**Scholarly Research Journal for Humanity Science & English Language, Online ISSN 2348-3083, SJ IMPACT FACTOR 2021: 7.278, www.srjis.com**
PEER REVIEWED & REFEREED JOURNAL, AUG-SEPT, 2022, VOL-10/53

# LOAD BALANCING IN CLOUD COMPUTING

**Meenal Sachdeva[1]& Reena[2], Ph.D.**

*[1]Assistant Professor, Hindu Institute of Management & Technology, Rohtak,*

*[1]meenalsachdeva1@gmail.com*

*[2]Assistant Professor Baba Mast Nath University, Rohtak,*

*[2]reena@bmu.ac.in*

**Paper Received On:** 25 SEPTEMBER 2022

**Peer Reviewed On:** 30 SEPTEMBER 2022

**Published On:** 1 OCTOBER 2022

## *Abstract*

*Cloud computing has proposed a new perspective for provisioning the large-scale computing resources by using virtualization technology and a pay-per-use cost model. Load balancing is taken into account as a vital part for parallel and distributed systems. It helps cloud computing systems by improving the general performance, better computing resources utilization, energy consumption management, enhancing the cloud services' QoS, avoiding SLA violation and maintaining system stability through distribution, controlling and managing the system workloads. In this paper we study the necessary equirements and considerations for designing and implementing a suitable load balancer for cloud environments. In addition we represent a complete survey of current proposed cloud load balancing solutions which according to our classification, They can be classified into three categories: General Algorithm-based, Architectural-based and Artificial Intelligence-based load balancing mechanisms. Finally, we propose our evaluation of these solutions based on Suitable metrics and discuss their pros and cons.*

*Index Terms—Cloud Computing, Load Balancing, Distributed Systems, Virtual Machine.*

***Keywords:*** *Cloud Computing, Load Balancing, Distributed Systems, Virtual Machine*

## Introduction

Cloud computing is known as a popular and important term in the IT society these days. It has emerged as a large scale distributed computing paradigm that is driven by economies of scale and provides the situation that services can be dynamically configured and delivered on demand . To emphasize that cloud computing has some featured goals, we refer to the definition of cloud computing provided by National Institute of Standards and Technology (NIST) that says: “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. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models” . As the main goal of cloud computing we can mention the better use of distributed resources and applying them to achieve a higher throughput, performance and solving large scale computing problems . Generally speaking, based on the NIST definition of cloud computing we can say best effort for offering the on demand services based on the best use of available shared resources is one of the most important goals of this model. To achieve these kinds of goals, improving the general performance of system, maintain stability, availability and some other features for a cloud computing data center, we need a mechanism which is called load balancing. Load balancing is one of the central issues and challenges in distributed systems like grid-based systems and cloud computing . It is still a new problem in the cloud computing that needs new architectures and algorithms to promote the traditional approaches. In cloud computing, scheduling and handling many jobs that their arrival pattern, type of service and other properties are so hard to predict cause of the dynamic on-demand network access feature of system, an efficient load balancing mechanism is so necessary to increase the Service Level Agreement (SLA), deliver a robust service and provide other essential system requirements. In addition load balancing is an important topic because it enables other important features such as scalability .

Many researches have been done in the field of cloud computing and different challenges that are related which includes studies about cloud computing security and privacy like researches that have been done in or researches related to modeling and performance analysis

in cloud environments , challenges related to the frameworks and architectures of cloud and other general challenges in cloud computing such as scheduling, energy efficient and green computing and etc Also there are many studies about load balancing in distributed and peer to peer networks , but a little comprehensive research about load balancing in the field of cloud computing has been done yet and we just can refer to some papers that there are in the field of load balancing in cloud computing . In this paper we present a survey of the algorithms, architectures and all techniques which have proposed for cloud computing load balancing.

We discuss their properties and parameters that they considered and also give a comparative view of the current real load balancing mechanisms. The rest of this paper is organized as follows. In section 2 we review the literature of load balancing and will discuss the general classifications, principles and mechanism of load balancing algorithms. We survey the challenges, issues and special points that should be taken into consideration during the designing and offering a load balancer in cloud computing in section 3. In section 4 we discuss the features, positive and negative point of current cloud load balancing approaches. In addition we will have a comparative look on some parameters which each of proposed algorithms have considered.

**Literature Review (Load Balancing)**

Technological progress in computer world and rapid developments of distributed systems over last few years caused that load balancing problem were took into consideration as a main challenge more than ever in these systems. This section presents some important concepts and approaches of load balancing mechanism. Load balancing is the process of redistributing the general system work load among all nodes of the distributed system (network links, disk drivers, central processing units…) to improve both resource utilization and job response time while avoiding a situation where some nodes are overloaded while others are under loaded or idle . Load balancing is a vital and inseparable part of cloud computing and elastic scalability . In order to avoid system failure, load balancing is often used by controlling the input traffic and stop sending the workload to resources which become overloaded and non-responsive. This is an inherited feature from grid-based computing which has been transferred to cloud computing . Here, there are some important goals of load balancing mechanism which have been mentioned in different researches:

- Reducing Job response time while keeping acceptable delay.
- Maintaining system stability.

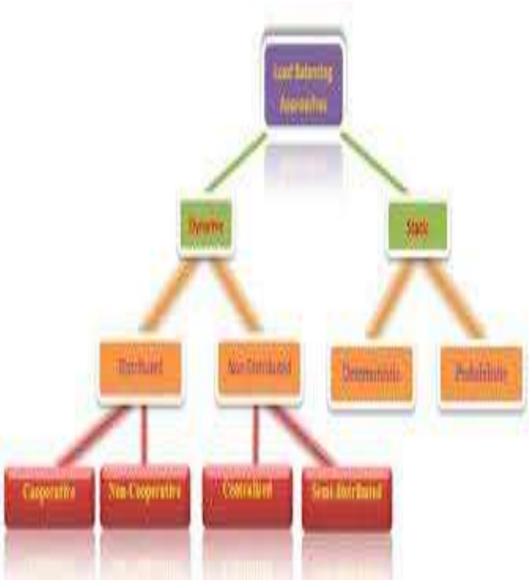
- Having fault tolerance ability using load balancing for implementing failover.
- Improving the general system performance for achieving optimum resource utilization.
- Improving and maintaining the availability of cloud computing.

To start designing load balancers, there are some considerations from the point of architectural which are pointed out by and we summarize them here:

- Design for providing scalability that could cover all different designing level such as CPU level, machine level, and network level or even could be at the application and data center level.
- Full tolerance for applications.
- Scalability of the request handling capacity in a self organized way.
- The ability of handling more higher and complex traffic

There are many various kinds of load balancing mechanisms and approaches which most of the studies have been classified as two main categories static and dynamic. In static techniques, there are usually prior knowledge and some assumptions about the global status of the system such as job resource requirements, communication time, processing power of system nodes, memory and storage devices capacity and so on. A static approach is a kind of assignment from a set of tasks to a set of resources which can take either a deterministic or a probabilistic form . In addition, this approach is defined usually in the design or implementation of systems . In deterministic assignments, the extra workload of a certain node will be transferred to another specific node all of the time, but inprobabilistic approaches each node sends its extra tasks with probability P to a node and with probability 1-P to another one. The main drawback of static load balancing algorithms is that the current state of the system is not considered when making the decisions and therefore it is not a suitable approach in systems such as distributed systems which most states of the system changes dynamically. Dynamic load balancing techniques take into consideration the current state of systems that their decisions are based on. In this technique tasks can move dynamically from an overloaded node to an under-loaded one and this is the main advantage of dynamic load balancing algorithms which can change continuously according to the current state of the system. .

Dynamic load balancing algorithms can be designed in two different ways: distributed and non-distributed. In distributed approaches. In addition, in this approach all nodes can communicate with each other for achieving a global goal in the system which is called cooperative or every node can work independently for just achieving a local goal that is non-cooperative form. But, in a non-distributed scheme , the responsibility of balancing the system workload would not be performed by all system nodes.



(Figure 1)

The dividing load balancing algorithms into two dynamic and static categories is based on the point that algorithms take into account the current state of system for making decision or not. But from another point of view that has mentioned in load balancing techniques can be divided into two other general categories based on some other factor

Based on the way that the system load is

- Based on the way that the system load is distributed and the resource is assigned to the tasks.
- Based on the system topology and available information..

The first category is designed as :

- Centralized approach.

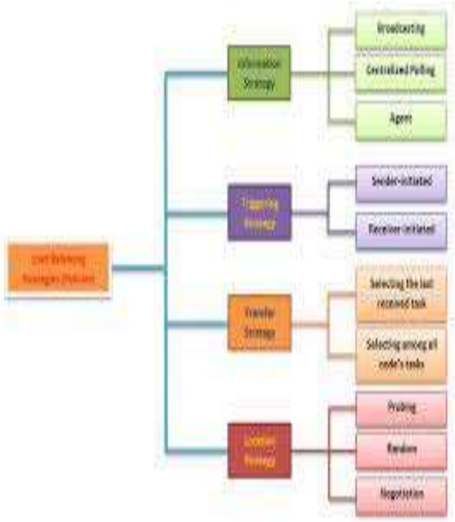
- Distributed approach.
- Mixed approach.

As we discussed earlier, this classification is based on the fact that who in charge of the load distributing is. The second category is designed as:

- Static Approach
- Dynamic Approach.
- Adaptive Approach.

Dynamic load balancing approaches are more suitable in large scale distributed systems like grid and cloud computing. For making decision based on the current workload of system, dynamic load balancers needs to know some information about the state of system. Therefore a dynamic load balancing mechanism requires some components for gathering and handling these types of information. Four main components of a dynamic load balancer have been discussed in detail in and we just introduce them here:

- **Information Strategy Component**: A Component of dynamic load balancing is called Information strategy component which is in charge of collecting information about status of resources in system
- **Triggering Strategy Component**: A component of dynamic load balancer which determines the appropriate time for starting a load balancing operation is called triggering strategy component
- **Transfer Strategy Component** : A component which in charge of selecting a task for transferring to another resource is called triggering strategy component.
- **Location Strategy Component**: it selects a destination resource and a component node for transferring a task. There are many approaches for selecting a destination such as: Probing, Random and Negotiation.



(Figure 2)

**Cloud Load Balancer Challenges and Considerations**

However cloud computing have many advances in theory and practice, researches in cloud are still in their early steps and one of the unsolved challenges is load balancing problem. Before reviewing the current load

balancing approaches, we discuss the main challenges and consideration while designing cloud load balancing algorithms which are affected by cloud characteristics. The challenges are in the following points A:Geographically Distributed Cloud Node A load balancing approach in cloud computing environment should take into account the graphical distribution of computing nodes for having an efficient performance . Some balancing algorithm are designed for closely located nodes and therefore they have no assumption about factors such as communication delay, network bandwidth in LANs and WANs.

B: Virtual Machine Migration:

By using virtualization technique in cloud computing, a physical server can contain several virtual machines that work as independent computing nodes. A common way to unload an overloaded physical server is virtual machine migration among load balancing approaches.

C: Heterogeneous nodes and self regulating:

In these types of solutions a combination of rules is needed for determining the conditions for load balancer. As the number of requests and active resources increase in size and complexity in cloud environment, the handling of these rules is not an easy task. In addition, these rules are static inherently and usually there is no provisioning for refinement or analysis of them Consequence of this, a load balancing mechanism should be able to self-regulates the system's workload among all cloud computing recourses.

D: Storage and Replication Management:

cloud's computing resources according to the tasks' or clients' requirements.

Development of cloud technology has solved the problem of traditional data storage methods which require high cost of hardware and personnel management through a resource-integrated heterogeneous system that can provide the best storage, the optimal performance and the load balancing.

E: Load Balancer Complexity:

In cloud computing systems that load balancing algorithm has high complexity and requires more information, the load balancing and system's performance will be impressed by higher communication cost and delays and it is not good condition for system's efficiency. As a result of this, we can say that a suitable cloud load balancer should be designed in the simplest possible form.

**Surveying of Existing Cloud Load Balancing Approaches**

In this section, we present and discuss some solutions and contributions that have been proposed in the scientific journals and conferences for cloud computing

Load balancing. We classify the load balancing approaches into three categories based on their perspective General Algorithm-based approaches Architectural-

- General Algorithm Based approaches
- Artificial Intelligence based Approach

**A: General Algorithm Based Approach in Cloud Computing :**

The General Algorithm-based load balancing abbreviated form in this paper are these mechanism which propose a complete algorithm and consider all parts of load balancing algorithm within. In this Category, a load balancing mechanism can be implemented based on proposed algorithm. In this section we will give a summary of the current proposed algorithms which are belongs to this category and represent a good comparison of their performance with one another. We will consider and list all them briefly. Rich Lee at el. proposed some most known GAL- based load balancing mechanism

Round-Robin: One of the most known and the simplest algorithm for dispatching workloads to servers is round-robin that usually have good performance in systems with low workload. Weighted round-robin: This algorithm which is derived from original round-robin has better performance compared to traditional round robin because it assigns higher weights to servers with higher performance. Therefore, the more capable computing nodes will get more incoming workloads. Least-connection: This algorithm counts the connections associated with each server dynamically and then based on that number it chooses the least count server and assigns new incoming workloads to the server with least connection. Weighted least-connection: This algorithm counts the associated connections of server too, but associated new incoming workloads based on a factor that calculated by multiplying sever weight by its connection number. Shortest expected delay: In this algorithm, the last response time for each server is considered and then the server with the least response time is selected as the next appropriate server proposed a new load proposed a new load balancing algorithm for better distribution of load and further enhancing the QoS. In this paper, a trust model has been presented which is based on current trust models and it uses initialization time, Machine Instruction Per Second (MIPS) and fault rate parameters to calculate trust value for each data center. In this algorithm, users and data centers are categorized into two groups: trusted and untrusted groups Osmoon Sarood stated that the visualization have some negative effects on HPC application. In this paper, a load balancing algorithm is proposed to achieve load balancing for tightly coupled parallel and HPC application execution in visualized and cloud computing environment. a load balancing Brototi Mandal proposed a load balancing algorithm that it was soft computing based. Stochastic hill climbing is a variant of hill climbing algorithm that is an incomplete approach for solving optimization problems. Because the represented load balancing algorithm is a centralized algorithm and therefore dealing with bottleneck problem, the solving optimization problem have taken into consideration for an efficient distribution of system
Workload.

**B: Architectural Based Load Balancing Approach in Cloud Computing**

Arch-based approaches are those mechanisms that focus on certain architecture and for achieving a load balancing, propose cloud architectures. In this category, algorithms usually not proposed explicitly and load balancing mechanism is represented through architecture components and the relations among them. Generally speaking, this solution usually is an Architectural-based solution and for catching proposed goals, a special cloud computing architecture should be taken into consideration. Here we survey some of these approaches which are proposed in our research literature proposed an architectural -based solution for load balancing in cloud computing. L3B is a centralized Arch-based load balancer which is placed between users and cloud heterogeneous nodes in the cloud infrastructure layer. L3B improves the overall cloud performance, reduces power consumption and customers cost through a mechanism that if the incoming workload exceeds beyond a certain threshold, a suitable VM instance will be initiated. In the other hand, if the workload decreases in compression with a certain threshold an active VM will be switched-off automatically. The proposed architecture consists of two main components: Resource Management Module (RMM) and Packet Management Module (PMM). In this architecture RMM manages the cloud resources and regularly gathers information of the active resource utilization and takes into consideration the customer requirement.

**Gaochao** Proposed a model for balancing workload in the cloud computing environment. Their solution is represented in an Arch-based which is presented through special cloud architecture. A public cloud is divided into some partitions. Partitioning helps to achieve better load balancing in the large-scale and complex environments. In the proposed solution, cloud has a main controller that communicate with the partitions balancers frequently to refresh the status information and each partition uses a local balancer that selects the best strategy for load balancing.

**Weanhong Tian** proposed a reference architecture for load balancing and scheduling algorithm. In this paper an algorithm which is called DAIRS is presented which balances the workload in cloud computing data centers by using of three parameters: CPU, memory and network bandwidth. The proposed load balancing algorithm is considered for both physical and virtual machines. In the algorithm some parameters such as, average CPU utilization, integrated load imbalance and integrated load balance are used. In addition, four queues such as: waiting queue, requesting queue, optimizing queue and deleting queue are used in DAIRS flowchart. **Jian Liane Chen** stated a problem that when a large number of users are accessing the cloud computing services, the computing capability of VM decreases. For solving the problem an optimal load balancing solution which is called EUQoS system for scheduling VMs is proposed. For realizing the EUQoS system, two cloud open platforms: Eucalypyus and Hadoop are used. The load balancer module in the proposed architecture consists of two components: load balancer and agent-based monitor. The load balancer module provides three mechanisms: balance triggering, EUQoS scheduling and VM control. The distribution of workload system is done by weighted round-robin load balancing algorithm

### C: Artificial Intelligence-based Load Balancing Approaches in Cloud Computing

AI-based load balancing mechanisms are the load balancing solutions that their main ideas are based on the Artificial Intelligence concepts. In the other word, AI-based approaches propose a solution for balancing the workloads in cloud computing environments through known artificial intelligence algorithms and methods by finding some similarity between them and cloud computing components and concepts. AI-based load balancing approaches can be proposed in an Arch-based form. Here we will introduce and consider some current AI-based load balancing mechanisms for cloud computing environments.

**Kamar Nishant** proposed version of Ant Colony Optimization (ACO). ACO is applied for cloud computing load balancing to help system for a proper functioning even at the peak usage hours. This AI-based solution has a head node which generates ants. Ants traverse the width and length of the cloud network in the way that they know about the location of under-loaded or over-loaded nodes. These ants along with their movements update a pheromone table for keeping the resource utilization information. In addition, movements of ants are proposed in two ways like the original ACO:

- Forward movement
- Backward movement

Based on the above movements and by using the pheromone table, the loads will be transferred from over-loaded nodes to under-loaded ones. **Seyad Mohssen Ghafari** represented a new power aware load balancing mechanism, named Bee-MMT for decreasing power consumption in cloud computing environments. This approach uses the artificial bee colony algorithm with the feature of minimal migration time. For finding over-loaded hosts the BCA is used. Then through MMT-VM selection, some VMs will be selected for migrating to a new host. In addition by using this method, after finding some under-loaded hosts, if it is possible, algorithm tries to migrate all VMs which allocated to these hosts to the other hosts while k eep them not over-loaded and when all the migration have been completed, switch host to the sleep mode. Proposed algorithm includes four phases:

1) Host overloading Detection
2) 2) VM selecting policy.
1) Host over-loading detection
2) VM selecting
3) Host under-loading detection
4) SLA

## Conclusion and Future Works

Load balancing in cloud computing data centers has been a main challenge and an active area of research in recent years. In this paper we have presented a survey on current load balancing techniques and solutions which have proposed only for cloud computing environments. According to our study, cloud load balancing mechanism can be categorized into three main groups based on their designing perspectives: General Algorithm-based, Architectural-based and Artificial Intelligence-based load balancing approaches. In addition, our survey on the current proposed cloud load balancing approaches has some contributions which we list the main ones here:

1) A new Classification of cloud load balancing in cloud computing environment based on the designing perspective has been proposed in this paper.

2) According to our study results the best load blanking approaches in cloud computing are those ones which have dynamic, distributed and non-cooperative features. Dynamic feature is

referred to dynamic and on demand property in cloud which is resulted in dynamic workload and services. Distributed designing can prevent the bottleneck and a single point of failure and therefore provides the better scalability and fault tolerant abilities. Finally non-cooperative approach can avoid unnecessary overheads while a load balancer is performing.

3) There are many architectural and designing time considerations for implementing a cloud load balancer that should be taken into account such as: virtual machine migration, elasticity, cost and energy consumption management which we proposed a complete discussion in section 3.

4) The result of this study in table 2 shows that the most of the proposed load balancing solutions are suffering from the bottleneck and single point of failure problems in the first place.

5) The recent cloud load balancing solutions are focusing on Green cloud topic in load balancing mechanisms by considering the challenges like: Reducing the energy and power consumption, Reducing carbon emission and also reducing cost customers as a result of energy efficiency topic.

As the future work of this paper, we will consider more cloud load balancing solutions according to our three levels classification and survey the load balancing solutions' trend.

## References

*Subashini, S. and V. Kavitha, A survey on security issues*
*in service delivery models of cloud computing. Journal of Network and Computer Applications, 2011. 34(1): p. 1-11.*

*Khan, A.N., et al., Towards secure mobile cloud computing: A survey. Future Generation Computer Systems, 2012.*

*1Lombardi, F. and R. Di Pietro, Secure virtualization for cloud computing. Journal of Network and Computer Applications, 2011. 34(4): p. 1113-1122.*

*Khorshed, M.T., A. Ali, and S.A. Wasimi, A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing. Future Generation Computer Systems, 2012. 28(6): p. 833-851.*

*Mauch, V., M. Kunze, and M. Hillenbrand, High performance cloud computing. Future Generation Computer Systems, 2012.*

*Ghosh, R., et al., Modeling and performance analysis of large scale IaaS Clouds. Future Generation Computer Systems, 2012.*

*Garg, S.K., S. Versteeg, and R. Buyya, A framework for ranking of cloud computing services. Future Generation Computer Systems, 2013. 29(4): p. 1012-1023.*

*Pérez-Miguel, C., A. Mendiburu, and J. Miguel-Alonso, Modeling the availability of Cassandra. Journal of Parallel and Distributed Computing, 2015.*

*Sousa, E., et al., A Modeling Approach for Cloud frastructure Planning Considering Dependability and Cost Requirements. IEEE Transactions on Systems, Man, and Cybernetics: Systems, , 2015. 45(4): p. 549-558.*

*Yang, X., et al., A business-oriented Cloud federation model for real-time applications. Future Generation Computer Systems, 2012. 28(8): p. 1158-1167.*

*Dukaric, R. and M.B. Juric, Towards a unified taxonomy and architecture of cloud frameworks. Future Generation Computer Systems, 2012. Fu, S., C.-Z. Xu, and H. Shen, Randomized load balancing strategies with churn resilience in peer-to-peer networks. Journal of Network and Computer Applications, 2011. 34(1): p. 252-261.*

*Wu, D., Y. Tian, and K.-W. Ng, Resilient and efficient load balancing in distributed hash tables. Journal of Network and Computer Applications, 2009. 32(1): p. 45-60.*