ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.4

Enhanced Load Balancing Mechanism in Cloud Computing Environment: A Review

Bindu Bala
Assistant Professor
Department of Computer Application
Shaheed Bhagat Singh State Technical Campus, Ferozepur, India

Abstract:
Cloud computing is a structured model that defines computing services, in which data as well as resources are retrieved from cloud service provider via internet through some well formed web-based tool and application. Cloud computing is the means of accessing a shared pool of configurable computing resources (including hardware, software, networks, servers, storage applications and services) that can be rapidly provided, used and released with minimal effort on the part of users or service providers. But it has some of the main concerns like load management and fault tolerance. In this paper we are discussing load balancing approach in cloud computing. Load balancing is helped to distribute the workload across multiple nodes to ensure that no single node is overloaded. It helps in proper utilization of resources .It also improves the performance of the system This paper focuses on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers. In this paper, we have reviewed several existing load balancing mechanisms and we have tried to address the problems associated with them.

Keywords: Cloud computing; Load balancing, Virtual machine, Host, Datacenter, Datacenter Broker

INTRODUCTION

Cloud is a computing framework which usually denotes storing and accessing data and programs over the internet, instead of your computer's hard drive. Cloud Computing is handled by means of the great prospective paradigm utilized for placement of applications taking place over Internet. It is green technologies which agree to accessing, computing and storing the assets by offering a variety of services. Cloud computing normally includes models like Infrastructure-as-a-service [1], Platform-as-a-service and Platform-as-a-service. To reduce the computation time and to conquer the storage space issues, most of the organization now a day's make regular use of cloud computing from the established process of calculation. It mainly focuses on allocating data and computations over a scalable information centers of network. Cloud computing [1] is an emerging paradigm in the computer industry where the computing is moved to a cloud of computers. It has become one of the buzz words of the industry. The Cloud Computing may be a term that describes the infrastructure, platform, services and different kind of applications. It reconfigure servers or applications where the server can be a physical machine or virtual display machines. Cloud computing is different from ancient computing paradigms because it is a customizable, scalable, encapsulated, abstract entity that gives totally different level of services, processes to the clients, driven by economies of scale and also the services area unit dynamically and totally configurable [2].

Cloud is a term used as a metaphor for the wide area networks (like internet) or any such large networked environment. It came partly from the cloud-like symbol used to represent the complexities of the networks in the schematic diagrams. It represents all the complexities of the network which may include everything from cables, routers, servers, data centers and all such other devices. Computing started off with the mainframe era.

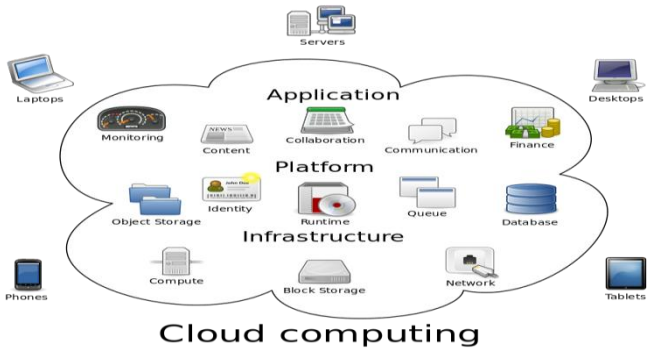

Figure 1. Cloud Computing Model

KEY CHARACTERISTICS

1. **On-demand self-service:** A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

2. **Broad network access:** Cloud computing provide the users with various capabilities over the network which are accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops etc.)

3. **Resource pooling:** The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer

demand. Examples of resources include storage, processing, memory, network bandwidth, and virtual machines

4. Rapid elasticity: Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out, and rapidly released to quickly scale in.

5. Measured Service: Cloud systems automatically control and optimize resource use by leveraging a metering capability1 at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts).

TYPES OF CLOUDS

Clouds are divided into 4 categories: -

1**. Public cloud computing**: It mainly depends on third individual to suggest services by paying them on regular basis according to the procedure. Public Cloud environment is made available to all unrestricted consumers who can subscribe the needed services [3].

2. Private Cloud Computing: The organization itself regulates the services. Usually administrations go for private cloud only in the case of involvement of sensible information. Scaling can be done very professionally by adding hardware and thus the environment can be expanded. The security will be more due to the control of internal structure contained in it and therefore data will be secured.

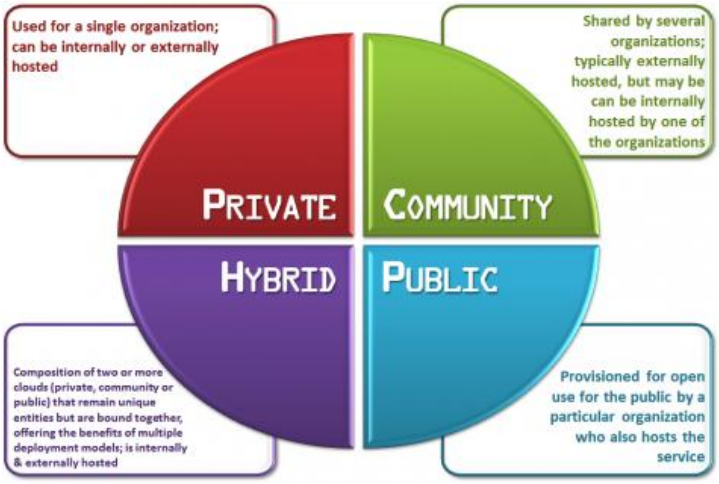

Figure 2. Types of Cloud

3. Hybrid Cloud Computing: It is the mixture of both public & private cloud computing. A less sensible data will be stored in public and all others in Private Cloud.

4) Community Cloud: -When cloud infrastructure construct by many organizations jointly, such cloud model is called as a community cloud. The cloud infrastructure could be hosted by a third-party provider or within one of the organizations in the community

SERVICES OF CLOUD MODEL

There are different types of service models provided by cloud like: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) which are deployed as public cloud, private cloud, community cloud and hybrid clouds.

1 Software As A Service (SaaS): - In SaaS capability provided to the consumer to use some applications which is running on a cloud infrastructure. The applications are accessible from many devices through an interface such as a web browser (e.g., web-based email).

2 Platform As A Service (PaaS): - PaaS provides all the resources that are required for implementation of applications and all services completely from the internet. PaaS provides all the resources that are required for building applications and services completely from the Internet, without downloading or installing software.Examples of PaaS include: AWS Elastic Beanstalk, OpenShift, Google App Engine, Windows Azure Cloud.

3 Infrastructure As A Service (IaaS): - The capability provided to the consumer is to access all the processing, storage, networks and other many fundamental computing resources. The consumer is able to deploy arbitrary software, which can include operating systems and applications. In the most basic cloud-service model, providers of IaaS offer hardware or (more often) virtual machines and other resources.

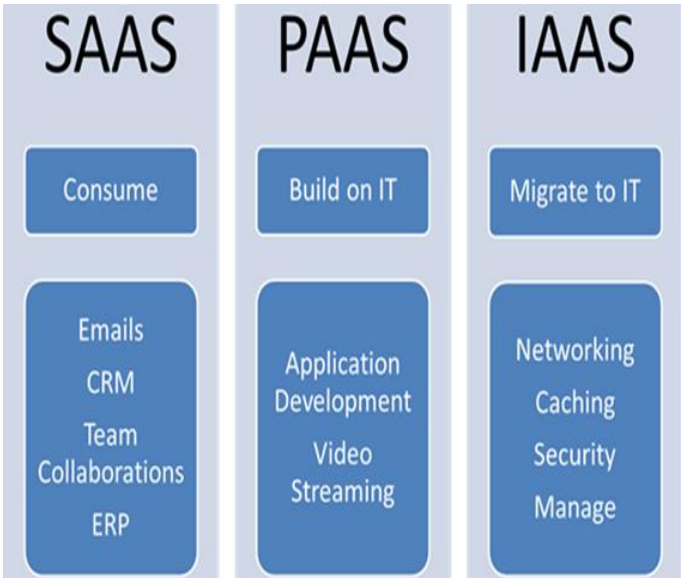

Figure 3. Models of Cloud

LOAD BALANCING

It is the mechanism of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. It also ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. Load Balancing is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server. They have the ability to handle the high speed network traffic whereas Software-based

load balancers run on standard operating systems and standard hardware components. Some load balancers provide a mechanism for doing something special in the event that all backend servers are unavailable. This might include forwarding to a backup load balancer, or displaying a message regarding the outage. Load balancing gives the IT team a chance to achieve a significantly higher fault tolerance. It can automatically provide the amount of capacity needed to respond to any increase or decrease of application traffic. It is also important that the load balancer itself does not become a single point of failure. Usually load balancers are implemented in high-availability pairs which may also replicate session persistence data if required by the specific application.

GOALS OF LOAD BALANCING

The goals of load balancing are:

• To improve the performance of the system.

• To have a backup of the load or entire server just in case the system fails or even partly fails.

• To maintain the system stability

• To accommodate future modification within the system

LOAD BALANCING CLASSIFICATION

This is chiefly divided into 2 categories: static load balancing mechanism and dynamic load balancing mechanism:

1. Static Load Balancing: In the static load balancing algorithm the decision of shifting the load does not depend on the current state of the system. It requires knowledge about the applications and resources of the system., Central Manager Algorithm, Threshold algorithm and randomized algorithm.

2. Dynamic Load Balancing: In this type of load balancing algorithms the current state of the system is used to make any decision for load balancing, thus the shifting of the load is depend on the current state of the system. An important advantage of this approach is that its decision for balancing the load is based on the current state of the system which helps in improving the overall performance of the system by migrating the load dynamically.

A) Centralized approach: - In centralized approach, solely one node is liable for managing and distribution among the complete cloud system model. Alternative all nodes aren't liable for handling the requests and providing the response.

B**) Distributed approach:** - In distributed approach, every node severally builds its own load vector. The work is divided among all the nodes of the server. They aggregate the load information of alternative nodes. Distributed approach is additional appropriate for complicated and very large systems inside the cloud computing.

RELATED WORK

Rajeshwari et al. (2015) has proposed a two stage scheduling algorithm Service Level Agreement (SLA) based scheduling algorithm determines the priority of the tasks and assigns the tasks to the respective cluster. In the second stage, the Idle-Server Monitoring algorithm balances the load among the servers within each cluster. The proposed algorithm is implemented using CloudSim simulator. The effectiveness of proposed model is evaluated under different loads and its results are compared with other two existing algorithms such as Throttled load balancing algorithm and Round Robin algorithm[2].

Saraswathi, Y.RA Kalaashri, Dr.S.Padmavath et al. (2015) This paper mainly focuses on allocation of VM to the user, based on analyzing the characteristics of the job. Main principle of this work is that low priority jobs (deadline of the job is high) should not delay the execution of high priority jobs (deadline of the job is low) and to dynamically allocate VM resources for a user job within deadline. The proposed algorithm suspends a low priority job and runs a high priority job in the VM from which low priority job was suspended. Resume the suspended job if any of the VM in which job have been completely executed. The method has less overhead in executing all jobs, when compared with creation of new VM [57]

Yi-Ju Chiang et al. (2014) discussed that cloud computing is a new service model for sharing a pool of computing resources that can be rapidly accessed and released based on a converged infrastructure. In the past, an individual or company can only use their own servers to manage application programs or store data. Thus it may cause the dilemma of complex management and burden in "own-and-use" patterns. To satisfy uncertain workloads and to be highly available for users anywhere at any time, providing more resources are required. Consequently, resource over provisioning and redundancy are common situations in a traditional operating system. However, most electricity dependent facilities will inevitably suffer from idle times or under-utilized for some days or months since there usually have off-seasons caused by the nature of random arrivals [5].

S.Yakhchi et al. (2015) discusses that the energy consumption has become a major challenge in cloud computing infrastructures. They proposed a novel power aware load balancing method, named ICAMMT to manage power consumption in cloud computing data centers. We have exploited the Imperialism Competitive Algorithm (ICA) for detecting over utilized hosts and then we migrate one or several virtual machines of these hosts to the other hosts to decrease their utilization. Finally, we consider other hosts as underutilized host and if it is possible, migrate all of their VMs to the other hosts and switch them to the sleep mode[6].

Surbhi Kapoor et al. (2015) aims at achieving high user satisfaction by minimizing response time of the tasks and improving resource utilization through even and fair allocation of cloud resources. The traditional Throttled load balancing algorithm is a good approach for load balancing in cloud computing as it distributes the incoming jobs evenly among the VMs. But the major drawback is that this algorithm works well for environments with homogeneous VMS, does not considers the resource specific demands of the tasks and has additional overhead of scanning the entire list of VMs every time a task comes. The issues have been addressed by proposing an algorithm Cluster based load balancing which

works well in heterogeneous nodes environment, considers resource specific demands of the tasks and reduces scanning overhead by dividing the machines into clusters[7].

Shikha Garg et al. (2015) aims to distribute workload among multiple cloud systems or nodes to get better resource utilization. It is the prominent means to achieve efficient resource sharing and utilization. Load balancing has become a challenge issue now in cloud computing systems. To meets the user's huge number of demands, there is a need of distributed solution because practically it is not always possible or cost efficient to handle one or more idle services. Servers cannot be assigned to particular clients individually. Cloud Computing comprises of a large network and components that are present throughout a wide area. Hence, there is a need of load balancing on its different servers or virtual machines. They have proposed an algorithm that focuses on load balancing to reduce the situation of overload or under load on virtual machines that leads to improve the performance of cloud substantially[8].

R. Panwa et al. (2015) focuses on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers. The proposed algorithm in this research work has been implemented using Cloud Analyst simulator and the performance of the proposed algorithm is compared with the three algorithms which are preexists on the basis of response time. In the cloud computing milieu, the cloud data centers and the users of the cloud-computing are globally situated, therefore it is a big challenge for cloud data centers to efficiently handle the requests which are coming from millions of users and service them in an efficient manner[9].

Ankit Kumar et al (2016) focuses on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers. The proposed algorithm in this research work has been implemented using Cloud Analyst simulator and the performance of the proposed algorithm is compared with the three algorithms which are preexists on the basis of response time. In the cloud computing milieu, the cloud data centers and the users of the cloud-computing are globally situated, therefore it is a big challenge for cloud data centers to efficiently handle the requests which are coming from millions of users and service them in an efficient manner. [21]

RESEARCH GAP

Cloud computing thus involving distributed technologies to satisfy a variety of applications and user needs. Sharing resources, software, information via internet are the main functions of cloud computing with an objective to reduced capital and operational cost, better performance in terms of response time and data processing time, maintain the system stability and to accommodate future modification in the system .So there are various technical challenges that needs to be addressed like Virtual machine migration, server consolidation, fault tolerance, high availability and scalability but central issue is the load balancing , it is the mechanism of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. It also ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. In the NBST algorithm, the jobs are present in the queue and we know the length i.e., number of instructions in request. The load balancing algorithm aims at reducing the load over resources. For achieving this, arrange all the virtual machines in order according to their execution speed that is in MIPS (Million instructions per second). After arrangement of machines, sorting of cloudlets is performed on the basis of their length (million instructions). Mid-point is taken of those sorted cloudlets list and sorted virtual machines list and then the divided cloudlet lists are mapped to the corresponding lists of virtual machines.

To make the final determination, the load balancer retrieves information about the candidate server's health and current workload in order to verify its ability to respond to that request. Load balancing solutions can be divided into software-based load balancers and hardware-based load balancers. Hardware-based load balancers are specialized boxes that include Application Specific Integrated Circuits (ASICs) customized for a specific use. They have the ability to handle the high-speed network traffic whereas Software-based load balancers run on standard operating systems and standard hardware components.The currently proposed work will work effectively and efficiently only in the homogeneous environment where all the machines are of same capacity. But as we know that in the cloud computing model, the configurations of the machines will be different from each other as the different users will have different requirements. The tasks of the user are allocated on the basis of availability of virtual machine. There is no checking of capacity of the virtual machine before allocating the request. There can be a scenario where a machine with high configuration is sitting idle and we have assigned the task to the low configuration machine. This may lead to overutilization and under-utilization of resources. There is no checking of the requirement of the user whether user wants to use the machine of high configuration or low configuration. No fault tolerance mechanism has been proposed in the current work.

CLOUD SIM

Cloud service providers charge users depending upon the space or service provided. In R&D, it is not always possible to have the actual cloud infrastructure for performing experiments. For any research scholar, academician or scientist, it is not feasible to hire cloud services every time and then execute their algorithms or implementations. For the purpose of research, development and testing, open source libraries are available, which give the feel of cloud services. Nowadays, in the research market, cloud simulators are widely used by research scholars and practitioners, without the need to pay any amount to a cloud service provider.

CONCLUSION

In present days cloud computing is one of the greatest platform which provides storage of data in very lower cost and available for all time over the internet. But it has more critical issue like security, load management and fault tolerance. In this paper we are discussing load balancing approaches. Resource scheduling management design on Cloud computing

is an important problem. Scheduling model, cost, quality of service, time, and conditions of the request for access to services are factors to be focused. A good task scheduler should adapt its scheduling strategy to the changing environment and load balancing Cloud task scheduling policy. Cloud Computing is high utility software having the ability to change the IT software industry and making the software even more attractive.

REFERENCES

[1] S. Yakhchi, S. Ghafari, M. Yakhchi, M. Fazeli and A. Patooghy, "ICA-MMT: A Load Balancing Method in Cloud Computing Environment," IEEE, 2015.

[2] B .S Rajeshwari and M.Dakshayini "Optimized Service Level Agreement Based Workload Balancing Strategy for Cloud Environment" International Advance Computing Conference (IACC), 2015 IEEE

[3] **S**araswathi , Kalaashri and Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing" ,Procedia Computer Science ,Volume 47, 2015,pp.30-36

[4] R. Panwar and D. B. Mallick, "Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm," IEEE, pp. 773-778, 2015.

[5] Y.-J. Chiang,, Y.-C. Ouyang and h.-H. Hsu, "An Efficient Green Control Algorithm in Cloud Computing for Cost Optimization," *IEEE,* pp. 1-14, 2013.

[6] S. Yakhchi, S. Ghafari, M. Yakhchi, M. Fazeli and A. Patooghy, "ICA-MMT: A Load Balancing Method in Cloud Computing Environment," IEEE, 2015.

[7] S. Kapoor and D. C, "Cluster Based Load Balancing in Cloud Computing," IEEE, 2015.

[8] S. Garg, R. Kumar and H. Chauhan, "Effcient Utilization of Virtual Machines in Cloud Computing using Synchronized Throttled Load Balancing," 1st International Conference on Next Generation Computing Technologies (NGCT-2015), pp. 77-80, 2015.

[9] R. Panwar and D. B. Mallick, "Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm," IEEE, pp. 773-778, 2015.

[10] M. P. V. Patel, H. D. Patel and . P. J. Patel, "A Survey On Load Balancing In Cloud Computing," International Journal of Engineering Research & Technology (IJERT), pp. 1-5, 2012.

[11] R. Kaur and P. Luthra, "LOAD BALANCING IN CLOUD COMPUTING," Int. J. of Network Security, pp. 1-11, 2013.

[12] Kumar Nishant, , P. Sharma, V. Krishna, Nitin and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," IEEE, pp. 3-9, 2012.

[13] Y. Xu, L. Wu, L. Guo,, Z. Chen, L. Yang and Z. Shi, "An Intelligent Load Balancing Algorithm Towards Efficient Cloud Computing," AI for Data Center Management and Cloud Computing: Papers from the 2011 AAAI Workshop (WS-11-08), pp. 27-32, 2011.

[14] A. K. Sidhu and S. Kinger, "Analysis of Load Balancing Techniques in Cloud Computing," International Journal of Computers & Technology Volume 4 No. 2, March-April, 2013, ISSN 2277-3061, pp. 737-741, 2013.

[15] O. M. Elzeki, M. Z. Reshad and M. A. Elsoud, "Improved Max-Min Algorithm in Cloud Computing," International Journal of Computer Applications (0975 – 8887), pp. 22-27, 2012.

[16] B. Kruekaew and W. Kimpan, "Virtual Machine Scheduling Management on Cloud Computing Using Artificial Bee Colony," Proceedings of the International Multi Conference of Engineers and Computer Scientists 2014 Vol I,IMECS 2014, 2014.

[17] R.-S. Chang, J.-S. Chang and P.-S. Lin, "An ant algorithm for balanced job scheduling in grids," Future Generation Computer Systems 25 (2009) 20–27, pp. 21-27, 2009.

[18] Z. Chaczko, V. Mahadevan, S. Aslanzadeh and C. Mcdermid, "Availability and Load Balancing in Cloud Computing," International Conference on Computer and Software Modeling IPCSIT vol.14 (2011) © (2011) IACSIT Press, Singapore, pp. 134-140, 2011.

[19] R. K. S, S. V and V. M, "Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud," Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012, pp. 31-35, 2012.

[20] Kumar Nishant, P. Sharma, V. Krishna, N. and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," IEEE, pp. 3-9, 2012.

[21] Ankit Kumar, Mala Kalra," Load Balancing in Cloud Data Center Using Modified Active Monitoring Load Balancer", IEEE pp. 1-5, 2016.

[22] Saraswathi AT, Kalaashri.Y.RA, Dr.S. Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing", ELSEVIER, pp. 30-36, 2015