# Load Balancing in Cloud Computing:Challenges & Issues

Ms. Shalini Joshi
Computer Science & Engineering,
Mody University of Science and Technology,
Lakshmangarh, India
shalini123.joshi@gmail.com

Dr. Uma Kumari
Computer Science & Engineering,
Mody University of Science and Technology,
Lakshmangarh, India
umasecd@gmail.com

***Abstract*— Cloud computing is latest emerging technology for large scale distributed computing and parallel computing. Cloud computing gives large pool of shared resources, software packages, information, storage and many different applications as per user demands at any instance of time. Cloud computing is emerging quickly; a large number of users are attracted towards cloud services for more satisfaction. Balancing the load has become more interesting research area in this field. Better load balancing algorithm in cloud system increases the performance and resources utilization by dynamically distributing work load among various nodes in the system. This paper presents cloud computing, cloud computing architecture, virtualization, load balancing, challenges and various currently available load balancing algorithms.**

***Keywords—Cloud computing; architecture; virtualization; load balancing***

## I. INTRODUCTION

A cloud introduces an IT environment which is invented for the motive of remotely provisioning measured and scalable resources [1]. Word "Cloud" in cloud computing is also known as "Internet". So cloud computing is called as internet based computing in which many different services such as server, storage, virtualization and various application are given to the users and organization over the internet[2]. Cloud computing uses the term "pay-per-usage" instead of traditional computing in which "own and use" technique is used. There are several issues in cloud computing paradigm but balancing the load is major issue (challenge) in cloud computing environment. Load balancing is a methodology which provides methods to maximize throughput, utilization of resources and performance of system [3]. As a part of its services, it gives easy and flexible process to keep data or files and make them available for large scale of users [4]. To make the use of resources most efficiently in cloud system, there are several load balancing algorithms.

## II. CLOUD COMPUTING

Cloud computing is an advanced technique which provides various online computing resources as well as storage [3]. Users that located anywhere in the world can access these resources on demand basis through the internet [5]. In cloud computing, the key analyses involve efficiently allocating tasks to several nodes in cloud system so that the request processing and effort is performed in a well-organized manner [3]. Cloud computing permits a large number of users to access virtualized, scalable, distributed hardware and software resources via the internet [6].

## III. CLOUD COMPUTING ARCHITECTURE

Cloud computing is fastest implementation technology. Amazon, Google, Microsoft all are using cloud computing and working towards providing powerful, reliable and logical platforms to its users[7]. A cloud computing architecture consists of three service models, five key characteristics and four cloud computing deployment models that are shown in following Fig. [8].

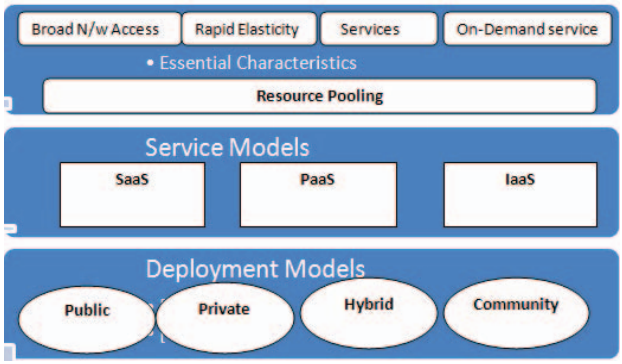


Fig 1 Cloud computing architecture

Fig. 1 illustrate the deployment models of cloud computing. It also defines the service models of cloud computing and its essential characteristics[1].

### *A. Cloud computing deployment models*

Cloud computing is an internet based service that allocates services to the users on demand at any instance of time. There are several types of clouds that provide computing services. These are:

- Public Cloud – The cloud infrastructure that is openly utilized by general public and always available in a "pay per usage" manner [9].

Examples are Amazon or Google clouds which are always open for all users [10].

- Private Cloud – The cloud infrastructure is made by individual organization for its exclusive use. The entire management is done by the organization in the area of cloud computing [11]. It is not open for all users and it will accessible within an organization [10].
- Hybrid Cloud – It is the combination of public cloud and private cloud. This type of cloud is influenced for commercial usage [10].
- Community Cloud – This cloud infrastructure is used by a particular community of users from organizations, that have common goals. An example of community Cloud is: a group of Universities that uses one cloud [11]. It is a hybrid form of private cloud.

### B. *Cloud Computing Service Models*

Cloud computing provides various types of services to its users:

- Infrastructure as a Service (IaaS) – It consists services that permits its consumers to request storage and computational resources on demand. And also enabling the so called "pay-per-use" paradigm. An example of IaaS is Amazon EC2 [12].
- Platform as a Service (PaaS) – It contains high levels of services that provides a platform to develop and manage the software infrastructure. In PaaS developer can built and deploy different types of applications using libraries, languages and tools handled by the cloud service providers. Google App Engine is an example of PaaS [12].
- Software as a Service (SaaS) – SaaS comprises end users applications that are delivered to consumers as a network services. So, this eliminates the need to install and run different applications on consumer's computers. An example of SaaS is SalesForce.com and Google mail [12].

## IV. Cloud Virtualization

In cloud computing, virtualization is a very important. Virtualization, as the name explains, is not real but provides all the facilities that actually exist. Virtualization is software implementation of different computers on which a number of distinct programs can be executed as a real machine. In other words, Amazon EC2 in which IT infrastructure is deployed in a cloud and provider's data centers in the structure of virtual machines. Virtual infrastructure management methods and tools for datacenters that have been around since before cloud computing became the industry's latest emerging buzzword [13]. A wide range of users can access multiple services of cloud computing. So all these services are given to many users by remote data centers that are based on virtualization [1]. Virtualization is divided into two parts that is as follows:

- Full Virtualization – In full virtualization, all the software that are available in actual server are also present in virtual system and this happens just because the complete(full) installation of one system has to be done on other system. Because of this virtualization, computer system shares among many users and emulates hardware located on multiple systems [1].
- Para Virtualization – In this kind of virtualization, by using system resources such as memory and the processor that allows multiple operating systems to operate on a single system. Partial services are given by this virtualization but complete services are not fully accessible. Migration, disaster recovery and capacity management are key features of this virtualization [1].

## V. Load Balancing

Load Balancing is a mechanism which distributes the workload on the resources of a node to respective resources on the other node in a network without eliminating any of the running task [10]. So balancing the load between various nodes of the cloud system became a main challenge in cloud computing environment. The load can be any type like network load, memory load, CPU load and delay load etc. Thus it is very important to share work load across multiple nodes of system for better performance and increasing resources utilization. Major goals of load balancing [14] are

- Establish fault tolerance system
- Maintain system stability
- Improve the performance and efficiency
- Minimizing the job execution time and waiting time in queue.
- To increase user satisfaction [14]
- To improve the resource utilization ratio.

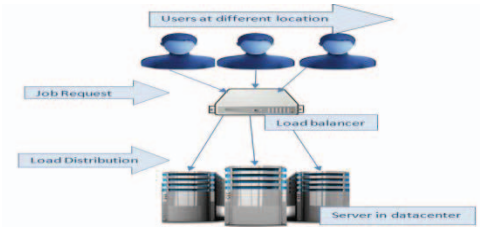


Fig. 2 Load Balancing

To balance the entire load, load balancing algorithm has been designed and two types of load balancing algorithms (environment) are introduced:

- Static Load Balancing – This algorithm requires prior knowledge about system resources. Therefore, the decision of load distribution does not depend on the current (present) state of the system [1]. In this environment performance of processors is explained at the starting of the execution and it does not change

the executing process at run time for making changes in the system load [15]. This algorithm is suitable for homogenous system environment [16].

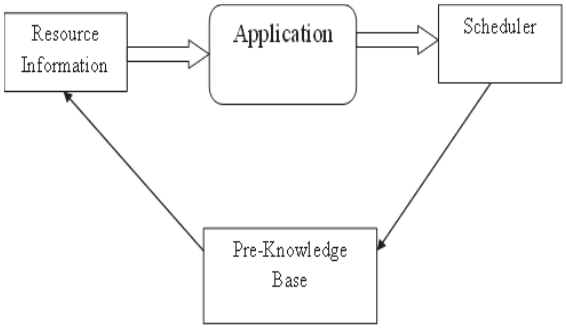


Fig. 3 Static Load Balancing

- Dynamic Load Balancing – This algorithm does not require any prior information about system resources because the load distribution decision is based on the current state of the system [1]. This is suitable for heterogeneous system. Dynamic load balancing make changes in load at run time [16]. This algorithm provides outstanding improvement in performance than static algorithm [15].

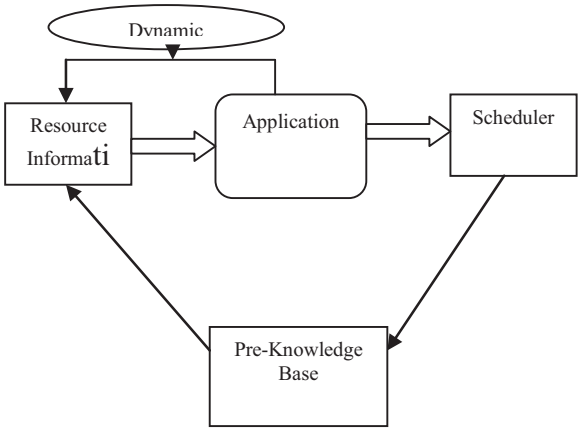


Fig. 4 Dynamic Load Balancing

## VI. CHALLENGES & ISSUES OF LOAD BALANCING

Even through cloud computing is widely used nowadays but still the research is in its starting stage [7]. So before explaining the load balancing algorithm for cloud computing it is required to identify some key challenges and issues that affect the performance of load balancing algorithms [17]. The Automated service provisioning – Elasticity is a major component in cloud computing due to which allocating and releasing of the resources take place as default. The challenge with using optimum resource is how cloud elasticity can be used and how work with traditional systems performance can be done simultaneously [7]?

- Virtual machine migration – The idea is to imagine a machine as a set of files or a file. It is possible to decrease the load on over loaded machine by moving the virtual machine among them in effective way. The main motto is to distribute the all type of load in a datacenter. The challenge is to remove and avoid drawbacks of cloud computing system when the load is dynamically distributed by virtual machine.
- Energy management – Energy Management is also a major point that permits users to use the resources from global center. This provides economy of scale and major advantage to cloud computing but a question emerges that by using just a portion of datacenter, how to meet better performance.
- Stored data management – Another key requirement is the storage of data. So how can data be distributed in the cloud system with most appropriate storage and fast access?
- Emergence of small different datacenters for cloud computing – A small size of data center can be more beneficial just because it will consume less electricity and cheaper than large one. And load balancing is showing as a global scale issue for certifying proper response time with optimal resource utilization and distribution.
- Spatial distribution of the cloud nodes – Some algorithms are proposed just for nearly located nodes in which communication delays are insignificant [17]. But still it is a issue to design an efficient load balancing algorithm which can be formulated properly for spatially distributed nodes[5].
- Storage and Replication – A full replication algorithm is not much beneficial for efficient storage utilization in a system. This is just because the same amount of data will be kept in all replicated nodes. Full replication algorithms cause unreasonable costs with requirements of large storage.
- Algorithm complexity – Load balancing algorithm is preferred to be less complicated in terms of operations and execution(implementation). A negative implementation complexity will lead to a extra complex process. Furthermore, for monitoring and controlling the implementation, algorithms require higher communication, more information and delays may cause more bottlenecks and then efficiency discards[17].
- Point of failure controlling – Some algorithms (centralized algorithms) provide effective mechanisms for processing load balancing in a particular pattern. But the issue is that there is only one controller for the entire system. In such condition, if the controller fails, then the entire system fails[17].

## VII. AVAILABLE LOAD BALANCING ALGORITHMS

For an efficient throughput, to reduce response time and avoiding the overload on a particular resource are achieved by load balancing algorithms (static and dynamic) [18]. These algorithms are compared in tabular form with their advantages and disadvantages that are given below:

TABLE I
COMPARISON OF CURRENT EXISTING LOAD BALANCING ALGORITHM.

| Algorithms | Static/Dynamic | Explanation | Pros | Cons |
|---|---|---|---|---|
| Round Robin and Randomized [6] | Static | 1. Processes are divided equally between all processors.<br>2. The process allocation order is maintained locally on each processor.<br>3. In this way user request are processed in circular way by using this algorithm. | 1. Works well with number of processes that is larger than number of processor.<br>2. Round Robin does not demand for inter-process communication. | In round robin, there are no expectations to obtain better performance. |
| Central Manager [7] | Static | 1. Central processor is responsible for selecting the host for all new process.<br>2. Minimum loaded processor depends on the entire load that is selected when process is established. | Load scheduler makes load balancing decisions that depends on the system load information. | High degree of inter-process communication which leads to bottleneck state. |
| Threshold [7] | Static | 1. Processes are allocated without delays upon creation to hosts.<br>2. Hosts for new processes are assigned regionally without sending remote messages. | 1. Threshold has low inter-process communication<br>2. A number of local process allocations are done. | When all remote processes are overloaded, then all processes are allocated locally. |
| Min-Min [19] | Static | 1. Minimum completion time is searched for all tasks.<br>2. From minimum times, the minimum value is taken which is minimum time among all the tasks on any type of resources in the system.<br>3. According to that minimum time, the task is assigned to the corresponding machine in the system. | Performs well with optimal number of resources. | It can take the condition of starvation. |
| Max-Min [19] | Static | Max-Min is approximately similar to the min-min algorithm but one thing is different that is as follows: After getting minimum execution times, the max value is chosen that is the maximum time among all the tasks on different type of resources. | Performs well with optimal number of resources. | It can also lead to starvation |

| | | | | |
|---|---|---|---|---|
| Honey Bee Foraging Behavior [20] | Dynamic | Using local server actions, it attains global load balancing. | Performing well as system diversity increases. | As the system size increasing, it does not increase throughput. |
| Biased Random sampling [20] | Dynamic | Achieve Load balancing among all system nodes utilizing random sampling of the system domain. | It is performing excellent when high and similar population of resources is provided. | Degrades as population diversity increases. |
| Active Clustering [20] | Dynamic | Optimizes job allocation by linking same services by local re-wiring. | 1. Performs good with high utilized resources.<br>2. Utilizing the increased system resources to increasing throughput. | Reduces as system diversity increases. |
| ACCLB(Ant Colony and Complex network Theory) [7] | Dynamic | It uses small-world and scale-free characteristics of complex network for achieving good load balancing. | 1. Control heterogeneity.<br>2. It is adaptive to dynamic environments.<br>3. Outstanding in case of fault tolerance.<br>4. Good Scalability Compare and Balance. | 1. Only utilized in Complex networks.<br>2. Low scalability and Performance. |
| Compare and Balance [7] | Dynamic | 1. Based on sampling.<br>2. It employs adaptive live migration of virtual machines (VMs). | 1. It can balance load among servers.<br>2. Assures about migration of VMs from upper-cost physical hosts to lower-cost host of the system. | Depending on assuming enough memory with each physical host. |
| Vector Dot [7] | Dynamic | Uses dot product to differentiate node relied on the item requirement. | 1. Monitor multi-dimension and hierarchical resource constraints.<br>2. Eliminates overloads on servers, switch and storage. | Only for complex networks. |
| GA Algorithm[3] | Dynamic | 1. Genetic algorithm is a search heuristic, deployed on evolutionary algorithm with natural selection.<br>2. GA mainly focuses on these three steps - selection, crossover, and mutation. | 1. Uses centralized balancing approach.<br>2. This algorithm gives efficient schedules. | It is not based on distributed balancing approach. |

## CONCLUSION

Cloud computing is a new paradigm in which different resources are access by multiple users over the internet in on demand basis. These resources are rapidly developing and also increasing uses of heterogeneous system in dynamic environment. But there are several research challenges in cloud computing.

Load balancing is major challenge (issue) in cloud computing. The key aim of load balancing is to satisfy user's needs by distributing work load among multiple nodes in system and maximize resource utilization and improves system performance. So efficient load balancing is vital for system performance, resource utilization, stability, maximizes the throughput and minimizes the response time that are the main objectives of this paper. To balance the load among multiple nodes in system, there are several load balancing algorithms could be introduced. This paper presents the overview of cloud computing, cloud computing architecture, virtualization, load balancing and some challenges related to balancing load in cloud computing.

## REFERENCES


[1] Desai, Tushar, and J. Prajapati. "A survey of various load balancing techniques and challenges in cloud computing." International Journal of Scientific & Technology Research 2.11 (2013): 158-161.

[2] L. Wang, J. Tao, M. Kunze,"Scientific Cloud Computing: Early Definition and Experience", the 10th IEEE International Conference Computing and Communications 2008.

[3] Rana, Madhurima, S. Bilgaiyan, and U. Kar. "A study on load balancing in cloud computing environment using evolutionary and swarm based algorithms." Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on. IEEE, 2014.

[4] Foster, Ian, et al. "Cloud Computing and Grid Computing 360-Degree Compared," IEEE Grid Computing Environments (GCE08) 2008, co-located with IEEE/ACM Supercomputing 2008." 2012 ACM/IEEE 13th International Conference on Grid Computing.

[5] Buyya, Rajkumar, R. Ranjan, and R. N. Calheiros. "Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services." International Conference on Algorithms and Architectures for Parallel Processing. Springer Berlin Heidelberg, 2010.

[6] Ray, Soumya, and A. D. Sarkar. "Execution analysis of load balancing algorithms in cloud computing environment." International Journal on Cloud Computing: Services and Architecture (IJCCSA) 2.5 (2012): 1-13.

[7] Katoch, Swati, and J. Thakur. "Load Balancing Algoritms in Cloud Computing Environment: A Review", International Journal on Recent and Innovation Trends in Computing and Communication,Vol 2, Aug 2014

[8] El-Gazzar, R. Fahim. "An Overview of Cloud Computing Adoption Challenges in the Norwegian Context." Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on. IEEE, 2014.

[9] Rajput, S. Singh, and V. S. Kushwah. "A Review on Various Load Balancing Algorithms in Cloud Computing." International Journal 6.4 (2016).

[10] More, S. R. Hiray. et al.,"Load balancing and resource monitoring in cloud." Proceedings of the CUBE International Information Technology Conference. ACM, 2012.

[11] Almubaddel, Majed, and Ahmed M. Elmogy. "Cloud Computing Antecedents, Challenges, and Directions." Proceedings of the International Conference on Internet of things and Cloud Computing. ACM, 2016.

[12] Lenk, Alexander. "What's inside the Cloud? An architectural map of the Cloud landscape." Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing. IEEE Computer Society, 2009.

[13] Sotomayor, Borja, et al. "Virtual infrastructure management in private and hybrid clouds." IEEE Internet computing 13.5 (2009): 14-22.

[14] Somani, Rajkumar, and J. Ojha. "A Hybrid Approach for VM Load Balancing in Cloud Using CloudSim." International Journal of Science, Engineering and Technology Research (IJSETR) 3.6 (2014): 1734-1739.

[15] Rajguru, Abhijit A., and Mrs Sulabha S. Apte. "Various Strategies of Load Balancing Techniques and Challenges in Distributed Systems."

[16] Raghava, N. S., and Deepti Singh. "Comparative study on load balancing techniques in cloud computing." Open Journal of Mobile Computing and Cloud Computing 1.1 (2014).

[17] Karuna G.Bakde, B .M. Patil , "Survey of techniques and challenges for load balancing in public cloud", International Journal of Technical Research and Applications, e-ISSN: 2320-8163, Volume 4, Issue 2 (March-April, 2016), pp.279-290

[18] Gupta, Ruhi. "Review on existing load balancing techniques of cloud computing." International Journal of Advanced Research in Computer Science and Software Engineering 4.2 (2014): 168-71.

[19] Gopinath, PP Geethu, and Shriram K. Vasudevan. "An in-depth analysis and study of Load balancing techniques in the cloud computing environment."Procedia Computer Science 50 (2015): 427-432.

[20] Randles, Martin, David Lamb, and A. Taleb-Bendiab. "A comparative study into distributed load balancing algorithms for cloud computing." Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on. IEEE, 2010.