IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI (Feb. 2014), PP 82-87
www.iosrjournals.org

Load Balancing In Public Cloud

Shrikant M. Lanjewar, Susmit S. Surwade,Sachin P. Patil,Pratik S. Ghumatkar, Prof Y.B. GURAV
Department of Computer Engineering,PVPIT,Pune,India
Under the Guidance of
(Comp Dept., PVPIT, Pune, India)

Abstract: *In present days cloud computing is one of the greatestplatform which provides storage of data in very lowercost and available for all time over the internet.But ithas more critical issue like security, load managementand fault tolerance. In this paper we are discussingLoad Balancing approach. Many types of load concernwith cloud like memory load, CPU load and networkload. Load balancing is the process of distributing loadover the different nodes which provides good resourceutilization when nodes are overloaded with job. Loadbalancing has to handle the load when one node isoverloaded. When node is overloaded at that time loadis distributed over the other ideal nodes. Manyalgorithms are available for load balancing like Staticload balancing and Dynamic load balancing.Load balancing in the cloud computing environment has an important impact on the performance. Good load balancing makes cloud computing more efficient and improves user satisfaction. This article introduces a better load balance model for the public cloud based on the cloud partitioning concept with a switch mechanism to choose different strategies for different situations. The algorithm applies the game theory to the load balancing strategy to improve the efficiency in the public cloud environment[6].*

Keywords: *Cloud Computing, Load balancing, Virtualization, load balancing model, public cloud, cloud partition, game theory*

I. Introduction

A) What is Cloud Computing?

The term "cloud" originates from the world of telecommunications when providers began using virtual private network (VPN) services for data communications. The definition of cloud computing[1]
provided by National Institute of Standards and Technology (NIST) says that: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage applicationsand services) that can be rapidly provisioned andreleased with minimal management effort or serviceprovider interaction." So through this cloudcomputing there is no need to store the data ondesktops, portables etc. You can store the data onservers and you can access the data through internet.Cloud computing provides better utilization ofdistributed resources over a largedata and they canaccess remotely through the internet.

II. Load Balancing

A) What is Load Balancing?

Load Balancing is a technique in which the workload on the resources of a node is shifts to respective resources on the other node in a network without disturbing the running task. A standard way to scale web applications is by using a hardware-based load balancer.The load balancer assumes the IP address of the web application, so all communication with the web application hits the load balancer first. The load balancer is connected to one or more identical web servers in the back-end. Depending on the user session and the load on each web server, the load balancer forwards packets to different web servers for processing. The hardware-based load balancer is designed to handle high-level of load, so it can easily scale .However, a hardware-based load balancer uses application specific hardware-based components, thus it is typically expensive. Because of cloud's commodity business model, a hardware-based load balancer is rarely occurred by cloud providers as a
service. Instead, one has to use a software based load balancer running on a generic server.[2]

B) 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.

C) Types of Load Balancing Algorithms

1) Static Algorithms

Static algorithms divide the traffic equivalentlybetween servers. By this approach the traffic onthe servers will be disdained easily and consequently it will make the situation more perfectly. This algorithm, which divides the traffic equally, is announced as round robinalgorithm. However, there were lots of problemsappeared in this algorithm. Therefore, weighted round robin was defined to improve the critical challenges associated with round robin. In this algorithm each servers have been assigned a weight and according to the highest weight they received more connections. In the situation that all the weights are equal, servers will receive balanced traffic.[4]

2) Dynamic Algorithms

Dynamic algorithms designated proper weights on servers and by searching in whole network a lightest server preferred to balance the traffic. However, selecting an appropriate server needed Real time **B) Types of cloud**

Public

Public clouds are made available to the general public by a service provider who hosts the cloud infrastructure. Generally, public cloud providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access over the Internet. With this model, customers have no visibility or control over where the infrastructure is located. It is important to note that all customers on public clouds share the same infrastructure pool with limited configuration, security protections and availability variances.Public Cloud customers benefit from economies of scale, because infrastructure costs are spread across all users, allowing each individual client to operate on a low-cost, " pay-as-you-go" model. Another advantage of public cloud infrastructures is that they are typically larger in scale than an in-house enterprise cloud, which provides clients with seamless, on-demand scalability. These clouds offer the greatest level of efficiency in shared resources; however, they are also more vulnerable than private clouds.

Private

Private cloud is cloud infrastructure dedicated to a particular organization. Private clouds allow businesses to host applications in the cloud, while addressing concerns regarding data security and control, which is often lacking in a public cloud environment. It is not shared with other organizations, whether managed internally or by a third-party, and it can be hosted internally or externally.

Hybrid

Hybrid Clouds are a composition of two or more clouds (private, community or public) that remain unique entities but are bound together offering the advantages of multiple deployment models. In a hybrid cloud, you can leverage third party cloud providers in either a full or partial manner; increasing the flexibility of computing. Augmenting a traditional private cloud with the resources of a public cloud can be used to manage any unexpected surges in workload. Hybrid cloud architecture requires both on-premise resources and off-site server based cloud infrastructure. By spreading things out over a hybrid cloud, you keep each aspect of your business in the most efficient environment possible. The downside is that you have to keep track of multiple cloud security platforms and ensure that all aspects of your business can communicate with each other.

Communicationwith the networks, which will lead to extra traffic added on system. In comparison between these two algorithms, although round robin algorithms based on simple rule, more loads conceived on servers and thus imbalanced traffic discovered as a result.[3]

Load balancing schemes depending onwhether the system dynamics are important can beeither static and dynamic.Static schemes do not usethe system information and are less complex whiledynamic schemes will bring additional costs for thesystem but can change as the system status changes. Adynamic scheme is used here for its flexibility. Themodel has a main controller and balancers to gather andanalyze the information. Thus, the dynamic control haslittle influence on the other working nodes. The systemstatus then provides a basis for choosing the right load balancing strategy.

D) Existing System

In single storage cloud system each cloud customer's data is stored on single higher configuration server. Even if that server has huge amount resources such as RAM, Hard disk, processing power, it has certain limit. If it crosses that limit then particular resource performance slows down

E) Proposed System

The load balancing model given in this article is aimed at the public cloud which has numerous nodes with distributed computing resources in many different geographic locations. Thus, this model divides the public cloud into several cloud partitions. When the environment is very large and complex, these divisions simplify the load balancing. The cloud has a main controller that chooses the suitable partitions for arriving jobs while the balancer for each cloud partition chooses the best load balancing strategy.

The load balancing strategy is based on the cloud partitioning concept. After creating the cloud partitions, the load balancing then starts. When a job arrives at the system, with the main controller deciding which cloud partition should receive the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, this partitioning can be accomplished locally. If the cloud partition load status is not normal, this job should be transferred to another partition.

III. Proposed System Architecture

There are several cloud computing categories with this work focused on a public cloud. A public cloud isbased on the standard cloud computing model, withservice provided by a service provider. A largepublic cloud will include many nodes and the nodes in different geographical locations. Cloud partitioning is used to manage this large cloud. A cloud partition is a subarea of the public cloud with divisions based on the geographic locations. The architecture is shown in Fig.1. The load balancing strategy is based on the cloud partitioning concept. After creating the cloud partitions, the load balancing then starts: when a job arrives at the system, with the main controller deciding which cloud partition should receive the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, this partitioning can be accomplished locally. If the cloud partition load status is not normal, this job should be transferred to another partition. The whole process is shown in Fig.2 [1].

A) Main Controller and Balancer

The load balance solution is done by the main controller and the balancers. The main controller first assigns jobs to the suitable cloud partition and then communicates with the balancers in each partition to refresh this status information [6]. Since the main controller deals with information for each partition, smaller data sets will lead to the higher processing rates. The balancers in each partition gather the status information from every node and then choose the right strategy to distribute the jobs. The relationship between the balancers and the main controller is shown in Fig.1.

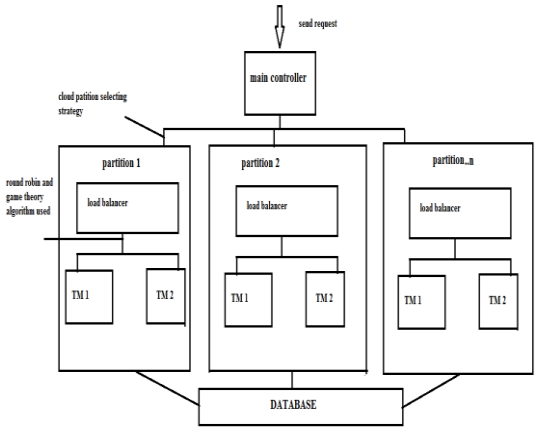

Fig.1 Relationship between Balancer and Main Controller

B) Assigning Jobs to the Cloud Partition

When a job arrives at the public cloud, the first step is to choose the right partition. The cloud partition status can be divided into three types:

(1) Idle: When the percentage of idle nodes exceeds α, change to idle status.

(2) Normal: When the percentage of the normal nodes exceeds β, change to normal load status.

(3) Overload: When the percentage of the overloaded nodes exceeds γ , change to overloaded status.

The parameters α, β, and γ are set by the cloud Partition balancers. [1]

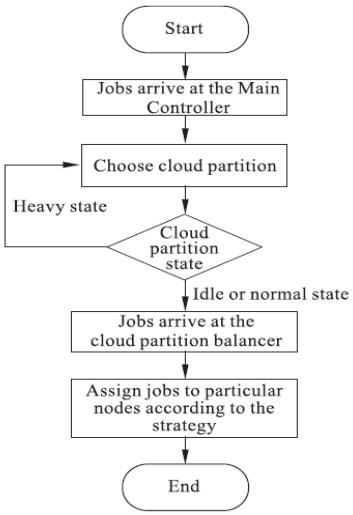

Fig.2 Flowchart for Job Assignment

C) Operational Environment

Operational environment means environment in which user interact with the application. For example, the DOSenvironment consists of all the DOS commands available to users The windows environment, on the other hand, is a graphical user interfaceuses icons and menus instead of commands. In proposed system user interacts with the application in following environment [10].

1) A Personal Computer.
2) Windows Operating System.

D) Techniques Used

Our system is based on the following techniques :

1. Partition Selection Technique

Best Partition Searching for Load Balancing:

Define a load parameter set

($F = \{F_1, F_2 \ldots\ldots, F_M\}$)With each Fi (1 <= i<= m; Fϵi [0, 1] parameter being either static or dynamic. m represents the total number of the parameters.

Let's, put it all into Turing machine…

Turing machine as a 7-tuple where M=(Q, Γ, b, Σ, δ, q_o, F)

1. Q is a finite, non-empty set of states
2. Γis a finite, non-empty set of the tape alphabet/symbols
3.b∈T is the blank symbol (the only symbol allowed to occur on the tape infinitely often at any step during the computation)
4.Σ ⊆Γ\{b} is the set of input symbols
5. q_o∈Q is the initial state
6.F⊆Q is the set of final or accepting states.
7. δ : Q\F x Γ →QxΓ x {L,R}is a partial function called the transition function, where L is left shift, R is right shift. (A relatively uncommon variant allows "no shift", say N, as a third element of the latter set.)

Anything that operates according to these specifications is a Turing machine.

Q={ A,B,C }
Γ={0,1}
b= -1 (“
q_0=A (initial State)
F=C
Process:
Turing Machine1: Calculating load degree

$$Load\ Degree = \sum_{k=0}^{m} \alpha i\ Fi$$

E) Mathematical Model

1.Compute Load Degree

Inputs:

The static parameters include the number of CPU's, the CPU processing speeds, the memory size, etc. Dynamic parameters are the memory utilization ratio, the CPU utilization ratio, the network bandwidth.

Process:-

1. Define a load parameter set: F= {F1,F2…Fm} with each Firepresents the total number of the parameters.
2.Compute the load degree as
Load Degree(N)= ∑ αiFi
Where i= 1…m
3. Average cloud partition degree from the node load degree statistics as:
Load degreeavg=∑i=1..n LoadDegree(Ni)

4. Three level node status are defined
Load_degree(N)=0 for***Idle***
0<Load_Degree(N)<Load_Degree(N)high for ***Normal***
Load_Degree(N)high <= Load_Degree(N) for ***Overloaded***

Output :-

Idle or Normal Or Overloaded

2.NoN Cooperative load balancing game

Input:-

Sji be the fraction of jobs that user j send to computer i
The vector sji =(Sj1,Sj2……Sjn) is called the *load balancing strategy* of user j.
The vector Sj = (S1,S2, ….Sm) is called the *strategy profile* of the load balancing game

Process :-

1. The expected response time at computer I is
Fi(S) = 1/ μi-∑ j=1 to m sjiϕk
2. The overall expected response time of user j is given by
Dj(S)= ∑i= 1 to n sji Fi(S) = ∑i=1 to n sji/μi-∑k=1 To m skiϕk
3. The goal of user j is to find a feasible load balancing strategy S ji such that Dj(S)is minimized.

Output:-

The decision of user j depends on the load balancing decisions of other users since Dj(S) is a function of S

Turing Machine 2:Get Avg. Load Degree & Classify Load

Q={ A,B,C,D,E } i.e. C= idle, D= Normal, E= Overloaded
Γ={0,1}
b= -1 ("
q_0=A (initial State)
F=C

a) Get Avg. Load Degree

$\text{Load\_degree}_{avg} = \frac{\sum_{i=1}^{n} load\_degree(Ni)}{n}$

b) Classify load

- **Idle**
  Load_degree(n)=0
- **Normal**
  0<load_degree(n)<=$\text{load\_degree}_{high}$
- **Overload**
  $\text{load\_degree}_{high}$<= load_degree(n)

The above Turing machines works for calculating load degree & classifies loads in three states ie. Idle, normal, overloaded. It just makes the summation of all static & dynamic weights of processes. Since machine has single

Finite State & No other halting condition this problem comes under NP type problem. Also, It gives result in polynomial time, So, it is P type.

2. Round Robin Scheduling

Assignment of exact resource to incoming request is **NP Hard** Problem. Round Robin Scheduling Algorithm makes slices of request to minimize overloads.

We assign the computing resource based on their load degrees, So problem gets reduced to **NP-Complete**.

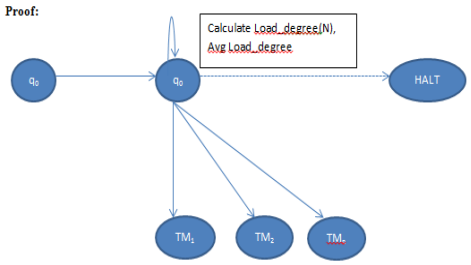

Fig.3 Processing through Turing Machine

IV. Advantages

- 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

V. Conclusion

- Cloud computing system has widely been adopted by the industry though there are many existing issues like load balancing, migration of virtual machine, server unification which have been not yet fully addressed.
- Load balancing is the most central issue in the system to distribute load in efficient manner. It also ensures that every computing resource is distributed efficiently and fairly.
- Existing load balancing technique have been studied mainly focus on reducing overhead, reducing migration time and improving performance.[9]

References

[1] Gaochao Xu, Junjie Pang, and Xiaodong Fu,*A LoadBalancing Model Based on Cloud Partitioningfor the Public Cloud*, IEEE TRANSACTIONS ON CLOUD COMPUTING YEAR 2013

[2] K. Nishant, P. Sharma, V. Krishna, C. Gupta, K. P. Singh,N. Nitin, and R. Rastogi, *Load balancing of nodesin cloud using ant colony optimization*, in Proc. 14thInternational Conference on Computer Modelling andSimulation (UKSim), Cambridgeshire, United Kingdom,Mar. 2012, pp. 28-30

[3] M. Randles, D. Lamb, and A. Taleb-Bendiab, *Acomparative study into distributed load balancingalgorithms for cloud computing*, in Proc. IEEE 24thInternational Conference on Advanced InformationNetworking and Applications, Perth, Australia, 2010,pp. 551-556

[4] Ms. Parin V. Patel, Mr. Hitesh. D. Patel, Pinal. J. Patel,*A Survey On Load Balancing In Cloud Computing*, International Journal of Engineering Research & Technology (IJERT)Vol. 1 Issue 9, November- 2012ISSN: 2278-0181

[5] Microsoft Academic Research, Cloud computing,http://libra.msra.cn/Keyword/6051/cloud-computing?query= cloud%20computing, 2012

[6] R. Hunter, The why of cloud, http://www.gartner.com/DisplayDocument?doc cd=226469&ref= g noreg, 2012.

[7] N. G. Shivaratri P. Krueger, and M. Singhal, Loaddistributing for locally distributed systems, Computer,vol. 25, no. 12, pp. 33-44, Dec. 1992. B. Adler, Load balancing in the cloud: Tools, tips and

[8] Techniques http://www.rightscale. com/infocenter/whitepapers/Load-Balancing-in-the-Cloud.pdf,2012

[9] D. MacVittie, *Intro to load balancing for developers The algorithms*, https://devcentral.f5.com/blogs/us/introto-load-balancing-for-developers-ndash-the-algorithms,2012

[10] Anthony T.Velte, Toby J.Velte, Robert Elsenpeter, *Cloud Computing A Practical Approach* TATA McGRAW-HILL Edition 2010.

[11] D. Grosu, A. T. Chronopoulos, and M. Y. Leung, *Loadbalancing in distributed systems: An approach usingcooperative games*, in Proc. 16th IEEE Intl. Parallel andDistributed Processing Symp., Florida, USA, Apr. 2002, PP. 52-61