IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 – 8791 (Impact Factor: 1.479)
www.ijreat.org

Enhanced Load Balancing Technique in Public Cloud

Sakthivelmurugan V[1], Saraswathi A[2], Shahana R[3]

[1, 2, 3]**Department of Information Technology, J.J College of Engineering and Technology, Trichy, Tamil Nadu, India**

Abstract

Load balancing in cloud computing is really a challenge now. It is a method for distributing workloads across system. It aims to optimize resource use, maximize throughput, minimize response time, and avoid overload of any one of the resources. This article improves the performance of the load balancer using round robin algorithm and reduces the time taken for completion of the task using priority queue scheduling.

Keywords: *load balancing model, round robin algorithm, priority queue scheduling algorithm*

1. Introduction

Cloud computing is a fascinating technology in the field of information technology. It is efficient and scalable but maintaining the stability of processing so many jobs in the cloud computing environment is a very complex problem with load balancing receives much attention for researchers. Cloud Computing is a framework for enabling a suitable, on-demand network, access to a shared pool of computing resources (e.g. networks, servers, storage applications, and services). It is Internet based computing, where by shared resources, software and information are provided to computers and other devices on-demand, like a public utility [1].

Load balancing is a relatively new technique that facilitates networks and resources by providing a maximum throughput with minimum response time [2]. Dividing the traffic between servers, data can be sent and received without major delay. Different kinds of algorithms are available that helps traffic loaded between available servers [3]. A basic example of load balancing in our daily life can be related to websites. Without load balancing, users could experience delays, timeouts and possible long system response.

The model has a main controller and balancer to gather and analyze the information. For load balancing problem work load control is vital to improve system performance and maintain stability.
Load balancing model given in this article is pointed at public cloud which has a collection of nodes with centralized main controller. Thus, this model handles different load balancer to assign a task to the efficient node to reduce the response time. The cloud has main controller that chooses the suitable load balancer for arriving job while the balancer chooses the best node using best algorithm.

2. Related works

There are many load balancing algorithm such as Connection Mechanism, Randomized Algorithm, Equally Spread Current Execution Algorithm, Throttled Load Balancing Algorithm, Min- Min algorithm, Max-Min algorithm, Round Robin, Priority Scheduling Algorithm.

Load Balancing Algorithm [4] can be based on least Connection Mechanism which is a part of dynamic scheduling algorithm it needs to count the number of connection for each server dynamically to estimate load. Randomized Algorithm [5] a process can be handled by a particular node n with a probability p. The process allocation order is maintained for each processor independent of allocation from remote processor.

Equally Spread Current Execution Algorithm [6] process handles with priorities. It distribute the load randomly by checking the size and transfer the load to that virtual machine which is lightly loaded or handle that task easy and take less time, and give maximum throughput.

Throttled algorithm [6] is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine which access the load easily and perform the operations which is given by the client or user.

Min – Min Algorithm [7] begins with a set of all unassigned tasks. First, minimum completion time for all task is found. Then among this minimum times the minimum value is selected which is the minimum time among all the tasks on any resources. Then according to that minimum time, the task is scheduled on the corresponding machine. Then the execution time for all other task is updated on that machine by adding the execution time of the assigned task to the execution time of other task on that machine and assigned task is removed from the list of the task that are to be assigned to the machine.

Max – Min [7] is almost same as the Min – Min algorithm except the following: after finding out minimum execution times, the maximum value is selected which is the maximum time among all the tasks on any resources. Then according to that maximum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines.

Round Robin [5] processes are divided between all processors. Each process is assigned to the processor in a round robin order. The process allocation order is maintained locally independent of the allocation from remote processors. Though the workload distributions between processors are equal but the job processing times for different processes are not same. Priority Queue Scheduling [8] is the process of sorting nodes in the efficient order then assigns the task to the sorted order of the node.

3. System Model

There are many cloud computing categories with this work focused on a public cloud. A public cloud is based on the standard cloud computing model, with service provided by a service provider [9]. Large public cloud will include many nodes and nodes in different geographical locations. It is a set of computers and computer network resources based on the standard cloud computing model in which service provider makes resources, such as applications and storage, available to the general public over the internet. In public cloud, load balancing is a computer networking for distributing workloads across multiple computing resources, such as computers, a computer cluster, network links, center processing units or disk drives. The architecture diagram is shown in figure 1.

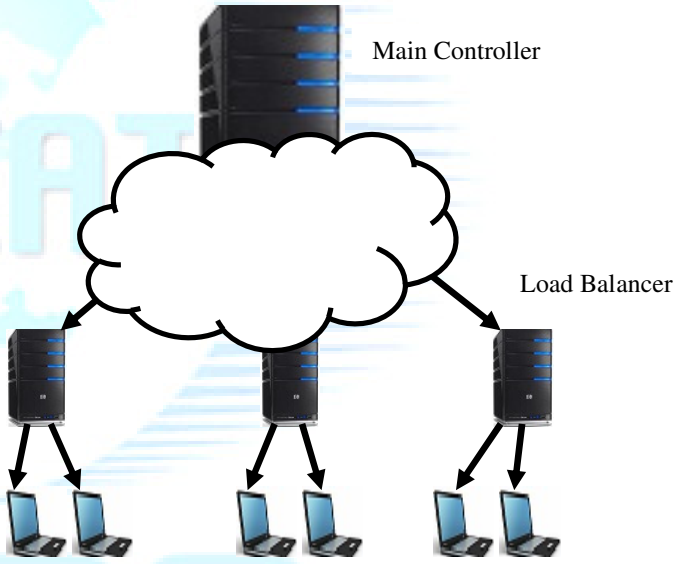

Fig. 1 System Architecture

The main controller deciding which load balancer should receive the job in public cloud. The load balancer then decides how to assign the jobs to the nodes. The main controller uses Round Robin algorithm to monitor the performance of all the load balancer then the load balancer uses priority queue scheduling to delegate the task to the efficient node based upon their priority.

3.1 Main controller and Load Balancer

The load balance problem is rectified by the use of main controller and best load balancing strategy. The main controller first assigns job to the suitable load balancer is possible. Since the main controller deals with information for each load balancer, smaller data set will leads to the higher processing rate. The load balancer gathers the efficiency information from every node and then chooses right strategy to distribute the job.

Load balancing is the process of improving the performance of the system by shifting of workload among the processors. Workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine. Load balancing is done so that every virtual machine in the cloud system does the same amount of work throughout therefore increasing the throughput and minimizing the response time.

Load balancing is one of the important factors to heighten the working performance of the cloud service provider. Balancing the load of virtual machines uniformly means that anyone of the available machine is not idle or partially loaded while others are heavily loaded.

3.2 Assigning jobs to the node

When a job arrives at the main controller, the first is to choose the load balancer.

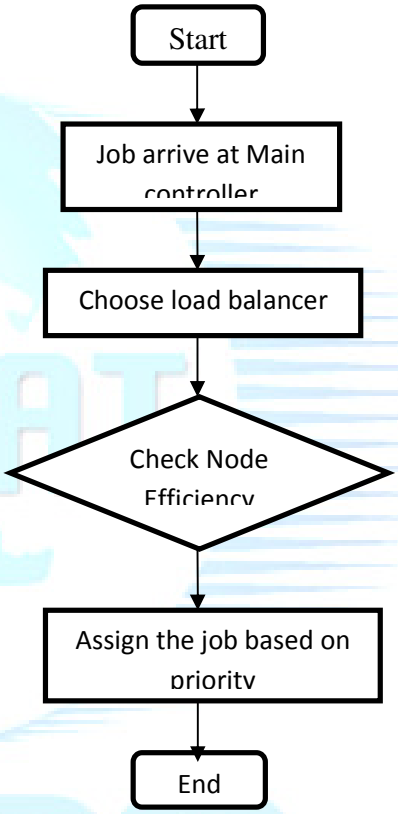

Fig. 2 job assignment strategy

The main controller has to communicate with the load balancer frequently to find the efficiency among the load balancer using Round Robin Algorithm.

Round Robin Scheduling Algorithm [5] follows:

- Time slice or time quantum is set at two seconds for each load balancer process before the next load balancer get controller of the CPU.

- Preemption is added between processes, if a load balancer process exceeds a time slice It is preempted and put back on the ready queue.
- Time slice setting may affect performance, if time slice is too large degenerates to FCFS (First Come First Serve) and can result in the convey effect where one process keeps the CPU until it releases.

The load balancer chooses the efficient node to assign the job based upon their Priority Queue Scheduling.

Priority Queue Scheduling algorithm follows:

- Insert_with_priority: add an node to the queue with an associated priority.
- Pull_highest_priority_node: remove the node from the queue that has the highest priority, and return it.
- Pull_lowest_priority_node: inspecting the first few highest or lowest priority node, clearing the queue, clearing subset of the queue, performing a batch insert, merging two or more queues in to one, incrementing priority of any node etc.,.

Advantages of the Round Robin include its simplicity and strict "First Come First Serve" nature. It means the each load balancer have equal chances to utilize the CPU.

4. Future Work

Find other best load balancer strategy with low cost and also to increase the efficiency still more.

References

[1] Manisha B. Jadhav, Vishnu J. Gaikwad, C. V. Patil, G. S.Deshpande,2010. CLOUD COMPUTING APPLICATIONS IN COMPUTATIONAL SCIENCE, International Journal of Advanced Computer and Mathematical Sciences.. Vol 1, Issue 1. Pp 1-6
[2] R. Shimonski. Windows 2000 & Windows Server 2003 Clustering and Load Balancing. Emeryville. McGraw-Hill Professional Publishing, CA, USA (2003), p 2, 2003
[3] A. Brian. Load Balancing in the Cloud: Tools, Tips, and Techniques. A Technical white paper in Solutions Architect, Right Scale.
[4] P.Warstein, H.Situ and Z.Huang(2010), "Load balancing in a cluster computer" In proceeding of the seventh International Conference on Parallel and Distributed Computing, Applications and Technologies, IEEE.
[5] Zhong Xu, Rong Huang,(2009)"Performance Study of Load Balanacing Algorithms in Distributed Web Server Systems", CS213 Parallel and Distributed Processing Project Report.
[6] Ms.NITIKA, Ms.SHAVETA, Mr. GAURAV RAJ; "Comparative Analysis of Load Balancing Algorithms in Cloud Computing", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May 2012.
[7] T. Kokilavani J.J. College of Engineering & Technology and Research Scholar, Bharathiar University, Tamil Nadu, India" Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing" International Journal of Computer Applications (0975 – 8887) Volume 20– No.2, April 2011.
[8] http://www.wikipedia.com/priority queue scheduling/

[9]A.Rouse,publiccloud,http://searchcloudcomputing.techtarget.com/definition/public-cloud.2012