Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 9, June-2019, ISSN 2230-7540

Dynamic on Demand Scaling and Load Balancing

Ravi Shanker Sharma[1]* Gireesh Kumar Dixit[2]

[1] Research Scholar

[2] Department of Computer Science, Shyam University, Dausa, Rajasthan

Abstract – Cloud computing is a well-defined computing paradigm in which data and assets are collected from cloud service providers over the Internet using well-designed Internet-based devices and software. Cloud computing is a collection of computing resources and services that make it available to customers at low cost. Sharing resources can lead to a stalemate due to lack of such resources. The process of distributing network traffic to different servers is called load balancing. As a result, a single server will not be overloaded. Load balancing improves application responsiveness by evenly distributing jobs. It also improves the usability of programs and websites. The purpose of this document is to help you better understand load balancing.

Keyword – Cloud Computing, Load Balancing, Static Load Balancing, and Dynamic Load Balancing.

----------X----------

1. INTRODUCTION

The main objective of the proposed Dynamic Load Balancer is to solve the problem of load balancing in an architecture agnostic way, machine learning techniques are used to identify the type of server and then do load balance accordingly. With the existing traditional load balancing systems, each of the job's operation specific resource requirement is identified and then the load balancing is done accordingly. Some of the algorithms even take the approach of random probability based balancing and it works fine with respect to homogeneous system. When it comes to heterogeneous systems, the load balancing is still restricted to the resources and the load on the server. This often leads to inefficient load balancing as the servers are not utilized to the maximum. Existing system needs the underlying architecture of each of the systems is to be known in prior. Also, if an additional system is added into the pool, then the load balancer should be informed of the type of server. This hinders the ability to add a server dynamically to the existing environment. Thus causing the decreased efficiency. The proposed solution tackles the above problems and helps to add servers dynamically to the system without the prior knowledge of underlying architecture.

II. BALANCE OF LOAD

Workloads and compute resources are distributed to one or more servers using cloud load balancing. This type of distribution guarantees the best results in the shortest amount of time. Isolate two or more servers, hard drives, network ports, or other computer resources to maximize resource utilization and reduce response time to system activity.

As a result, effective cloud load balancing ensures that busy websites continue to function. The term "load" refers not only to website traffic, but also to CPU load, network load, and server storage capacity. Load balancing ensures that all machines on your network are constantly exposed to similar loads.

This indicates that they are overworked or underutilized. The load balancer distributes the data based on the busy state of each server or node. Customers have to wait for the operation to complete without a load balancer. This can be too tiring and discouraging.

III. OBJECTIVES FOR BALANCED LOADING

The key issues that data centres faces are as follows:

1. In order to accommodate incoming service requests, data centre resources must be regularly supplied.

2. The efficient assignment of various tasks to available resources such that the

burden is distributed evenly across all resources globally.

3. Cloud data centre server distribution: Cloud data centre servers are spatially scattered around the globe, necessitating the development of efficient load balancing techniques for datacenter servers that are separated by networks that are prone to network delays.

4. Data storage and replication: Data centres replicate data so that it is always available, even if there are severe catastrophic failures. The cost of fully duplicating data into servers will be higher because the amount of storage required would be substantial. As a result, partial data replication is performed at various data centre servers dependent on their processing power and storage capacity. 8. To increase the user's delighted.

IV. BENEFITS OF LOAD BALANCE

High Performance Applications: Cloud load balancing technology is cheaper and easier to implement than traditional on-premises technology. Enterprises can run their customers' apps faster and improve performance than they can at low cost.

Increased scalability: Cloud balancing supports the scalability and agility of website traffic. A good load balancer makes it easy to scale and distribute increasing user traffic across multiple servers or network devices. This is very important for online sites that handle thousands of web accesses per second. A load balancer of this caliber is needed to distribute the workload during sales or other advertising services.

Ability to handle traffic spikes: During the announcement of results, a functioning university site could go down altogether. This is because you may receive a large number of requests. If you use a cloud load balancer, you don't have to worry about traffic spikes. Regardless of the size of your request, you can carefully distribute your requests across a large number of servers for best results in the shortest amount of time.

Full flexibility continuity: The basic purpose of setting up a load balancer is to rescue or protect your site from sudden failures. If the load is distributed across multiple servers or network units, jobs can be offloaded to another active node even if one node fails.

V. THE DEMAND FOR LOAD BALANCE

The scalability of a cloud is also dependent on load balancing. Cloud infrastructures should be easily expandable to accommodate traffic spikes. When a cloud "scales up," it usually spins up a bunch of virtual servers and runs a bunch of apps. The load balancer is the main network component that distributes traffic across these new instances.

Virtual servers that are created from scratch without load balancers may not be able to accept incoming traffic in a coordinated manner, if at all. Some servers are unaffected by traffic, while others are overburdened. Load balancers can also detect downed servers and redirect traffic to those that are still operational. Load balancers can even analyse if a given server (or server set) is likely to be overrun sooner and divert traffic to other nodes deemed to be more healthy, depending on load balancing algorithms. Such preemptive abilities can considerably reduce the chances of your cloud services becoming unavailable.

In order to achieve green cloud computing, load balancing is also necessary. The following are the reasons for this:

1. Reduced power consumption: Load balancing can minimise power consumption by spreading out heavy workloads among core nodes or virtual machines.

2. Carbon Emissions Reduction: Carbon emissions and energy consumption are opposite sides of the same coin. They're both proportional in the same way. Load balancing aids in the reduction of energy use, which naturally minimizes carbon emissions and results in Green Computing.

VI. NEED FOR FAULT TOLERANCE IN DATA CENTER NETWORKS

Data centres are regarded as the information and communication's central hub. Due to the vast increase in application size, applications became business-critical, resulting in unacceptable service downtime. This, combined with the growing demand for cloud computing, forced scientists and developers to focus more on designing fault-tolerant and scalable data centre network architectures.

Today, with recent breakthroughs in the development of cloud-based data storage centres, the need for additional cloud services has expanded, resulting in data centre failures owing to massive scales of data storage. As the number of nodes in data centres increases, data size and access get more difficult, requiring distinct degrees of access to retrieve each application or data item. In the face of growth, the goal of fault tolerance is to establish robustness and reliability in any system.

Existing fault-tolerance solutions in cloud computing take into account a variety of factors, including fault-tolerance type (proactive, reactive, and adaptive), performance, response time, scalability, throughput, dependability, availability, usability, security, and overhead. The purpose of a data centre network is to keep data centre servers connected while also providing efficient and fault-tolerant routing techniques.

The importance of having a fault-tolerant data centre cannot be overstated, especially now that big data traffic, the internet of things, and other on-demand internet applications are on the rise. The velocity at which these data are transported across the internet is alarming, and data centre developers are concerned.

VIII. LOAD BALANCING DIFFICULTIES

Virtual machine migration: The idea is to create a machine as a file or a file collection. The strain on a loaded computer can be lessened by moving the virtual machine around effectively. When load is distributed dynamically on the system, the goal is to eliminate and minimise strain on cloud computers.

Energy management: One of the benefits of adopting the cloud is the economies of scale. Conservation of energy is a critical issue for the world economy. Because a variety of global assets are supported by decreased suppliers and each one has its own assets. How can the data centre component be utilised while ensuring adequate throughput

Data storage and management: Information storage is another critical requirement. So, with a cloud system, how can data be dispersed with the best storage and access the spatial distribution of cloud nodes: Some methods are only available for nodes that are close together and have low communication latency. Designing an effective load balancing mechanism that can be articulated for spatially separated nodes, on the other hand, remains a challenge.

LB Scalability: Guests can access resources for rapid scaling at any moment thanks to accessible and on-demand scalable cloud services. A good load balancer should adjust for rapidly changing processing conditions, memory, device architecture, and other factors.

IX. CONCLUSION

Simply described, cloud computing is a technique allowing multiple users to access a variety of resources over the internet on a need-to-know basis. Cloud computing, on the other hand, faces considerable challenges.

In cloud computing, load balancing is a significant challenge. This paper discusses a variety of static and dynamic algorithms. Clouds, as is well known, are diverse in nature. Static algorithms make modelling and monitoring of the environment simple, but they don't emulate the complexities of cloud computing. Dynamic load balancing methods are difficult to model, but they are ideally suited to the many different types of cloud environments.

This paper provides an overview of load balancing, its benefits, requirements, and challenges, as well as a discussion of some existing load balancing solutions.

REFERENCES

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

[2] "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach," by Muhammad Asim Shahid, Noman Islam, Muhammad Mansoor Alam, Mazliham Mohd Su'ud, and Shahrulniza Musa. 10.1109/ACCESS.2020.3009184 DOI: 10.1109/ACCESS.2020.3009184

[3] Foram F Kherani and 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). https://doi.org/10.1186/s13677-019-0146-7/s13677-019-0146-7/s13677-019-0146-7/s13677-019-0146-7/s136

[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 (June 2013). (June- 2015)

[6] Bhawesh and Rekha Kumawat, "A Comparative Study of Load Balancing Algorithms in a Cloud Computing

Environment Using Cloud Analyst," IJESC Volume 7 Issue No.3.

Corresponding Author

Ravi Shanker Sharma*

Research Scholar