Load Balancing Techniques In Cloud Environment

R.Sandeep Kumar[1], G.Shaheen Firdous[2]
Department of Computer Science and Engineering
[1].Assistant Professor, G.Pullaiah College of Engineering and Technology, Kurnool
[2]M.Tech Student, G.Pullaiah College of Engineering and Technology, Kurnool

Abstract-*Typically applications in cloud have different structure, configuration, and deployment requirements. Since cloud is an internet based dynamic computing, cloud resources and their loads possess dynamic characteristics and suffers from overloading of requests. Computing the performance of resource allocation at the lowest level under varying load and size is a difficult problem to address. In this paper we presented load balancing techniques to address the issues with overloading of requests .Defined as process to evenly distribute the workload to a set of servers to maximize the throughput, minimize the response time, and increase the system resilience to faults by avoiding overloading the systems improving resource utilization, decreasing the traffic, and increasing the efficiency and throughput" of the cloud along with the system performance.*

Keywords-Load balancing, Cloud Computing, CPU Workload,

I. INTRODUCTION

Cloud computing has created a reverberation in the IT landscape and also, its adoption rate is increasing dramatically. Various concepts and technologies have laid the foundation as depicted in figure 1. For cloud environment. It is a model that is constantly evolving. In tracking the evolution of cloud, we have found that from centralized "Mainframe" to distributed "Grid computing" all have played important role for the realization of the cloud system.

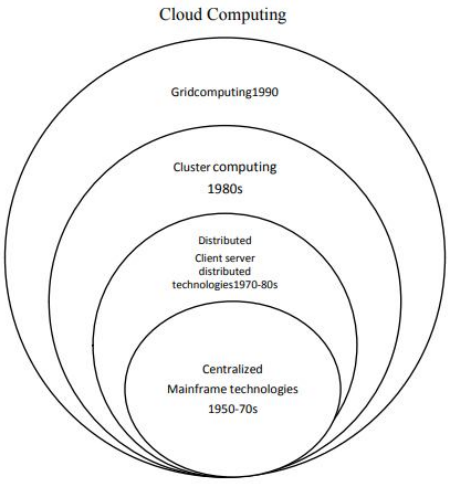

Figure 1. Cloud computing environment

A platform to develop distributed applications, cloud not only superimposes the grid computing but has evolved too, from it. It provides backbone and infrastructure support to cloud computing where the concept of virtualization brings the difference between the two. Mondal R. and Sarddar D [4] has established relationship with conventional client-server model where a user connects with a server for execution of a job. Cloud is a "Internet based Development and Computing platform, since it revolves around internet based acquisition and release of resources from a data center" [5]. It is a form of the "network of networks" and conceals the underline infrastructure. Prior to cloud computing, the symbol of a cloud was used to represent the network in a various specification or as an abstraction of the networks in system diagram as depicted in figure 1. Cloud represented a virtual and distinct environment for the purpose of remotely leveraging scalable and measured computing resources.

Components

"Cloud computing architecture refers to the components and subcomponents required for realization of cloud system" Cloud model consists of front end and a back end supported by the network where the network acts as a communication link between the two. With the "front-end" user interacts with the system; the "back end" does the processing.

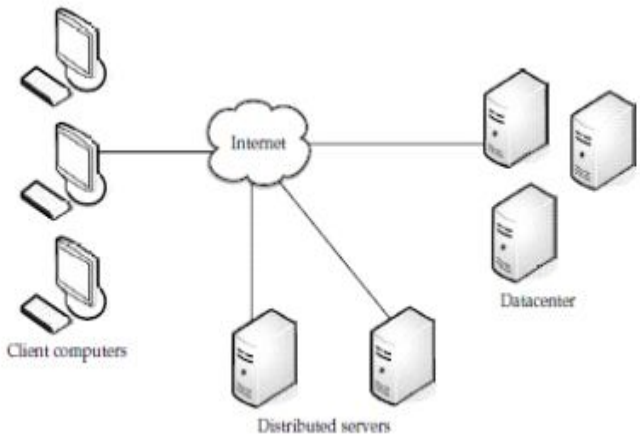

Figure 2: Component level architecture of cloud computing

In the view of Herewith J. et.al [1] "As a whole cloud environment is made up of 3 major components such as end clients, datacenter and dispersed servers with a definite roles and purpose" as described below:

Clients Compromises of end users and manage cloud information. They can be further classified as a Mobile, Thin and Thick clients.

Mobile: Location is not fixed. Cellular phones falls under the category of mobile client.
Thin: These clients do not have any internal memory. Only function is to display the information and is supported by servers, which do all the works.

Thick: They are also named as Fat clients. Thick client possesses the ability to perform many functions without having constant connection to server. Datacenter Collection of physical and virtual servers hosting several types of diverse applications. It is situated at remote location and can host both physical and virtual infrastructure used by the enterprises.

Data centers are generally used for processing enormous data. Construction of cloud is based on one or more data centers. Mostly the virtualization of hardware occurs to for use of Infrastructure. With respect to specific services, delivered to clients, discrete layers are stacked on top of this virtual infrastructure

Servers are the parts of a cloud, distributed at dispersed location. And appears to user as single coherent system [1]. The entire cloud system can be broadly categorized with respect to the deployment and also in the way of services to be provided [2]. According to NIST [3] there are three major deployment or accessibility model, they are 1) public cloud, 2) private /enterprise cloud and hybrid model for cloud. Since the cloud is defined as "utilities" therefore services and accessibility in cloud is provided as- XaaS

II. A MATHEMATICAL MODEL FOR LOAD BALANCING

A mathematical model for describing load balancing problems is described below. Several new notations are introduced for scheduling systems/Load balancing in cloud environment. Such as:

2.1 Jobs, Tasks, Operations

Let J = {J1, J2, ..., Jn} be the set of jobs. (Defined as a Workloads)

T= {T1, ..., Tk} is set of tasks.

Each job Ji is made of tasks {T1,T2,…….Tk } and each Ti is be made up of several operations {Oi1, Oi2, ..., Oini }.

2.2 Machines, Resources, and Queues

Let M = {M1, M2, ..., Mm} the set of machines. The machines here can be virtual or real nodes or it can be a datacenter.

Each M_j has MC_j = {a, p, r, k, x, o} attributes of machine M_j
Further Let R = {R1, R2, ..., Rn} be the set of resources (Rk = {U Mj | Mj ∈ M} ⊆ M}, Each resource Ri are compute resources.

Load balancing is represented as a "function f: T → R which maps every task Ti ∈ T on a resource Rj ∈ R" depending upon the availability of resources.

In or work we have considered that Oij symbolizes an workload processed by capable machine.

Workload O_{ij} can simultaneously use all machines Mij which represents the distributed/geographical dispersed nodes in cloud environment.

In a cloud environment, the traditional load balancing mechanism faces the problems at several stages like:

- The node load condition after the definition
- Acquisition of load information,
- Selection algorithm.

III. CHALLENGES

Load balancing in cloud is important and challenging issue since:

- Requests prediction issued to the server is difficult. [6]
- Scale up to the increasing demands.

3. Heterogeneous nature of cloud environment makes allocation decision troublesome.

3.1 Objective of load balancing:

- High user satisfaction and efficient resource utilization ratio.
- Optimum allocation of services onto a set of machines so the machine usage can be maximized.
- To scale up to increasing demands.
- Handling of dynamic service demands.
- Increase throughput.
- Allocation with environment constraints

3.2 Types of load balancing

Load balancing is a mechanism to handle sudden variation in demand. It is the prerequisite for increasing the cloud performance and utilizing the resources. The load balancing mechanism is classified as described in below figure 3.

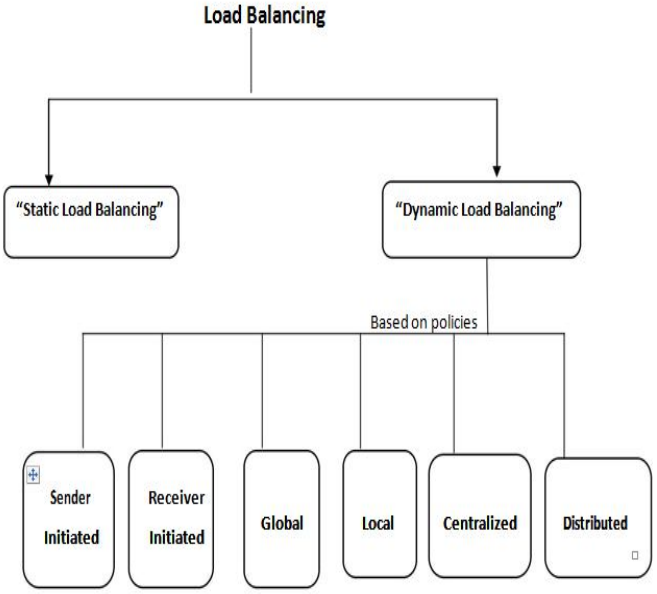

Figure 3: Classification of Load balancing algorithm

On the broader view the algorithm of load balancing can be categorized on the basis of initiation of process such as:

- **Sender Initiated:** Initializing of balancing process is done by the sender.
- **Receiver Initiated:** It is the receiver initiated load balancing process. The receiver sends request messages till it finds a machine to provision the load.
- **Symmetric:** When the mechanism is comprised of sender and receiver initiated technique then Symmetric process takes place.

Depending on moment and state of allocation, the process can be further divided into two categories as:

a) **Static load balancing:** Migration decisions are independent of present system state. Prior knowledge of the computing environment is required.

b) **Dynamic load balancing**: distributed at runtime. It distributed. [6][7] Present state of the system is considered. Work load is is further categorized in two forms: distributed and non-Load balancing can also be categorized based on the manipulation of ID and virtual server too. Along with the centralized or distributed approach [8][9][10].

3.3 Policies in dynamic load balancing:

Policies decide the rules with respect to various operations resulting into enablement of balancing.

There are different policies applied during the process of load balancing such as:

S.no	Policies	Description
1	Transfer	It determines migration process from overloaded to under loaded machines
2	Selection	Processor matching process w.r.t to the configuration and availability.
3	Location	Specification for location Determination.
4	Information	Used to determine the Information determination and collection strategy.
5	Load estimation	Used for Estimation of workload.
6	Process transfer	Local or global process execution policy
7	Priority assignment	Priority of load to be executed.
8	Migration limiting	Transfer constraints

3.4 Types of workload:

Because of features such as virtualization abstraction of workload is possible. The abstracted workload is the best way to think about a Cloud workload along with it also helps in creation of distributed virtualized cloud environment.

Several types of workload exist depending upon compute resources and functionality they perform, a few specific types of workload that applies to cloud systems includes:

Table.2: Types of Workloads

Balancing of these workloads is an important concern in cloud system

S.NO	Types of workload	Description	Effect
1	Memory Workload	Used memory for a given period of time or at a specific instant in time.	Bottleneck to the application performance.
2	CPU Workload	Executed number of instructions at a particular instant of time.	Decreases in processing capabilities.
3	I/O Workload	Defined as the number of inputs and outputs for an application.	Decides load performance
4	Batch Workloads	Background jobs.	Process huge volumes of data on a regular schedule.
6	Network Workloads	It is a mix of many and complex sources.	Performance of networks and network devices are affected along with the availability and QOS
7	Application server Workloads	Work load related to the applications	

The current, load balancing techniques have solved the various problems such as:

(i) Load balancing after a server was overloaded.

(ii) Load balancing and load forecast for the allocation of resources;

(iii) Improving the parameters affecting to load balancing in cloud.

The study of improving the process and corresponding parameters have great significance in increasing the overall cloud performance. On the basis of it, we can propose more effective methods of load balancing, in order to increase system performance. In the next chapter, we have studied the parameters that affect the performance of load balancing on cloud, which forms the basis to propose our frame work for load balancing.

IV. CONCLUSION

In this paper we have focused on types of workloads, the needs of Load balancing, its concern and challenges in a multi-tenant environment like cloud. With various concerns and needs a proper load balancing framework is need to be developed.

REFERENCES

[1] Herewith J.et.al, “Cloud Computing for Dummies”,Wiley Publishing, special edition, 2010.

[2] Buyya R. et.al., “Mastering Cloud computing”, Tata McGraw-Hill Education,ISBN:1259029956,9781259029950,2013

[3] Mell P. and Grance T.”National Institute of Standards and Technology Special Publication”, 800-145 , pp.0-7 , 2011.

[4] Mondal R. and Sarddar D. “Node Designing of Complex Networks in Cloud Computing Environment”, International Journal of Hybrid Information Technology, Vol.8, issue 7 ,2015

[5] Pan J. et.al, “Research on Heuristic Based Load Balancing Algorithms in Cloud Computing”, Intelligent Data Analysis and Applications”Advances in Intelligent Systems and Computing 370, Springer International Publishing Switzerland 2015, DOI 10.1007/978-3-319-21206-7_35, pp417- 426,2015.

[6] Sangwan A. et.al , “ATO CONVALESCE TASK SCHEDULING IN A DECENTRALIZED CLOUD COMPUTING ENVIRONMENT”, Review of Computer Engineering Research, Volume 3 , issue 1,pp.25-34 ISSN(e): 2410-9142/ISSN(p): 2412-4281,2016

[7] Paya A. and Marinescu A., “Energy-aware Load Balancing and Application Scaling for the Cloud Ecosystem”, IEEE TRANSACTIONS ON CLOUD COMPUTING, Issue: 99, ISSN: 2168-7161,2015.

[8] Gopinath G. et.al, "An in-depth analysis and study of Load balancing techniques in the cloud computing environment" ,2nd International Symposium on Big Data and Cloud Computing, doi: 10.1016/j.procs.2015.04.009 Published by Elsevier B. P.P. Procedia Computer Science, Vol.50, pp.427 – 432,2015.

[9] Bhaskar R. and Dr. Shylaja B. S, "KNOWLEDGE BASED REDUCTION TECHNIQUE FOR VIRTUAL MACHINE PROVISIONING IN CLOUD COMPUTING", International Journal of Computer Science and Information Security, Vol. 14, Issue.7, pp 472-475, July 2016.

[10] Chien N.et.al, "Load Balancing Algorithm Based on Estimating Finish Time of Services in Cloud Computing" , 18th International Conference on Advanced Communication Technology,ISBN: 978-8-9968-6506-3, DOI: 10.1109/ICACT.2016.7423340,IEEE,2016.