JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

OPTIMIZING CLOUD COMPUTING: UNDERSTANDING LOAD BALANCING FOR ENHANCED RESOURCE ALLOCATION AND APPLICATION PERFORMANCE

[1]T.Manigandan , [2]S.Indhumathi, [3]S.T.Saravanan,
[1]Assistant Professor,[2] Assistant Professor, [3]Assistant Professor
[1]Department of CSE, [2]Department of CSE, [3]Department of CSE,
[1]SRS College of Engineering and Technology, Salem, India,
[2]Bharathiyar Institute of Engineering for Women, Salem, India
[3]Bharathiyar Institute of Engineering for Women, Salem, India

Abstract :A well-defined paradigm for computer services, cloud computing involves using internet-based equipment and packages that are specifically built to gather data and resources from cloud service providers. To put it simply, cloud computing is the economical provision of computer resources and services to users. Sharing resources may result in a problem with their availability, which creates a situation where there is a standoff. Load balancing is the process of distributing network traffic among numerous servers. This ensures that no server will be overworked. Load balancing distributes jobs evenly, which improves application responsiveness. Additionally, it improves user accessibility for websites and programs. This paper aims to provide an understanding of load balancing.

Keywords: dynamic load balancing, static load balancing, cloud computing, load balancing.

I. INTRODUCTION

The distribution of various services, including storage, servers, networking, software, analytics, and intelligence, over the internet is known as cloud computing. This allows for speedier innovation, more flexible source options, and economies of scale. Consider an example of a website that is accessible to all. A large number of users can access a website or online application at any time. The ability of a web application to promptly handle these user requests becomes critical. It may even lead to system malfunctions. A website owner whose entire business depends on his portal may also lose out on potential customers as a result of the awful perception that his website is unavailable or down. In this case, load balancing is vital.

There are two load balancing layers in cloud systems.

Basic level: The load balancer sends the necessary instances to physical machines when an application is running in an effort to balance the computational load of many applications across physical computers.
Second level: To distribute the computer load among a number of instances, each incoming request that an application receives should be assigned to a specific instance of the program.

1.1 Load Balancing

Workloads and computer resources are dispersed among one or more servers using cloud load balancing. The best performance in the shortest amount of reaction time is guaranteed by this kind of distribution. To optimize resource usage and reduce response time, two or more servers, hard drives, network ports, or other computer resources are divided. Therefore, a website with a lot of traffic may be guaranteed to continue operating with effective cloud load balancing.
In addition to website traffic, "load" also refers to CPU load, network load, and server storage capacity. Every machine in the network is given an equal amount of work at random times thanks to load balancing. This indicates that none of them are in any anyway overworked or underutilized. The load balancer distributes data based on how busy each server or node is. The customer may find it too taxing and demoralizing to wait for his operation to be finished in the absence of a load balancer.

1.2 Load Balancing Objectives

The primary goals of load balancing are as follows:
1. Handling and regulating spikes in traffic on a single server
2. Improving user request response times
3. To raise the resource utilization ratio.
4. A notable boost in performance.
5. Preserving the stability of the system.
6. Make the system more adaptable to handle the changes.
7. There is less waiting in a queue and less time spent on workflow.
8. To encourage user enjoyment.

1.3 Advantages of Load Balancing

High-Performance Applications: Cloud load balancing techniques are less expensive and simpler to implement than their traditional local counterparts. Businesses are able to improve the speed and functionality of their client apps at potentially lower costs.

Improved scalability: Cloud balancing contributes to the scalability and agility of website traffic. Using efficient load balancers, you can easily match rising user traffic and disperse it among multiple servers or network devices. It is especially important for websites that handle thousands of visitors every second. It takes these kinds of efficient load balancers to distribute workloads during promotions or other promotional activities.

Ability to handle traffic surges: A normally operational University website may completely crash during any result announcement. This is a result of the potential for large-scale request receipt. They can employ cloud load balancers and not worry about traffic surges. Regardless of the request's magnitude, it might be distributed amongto get the greatest results in the quickest amount of time, use many servers.

Full flexibility and continuity: The major objective is to protect a website against sudden outages by using a load balancer. When the burden is split among multiple servers or network units, it can be transferred to another active node even in the event of a node failure.

1.4 Need of Load Balancing

Another essential component of cloud scalability is load balancing. Cloud infrastructures must to be easily scalable to accommodate fluctuations in traffic volume. When a cloud "scales up," it typically runs multiple apps and spins up multiple virtual servers. The load balancer is the main network element that divides traffic among these other instances. In the absence of load balancers, newly spun-out virtual servers might not be able to receive incoming traffic at all or in a coordinated manner. While some servers are overloaded, others are left with no traffic handled. Additionally, load balancers have the ability to detect unavailable servers and direct traffic to those that are still operational. The algorithms used by load balancers can even determine whether a particular server (orserver set) is probably going to get overloaded sooner and divert traffic to other nodes that are deemed to be in better condition. Having such preventative abilities can significantly reduce the chance that your cloud services won't function.
Moreover, load balancing is necessary to achieve green cloud computing. The following justifies this:
1. Limited power consumption: By distributing an excessive amount of work among core nodes or virtual machines, load balancing can lower power consumption.
2. Reducing Carbon Emissions: Energy usage and carbon emissions are similar to a coin's heads and tails. Both of them have a direct proportionality. Load balancing contributes to energy efficiency, which in turn leads to automated reductions in carbon emissions and green computing.

1.5 Classification of Load Balancing Algorithms

They are divided into two categories based on the system's current state:
1. 5.1 Static Load Balancing: A load balancing technique is "static" if it distributes jobs without taking the system's current condition into account. The system status includes things like how much each CPU is loaded, and sometimes even excess). Rather, presumptions about the entire system—such as arrival times and incoming resource needs—are made beforehand. Additionally known are the quantity, capacity, and communication rates of CPUs. Therefore, the goal of static load balancing is to minimize a given performance function by matching a given set of workloads with processors that are available. Typically, routers, or Masters, that distribute and optimize loads are the centre of attention for static load balancing systems. Predicted runtime and details about the jobs to be distributed can be factored into this reduction. Static algorithms have the advantage of being very effective and simple to setup for operations that occur very frequently.
1.5.2. Dynamic Load Balancing: Unlike static load distribution strategies, dynamic algorithms take into account the current load of each computer unit, or node, on the system. In order to complete tasks more rapidly, they can therefore be dynamically moved from an overloaded node to an underloaded node. These algorithms can produce excellent results even if they are much more difficult to construct, especially if the execution periods of the various tasks differ significantly. Dynamic load balancing designs may be more adaptable since they do not necessitate a separate node for work allocation. When tasks are assigned to a processor according to its current state at a specific moment, it is a special kind of assignment. Conversely, dynamic assignment describes the capacity to continuously reallocate responsibilities in accordance with the circumstances ofthe system and its development. It goes without saying

that a load balancing algorithm that necessitates a lot of communication in order to reach its conclusions runs the risk of postponing the solution to the problem as a whole.

2 .LOAD BALANCING ALGORITHMS

TABLE I LOAD BALANCING ALGORITHMS

Algorithm	State of Algorithm	Job Distribution	Advantages	Disadvantages
Round Robin and Randomized	Static	1. All processors share the same amount of work. 2. Each processor keeps track of the sequence in which processes are assigned. 3. This method is used to handle user requests in a circular manner.	1. It is good to use it when no of processors are significantly less when compared to no of processes. 2. Inter-process communication is not required in Round Robin.	Because various processes take various amounts of time to complete, some nodes may be fully occupied while others are idle and underutilised at any same moment.
Central Manager	Static	1. The central processor has to pick the host for all new processes. 2. The loading processor minimum relies on the total load determined during the process setup.	The load scheduler decides on load balancing based on the system load statistics	A high level of interprocess communication is required which may be expensive..
Min-min	Static	1. For all jobs, the shortest possible completion time is sought. 2. The minimum value is determined from the minimum times.	Good performance with the best number of resources.	Starvation scenario is possible here.

		3. The work is allocated based on that minimal time.		
Max-Min	Static	This method is quite similar to the above method. However there is one major difference: After obtaining the shortest execution times, the one which is maximum is picked.	Good performance with the best number of resources.	Starvation scenario is possible here.
Honey Bee Foraging Behavior	Dyna mic	Based on the behaviour and approach of the honeybees to harvest honey. The global load balancing is achieved by local server activities.	The virtual machine has reduced response times and waiting time.	Throughput is indirectly proportional to resource number.
Biased Random sampling	Dyna mic	Each server is considered a node's vertex, and the indegree symbolises the nodes' available free resources. The work is assigned based on the in degree. If each node has more than one degree, work will be assigned to that node. When a work is given to a node, the degree of the node is decreased by one, and it is increased after the work is completed.	Performs better with a large and similar resource population.	Degrades with increasing population variety.
Active Clustering	Dyna mic	This algorithm's fundamental premise is to group similar nodes together and operate with those grouped nodes. The resources can enhance throughput more efficiently by grouping nodes together.	1. When there are a lot of resources being used, performance is good. 2. Using additional system resources to boost performance.	Degrades with increasing population variety.
ACCLB(Ant Colony and Complex	Dyna mic	Small-world and size-free features of a complex network make it possible to	This methodology eliminates heterogeneity, is	1. Used in networks wich are complex only.

3. CHALLENGES OF LOAD BALANCING

3.1 Virtual machine migration: The idea is to create a machine as a file or a collection of files. The virtual machine can be moved effectively to lighten the strain on a loaded computer. The goal is to reduce or eliminate strain on cloud computers when the machine's burden is distributed dynamically.

3.2Energy management: Using the cloud offers benefits such as scale economies. For a global economy, energy conservation is an essential concern. Since each person has their own assets and because reduced providers support a variety of global assets. How can the data centre component be used while still having a reasonable throughput?

3.3 Data management and storage: Information storage is yet another essential requirement. Thus, in a cloud system, how can data be shared while ensuring optimal storage and quick access? Spatial distribution of cloud nodes: Some techniques are available only for nearby nodes with slight lag in communication. It is still difficult to create an efficient load balancing strategy that works for geographically dispersed nodes.

3.4 LB Scalability: Guests can access resources for rapid scalability at any moment with cloud services that are both accessible and scalable on-demand. Robust load balancers should accommodate dynamic demands in memory, device architecture, processing conditions, and other areas.

4. CONCLUSION

To put it simply, cloud computing is a way for multiple users to have on-demand access to a variety of online resources. But there are major obstacles when it comes to cloud computing. One major obstacle in cloud computing is load balancing. This paper discusses numerous algorithms, both static and dynamic. As it is well known that clouds have a diverse composition. Staticalgorithms make environment monitoring and modelling simple, but they cannot replicate the varied character of clouds. Although dynamic load balancing methods are challenging to model, they work well in cloud systems because of their diversity. This essay describes load balancing, its benefits, requirements, and challenges. It also examines various load balancing techniques now in use.

REFERENCES

[1]. Ms. Shalini Joshi, Dr. Uma Kumari "Load Balancing in Cloud Computing: Challenges& Issues" , Conference: 2016 2nd International Conference on Contemporary Computing and Informatics, DOI:10.1109/IC3I.2016.7917945.

[2]. Muhammad AsimShahid, Noman Islam, Muhammad MansoorAlam, MazlihamMohdSu'ud, ShahraniMusa, "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach " . DOI: 10.1109/ACCESS.2020.3009184

[3]. Foram F Kherani, Prof.Jignesh Vania, "Load Balancing in cloud computing", 2014 IJEDR | Volume 2, Issue 1 | ISSN: 2321-9939

[4]. Shahbaz Afzal and G. Kavitha, "Load balancing in cloud computing – A hierarchical taxonomical classification", Journal of Cloud Computing volume 8, Article number: 22 (2019). DOI: https://doi.org/10.1186/s13677-019-0146-7

[5]. Abhijit Aditya, Uddalak Chatterjee and Snehasis Gupta, "A Comparative Study of Different Static and Dynamic Load Balancing Algorithm in Cloud Computing with Special Emphasis on Time Factor ",International Journal of Current Engineering and Technology, Vol.5, No.3 (June2015)

[6]. Bhaweshkumawat ,Rekhakumawat,"A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment using Cloud Analyst", IJESC Volume 7 Issue No.3.