Dynamic Resource Allocation Technique in Cloud

Ishita Patel[1] Brona Shah[2]

[1,2]Silver Oak Collage of Engineering & Technology, Gujarat Technical University, Ahmedabad, India

Abstract— "Cloud computing" is a term, which includes virtualization, dispersed figuring, systems administration, programming and web administrations. A cloud comprises of a few components, for example, customers, datacentre and conveyed servers. It incorporates adaptation to internal failure, high accessibility, versatility, adaptability, lessened overhead for clients, decreased expense of possession, on interest administrations and so forth. Vital to these issues lies the foundation of a viable load balancing calculation. The heap can be CPU load, memory limit, defer or arrange load. Load balancing is the procedure of circulating the heap among different hubs of a disseminated framework to enhance both asset use and employment reaction time while additionally dodging a circumstance where a portion of the hubs are vigorously stacked while different hubs are sit without moving or doing next to no work. Load balancing guarantees that all the processor in the framework or each hub in the system does around the equivalent measure of work at any moment of time. This system can be sender started, beneficiary started or symmetric sort. Our goal is to build up a viable load balancing calculation utilizing divisible load planning hypothesis to expand or minimize diverse execution parameters for the billows of various sizes.

Key words: Cloud Computing, Load Balancing, Dynamic, Resource Allocation

I. INTRODUCTION

A. Cloud Computing

In case of Cloud computing services can be used from diverse and widespread resources, rather than remote servers or local machines. There is no standard definition of Cloud computing. Generally it consists of a bunch of distributed servers known as masters, providing demanded services and resources to different clients known as clients in a network with scalability and reliability of datacenter. The distributed computers provide on-demand services. Services may be of software resources (e.g. Software as a Service, SaaS) or physical resources (e.g. Platform as a Service, PaaS) or hardware/infrastructure (e.g. Hardware as a Service, HaaS or Infrastructure as a Service, IaaS). Amazon EC2 (Amazon Elastic Compute Cloud) is an example of cloud computing services [6].

1) Cloud Components:

A Cloud system consists of 3 major components such as clients, datacenter, and distributed servers. Each element has a definite purpose and plays a specific role.

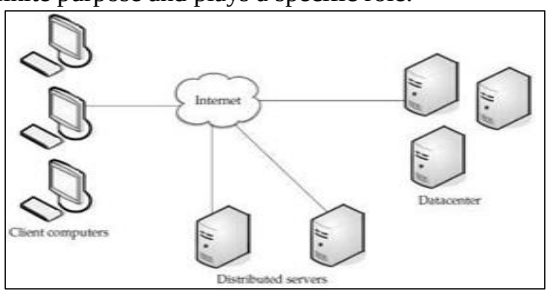


Fig. 1: Cloud Component [6]

a) Clients:

End users interact with the clients to manage information related to the cloud. Clients generally fall into three categories as given in [6]

Mobile: Windows Mobile Smartphone, smartphones, like a Blackberry, or an iPhone. Thin: They don't do any computation work. They only dispaly the information.

Servers do all the works for them. Thin clients don't have any internal memory.

Thick: These use different browsers like IE or mozilla Firefox or Google Chrome to connect to the Internet cloud. Now-a-days thin clients are more popular as compared to other clients because of their low price, security, low consuption of power, less noise, easily replaceble and repairable etc.

b) Datacenter:

Datacenter is nothing but a collection of servers hosting different applications. An end user connects to the datacenter to subscribe different applications. A datacenter may exist at a large distance from the clients. Now-a-days a concept called virtualization is used to install software that allows multiple instances of virtual server applications.

c) 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.

B. Services provided by the cloud

Service means different types of applications provided by different servers across the cloud. It is generally given as "as a service". Services in a cloud are of 3 types as given:

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Hardware as a Service (HaaS) or Infrastructure as a Service (IaaS)

a) Software as a Service (SaaS):

In SaaS, the user uses different software applications from different servers through the Internet. The user uses the software as it is without any change and do not need to make lots of changes or doen't require integration to other systems. The provider does all the upgrades and patching while keeping the infrastructure running.

The client will have to pay for the time he uses the software. The software that does a simple task without any need to interact with other systems makes it an ideal candidate for Software as a Service. Customer who isn't inclined to perform software development but needs high-powered applications can also be benefitted from SaaS.

b) Platform as a Service (PaaS):

PaaS provides all the resources that are required for building applications and services completely from the Internet, without downloading or installing software. PaaS services are software design, development, testing, deployment, and hosting. Other services can be team collaboration, database

integration, web service integration, data security, storage and versioning etc.

c) Infrastructure as a Service (IaaS)

It is also known as Hardware as a Service (HaaS). It offers the hardware as a service to an organization so that it can put anything into the hardware according to. HaaS allows the user to "rent" resources as Server space, Network equipment, Memory, CPU cycles, Storage space.

C. Load Balancing

It is a process of reassigning the total load to the individual nodes of the collective system to make resource utilization effective and to improve the response time of the job, simultaneously removing a condition in which some of the nodes are over loaded while some others are under loaded. A load balancing algorithm which is dynamic in nature does not consider the previous state or behavior of the system, that is, it depends on the present behavior of the system. The important things to consider while developing such algorithm are : estimation of load, comparison of load, stability of different system, performance of system, interaction between the nodes, nature of work to be transferred, selecting of nodes and many other ones. This load considered can be in terms of CPU load, amount of memory used, delay or Network load

1) Goals of Load balancing:

The goals of load balancing are

- To improve the performance substantially
- To have a backup plan in case the system fails even partially
- To maintain the system stability
- To accommodate future modification in the system

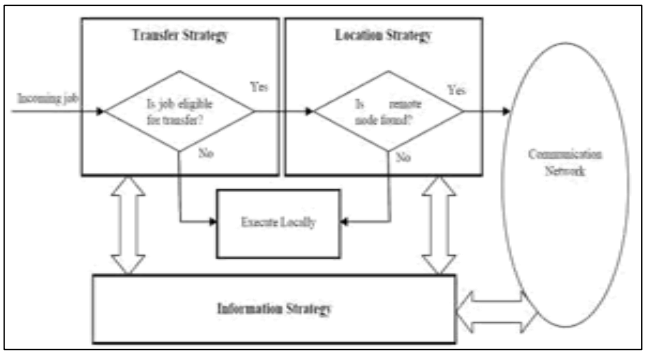


Fig. 2: Interaction among components of a dynamic load balancing algorithm [6]

2) Types of load balancing algorithm:

Types of Load balancing algorithms Depending on who initiated the process, load balancing algorithms can be of three categories as given in:

- Sender Initiated: If the load balancing algorithm is initialized by the sender
- Receiver Initiated: If the load balancing algorithm is initiated by the receiver
- Symmetric: It is the combination of both sender initiated and receiver initiated Depending on the current state of the system, load balancing algorithms can be divided into 2 categories
- Static: It doesn't depend on the current state of the system. Prior knowledge of the system is needed.
- Dynamic: Decisions on load balancing are based on current state of the system. No prior knowledge is needed. So it is better than static approach. Here we will discuss on various dynamic load balancing algorithms for the clouds of different sizes.

II. LITERATURE SURVEY

Paper	Method	Algorithm	Problem	Solution
Paper 1	Dynamic	Central Load Balancer	The test for cloud Data centers is how to manage and service trillions of requests that are coming frequently from end client efficiently and correctly.	"Central Load Balancer" a Load adjusting calculation to balance the load between VM at cloud data center.
Paper 2	Dynamic	Agent based Dynamic algorithm	Requires proper Assets distribution Between the tasks, Otherwise in some Situations assets may higher-utilized or lower-utilized.	Agent Based Dynamic approach in which mobile agent plays very important role to manage assets.
Paper 3	Dynamic	Prediction based load balancing	Only supporting template-based deployment of new VM, not supporting the trend prediction, failing to gain assets dynamically, and not effectively providing the elastic management of assets.	Dynamically manage processing capacity of back-end server cluster with the applied burden.
Paper 4	Dynamic	Weighted Signature Based load Balancing	Response time	Algorithm is proposed to minimize users response time.
Paper 5	Dynamic	Stable Matching theory	Cloud Providers are concerned with the Assets usage in a multi-type assets sharing environment while cloud customers desire higher quality of services.	The stable matching theory is given to generate an optimal mapping from containers to physical servers. Simulations are implemented to evaluate asset scheduling approach.

III. PROPOSED WORK

- We will implement central manager as well as data controller.
- we will find {memory_usage, Load factor} + {priority}
- we will implement BACKMON i.e. Background Monitoring process which will periodically finds above parameters
- Based on above mentioned parameters we will optimally allocate the requests
- Objective: Effective allocation and optimal usage of resources
  - minimizing resource consumption
  - fair allocation

For better performance of our algorithms 2 parameters are taken in to thought.

- Load on the server.
- Current performance of server.

- Load on the server:
  - We are additional interested in free resources accessible on node. The node having additional free resources can in a position to handle more requests simply while not degrading its performance.
- Current performance of Sever:
  - A request will be send to the instances at regular interval and get the response time of every instances, and on the basis of that the performance parameter is measured.
- It might be the case that the latent period of node may modification on every occasion counting on the shopper usage of its resources.
- So, the above 2 parameters are used to engineered a brand new queue for future allocation.

This information of the entire node is calculated by a function to count c-parameter worth for every node.

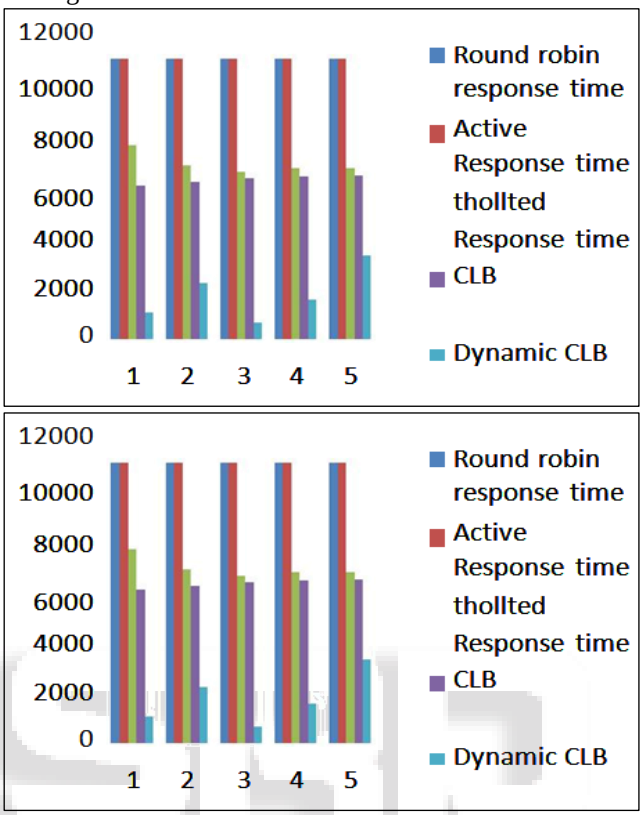
A. Algorithm

- Step 1: [Calculate Load Factor a]
  a = (Total _Resources – Used _Resources)
  // where a is free memory in terms of percentage.
- Step 2: [Calculate Performance Factor b]: b1 = average (current _response _time)
  b = b1 - (previously calculated b1)
  b = b/(previous b1)*100
  //counting b in terms of previously counted b1.
- Step 3: [finding c] c = a - b;
  If (c < 0) c = 0;
- Step 4: [Find minimum of all c except the nodes with c value 0]
  - Min _c= min (all c's)
- Step 5: [Find min _factor and divide all c by that factor]
  - Min _factor = min _c
  - C = c / min _factor
- Step 6: [Generate Dynamic Queue on base of c]
  - In the above algorithm a is considered as a free load on server, b for the performance on the server and b1 is the current response time.

IV. PERFORMANCE EVALUATION

Case: For testing the proposed approach we kept load constant and increase the number of virtual machine. And compare the different algorithm's response time.

From below calculation we can observed that response time of dynamic CLB is less as compared to other four algorithms.



V. CONCLUSION

Central load balancer whose sole responsibility is to parse the table for VM having highest priority as well as that vm must not be busy. if not busy then Data Controller will be informed and New request is to be mapped to that vm id means vm and table will be updated through central manager. Here it used utilization parameter like CPU usage, memory usage along with priority parameter So in our proposed work, we will extend it by take memory usage and load factor parameters in to consideration along with priority.

REFERENCE

[1] Gulshan Soni, Mala Kalra "A novel Approach for Load Balancing in Cloud Data Centre": Advance Computing Conference(IACC),Feb 2014 IEEE , ISBN: 978-1-4799-2571-1

[2] Jitender Grover,Shivangi Katiyar "Agent Based Dynamic Load Balancing in Cloud Computing": Human Computer Interaction(ICHCI),2013 International Conference, Aug 2013 IEEE

[3] He-Sheng WU,Chong-Jun WANG,Jun-Yuan XIE "Terascaler ELB-an Algorithm for prediction-based Elastic load Balancing Resource Management in Cloud Computing" : 2013 27th international conference on advanced information networking and application workshop, March 2013,ISBN: 978-1-4673-6239-9

[4] M. Ajit, G.Vidya "VM Level Load Balancing in Cloud Environment": Computing, Communication And Networking Technolgies (ICCCNT), 2013 Fourth International Conference,July 2013 IEEE, ISBN:978-1-4799-3925-1
[5] Xin Xu, Huiqun Yu, Xin Pei "A Novel Resource Scheduling Approach in Container Based Clouds" Computational Science and Engineering (CSE), 2014 IEEE 17th international Conference, Dec 2014 IEEE, ISBN: 978-1-4799-7980-6
[6] Rajesh George, V. Jeyakeishnan "A survey on load balancing in cloud computing environment" vol.2, Issue 12, Dec 2013 IJARCCE
[7] Tushar Desai, Jignesh Prajapati "a survey of various load balancing techniques and chalanges in cloud computing" Vol 2, Issue 11, Nov 2013 IJSTR

