International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 16 (2014), pp. 1605-1610

http://www. irphouse.com

A Load Balancing Model Based on Cloud Partitioning for the Public Cloud

Azizkhan F Pathan[1], S. B. Mallikarjuna[2]

M. Tech (Dept. of CS&E),
Bapuji Institute of Engineering And Technology, Davangere, Karnataka, India.
[2]Dept. of CS&E,
Bapuji Institute of Engineering And Technology, Davangere, Karnataka, India

Abstract

The increasing cloud computing services offer great opportunities for clients to find the maximum service and finest pricing, which however raises new challenges on how to select the best service out of the huge group. Cloud computing employs a variety of computing resources to facilitate the execution of large-scale tasks. Therefore, to select appropriate node for executing a task is able to enhance the performance of large-scale cloud computing environment. Also the numbers of users accessing the cloud are rising day by day. As the cloud is made up of datacenters; which are very much powerful to handle large numbers of users still then the essentiality of load balancing is vital. 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 Job Seekers Web Portal based on the cloud partitioning concept in which the jobs are partitioned based on the arrival date and the Main Controller (Admin) balances the load.

Keywords: load balancing model; public cloud; cloud partition; game theory**.**

Introduction

Cloud computing is an attracting technology in the field of computer science. In Gartner's report [1], it says that the cloud will bring changes to the IT industry. The cloud is changing our life by providing users with new types of services. Users get service from a cloud without paying attention to the details [2]. 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 (e. g.,

networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [3].

The five important characteristics of cloud computing defined by NIST includes *On-demand self-service, Global network access, Distributed resource pooling, Scalable, Measured service*[4].

Based on the domain or environment in which clouds are used, clouds can be divided into 3 categories Public Clouds (which are available to clients from a third party service provider via the Internet. e. g., Amazon, Google and IBM), Private Clouds(a business essentially turns its IT environment into a cloud and uses it to deliver services to the users), Hybrid Clouds(Hybrid cloud means either two separate clouds joined together (public, private, internal or external or a combination of virtualized cloud server instances used together with real physical hardware)[5].

In the whole, cloud computing provides us the attracting conventional services like : Software as a Service (SaaS), where the user uses different software applications from different servers through the Internet which can be consumed using web browsers without purchasing or maintaining overhead. Some examples are Gmail, Salesforce. com, etc. Platform as a Service (PaaS), where Application development framework offered as a service to developers for quick deployment of their code. Some examples for PaaS include Google App Engine, Heroku, Cloud Foundry, etc. Last but not the least Infrastructure as a Service (IaaS) where Shared infrastructure such as servers, storage and network are delivered as a service over the internet. Some examples include Amazon Web Services, Rackspace Cloud, etc[6].

Cloud computing architectures are inherently parallel, distributed and serve the needs of multiple clients in different scenarios. This distributed architecture deploys resources distributively to deliver services efficiently to users in different geographical locations [7]. Clients in a distributed environment generate request randomly in any processor. So the major drawback of this randomness is associated with task assignment. The unequal task assignment to the processor creates imbalance i. e., some of the processors are overloaded and some of them are under loaded [8].

The objective of load balancing is to achieve a high user satisfaction and resource utilization ratio, and to avoid the situation where nodes are either heavily loaded or under loaded in the network, hence improving the overall performance of the system. Proper load balancing can help in utilizing the available resources optimally, thereby minimizing the resource consumption [9].

We have seen the problems in Job Seeker’s Web Portal’s. i. e., they contain information about all the jobs. They contain outdated jobs, currently running jobs and also upcoming jobs. So it will be difficult for the Job Seeker’s to search for a job since it shows all the jobs. So in order to avoid this problem we are proposing a new load balance model for the Job Seeker’s portal which is based on cloud partitioning concept.

System Architecture

There are several 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 [10]. A large public 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 Figure 1.

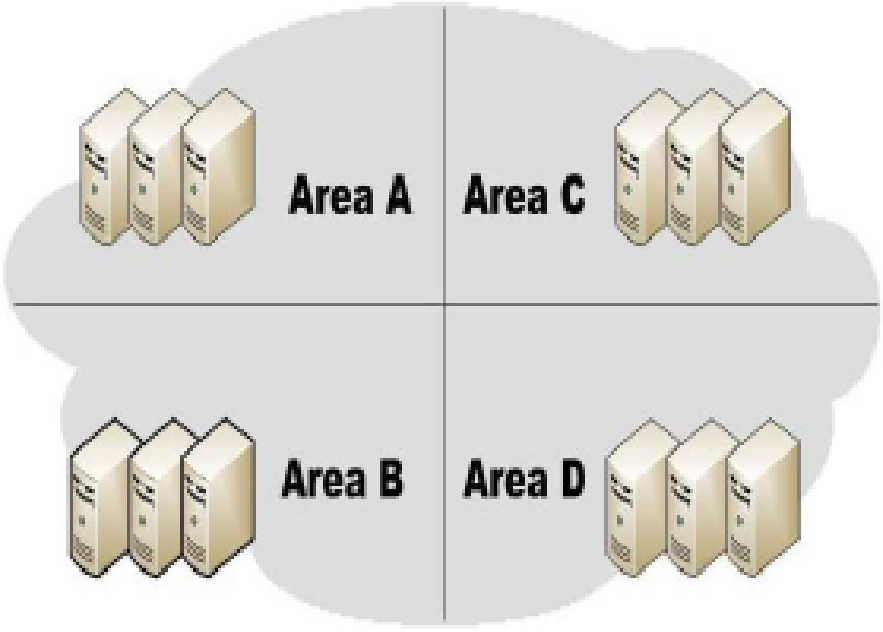

Figure 1 – Typical Cloud Partitioning.

When job i arrives at the system, the main controller (Admin) decides to which partition the job should be assigned. If this is the last updated job, then the job is assigned to Partition1. If it is an upcoming job, then it is assigned to Partition2. If it's a currently running job then it is assigned to Partition3. If it is an outdated job then it is assigned to Partition4. The Best Partition Searching algorithm is shown in Algorithm 1.

Algorithm 1 Best Partition Searching

```
begin
while job do
 searchBestPartition (job);
 if Update(job) then
Send Job to Partition1;
else if EndDate > CurrentDate then
 Send Job to Partition2;
else if ArrivalDate<=CurrentDate && EndDate =CurrentDate then
Send Job to Partition3;
else if EndDate<CurrentDate then
```

Send Job to Partition4;
end if
end while
end

Load Balancing Algorithms

There are many simple load balancing algorithm methods such as the First Come First Served (FCFS), Round Robin algorithm, Equally spread current execution algorithm and Throttled algorithm. The FCFS and Throttled algorithms are used here for their simplicity and also they provide good response time compared to other algorithms.

First Come First Serve

FIRST COME FIRST SERVE ALGORITHM

Main Controller (Admin) maintains an index table of job requests.
The job requests are stored in the table on the basis of their arrival time.
The Main Controller (Admin) scans the index table from top to bottom.
The first job request according to the arrival time is allocated the grant by the Main Controller (Admin).
The HR receives the response to the request sent and then posts jobs by providing details about the interview.
In this way all the jobs are processed in the first come first serve basis.

Throttled

THROTTLED ALGORITHM

The Main Controller (Admin) maintains an index table of job requests.
The job requests are stored in the table based on the arrival time.
The Main Controller (Admin) scans the index table from top to bottom.
The Main Controller (Admin) grants the permission to post jobs and changes the REQUEST_NEED flag to GRANTED.
The HR receives the response to the request sent and then posts jobs by providing details about the interview.
In this way only one job interview details is posted by a company at a time and if a Company HR wants to post another job then he should send job request again.

Experimental Results

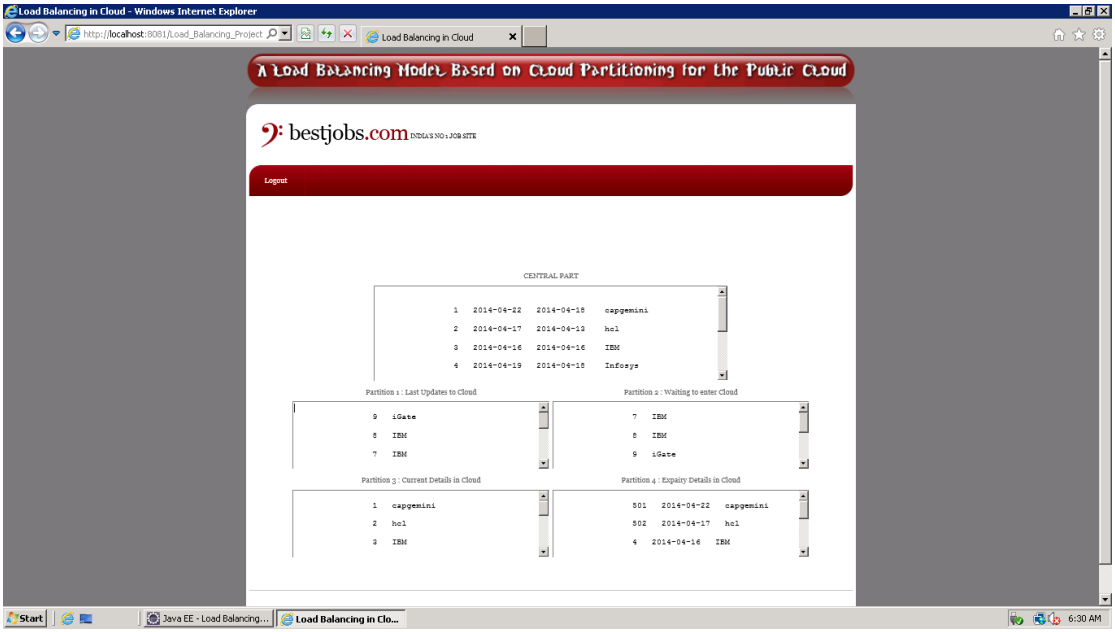

Figure 2-Results of Cloud Partitioning for the Job Seeker's Portal.

We can see from the figure 2 that how the jobs are partitioned based on arrival date. The first partition contains the last updated jobs. The second partition contains the jobs which are waiting to enter cloud. The third partition contains the jobs which are currently running in cloud and the fourth partition contains the expired jobs.

The expired or outdated jobs will be saved in a table called balanced data table. The Main Controller (Admin) scans the balanced data table and deletes the outdated jobs. In this way the outdated jobs are deleted and the Job Seeker gets only upcoming jobs.

Conclusion

The response time and data transfer cost is a challenge of every engineer to develop the products that can increase the business performance and high customer satisfaction in the cloud based sector. The several strategies lack efficient scheduling and load balancing resource allocation techniques leading to increased operational cost and give less customer satisfaction. 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. In this paper we have proposed a better load balance model for the Job Seeker's Web Portal based on the cloud partitioning concept with a switch mechanism to choose different strategies for different situations. 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 based on arrival date. Thus with cloud partitioning concept it is possible to provide good load balancing and hence improving the overall performance of cloud environment and user satisfaction.

References

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

[2] M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, and A. Vakali, Cloud computing: Distributed internet computing for IT and scientific research, Internet Computing, vol. 13, no. 5, pp. 10-13, Sept.-Oct. 2009.

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

[4] P. Mell and T. Grance "The NIST definition of Cloud Computing" version 15. National Institute of Standards and Technology (NIST), Information Technology Laboratory (October 7, 2009).

[5] AlexaHuth and JamesCebula"The Basics of Cloud Computing", www. us-cert. gov/reading... /USCERT CloudComputing HuthCebula. ”

[6] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic, “Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility, Future Generation Computer Systems”, Volume 25, Number 6, Pages: 599-616, ISSN: 0167-739X, Elsevier Science, Amsterdam, The Netherlands, June 2009.

[7] M. D. Dikaiakos, G. Pallis, D. Katsa, P. Mehra, and A. Vakali, “Cloud Computing: Distributed Internet Computing for IT and Scientific Research”, IEEE Journal of Internet Computing, Vol. 13, No. 5, September/October 2009, pages 10-13.

[8] A. Khiyaita, H. EI Bakkli, M. Zbakh, Dafir EI Kettani, ” Load Balancing Cloud Computing: State Of Art”, 2010, IEEE.

[9] Ram Prassd Pandhy (107CS046), P Goutam Prasad rao (107CS039). “Load balancing in cloud computing system” Department of computer science and engineering National Institute of Technology Rourkela, Rourkela-769008, Orissa, India May-2011.

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