ISSN:0975-9646

Load Degree Calculation for the Public Cloud based on Cloud Partitioning Model using Turnaround Time

Priti Singh, Pankaj Sharma

- *Priti Singh, pursuing masters of technology degree program in computer science from G.B.T.U. India.*
- *Pankaj Sharma, Sr. Asstt. Prof., Department of Information Technology, ABES Engineering College, Ghaziabad, India.*

Abstract—"Cloud Computing" is a growing technology and is becoming popular because of its great features as it provides almost everything- hardware, software, infrastructure and platform as a service. Clients are becoming more demanding and expect better service, so LOAD BALANCING has become an inevitable requirement for the Cloud Service Providers. In this paper we have proposed a different solution to calculate Load degree of a node in the public cloud based on the Turn Around Time so that the Load Balancers can improve the Load Balancing strategy of the load balancing model in the public cloud.

Index Terms—Load Balancing, Load Balancing in Cloud, Load Balancing in Public Cloud, Load Degree Calculation in Cloud, Load Degree Calculation using Turn Around Time in Cloud.

1 Introduction

The term "Cloud Computing" is all about providing the computing efficiency to the users in terms of software, platform and infrastructure through the internetworked connection, without the hassle of knowing the details. NIST gave a definition of cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources such as networks, servers, storage, applications and services, that can be rapidly provisioned and released with minimal management effort or service provider interaction[1]. A Cloud comprises of 3 basic elements: clients, data-center, and distributed servers. Each element has a definite purpose and plays a specific role. Based on the domain or environment in which clouds are used, clouds are of three types -Public Clouds, Private Clouds and Hybrid Clouds (combination of both private and public clouds). Cloud is related to virtualization, because using virtualization an end user can use different services of a cloud. Virtualization means something which isn't real, but gives all the facilities of a real. It is the software implementation of a computer which will execute different programs like a real machine. The remote data-center will provide different services in a full or partial virtualized manner.

2 Load Balancing

Load balancing is a process of reassigning the jobs 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[2]. The purpose of load balancing is to enhance the efficiency of the system by distributing the work load among the available resources such as network link, central processing unit, disk drive etc. to achieve optimal resource utilization, maximum throughput and avoiding overload. Load Balancing algorithms are classified into two major types-static and dynamic. Static algorithms are mostly suitable for homogeneous and stable environments and can produce very good results in these environments. However, they are usually not flexible and cannot match the dynamic changes to the attributes during the execution time. Dynamic algorithms are more flexible and take into consideration different types of attributes in the system both prior to and during run-time[3].

3 Related Work

The load balancing model described by Gaochao Xu [4] in the article "**A Load Balancing Model Based on Cloud Partitioning for the Public Cloud**" is modeled at the public cloud which has multiple nodes with different computing resources at different locations. This model distributes the public cloud into a number of partitions, which are referred as "cloud partitions". A cloud partition is a subarea of the public cloud with divisions based on the geographic locations[4].

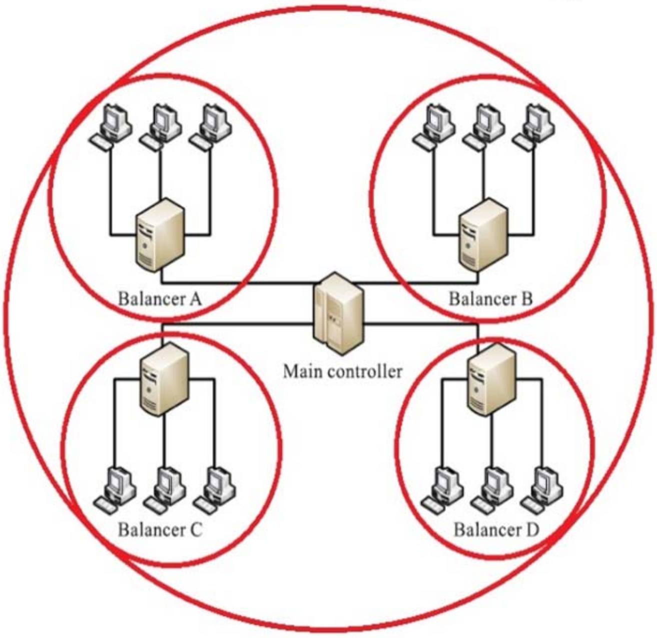

Fig. 1.Public cloud with main controller and balancers in the partitions.

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 balance solution is done by the main controller and the balancers[4].

As described in paper [4] the cloud partition load status can be divided into three types:

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

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

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

The parameters X, Y, and Z are set by the cloud partition balancers.

Job Assignment Strategy:

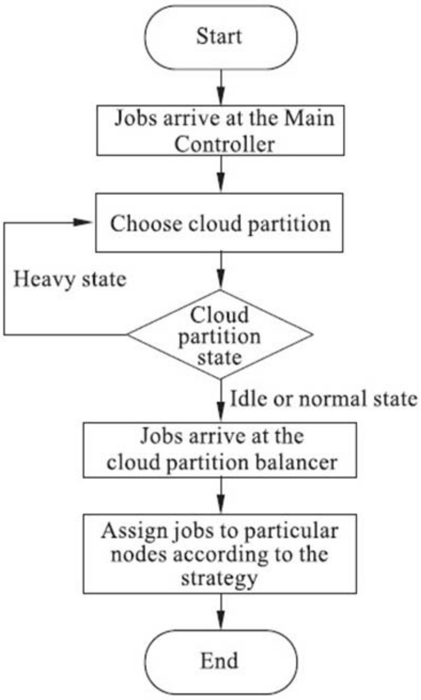
Start
Jobs arrive at the Main Controller
Choose cloud partition
Heavy state
Cloud partition state
Idle or normal state
Jobs arrive at the cloud partition balancer
Assign jobs to particular nodes according to the strategy
End

Fig. 2.Job assignment strategy.

When job I arrives at the system, the main controller queries the cloud partition where job is located. If the load status of this partition balancer is idle or normal, the job is handled locally. If not, another cloud partition is found that is not overloaded[4].

Load Degree Calculation:

The Load Status of a Cloud Partition is dependent on the load degrees of the individual nodes in that partition[4].

Step 1: Define a load parameter set: F = {F1, F2,...Fm}, with each Fi (1 <= i<= m; Fi $\in$ [0; 1]) parameter being either static or dynamic, m represents the total number of the parameters.

Step 2: Compute the load degree as:

Load Degree(N) =$\sum_{i=0}^{m} \alpha_i F_i$

whereα_i are weights that may differ for different kinds of jobs. N represents the current node.

Step 3: Define evaluation benchmarks. Calculate the average cloud partition degree from the node load degree statistics as:

Load Degree$_{avg}$=$\left(\sum_{\mathrm{i=1}}^{\mathrm{m}} \text{Load\_Degree}(N_i) \right) /\mathrm{n}$

The bench mark Load degree$_{high}$ is then set for different situations based on the Loaddegree$_{avg}$.

Step 4: Three node load status levels are then defined as:

Idle When

Load Degree(N)=0;

there is no job being processed by this node so the status is charged to Idle.

Normal For

0< Load Degree(N)<=Load degree$_{high}$;

the node is normal and it can process other jobs.

Overloaded When

Load degree$_{high}$<Load degree(N);

the node is not available and cannot receive jobs until it returns to the normal state. The load degree results are input into the Load Status Tables created by the cloud partition balancers[4].

The load degree results are input into the Load Status Tables created by the cloud partition balancers. The table is then used by the balancers to calculate the partition status. Each partition status has a different load balancing solution. When a job arrives at a cloud partition, the balancer assigns the job to the nodes with low load degrees[4].

4 Proposed Work

With the increasing popularity of cloud computing, the amount of processing that is being done in the clouds is increasing drastically. A cloud is constituted by various nodes which perform computation according to the requests of the clients. As the requests of the clients can be random to the nodes they can vary in quantity and thus the load on each node can also vary. This phenomenon can drastically reduce the working efficiency of the cloud as some nodes which are overloaded will have a higher task completion time compared to the corresponding time taken on an under loaded node in the same cloud .Since the job arrival pattern is not predictable and the capacities of each node in the cloud differ, for load balancing problem, workload control / Load balancing is crucial to improve system performance and maintain stability[5].

In the current model[4], the node load degree is related to various static parameters and dynamic parameters. 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, etc.

The load degree of a node is calculated by the load balancer based on different parameters and their weights which may differ for individual node, calculation of the load degree is a time taking and complex task for a balancer, since the nodes may be heterogeneous, in different networks, having different workloads with diverse resources. So, we propose a solution for calculating the load degree of a node by using the "Turn Around Time" of the last process in the process queue of a node using FCFS Scheduling, instead of taking different parameters and their weights. This calculation of the TAT will be done by the node itself. The TAT of the node will be directly proportional to the load degree of a node. Theoretically, the TAT can be best calculated by the node itself. The nodes can then update the Load Status Table of the balancers, which eventually will decrease the workload from the balancers. Also the overloaded nodes can distribute their load to sub nodes which are in Normal or Idle state, so that there are no overloaded nodes.

Calculating the load degree of a node:

Step 1: Compute the load degree of a node N as:
Loaddegree (N) = TAT (P)
where P is the last process in the processing queue of the Node N. The Turn Around Time (TAT) of the process is calculated using FCFS Scheduling.
Step 2: Define evaluation benchmarks. Calculate the average cloud partition degree from the node load degree statistics as:

$$\text{Load Degree}_{avg}=\left(\sum_{i=1}^{m}\text{Load\_Degree}(N_i)\right)/n$$

The bench mark Load degree$_{high}$ is then set for different situations based on the Loaddegree$_{avg}$.

Assigning jobs to nodes in the cloud partition

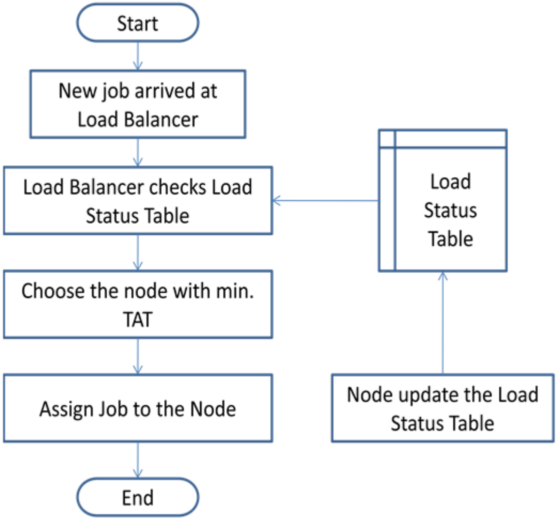
Start
New job arrived at Load Balancer
Load Balancer checks Load Status Table
Choose the node with min. TAT
Assign Job to the Node
End
Load Status Table
Node update the Load Status Table

Fig. 3.Jobs assignment to nodes in cloud partition.

When a new job arrives at the load balancer, it will check its Load Status Table. Load Status Table will have information about the Turn Around Time of each node. The Load Status Table is sorted periodically by the Load balancer in ascending order of Turn Around Time of the nodes, placing the node with the shortest TAT on the top. When the Job arrives at the system, it is assigned to the node on the top of the table. Every node periodically updates the Load Status Table with their TAT status.

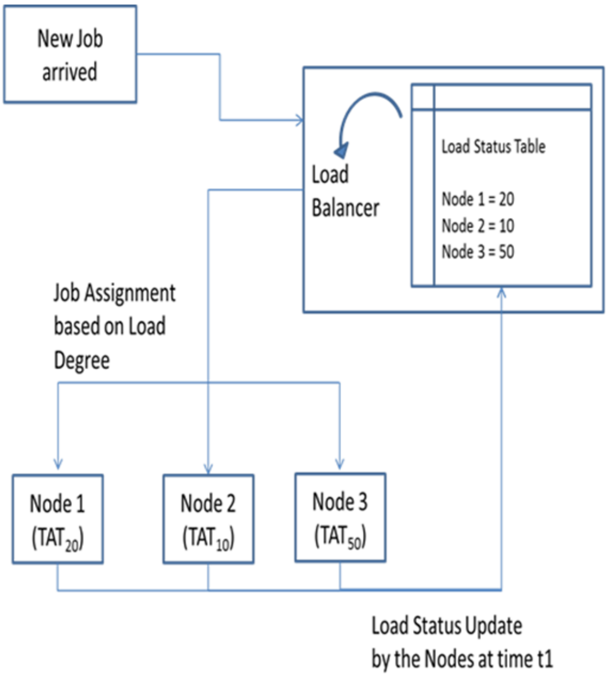

Fig. 4.Nodes updating the load status table.

5 Conclusion

In this paper we have provided a different solution to calculate the Load Degree of a node, which theoretically provides a more accurate calculation of Load Degree of a node in order to improve the Load Balancing strategy of the load balancing model described by Gaochao Xu [4] in the article "**A Load Balancing Model Based on Cloud Partitioning for the Public Cloud**". This model is still a conceptual framework and more work is needed to implement this framework.

References

[1] P. Mell and T. Grance, "The NIST definition of cloud computing", http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf, 2012.

[2] Kaushal Kulkarni , Sayali Mahajan, AshutoshKatkar, AkshayWalvekar,"Load balancing" , 2012 International Conference on Education and e-Learning Innovations, 978-1-4673-2225-6/12/$31.00 ©2012 IEEE

[3] Rimal, B. Prasad, E. Choi and I. Lumb, "A taxonomy and survey of cloud computing systems." In proc. 5th International Joint Conference on INC, IMS and IDC, IEEE, 2009.

[4] Gaochao Xu , Junjie Pang and Xiaodong Fu, "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud" , TSINGHUA SCIENCE AND TECHNOLOGY ISSNl l1007-0214l l04/12l lpp34-39 Volume 18, Number 1, February 2013

[5] N. G. Shivaratri, P. Krueger, and M. Singhal, Load distributing for locally distributed systems, Computer, vol. 25, no. 12, pp. 33-44, Dec. 1992.

[6] A. khiyaita , M. zbakh , H. el bakkali , Dafir el kettani, "Load Balancing Cloud Computing : State of Art", 978-1-4673-1053-6/12/$31.00 ©2012 IEEE

[7] Mrs. SharadaPatil and Prof.Dr.ArpitaGopal , "Cluster performance evaluation using load balancing algorithm" .

[8] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al-Jaroodi," A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms", 2012 IEEE Second Symposium on Network Cloud Computing and Applications