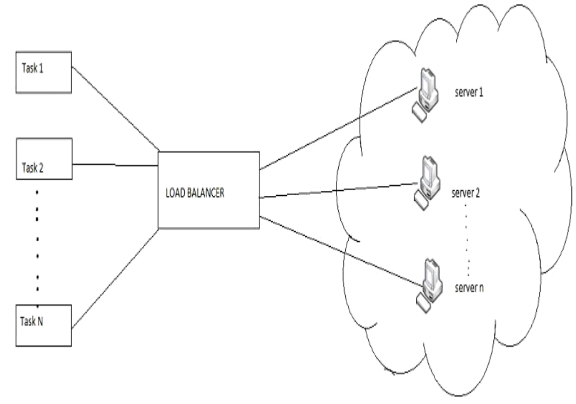

End Users
Resources Managed at Each Layer
Example:
SaaS
Business Applicaion, Web Services, Multimedia
Application
Google Apps
Software Framework (Java, .Net), Storage(DB)
Platform
PaaS
App Engine
Computation(VM) Storage (Block)
Infrastructure
Amazon EC2
GoGrid
Flexiscale
IaaS
CPU, Memory, Disk, Bandwidth
Hardware
Data Center

Fig. 3 Cloud computing Architecture

There is a hardware layer at the lowest level of hierarchy, which is organized for managing the physical resources of the cloud system, like storage, servers, new devices, cooling and power system. At the top of hardware layer, there exists infrastructure layer that is responsible for providing a pool of storage and computing resources by partitioning the physical resources of hardware layer by the virtualization techniques [8]. The platform layer that is built on the top of the infrastructure layer that involves a set of operating system and application frameworks. The objective of this layer is to minimize overload of deploying application directly into infrastructure resources by providing support for implementing data base storage and business strategies of cloud application. At the end, highest level of hierarchy is application layer that provides cloud application.

5. LOAD BALANCING

Load balancing plays very essential role in cloud environment fo maintaining the rhyme of cloud computing. The main goal of load balancing is to obtain optimal resource utilization, maximize throughput, minimum response time and avoids the system overload. To achieve this goal, there are various load balancing algorithm. In these load balancer distributes workload among various nodes of the system in the network. Load balancing algorithms ensure that there is no node in the system that is overloaded and under loaded (idle) [5]. Which means these algorithms are used to assign same volume of work among all the available nodes [10]. However, with emergence of some existing cloud computing platforms, such as Amazon S3 and Windows Azure Platform, it will be very common for cloud service providers to publish their web services at different cloud platforms, which has distinctive features such as scalability, adaptability and transparency of load scheduling [11].The reason behind using these algorithms (techniques) is to serve better to the user without service breaking [9].



Fig, 4 Diagram for load balancing [12]

Load balancing is based on two categorize that are:

Static Load Balancing: In static load balancing algorithm, performance does not depend upon the current state of the system. In static environment, the resources in the cloud are not flexible and also require prior knowledge regarding system resources and details of all tasks in an application [13]. User requirements can't change at run time. It is well suited for homogenous cloud environment.

Dynamic Load Balancing: In this algorithm performance depends on the current state of the system [14]. In dynamic environment, resources are flexible in nature and do not require prior knowledge about the system. It is well suited for heterogeneous cloud environment and also balances the load in very efficient way.

6. LOAD BALANCING ALGORITHM FOR CLOUD COMPUTING

Following load balancing algorithms are currently used in clouds.

A. *Round Robin Algorithm*

This is a static load balancing algorithm that uses the round robin fashion for allocating jobs. Round-Robin scheduling is known as very effective and efficient time triggered scheduling policy[7]. This algorithm is based on random selection that means it select the nodes(processors) randomly for balancing load. Here, Data Centers plays very important role to handle the load balancing process in cloud. When the data center controllers receives the request from user then it pass the request to round robin algorithm. In this algorithm time is divided into small units that is called time slice(time slot) and quantum. So this algorithm is designed specially for time sharing.

Firstly, all the processors that is runnable are stored in circular queue. For the defined time slot, scheduler allocates server to all processes in the queue. Whenever, new processes come, it will be added at th end of queue that is called tail of the queue. Scheduler selects the first process randomly from the queue. As the time slot for the process is over, the process is forwarded from the server and attached at the tail of the queue. And if the process is completed before the time slot, then the process is released voluntary by itself. And the scheduler assigns the server to the ready process in queue. In this way user request are processed in circular way by using this algorithm. But due to random selection of server multiple times some servers are overloaded and others are under loaded(idle) which results in decreasing the performance of the load balancing. To overcome this type of bottleneck, a better allocation technique is introduced that is known as weight round robin load balancing algorithm[15].

B. *Opportunistic Load Balancing Algorithm*

This algorithm does not analyze the current state of the virtual machine because it is a static load balancing algorithm. It makes an effort to keep each node busy. This algorithm deals rapidly with the unexecuted tasks randomly to the currently available nodes in the system. Each task can be assign randomly to a node. Hence, this algorithm does not provide load balance with good result. Due to this reason, it does not calculate the current execution time of the node, so task will process slow in manner[8].

C. *Min-Min Load Balancing Algorithm*

It is a simple or fast algorithm that provides improved performance [14]. This algorithm consists of a task set. Initially, no tasks are assigns to any of the nodes. So the minimum completion time is calculated for all the available nodes in the system [5]. After the calculation had done the task is selected that have minimum completion time and then, assign to the respective node. The currently available execution time is updated and the task gets removed from the available task set. This process is done time to time until all the tasks will be allocated to the equivalent machines. This algorithm works much better in situations where the small tasks are greater than large task. A disadvantage of this algorithm is that it leads to starvation because it assigning small number of task firstly, while large tasks remaining in the waiting stage [12].

D. *Max Min Load Balancing Algorithm*

This algorithm is similar as Min-Min load balancing algorithm. At the starting all the available tasks are submitted to the system and then, minimum completion time for all available tasks is calculated. After this calculation, select a task that have maximum completion time and that task is assigned to the corresponding machine[14]. This algorithm performs better than Min-Min algorithm because if there is only one long(large) task in a task set then Max-Min algorithm runs short tasks parallelly with long task[12].

E. *Active Monitoring Load Balancing Algorithm*

This is a dynamic load balancing algorithm in which the load is assigned to the virtual machine by finding out the idle VM (Virtual Machine) or the least loaded VM in the list. Initially, the null VM is searched if there is no null VM. Further least loaded one is selected. Here an index table of all the servers and requests that are currently assigned to the servers is maintained by the load balancer. Whenever a new request comes, data center does scan the index table of the servers which is idle or least loaded. This algorithm uses the FCFS(first come first serve) concept for assigning load to the servers when more than two servers are kept with least index number. Then using server id(task assigned id) allocate load to the server and the server's index table is incremented. Whenever, the task is completed. The information is forwarded to datacenter and balancer decrements the server index table [15]. And when again a new request arrives, load balancer re-scans the index table and process allocation is carried out.

F. *Equally Spread Current Execution Algorithm*

This is dynamic load balancing algorithm. In which load balancer makes effort to distributes almost equal amount of load among all the servers that are available at the data centers. At the starting of this algorithm assign priority to all the processes, then it checks the size and capacity to transfer that load to a server which can handle that load in less time and with maximum throughputor which is lightly loaded [15]. At this point, the capacity of the VM is measured and also load is estimated. Load is assigned according to the size and capacity of that matching VM.

G. *Throttled Load Balancing Algorithm*

This algorithm is based on virtual machine(VM). A TLB(Throttled Load Balancer) is that which maintains all the processes and also monitoring the work on the servers[14]. So in this algorithm, load balancer finds the best VM for the client request that can handle the load very easily and in effective way. Different VM's has different capacity and propertities to handle different loads. Thus according to the load, the right VM is selected for that load. Here an index table is maintained for all the servers and when client send a request to data center, data center controller forward the request to TLB. For finding the available idle server, TLB scans the index table and then send back the server id(idle id) to the datacenter and the task is assigned to that servers. After allocation, index table is updated. And whenever data center controller gets the information about completion of task, the index table is decremented again. In this algorithm if there is no server found in idle state, the request is remaining in queue[15].

H. *Active Clustering Algorithm*

This algorithms defines the clustering of VM (Virtual Machine) for balancing the load in cloud computing. In this algorithm clustering means "grouping of objects together which have same type of properties" [15]. So VM which have same properties are grouped together in cluster to handing the type of load.

7. CONCLUSIONS

Table 1gives a comparative analysis of different load balancing techniques (algorithms) with respect to various performance parameters [13]. Cloud computing is a new emerging trend in IT era with large requirements of resources, storage and infrastructures. Load balancing is an essential aspect of cloud computing to balance the load in the system. This paper explains the concept of cloud computing, cloud types, architecture, load balancing and efficient load balancing algorithm. Each algorithm has its pros and cons. Paper ended by comparing these algorithms by different parameters like throughput, overhead, fault tolerance, response time, resources utilization, scalability and performance. These load balancing algorithms ensures resources utilization by distributing the load among various nodes in the system using task scheduling. The level of implementation in static and dynamic algorithms plays very important role in deciding the efficiency and effectiveness of the algorithms. Paper focuses on minimization of overhead, service response time but maximization of throughput and performance qualitative analysis on VMs. In future, multiple load balancing algorithms can be combined which will maintain a better trade-off among various performance criteria.

Table-1: A comparative analysis of different load balancing techniques

Load Balancing algorithms	Through Put	Overhead	Fault tolerance	Response time	Resources Utilization	Scalability	Performance
Round Robin	yes	yes	no	yes	yes	yes	yes
Opportunistic	no	no	no	no	yes	no	yes
Min-Min	yes	yes	no	yes	yes	no	yes
Max-Min	yes	yes	no	yes	yes	no	yes
Active Monitoring	yes	yes	no	yes	yes	yes	no
ESCE	no	no	no	yes	yes	yes	no
Throttled	no	no	yes	yes	yes	yes	yes
Active Clustering	no	yes	no	no	yes	no	no

REFERENCES

[1] Randles, Martin, David Lamb, and A. Taleb-Bendiab. "A comparative study into distributed load balancing algorithms for cloud computing." *Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on*. IEEE, 2010.

[2] Ray, Soumya, and Ajanta De Sarkar. "Execution analysis of load balancing algorithms in cloud computing environment," International Journal on Cloud Computing: Services and Architecture (IJCCSA) 2.5, 2012, pp. 1-13.

[3] Panwar, Reena, and Bhawna Mallick. "A Comparative Study of Load Balancing Algorithms in Cloud Computing," *International Journal of Computer Applications* 117.24, 2015.

[4] More, Nitin S., and Swapnaja R. Hiray. "Load balancing and resource monitoring in cloud," *Proceedings of the CUBE International Information Technology Conference*. ACM, 2012.

[5] Mesbahi, Mohammadreza, and Amir Masoud Rahmani. "Load Balancing in Cloud Computing: A State of the Art Survey," *International Journal of Modern Education and Computer Science* 8.3 , 2016, p.64.

[6] Cardellini, Valeria, Michele Colajanni, and S. Yu Philip. "Dynamic load balancing on web-server systems," *IEEE Internet computing* 3.3, 1999, p. 28.

[7] Jena, Soumya Ranjan, and Zulfikhar Ahmad. "Response time minimization of different load balancing algorithms in cloud computing environment."*International Journal of Computer Applications* 69.17, 2013.

[8] Haridas Kataria, Vipul Pant. " Review of Load Balancing Types, Services and Algorithms in Cloud Computing Network,"International Journal of Advanced Research in Computer Science and Software Engineering Research Paper. Volume 5, Issue 10, October-2015.

[9] Katyal, Mayanka, and Atul Mishra. "A comparative study of load balancing algorithms in cloud computing environment." *arXiv preprint arXiv:1403.6918*, 2014.

[10] Moharana, Shanti Swaroop, Rajadeepan D. Ramesh, and Digamber Powar. "Analysis of load balancers in cloud computing." *International Journal of Computer Science and Engineering* 2.2 , 2013, pp. 101-108.

[11] Nitika, Ms, Ms Shaveta, and Mr Gaurav Raj. "Comparative analysis of load balancing algorithms in cloud computing," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 1.3 , 2012, p. 120.

[12] Gopinath, PP Geethu, and Shriram K. Vasudevan. "An in-depth analysis and study of Load balancing techniques in the cloud computing environment," Procedia Computer Science 50, 2015, pp.427-432.

[13] Raghava, N. S., and Deepti Singh. "Comparative study on load balancing techniques in cloud computing," *Open Journal of Mobile Computing and Cloud Computing* 1.1, 2014.

[14] Rajeshkannan, R., and M. Aramudhan. "Comparative Study of Load Balancing Algorithms in Cloud Computing Environment," *Indian Journal of Science and Technology* 9.20, 2016.

[15] Patel, Jay, Chirag S. Thaker, and Hardik Chaudhari. "Task Execution Efficiency Enrichment in Cloud Based Load Balancing Approaches,"*Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*. ACM, 2014.