Survey Paper for Load Balancing in Cloud Computing Systems

Karishma Sethi
M.E., IV Sem, Maharana Pratap College of technology, Gwalior

Abstract-*Cloud computing is an on demand model for delivering information technology services in which resources are retrieved from the internet through web based tools and applications, rather than a direct connection to a server. It is the next generation of computation which allows people to have everything they on cloud. It provides resources to clients on demand. The purpose of cloud computing is to provide efficient services at the right time by consuming fewer resources. The servers are loaded due to concurrent access of services at the same time and slows the system which is not the aim of cloud computing.*

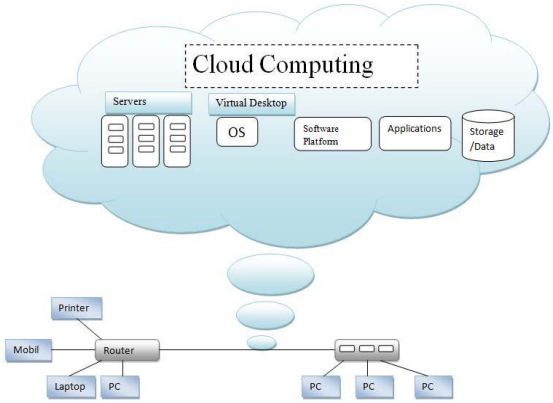
It demanded the need for a load balancer which balances the services by analyzing the resources of each virtual machine and then allocating the load to a particular virtual machine.

Balancing of load is a challenging and core issue in cloud computing. In order to utilize the benefits of cloud computing resources on internet load balancing algorithms are required. The aim of balancing the load is to maximize the throughput, minimize response time and avoid overload of any single machine.

Keywords*-*concurrent, virtual machine, throughput, load balancer

I. INTRODUCTION

Cloud computing is an emerging trend in the web world. It uses internet and centralized servers located at remote locations to maintain various data and applications of the users.



II. CLOUD COMPONENTS

There are three main components of a cloud- client, distributed server and datacenter. Each component has a definite purpose and plays significant role.

2.1 Clients

Clients generally fall in to three categories. [1]

Mobile:- Windows Mobile Smartphones, like a Blackberry and an iphone.

Thin:- They don't do any computation work. They only display the information. Servers do all the works for them. Thin clients don't have any internal memory.

Thick:- These use different browsers like internet explorer, mozillla firefox to connect to the internet cloud.

2.2 Datacenter

It is a collection of servers hosting different applications. A user connects to the datacenter to subscribe different applications. A datacenter may exist at a large distance from the clients.

2.3 Distributed Servers

Distributed Servers are the parts of a cloud which are present throughout the internet hosting different applications. But while using the application from the cloud, the user will feel that he is using this application from its own machine.

III. TYPES OF CLOUDS

3.1 Public clouds

Public clouds [2] are used by individuals or organizations based on their requirements and necessities. They offer greatest level of efficiency in shared resources. They follow the "Pay-as–you-go" model. Confidentiality is the major security issue in public cloud. Vulnerability is higher in public clouds as compared to that in private clouds. Amazon Web Services, Google Compute Engine, Microsoft Azur are some of the public clouds.

3.2 Private clouds

Private clouds [2] are owned by enterprises or business for their internal use. They are used as repositories to store and manage large data of the organizations and provide resources on demand basis to the clients. It has higher level of security. Openstack, VMware are some of the private clouds.

3.3 Hybrid clouds

Hybrid cloud [2] is a combination of public and private cloud. It allows organizations / enterprises to manage some resources externally and some within in the organization.

IV. TYPES OF CLOUD SERVICES

Cloud providers offer services which are grouped into three categories [3]:-

4.1 Software as a Service (Saas)

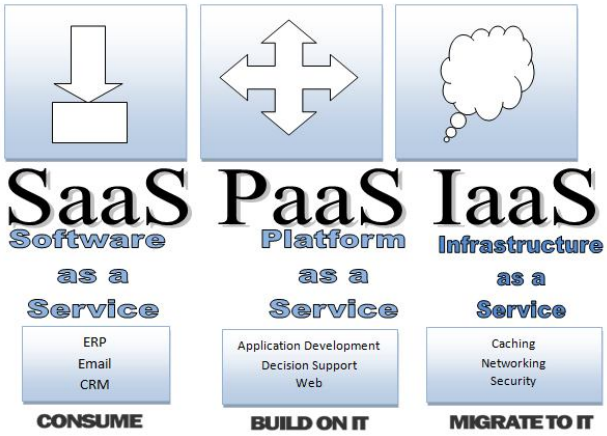
A complete application, in the form of software is offered to the customer, as a service on demand. On the cloud, a single instance of the service runs and multiple end users are served. It is beneficial for customer end as there is no need for investment in servers or software licenses, while for the provider, the costs are lowered, as only a single application needs to be hosted and maintained.

4.2 Platform as a Service (Paas)

A development environment or a layer of software is encapsulated and offered as a service; other higher level of services can be built as per the need of the client using this environment as a platform.

4.3 Infrastructure as a service (Iaas)

It provides basic storage and computing capabilities, as services over the Internet.



V. LOAD BALANCING

Load Balancing is used for distributing a larger processing load to smaller processing nodes for enhancing the performance of the system. It helps in fair allocation of computing resources. It is the process of reassigning the total loads to the individual nodes of the system to make the best response time and also good utilization of the resources and processes. The load balancer system allows creating an infrastructure, which is able to distribute the workload balancing between two or more cloud servers.

5.1 Load balancing classification:

Load balancing algorithms [4] are mainly classified into two categories: static load balancing algorithm and dynamic load balancing algorithm. The figure represents various load balancing algorithms:

1) Static approach: - This approach is pre defined in the design of the system. This does not depend on the current state of the system. Prior knowledge of the system is required. Static load balancing algorithms divide the traffic equivalently between all servers.

2) Dynamic approach:- In this approach only the present state of the system, during load balancing decisions, is considered. Dynamic approach is preferable for widely distributed systems. Dynamic load balancing approaches are of two types viz., distributed approach and non-distributed (centralised) approach. They are:

a) Centralized load distribution- in this kind of approach, the load balancing algorithm is executed only by a unique node in the whole system and this node is represented as "central node." The other nodes, in the system communicates, only with the central unique node. In this approach the number of interaction between the nodes decreases as the central unique node manages the interaction among all the nodes, however it

can cause a bottleneck if the central unique node fails to perform. Hence this approach is preferred for small sized networks.

b) Distributed Approach: Load balancing in distributed system is executed by all nodes, present in the system. This allows the task to be shared among all the nodes present in the system. The communication process is longer, as the number of messages needed to accomplish a task increases due to sharing of messages amongst the nodes. The advantage of this approach is that even if one of the node fails the system is not hampered, however it affects the system performance to some extent. It increases the stress on every node as they need to share status among every other node in the system

5. 2 Matrices for Load Balancing [4]:

1. Throughput: - It is the rate, which is calculated by the maximum number of tasks whose execution has been completed .The performance of any system is improved if throughput is high.

2. Fault Tolerance: -It is the ability of a system to recover from the faults .

3. Migration time: -It is the time to migrate the tasks or resources among nodes. The performance of the system depends on migration time , the smaller the migration time between the nodes the higher is the performance.

4. Response Time: - The time taken by the nodes in the system to respond to a particular task is its response time. The value of this parameter should be minimized to maximize the performance of the system.

5. Scalability: - It is the ability of any algorithm to balance the load for the finite number of nodes in the system . The higher the scalability of the system higher is the performance of the system.

5.3 Policies of load balancing algorithm

There are many policies are used in load balancing algorithms:

- Information policy: It defined that what information is required and how this information is collected. This is also defined that when this information is collected from one node to other.

- Triggering policy: This policy defined that time period when the load balancing operation is starting to manage the load.

- Resource type policy: This policy defined the all types of resources which are available during the load balancing. Location policy: This uses all the results of the resource type policy. It is used to find a partner for a server or receiver.

- Selection policy: This policy is used to find out the task which transfers from overloaded node to free node.

5.4 Major goals of load balancing algorithms

1. Cost effectiveness:

Load balancing help in provide better system performance at lower cost

2. Scalability and flexibility:

The system for which load balancing algorithms are implemented may be change in size after some time. So the algorithm must handle these types' situations. So algorithm must be flexible and scalable

3. Priority:

Prioritization of the resources or jobs needs to be done. So, higher priority jobs get better chance to execute.

VI. PROPOSED WORK

Introduction about Load Divisibility Theory

In 1988 the first article about Divisible Load Theory (DLT) was published [6]. Based on DLT, it is assumed that the computation can be partitioned into some arbitrary sizes, and each partition can be processed independently by one processor. In the past two decades, DLT has found a wide variety of applications in parallel processing area such as data intensive applications [3], data grid application [5], image and vision processing [4] and so on. Also it was applied for various network topologies including chain, star, bus, tree, three-dimensional mesh.

6.1 Divisible Load Scheduling

In general, DLT assumes that the computation and communication can be divided into some parts of arbitrary size and these parts can be independently processed in parallel by processors as bellow figure. 2. DLT assume that initially amount of load is held by the originator P0. The originator does not do any computation. It only distributes $\alpha 1, \alpha 2, \alpha 3, \ldots, \alpha m$ fractions of load on worker processors P1,P2,...,Pm.

condition for optimal solution is that all the processor stop processing at the same time. This fraction of load must be allocated based on criteria and priorities.

REFERENCES

[1] Anthony T.Velte, Toby J.Velte, Robert Elsenpeter, Cloud Computing A Practical Ap- proach, TATA McGRAW-HILL Edition 2010.
[2] Zhang, Q., Cheng, L. & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. Journal of Internet Services and Applications, 1(1), 7-18. DOI 10.1007/s13174-010-0007-6.
[3] Jenia Sagar, Lekha Bhambhu, Impact Factor (2012): 3.358. International Journal of Science and Research (IJSR)Implementation of Load Balance Algorithm in Cloud Computing
[4] Amandeep, Vandana Yadav et. al., International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 3 Issue 1, April 2014, Different Strategies for Load Balancing in Cloud Computing Environment: a critical Study.