International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Survey on Resource Allocation Technique in Cloud

Ishita Patel[1], Brona Shah[2]

Silver Oak Collage of Engineering And Technology, Gujarat Technical University, Ahmedabad, India

Abstract: ***"Cloud computing" is a term, which involves virtualization, distributed computing, networking, software and web services. A cloud consists of several elements such as clients, data centre and distributed servers. It includes fault tolerance, high availability, scalability, flexibility, reduced overhead for users, reduced cost of ownership, on demand services etc. Central to these issues lies the establishment of an effective load balancing algorithm. The load can be CPU load, memory capacity, delay or network load. Load balancing is the process 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. Load balancing 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. This technique can be sender initiated, receiver initiated or symmetric type. Our objective is to develop an effective load balancing algorithm using Divisible load scheduling theorem to maximize or minimize different performance parameters for the clouds of different sizes.***

Keywords: Cloud Computing, load balancing, dynamic, resource allocation

1. Introduction

1.1 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].

1.1.1 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.

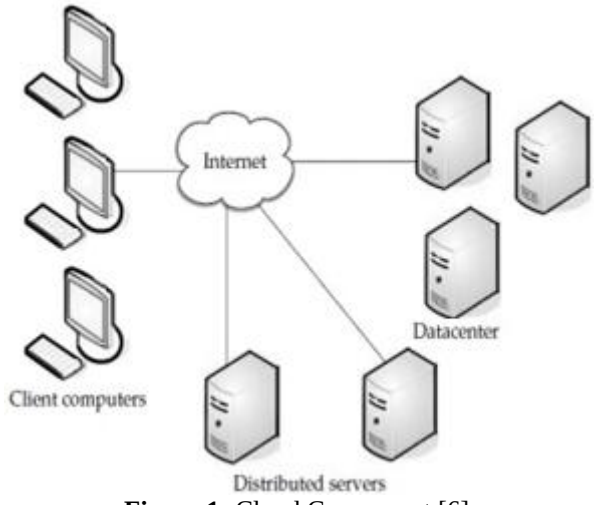

Figure 1: Cloud Component [6]

1.1.1.1Clients

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.

1.1.1.2 Datacenter

Datacenter is nothing but a collection of servers hosting different applications. A 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 virtualisation is used to install a software that allow multiple instances of virtual server applications.

1.1.1.3 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.

1.1.2 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)

1.1.2.1 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.

1.1.2.2 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 a 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.

1.1.2.3 Infrastructure as a Service (IaaS)

It is also known as Hardware as a Service (HaaS). It offers the hardware as a service to aorganisation 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.

1.2 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

1.2.1 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

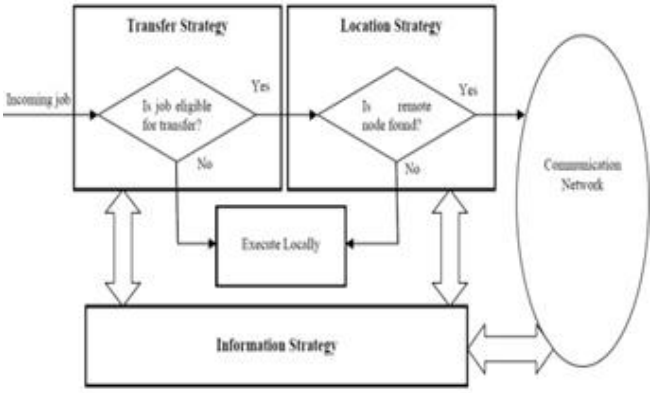

Figure 2: Interaction among components of a dynamic load balancing algorithm [6]

1.2.2 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 doesnt 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.

2. Literature Survey

[1]The proposed load balancing algorithm "Central Load Balancer" will balance the load among virtual machines having different hardware configurations and will distribute the load based on hardware configuration and states of virtual machines in data center. The proposed technique will be able to perform quick and reliable load balancing in cloud computing environment through utilization of all virtual machines according to their computing capacities. In the proposed technique, every request from user bases arrives at Data Center Controller. Data Center Controller queries the Central Load Balancer for allocation of requests. Central Load Balancer maintain a table that consist of id, states and priority of virtual machines. Central Load balancer parses the table and find out highest priority virtual machine, then check its states and if its states available then return that virtual machine id (VMid) to Data Center Controller. If the states of virtual machine is Busy then it chooses next less high priority virtual machine. Finally Data Center Controller assigns the request to that VMid that is provided by Central Load Balancer (CLB). The Central Load Balancer (CLB) is connected to all users and virtual machines present in cloud data center through Data center Controller as shown in Figure 3.1. The Central Load Balancer calculates the

priorities of virtual machines based on their CPU speed (MIPS) and memory.

[2] Figure 3, shows the proposed system architecture which consists n number of clients connected with cloud service providers via internet and service provider consists virtual machines, management unit, and m number of shared pool of resources which we are considering as servers. At the shared pool of servers, agent complete one cycle in two walks:

In First walk it moves from initiation server to last server and gathers information from all servers, for making appropriate decision for load balancing and in second walk it balances the server's load on the basis of average load of the cloud.

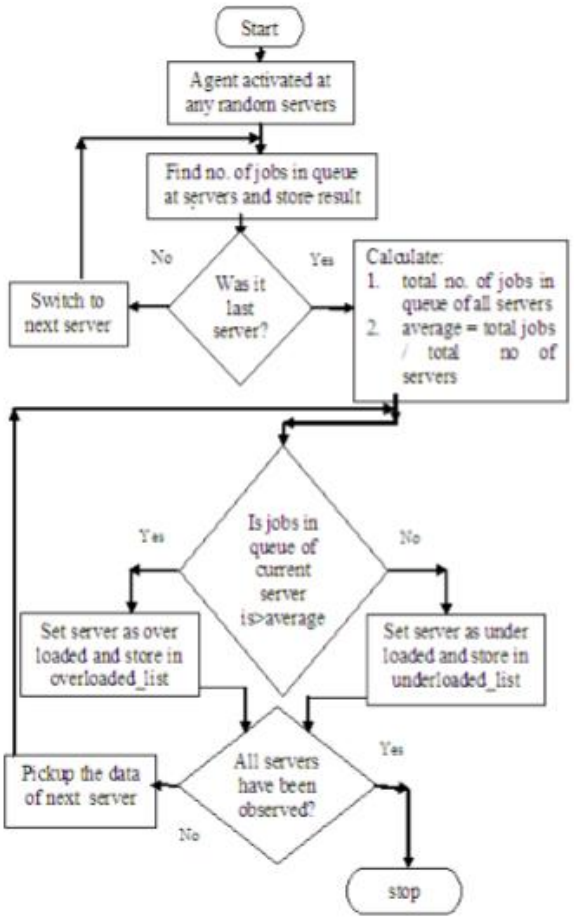

Figure 3: Agent Walk 1[3]

[3] ELB resource management algorithm is accomplished through two regular events:

- The load balancer collects the resource information from back-end server on a regular basis.
- The load balancer determines whether or not to applyfor /delete back-end virtual machine based on resource information collected from back-end server within the previous time.

In event 2, how the load balancer judges whether to add/delete back-end server is key to elastic resource management algorithm. Compare with traditional algorithm described as Algorithm 1, the proposed TeraScaler ELB is described as Algorithm 2. Algorithm 1 decides whether to increase/delete virtual machines based on the current load, while Algorithm 2 decides whether to add/delete virtual machines according to the analysis & prediction of current load and historical load traces.

[4] Cloud analyst is used to test the performance of the algorithms and compare them with proposed one with respect to the response time. Cloud analyst simulation tool is based on cloudsim library written in java and provides a GUI interface to configure various parameters to perform the experimental work. This research work considers Datacenter, VM, host and Cloudlet components from CloudSim for implementation of a proposed algorithm. Datacenter component handles service requests. VM consist of application elements which are connected with these requests, so Datacenters host should allocate VM requested by user. Cloud Analyst can evaluate any algorithm or application deploying in the cloud. VM life cycle starts from provisioning of a host to a VM, VM creation, VM destruction, and VM migration.

[5] The virtualized infrastructure in cloud data center is considered as using the container-based virtualization technology. Cloud provides should manage all the available resources in data center through an adaptive and flexible mechanism to deal with the dynamic workload. In our work, a cooperative and centralized controlling architecture of cloud resource manager is proposed. Multiple cloud customers send their requests to the resource manager to implement their jobs. The resource manager is responsible for collecting available resources in data center and analyzing requests from customers. More important, the resource manager should make an intelligent decision on scheduling the available resources to requests. To ensure a scalable and efficient on-demand resource provisioning, the resource manager is designed to deliver a feasible and optimal resource scheduling scheme for customers. The resource manager consists of four core components, including the Register, Monitor, Infrastructure Driver and Decision Maker.

3. Conclusion

Cloud computing is a computing service paradigm that charges under the basis of the amount of resources consumed i.e. pay per use constraint. Primary advantage in cloud environment is that IaaS controls the user and manages the systems in terms of bandwidth, response time resource expenses, and network connectivity, but do not concentrate on infrastructure. These papers discuss about the various types of resources allocation and task scheduling algorithm. Although, there are various algorithms and methods were existing to solve the problem of resource allocation but none of these algorithms could be extended. Efficiency of cloud depends on the type scheduling algorithm used in environment. All above discussed algorithm used for resource allocation completely depends on types of task to be scheduled.

Reference

[1] GulshanSoni, 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] JitenderGrover,ShivangiKatiyar "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] XinXu, 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, JigneshPrajapati "a survey of various load balancing techniques and chalanges in cloud computing" Vol2 , Issue 11, Nov 2013 IJSTR

Author Profile

Ms. Ishita Patel received the B.E degree in Computer Engineering from Sabar Institute of Technology for Girls in 2014. She will complete her M.E in Computer Engineering from Silver Oak College of Engineering and Technology in 2016.

Ms. Brona Shah received the B.E. degrees in Computer Engineering and received Master Degree in Information Technology from L.J Institute of Engineering and Technology under Gujarat Technological University. Currently she is working as assistant professor at Silver Oak College of Engineering and Technology.