Comparative Study of Load Balancing Algorithms in Cloud ComputingEnvironment

Meenakshi

Assistant Professor,

Department of Computer Science and Engineering, Nitte Meenakshi Instituteof Technology, Bangalore 64.

Abstract

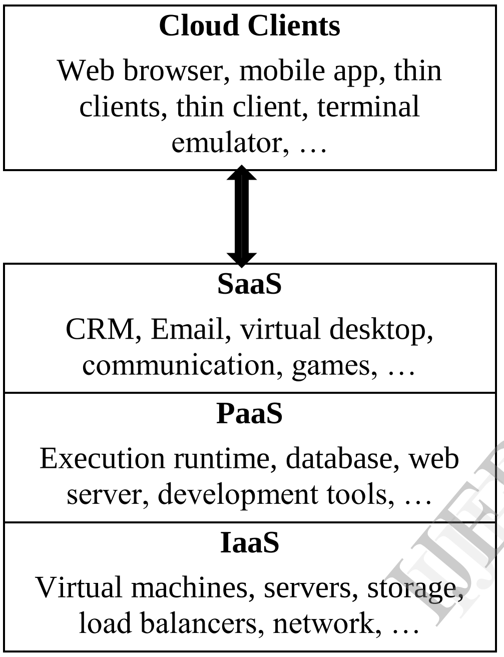
The Cloud computing provides on-demand network access to a shared pool of scalable and often virtualized resources (e.g., networks, servers, storage, applications, and services) that can be quickly provisioned and released.Clouds are high configured infrastructure delivers platform, software as service, which helps customers to make subscription for their requirements under the pay as you go model. Generally cloud is based on data centers which are powerful to handle large number of users. The reliability of clouds depends on the way it handles the loads, to overcome such problem clouds must be featured with the load balancing mechanism. Load balancing in cloud computing will help clouds to increase their capability, capacity which results in powerful and reliability clouds. This paper gives an overview of classification of different load balancing algorithms and survey on few load balancing algorithms.

Keywords: Cloud computing, algorithm, load balancing, virtualizationCloudSim, static, dynamic.

1. Introduction

Cloud computing is an abstraction based on the notion of pooling resources and presenting them as avirtual resource.Cloud computing takes the technology, services, and applications that are similar to those on the Internet and turns them into a self-service utility. Abstraction and virtualization are two important concepts of cloud. From users and developers cloud computing abstracts the details of the system implementation.Figure1 gives the clear view of cloud computing layers. Applications run on physical systems that are not specified, data is stored in locations that are unknown, administration of systems is outsources to others, and access by users is ubiquitous.

Figure1: Cloud Computing Architecture


Cloud Clients
Web browser, mobile app, thin clients, 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, …

Google, Azure platform and amazon web services are some of examples of cloud computing.

2. Load Balancing and Virtualization

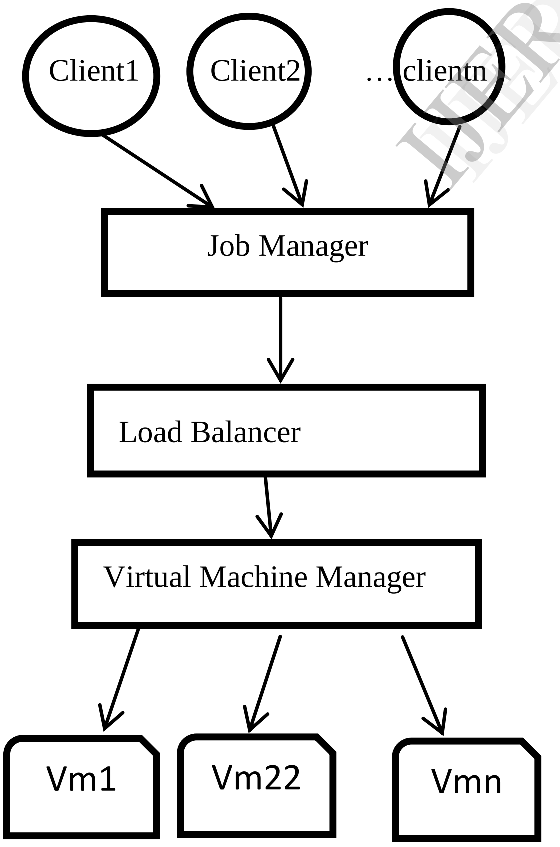
Virtualization assigns a logical name for a physical resource and then provides a pointer to thatphysical resource when a request is made.Mapping of virtual resources to physical resources can be bothdynamic and facile. Virtualization is dynamic in that the mapping can be assigned based on rapidly changing conditions and it is facile because changes to a mapping assignment can be nearly instantaneous.

Characteristic of cloud computing are storage, application, CPU and access.Data is stored across storage devices and often replicated for redundancy. A cloud has multiple application instances and directs request to an instance based on conditions. Computers can be partitioned into a set of virtual machines with each machine being assigned a workload. Alternatively systems can be virtualized through load balancing technologies. A client can request access to a cloud service from any location. No matter where we access the service, we are directed to the available resources. The technology used to distribute service requests to resources is referred to as load balancing. Figure shows the layers of cloud computing architecture.

The following network resources can be load balanced: network interfaces and services such as DNS, FTP,and HTTP; connections through intelligent switches; processing through computer system assignment; access to application instances.

Without load balancing, cloud computing would very difficult to manage. Load balancing provides the necessary redundancy to make an intrinsically unreliable system reliablethrough managed redirection. It also provides fault tolerancewhen coupled with a failover mechanism. Load balancing is nearly always a feature of server farms and computer clusters and for high availably applications.A load-balancing system can use different mechanisms to assign service direction. In the simplest load balancing mechanisms, the load balancer listens to a network port for service requests. When a request from the client or a service requester arrives, the load balancer uses a scheduling algorithm to assign where the request is sent.Figure2 shows the steps involved in algorithm execution.

Figure2:Load balancing Algorithm Execution



A session ticket is created by the load balancer so that subsequent related traffic from the client that is part of that session can be properly routed to the same resource. Without this session record or persistence, a load balancer would not be able to correctly failover a request from one resource to another. Persistence can be enforced using session data stored in a database and replicated across multiple load balancers. Other methods can use the client's browser to store a client-side cookie or through the use of a rewrite engine that modifies the URL. Of all these methods, a session cookie stored on the client has the least amount of overhead of a load balancer because it allows the load balancer an independent selection of resources.

The algorithm can be based on a simple round robin system where the next system in a list of systems gets the request. Round robin DNS is a common application, where IP addresses assigned out of a pool of available IP addresses. Google uses round robin DNS.

Load balancing is also needed for achieving Green computing[1]inclouds. The factors responsible for it are:

a) **Limited Energy Consumption:** Load balancing can reduce the amount of energy consumption by avoiding over hearting of nodes or virtual machines due to excessive workload.

b) **Reducing Carbon Emission:** Energy consumption and carbon emission are the two sides of the same coin. Both are directly proportional to each other. Load balancing helps in reducing energy consumption which will automatically reduce carbon emission and thus achieve Green Computing.

3. Classification of load balancing algorithms

Load Balancing Algorithm broadly can be classified as follows [2]:

A) Depending on system state
   - i. Static
   - ii. Dynamic
     - a) Distributed: cooperative and non-cooperative
     - b) Non Distributed: centralized and semi distributed.

B) Depending on who initiated the process
   - i. Sender initiated
   - ii. Receiver initiated
   - iii. Symmetric

Based on process origination, load balancing algorithms can be classified as [3]:

a) **Sender Initiated:** In this type of load balancing algorithm the client sends request until a receiver is assigned to him to receive his workload i.e. the sender initiates the process.

b) **Receiver Initiated:** In this type of load balancing algorithm the receiver sends a request to acknowledge a sender who is ready to share the workload i.e. the receiver initiates the process.

c) **Symmetric:** It is a combination of both sender and receiver initiated type of load balancing algorithm.

Based on the current state of the system there are two other types of load balancing algorithms.

3.1. Static Load Balancing

Static load balancing algorithms require knowledge about the applications and resources of the system. The decision of shifting the load does not depend on the current state of the system. The performance of the virtual machines is determined at the time of job arrival. The master processor assigns the workload to other slave process.

rs according to their performance. The assigned work is thus performed by the slave processors and the result is returned to the master processor.

Static load balancing algorithms are not preemptive and therefore each machine has at least one task assigned for itself. Its aims in minimizing the execution time of the task and limit communication overhead and delays. This algorithm has a drawback that the task is assigned to the processors or machines only after it is created and that task cannot be shifted during its execution to any other machine for balancing the load. The four different types of Static load balancing techniques are Round Robin algorithm, Central Manager Algorithm, Threshold algorithm and randomized algorithm.

3.2. Dynamic Load Balancing

In dynamic approach of load balancing algorithms the current state of the system is important factor because it is used to make any decision for load balancing. It allows for processes to move from an over utilized machine to an underutilized machine dynamically for faster execution. This means that it allows for process preemption which is not supported in Static load balancing approach.

In the distributed one, all nodes present in the system will take part in load balancing activity by executing load balancing algorithm. So there is a share of the task of load balancingamong them.The interaction among nodes to achieve load balancing can take two forms: cooperative and no cooperative. In the first one, the nodes work side-by-side to achieve a common objective, forexample, to improve the overall response time, etc. In the second form, each node worksindependently toward a goal.

Dynamic load balancing algorithms of distributed nature, may lead to produce more messages thanthe non-distributed ones because, each of the nodes in the system needs to interact with everyother node. A benefit, of this is that even if one or more nodes in the system fail, it will not causethe total load balancing process to halt; it instead would affect the system performance to someextent.

In non-distributed type, instead of all nodes, the task of loadbalancing is given to either one node or to a group of nodes. Nondistributeddynamic load balancing algorithms can take two forms: centralized and semidistributed.In the centralized non distributed dynamic load balancing algorithm, among the nodes, there is a specially designated node called a central node where the load balancing algorithm is executed in thewhole system. This central node is responsible for load balancing of the wholesystem. All the other nodes need to interact with the central node. In semi-distributed form, cluster formation is required. Clusters are nothing but group of nodes. The load balancing in each cluster is of centralizedform. A central node is elected in each cluster by appropriate election technique which takes careof load balancing within that cluster.Hence, the load balancing of the whole system is done via the central nodes of each cluster.

Among these two approaches, centralized dynamic load balancing can be applied for networks with small size only. Centralized algorithms can cause a problem because only a single node that is central node takes a major roll . The load balancing process may be disturbed in case of central node crashes.

Advantage part of centralized dynamic load balancing is it takes fewer messages to reach a decision, as thenumber ofoverall interactions in the system decreases drastically as compared to the semi distributed case.So there is a less chance of congestion.

3.3. Qualitative metrics for load balancing

The different qualitative metrics or parameters that are considered important for load balancing in cloud computing are discussed as follows:

1. **Throughput:** The total number of tasks that have completed execution is called throughput. A high throughput is required for better performance of the system.

2. **Associated Overhead:** The amount of overhead that is produced by the execution of the load balancing algorithm. Minimum overhead is expected for successful implementation of the algorithm.

3. **Fault tolerant:** It is the ability of the algorithm to perform correctly and uniformly even in conditions of failure at any arbitrary node in the system.

4. **Migration time:** The time taken in migration or transfer of a task from one machine to any other machine in the system. This time should be minimum for improving the performance of the system.

5. **Response time:** It is the minimum time that a distributed system executinga specific load balancing algorithm takes to respond.

6. **Resource Utilization:** It is the degree to which the resources of the system are utilized. A good load balancing algorithm provides maximum resource utilization.

7. **Scalability:** It determines the ability of the system to accomplish load balancing algorithm with a restricted number of processors or machines.

8. **Performance:** It represents the effectiveness of the system after performing load balancing. If all the above parameters are satisfied optimally then it will highly improve the performance of the system.

4. Survey of load balancing algorithms

Distributing the workload of multiple network links results into achieve maximum throughput, minimize response time and to avoid overloading. For example to distribute the load, one can use three algorithms namely Round robin, equally spread current execution load and Throttled Load balancing. The performance of these three load balancing algorithms is studied in [4].

In Round robin algorithm circular order is maintainedto handle the process butwithout priority. Round robin algorithm is random sampling based. It means it selects the load randomly in case that some server is heavily loaded or some are lightly loaded.

Butequally spread current execution handle the processwith priorities. It distribute the load randomly by checking the size and transfer the load to that virtual machine which is lightly loaded or handle thattask easy and take less time, and give maximize throughput. It is spread spectrum technique in whichthe load balancer spread the load of the job in hand into multiple virtual machines.

In Throttled algorithm the client firstrequests the load balancer to find a suitable VirtualMachine to perform the required operation.

Task is to find which of the three algorithms is the most efficient in terms of cost of usage.Important point to be observed is for the experiment conducted in [4],request time was the same for all three algorithms that means there is no effect on data centers request time after changing the algorithms. The cost analysis showed for each algorithm is calculated in the experimental work. The experiment showed that the cost calculated for virtual machine usage per hour is same for two algorithms Round Robin, Equally spread current execution load, but Throttled Load balancing algorithm reduces the cost of usage, so conclusion was that Throttled Load balancing algorithm works more efficiently in terms of cost for load balancing on cloud data centers.

Load balancing algorithms also must consider the careful consumption of power. So there should be a mechanism to keep idle nodes off. Along with this point to be considered, there should be a mechanism to find minimum number of active nodes required. Power Aware Load Balancing algorithm[5], PALB is one such approach where the state of all compute nodes are maintained, and based on utilization percentages, it decides the number of compute nodes that should be operating. It balances resources across available compute nodes in a cloud with power savings in mind. Depending on the job schedule distribution and virtual machine request size, organizations can save 70% - 97% of the energy consumed compared to using load balancing techniques that are not power aware.

A good task scheduler should adapt its scheduling strategy to the changing environment and the types of tasks. Therefore, a dynamic task scheduling algorithm is appropriate for clouds. A cloud task scheduling policy based on LoadBalancing Ant Colony Optimization (LBACO) algorithm[6] is an example. It is distributed in nature. Task is to balance the entire system load while trying to minimizing the makespan of a given tasks set. The idea of developing this algorithms raised by one aspect of ant behavior, the ability to find what computer scientists would call shortest paths, has become the field of ant colony optimization (ACO), the most successful and widely recognized algorithmic technique based on ant behavior. This scheduling strategy was simulated using the CloudSim toolkit package. Experiments results showed the proposed LBACO algorithm outperformed FCFS (First Come First Serve) and the basic ACO (Ant Colony Optimization).

The overall performance of the cloud Environment can be increased and also the average response time can be decreased by the selection of an efficient virtual machine A new VM load balancing algorithm[7] has been proposed which is implemented in CloudSim, an abstract cloud computing environment using java language. This new algorithm came up with little modification to the throttled load balancing; in order to achieve better response time, processing time and cost. Proposed algorithm has to find the expected response time of each resource (VM). Next step is to send the ID of virtual machine having minimum response time to the data center controller for allocation to the new request.

One of the significant features of Long-connectivity application is that the users' requests maintain a long connection with web server in a period of time, but they take up very little CPU and memory. An improved algorithm is proposed based on the weighted Least Connection algorithm. In the new algorithm, load and processing power are quantified, and singleExponential smoothing forecasting mechanism [8] isadded. Finally, the article proves by experiments that the new algorithm can reduce the server load tilt, and improve client service quality effectively.

Since, cloud has inherited characteristic of distributed computing and virtualization, there is a possibility of occurrence of deadlock. Deadlock can be explained as follows: there is a queue of requests waiting for their turn to access resources which are shared among them. Further these requests cannot be serviced as the resources required by each of these requests are held by another process or request by virtual machines.Communication between the Load Balancer and the DataCenterController for updating the index table may lead to deadlock. So it is difficult to get the response by the system for incoming requests. Hence, in paper [9] a new load balancing algorithm has been proposed to avoid deadlocks among the Virtual Machines (VMs) while processing the requests received from the users by VM migration. The deadlock avoidance enhances the number of jobs to be serviced by cloud service provider and thereby improving working performance and the business of the cloud service provider.

5. Conclusion

The purpose of this paper is to focus on one of the major concerns of cloud computing that is load balancing.Cloud computing is a promising technology, which is a pay-go model that provides the required resources to its clients. Since, virtualization is one of the core characteristics of cloud computing it is possible to virtualize the factors that modulates business performance such as IT resources, hardware, software and operating system in the cloud-computing platform.

The aim of this paper was to briefly discuss about various efficient and enhanced load balancing algorithmsalong with concepts of virtualization and cloud computing. It is required to distribute the dynamic local workload evenly across all the nodes to achieve a high user satisfaction resource utilization ratio by making sure that every computing resource is distributed efficiently and fairly. With proper load balancing, resource consumption can be kept to a minimum which will further reduce energy consumption and carbon emission rate which is a need of today'scloud computing. A lot of research work need to be done in order to focus on energy consumption and carbon emission factors along with other factors like reducing associated overhead, service response time and improving performance etc.

6. References

[1] Nidhi Jain Kansal1, Inderveer Chana, "Cloud Load Balancing Techniques: A Step Towards GreenComputing",IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012.

[2] Nayandeep Sran, NavdeepKaur, "Comparative Analysis of Existing Dynamic Load Balancing Techniques", International Journal of Computer Applications (0975 – 8887) Volume 70– No.26, May 2013 25.

[3] AartiKhetan,Vivek Bhushan, Subhash Chand Gupta, "A Novel Survey on Load Balancing in Cloud Computing", International Journal of Engineering Research & Technology (IJERT) , ISSN: 2278-0181, Vol. 2 Issue 2, February- 2013.

[4] Dr. Hemant S. Mahalle Prof. Parag R. KaveriDr.VinayChavan,"Load Balancing On Cloud Data Centres", International Journal of Advanced Research in Computer Science and Software Engineering", ISSN: 2277 128X, Volume 3, Issue 1, January 2013

[5] Jeffrey M. Galloway, Karl L. Smith, Susan S. Vrbsky, "Power Aware Load Balancing for Cloud Computing", Proceedings of the World Congress on

Engineering and Computer Science 2011 Vol I, October 19-21, 2011, San Francisco, USA.

[6] Kun Li, Gaochao Xu, Guangyu Zhao, Yushuang Dong, Dan Wang, “Cloud Task scheduling based on Load Balancing Ant Colony Optimization”, 2011 Sixth Annual ChinaGrid Conference, Published in IEEE computer society 2011.

[7] Prof.Meenakshi Sharma, Pankaj Sharma, “Performance Evaluation of Adaptive Virtual Machine Load Balancing Algorithm”, International Journal of Advanced Computer Science and Applications, Vol. 3, No.2, 2012.

[8] XiaonaRen, RonghengLin,HuaZou, “A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast”,Proceedings of IEEE CCIS2011.

[9]Rashmi. K. S, Suma. V, Vaidehi. M, “Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud”, Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012.