International Journal of Engineering and Techniques - Volume 9 Issue 2, July 2023

RESEARCH ARTICLE OPEN ACCESS

Load balancing in Cloud Computing

[1]Shreya Bhar, [2]Rupsa Das, [3]Sudip Das

[1]Student, [2]Student, [3]Assistant Professor

[1]Department of Computer Application,

[1]Narula Institute of Technology, India

[1]shreyabhar3@gmail.com, [2]rupsadas334@gmail.com, [3]sudip.das.mtech15@gmail.com

Abstract:

Load balancing is a crucial technique in computer systems and networks to distribute workload efficiently among available resources with the increasing demand for high-performance and scalable applications, load balancing algorithms have become essential for maintaining optimal performance and maximizing resource utilization. This abstract provides an overview of load balancing, its significance, and various approaches employed to achieve load balancing in different contexts. this abstract provides a comprehensive overview of load balancing, its significance, various strategies and algorithms, challenges, and its application in diverse computing environments. It serves as a foundation for further exploration and understanding of load balancing techniques and their role in achieving optimal resource utilization and system performance.

Keywords **— Cloud Computing, Load Balancing, Static Load Balancing, Dynamic Load Balancing, Load Balancing Algorithms and techniques, Challenges.**

I. Introduction

Cloud computing is the next generation technology that's rapidly expanding as clients demand more services and better results. Cloud Load balancing is the process of dividing workloads and computing resources across multiple servers to ensure maximum throughput in minimum response time. This distribution helps meet the increasing demand for services and better results while maintaining efficiency.

II. Load balancing

Load balancing distributes network traffic across multiple instances to ensure peak performance without overburdening any single instance, traditionally using a dedicated physical network device or application, but increasingly through a software-based virtual appliance or network service provided by public cloud providers.

III. Load balancing objectives

Load balancing aims to achieve the following objectives:

Optimize resource utilization

Load balancing distributes workloads evenly across resources to maximize utilization, reduce wastage, and achieve cost savings by avoiding overburdened/underutilized resources.

Improve system performance

By distributing workloads across multiple resources, load balancing helps to prevent system overload and reduce response times, thereby improving system performance and ensuring that services are delivered with optimal speed and efficiency.

Enhance system availability

Load balancing maintains system availability and prevents service disruptions by directing traffic to alternative resources in the event of failures or outages, ensuring that resources are always accessible to users.

Facilitate scalability

Load balancing supports system scalability by dynamically adding or removing resources as demand changes, ensuring scalability to meet changing requirements.

IV. Advantages of load balancing

Performance and Speed

Load balancing improves cloud application performance by evenly distributing workloads across multiple servers, preventing overburdening and resulting in faster response times and better user experience as slow websites are a turnoff for customers who expect fast results.

Redundancy

Load balancers provide a backup and redundancy strategy for servers that may go down due to routine maintenance or hardware failure. Let's say the cluster that powers your website has four servers — A, B, C, and D, if one server fails, the load balancer redirects traffic to other servers to prevent the website from going down and ensure uninterrupted service. This makes load balancers an effective solution for both small and large companies.

Scalability

Load balancing enables scalability by adding servers to handle increased traffic without performance loss.

Cost savings

Load balancing reduces costs by maximizing the potential of existing resources, eliminating the need for expensive upgrades or additional servers.

Security

Load balancing can improve security by providing a layer of defense against DDoS attacks. By distributing traffic across multiple servers, it becomes harder for attackers to target a single server.

V. Need of load balancing

Improve scalability

By distributing workloads across multiple servers or instances, load balancing enables organizations to scale their infrastructure seamlessly without any disruption to their services.

Increase high availability

Load balancing ensures that resources are available to handle incoming requests even if one or more servers or instances fail. By redirecting traffic to healthy servers, load balancing helps to maintain service continuity and minimize downtime.

Increase performance

Load balancing enables organizations to achieve high performance by ensuring that incoming requests are processed by the most available and least busy server or instance, reducing response time and increasing throughput.

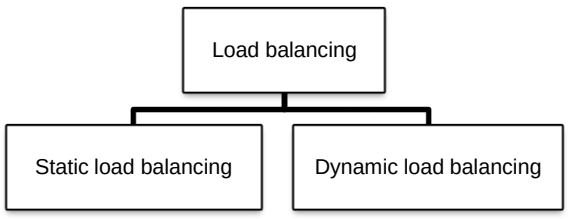
Advance resource utilization

Load balancing ensures that computing resources are used optimally by distributing workloads evenly across all available servers or instances

VI. Classification of Load balancing in cloud computing

Basically, load balancing can be divided into two parts. Those two parts are,

- Static load balancing
- Dynamic load balancing



Static load balancing

Static load balancing divides incoming server loads using predetermined algorithms based on existing servers in the distributed network and a pre-planned schedule.

Dynamic load balancing

Dynamic load balancing in cloud computing is a software tool that allows each parallel job to do its application level load balancing while ensuring that system load is balanced.

VII. Load balancing algorithm

Load balancers (reverse proxy) use some common load-balancing algorithms. These algorithms include the following:

Round robin

This process divides incoming traffic requests equally over all the workload instances (nodes). For example, if there are two workload instances, the load balancer will turn traffic to each instance orderly -- request 1 to server 1, request 2 to server 2, request 3 to server 1 and so on.

Weighted round robin

Some workload instances use servers with distinct computing proficiency. Weighted round robin can move the percentage of traffic to distinct nodes by assigning a "weight" to each node.

Least connection

Traffic is directed to less busy workload instances with the fewest connections or shortest queue, relieving demands on instances with complex processing needs.

Weighted least connection

A "weight" is given to each node, allowing administrators to adjust traffic distribution based on connection activity, resembling round robin or weighted round robin when nodes are similar.

Resource-based

VIII. This method employs a software agent on each node to assess computing load and report accessibility to the load balancer, which makes dynamic traffic routing decisions.

Request-based

IX. Cloud load balancers distribute traffic based on request fields such as HTTP, source and destination IP addresses, to route traffic from specific sources to required purposes and maintain periods that were separated.

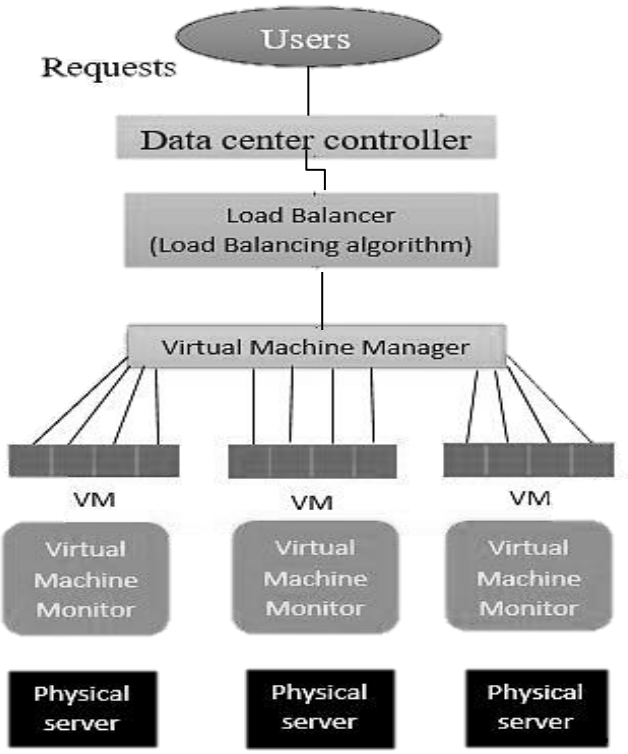


Fig.1 Load Balancing algorithms execution
Source: [9] Journal of Network and Computer Applications Volume 88, 15 June 2017

X. Challenges of Load Balancing

Following are some load balancing problems that are mentioned below:

Geographical Distributed Nodes

Geographically scattered data centers are strategically distributed based on the geographical features of an area or a place to create a cohesive

system for efficient computation and carrying out user requested operations.

Single Point of Failure

In non-distributed dynamic load balancing algorithms, the master node is responsible for making decisions related to load balancing. However, if the master node crashes, the entire computing domain is disrupted.

Virtual Machine Migration

Virtualization involves creating or combining multiple individualistic VMs with varying configurations on a single physical system, and in the event of overburdening, Cloudlet migration techniques can be used to relocate certain VMs to a different location.

Algorithm Complexity

Algorithm design should prioritize simplicity and ease of implementation as increased complexity can result in decreased performance and efficiency in a cloud environment.

Load Balancer Scalability

Cloud services allow for easy accessibility and resource scalability to meet user demand, and a reliable load balancing algorithm must be capable of adapting quickly to changes in network topology, power, and other factors to ensure efficient system performance.

XI. Dynamic load balancing policies and strategies

The different policies as described in [2], [3] are as follows:

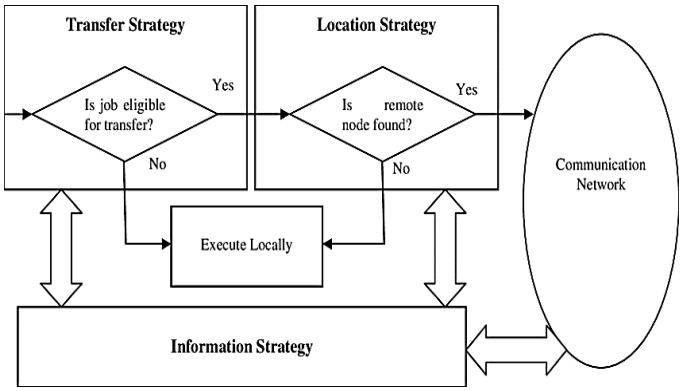


Fig.2 Interaction between different components of Dynamic Load Balancing Algorithm
Source: [2] Ali M Alakeel, "A Guide to Dynamic Load Balancing in Distributed Computer Systems", International Journal of Computer Science and Network Security, Vol. 10 No. 6, June 2010.

Location Policy

The policy used by a processor or machine for sharing the task transferred by an over loaded machine is termed as Location policy.

Transfer Policy

The policy used for selecting a task or process from a local machine for transfer to a remote machine is termed as Transfer policy.

Selection Policy

The policy used for identifying the processors or machines that take part in load balancing is termed as Selection Policy.

Information Policy

The policy that is accountable for gathering all the information on which the decision of load balancing is based id referred as Information policy.

Load estimation Policy

The policy which is used for deciding the method for approximating the total work load of a processor or machine is termed as Load estimation policy.

Process Transfer Policy

The policy which is used for deciding the execution of a task that is it is to be done locally or remotely is termed as Process Transfer policy.

Priority Assignment Policy

The policy that is used to assign priority for execution of both locals and remote processes and tasks is termed as Priority Assignment Policy.

Migration Limiting Policy

The policy that is used to set a limit on the maximum number of times a task can migrate from one machine to another machine.

XII. Comparison Chart

Comparative study on different factors							Time analysis	
Metrics/ Techniques	Throughput	Overhead	Fault tolerance	Resource Utilization	Scalability	Performance	Migration Time	Response Time
Round robin	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Olb+lbmm	No	No	No	Yes	No	Yes	No	No
Min-min	Yes	Yes	No	Yes	No	Yes	No	Yes
Max-min	Yes	Yes	No	Yes	No	Yes	No	Yes
Olb	No	No	No	Yes	No	Yes	No	No

Fig.3 Comparison of various algorithms

Source: [10] Jitendra Bhatia, Tirth Patel, Harshal Trivedi, Vishrut Majmudar,” HTV Dynamic Load Balancing Algorithm for Virtual Machine Instances in Cloud”,18 Dec2 012 Pages 15-20 IEEE.

XIII. Conclusion

Load balancing is a critical aspect of cloud computing that ensures optimal resource utilization, improved system performance, enhanced system availability, and facilitated scalability. By distributing workloads and computing resources across multiple servers, load balancing maximizes performance, reduces costs, and enhances security. However, it also poses challenges such as geographical distribution, single points of failure, and scalability. Effective load balancing policies and strategies, coupled with appropriate load balancing algorithms, are essential for efficient resource allocation and workload distribution in cloud computing environments.

References

[1] "Load Balancing in Cloud Computing" by Avi Networks

[2] "Dynamic Load Balancing Policies and Strategies" by Google Cloud Platform

[3] "Load Balancing Algorithms and Techniques" by Amazon Web Services

[4] "Load Balancing: A Beginner's Guide" by Cloud Academy

[5] "How to Implement Load Balancing in Your Cloud Environment" by Cloudflare

[6] "The Ultimate Guide to Load Balancing" by Nexcess

[7] Jitendra Bhatia, Tirth Patel, Harshal Trivedi, Vishrut Majmudar,” HTV Dynamic Load Balancing Algorithm for
Virtual Machine Instances in Cloud”,18,Dec2 012,Pages 15-20 IEEE.

[8] Ali M Alakeel, “A Guide To Dynamic Load Balancing In Distributed Computer Systems”, International
Journal of Computer Science and Network Security, Vol. 10 No. 6, June 2010.

[9] Journal of Network and Computer Applications Volume 88, 15 June 2017