IJCAT - International Journal of Computing and Technology, Volume 1, Issue 11, December 2014
ISSN : 2348 - 6090
www.IJCAT.org

A Novel Approach for Efficient Load Balancing in Cloud Computing Environment by Using Partitioning

[1] P. Vijay Kumar, [2] R. Suresh

[1] M.Tech 2nd Year, Department of CSE, CREC
Tirupati, AP, India

[2] Professor & HOD, Department of CSE, CREC
Tirupati, AP, India

Abstract - Cloud computing is usually a novel trend emerging in IT environment together with huge requirements of infrastructure as well as resources. Load Balancing is a crucial part of cloud computing environment. The technique of load balancing is essential in cloud computing to further improve the efficiency from the cloud. Efficient load balancing strategy ensures efficient resource utilization by simply provisioning of resources to be able to cloud user's on-demand basis in pay-as-you-say-manner. Load Balancing could even assistance prioritizing users by making use of appropriate scheduling criteria. Good load balancing can make cloud computing extremely effective and improves user satisfaction. This particular paper gives a technique for balance the actual incoming load in cloud environment by looking into making partitions from the public cloud.

Keywords **- Outsourcing data storage, Dynamic environment, Mutual trust, CSP, Cheating detection module, Access control, TTP.**

1. Introduction

Cloud computing can be a novel paradigm during which computing resources for instance processing, memory, as well as storage usually are not physically present with the user's location. Instead, the service provider possesses and manages these kinds of resources, and also user accesses all of them through the Internet. Cloud computing is really a general term with regard to anything which involves delivering hosted services on the internet. These types of services are broadly split into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a- Service (PaaS) and Software-as-a-Service (SaaS) A cloud might be private or even public. A public cloud sells services in order to anyone on the net. A non-public cloud is usually a proprietary network or even a data center in which supplies hosted services with a limited number of people. Whenever a service provider uses public cloud resources to produce their particular private cloud, the end result is termed a virtual private cloud. Private or even public, the objective of cloud computing is usually to provide easy, scalable usage of computing resources also it services. As an example, Amazon Web Services let us users store personal data via their Simple Storage Service (S3) as well as perform computations upon stored data while using the Elastic Compute Cloud (EC2). This sort of computing gives several benefits for businesses which includes low initial capital investment, shorter start-up time for brand new services, lower maintenance as well as operation costs, higher utilization by means of virtualization, and simpler disaster recovery which make cloud computing an attractive option.

Reports declare that there are numerous benefits within shifting computing through the desktop towards the cloud. Load balancing is really a strategy of reassigning the entire load towards the individual nodes on the collective system to produce resource utilization effective in order to enhance the response time from the job, simultaneously removing the condition during which a few of the nodes are gone for good loaded even though some others tend to be under loaded. A load balancing algorithm which can be dynamic in nature will not think about the previous state or even behavior from the system, which is, this will depend about the present behavior from the system. The particular important facts to consider even though developing such algorithms are usually estimation of load, comparison of load, stability of

various system, performance associated with system, interaction involving the nodes, nature of work to become transferred, selecting associated with nodes and several other ones. This particular load considered could be with regards to CPU load, quantity of memory used, delay or even Network load.

1.1 Goals of Load Balancing

- 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

1.2 Categories of Cloud

Public cloud: The public cloud can be obtained towards the clients through the third party service provider through the internet. A service provider tends to make resources for example applications as well as storage offered to the general public on the internet. Public services could possibly be free or even offered on the pay per usage model.

Private cloud: Inwards private cloud the particular computing resources is actually fully specialized in a specific firm or organization. Not one other organization are able to use the infrastructure.

Hybrid Cloud: Hybrid cloud is really a composition of a couple of clouds that continues to be distinct however bound together, providing the advantages of multiple deployment models.

Cloud comprises of huge resources. Management of those resources needs proper planning as well as proper layout. Cloud computing is actually efficient as well as scalable however maintaining the stability of processing a lot of requests within the cloud computing is an extremely complex problem. Cloud computing is extremely significant. This improves the particular efficiency and satisfaction of cloud computing. Because there are tremendous improvement in traditional utilization of internet which means that uneven distribution associated with workload may appear, also it might cause some server overloaded along with other under loaded which often might cause server crash. Workload distribution problem within cloud computing is extremely crucial as well as complex task till today, since the obtain arrival pattern within the cloud just isn't predictable and also the convenience of different servers within the cloud differ. Within this paper we give a technique for manage the load in cloud environment utilizing the idea of load balancing as well as cloud partitioning, which usually makes simpler the load balancing concept.

2. Related Work

Load balancing is the procedure of distributing the particular load between various nodes of the distributed system to enhance both resource utilization as well as job response time. Load balancing makes sure that all of the processor within the system or even every node within the network will just about the equal amount of work on any instant of their time. Load balancing within cloud computing systems is often a challenge now. Load Balancing is completed by making use of load balancers exactly where each incoming request is redirected and is also transparent in order to client which makes all the request. According to predetermined parameters, for example availability or even current load, the particular load balancer makes use of numerous scheduling algorithm to find out which usually server need to handle as well as onwards the request about the selected server. Load balancing algorithms are usually two sorts depending on the cloud computing environment, whether it be static or dynamic:

1. Static load balancing algorithm
2. 2. Dynamic load balancing algorithm

Throughout static cloud environment, the particular cloud service provider loads all of the homogeneous resources. In this instance, the particular cloud will need to have prior information regarding the nodes capacity, memory, processing speed, performance from the nodes, etc. These kinds of requirements is not changed with the run time. Every one of the information in regards to the method is known within advance, as well as the load balancing technique is manufactured simply by load balancing algorithm from compile time.

Throughout dynamic cloud environment, the particular cloud service provider loads the actual heterogeneous resources. Here the particular cloud system cannot be determined by the prior information. The consumer requirement can change throughout the run-time. Load balancing isn't going to look at the previous state as well as behavior from the system; this will depend around the present behavior from the system. The actual important facts to consider while developing such algorithm are: estimation of load, comparison of load, stability of various system, performance associated with system, interaction

relating to the nodes, nature of work to become transferred, choosing of nodes, etc. Dynamic algorithm is actually implemented from run time.

At present, there is a lot of research underway in the field of load balancing. However, load balancing is not an easy task. There are a number of policies available for load balancing. Some of the policies are static, while others are dynamic. In spite of the many load balancing schemes available, if these load balancing schemes are not effectively used, it becomes impossible to accomplish the definition of cloud computing. This paper discusses how this can be done effectively and efficiently. The tasks provided by the user go through different levels. Only if load balancing is achieved at each and every level, will there be any benefit to the user. Data centres in association with a cloud computing system could be located anywhere in the world. The first priority involves selecting the correct data centre. If this is done effectively, we can say that almost 10% of the work is done and we are one step closer to reaching the definition of a cloud computing system. Similarly, load balancing schemes have to be correctly used in the remaining levels.

The decision regarding which algorithm to use at each level has to be made correctly. As mentioned before, there are lots of load balancing schemes, but their effectiveness varies across levels. Using a particular scheme in one level may be optimal, but using the same algorithm in another level may not be as effective as using an alternative technique. On inspection of each task, it becomes obvious that each node may not be equally capable of handling each task. Each node has its own capacity. Understanding the capacity of each node is necessary in assigning tasks to each node. However, if these assignments are not done in a controlled fashion, it may lead to an imbalance situation. Brokers that act between the user and the data centre should take this into consideration. Many companies provide cloud computing services. Satisfying the requirements of the user is a prime factor for every cloud provider. So, the basic objective of each cloud provider is to correctly complete the users task within the stipulated time. Thus it becomes a primary task for developers to make headway in this field.

The main focus is on the efficient utilization of the virtual machines and balancing the virtual machines with the incoming request. Load balancing is defined as a process of making effective resource utilization by reassigning the total load to the individual nodes of the collective system and thereby minimizing under or over utilization of the available resources or virtual machines. Have developed Modified Throttled algorithm which maintains an index table of virtual machines and also the state of VMs similar to the Throttled algorithm. There has been an attempt made to improve the response time and achieve efficient usage of available virtual machines. Priority determines the importance of the element with which it is associated. In terms of task scheduling, it determines the order of task scheduling based on the parameters undertaken for its computation. In the present framework, the deadline based tasks are prioritized on the basis of task deadline. The tasks with shorter deadline need to be executed first. So they are given more priority in scheduling sequence. The task list is rearranged with tasks arranged in ascending order of deadline in order to execute the task with minimum time constraint first. The cost based tasks are prioritized on the basis of task profit in descending order. This is appreciable as tasks with higher profit can be executed on minimum cost based machine to give maximum profit. 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.

3. Proposed System

Our proposed system implements the particular load balancing idea within the public cloud environment. The particular public cloud provides numerous nodes in various geographical locations. Within our system, we divide the public cloud directly into several partitions to handle as well as simplify the load balancing within cloud environment, which can be huge as well as complex. The actual partitioning will depend on the geographical locations. Therefore the cloud partition is actually subarea associated with public cloud and every cloud partition provides many nodes associated with particular geographical region.

The proposed model includes a Main Controller as well as Load Balancers. Main Controller selects the proper cloud partition with regard in order to processing the particular inward requests depending on the status from the cloud partition. The Load Balancers can be present in each and every cloud partition, which usually selects the load balancing strategy. In the event the request arrives within the cloud; the primary Controller chooses which usually cloud partition need to have the request. Then within the cloud partition the Load Balancer decides which usually nodes within the cloud partition need to process the particular requests. This particular decision is performed through the load balancing algorithm. The particular

loud partition might have three status depending on that the Main Controller chooses a specific partition**.**

IDLE: In this status, most of the nodes are in idle state.

NORMAL: In this status, some of the nodes are in idle status while some others are overloaded.

HEAVY: In this status, most of the nodes are overloaded.

The main controller will choose the partition that is idle or normal in status. If the cloud partition status is heavy, the Main Controller will forward the request to another cloud partition that has normal status.

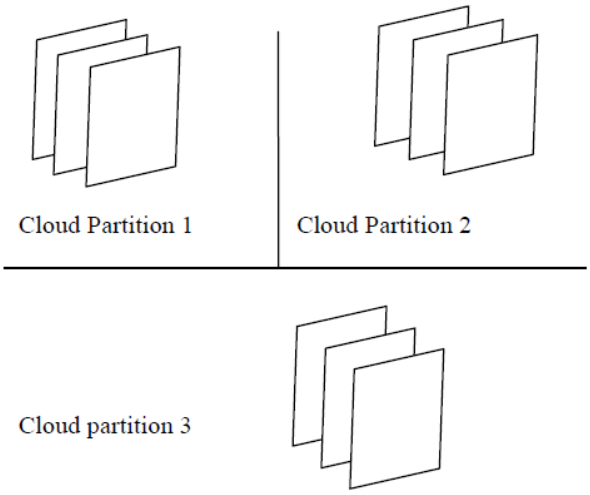


Figure 1: cloud Partition Model

Algorithm 1: Best Partition Algorithm

Begin
While job do
SearchBestPartition(request);
If PartitionState==idle||partitionstate==normal then
Send Job to Partition;
else
Search for another Partition;
End if
End while
End

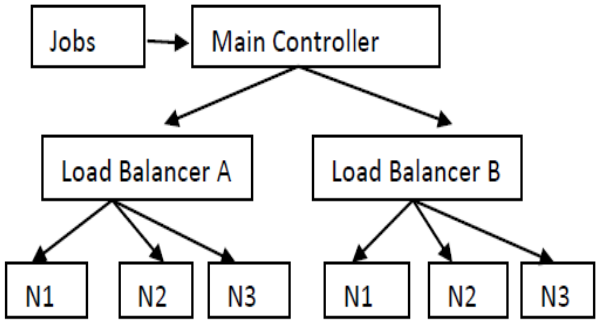


Figure 2.System Architecture

The Load Balancer gathers the load information through every node in order to evaluate the particular cloud partition status. The initial step is to locate the load level of each node. The load level of node relates to various static as well as dynamic parameters. The load status of the node could be:

IDLE: when LD (N) =0
NORMAL: 0<LD (N) <= High_LD
OVERLOADED: High_LD<=LD (N)

In which, LD (N) may be the load degree of node N. The Load Balancer will certainly select the node that may be either idle or even normal within status. After that, it's going to allocate the request to that particular node for more processing. The load degree results tend to be input towards the load status table developed by the Load Balancer. Each and every Balancer includes a load status table, which is often used in order to calculate the particular cloud partition status.

Algorithm 2: Request allocation to Node

Set counter=0;
Select node sequentially;
If counter<max load capacity of node
Assign request to node &&counter++;
else select next node
Check for node
If counter<max capacity of node
Assign request to node &&counter++;
If counter ==max capacity of node
Request assignment not possible;
Wait for some nodes to become free;
end if;
end else
end if

This particular algorithm explains that every node within the cloud partition are designed for approximately certain

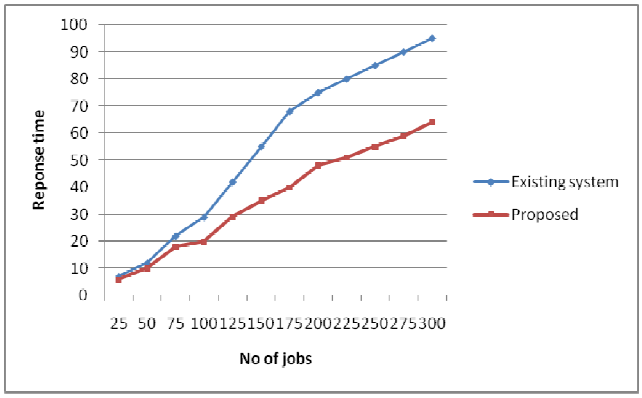
quantity of requests. The Load Balancer chooses the particular nodes sequentially. Each node is assigned to a counter. The Load Balancer will certainly check the node whether or not the counter is actually less than the absolute maximum variety of requests it may handle. Whether it is so, then it is going to assign the requests compared to that node. In the event the counter is equivalent to the absolute maximum variety of requests, then this Load Balancer is not going to allocate the requests with it and definately will look into the next node and performance analysis as shown in below figure 3.

4. Performance Analysis

The particular resulting load balancing model continues to be implemented along with a graph has become plotted. The particular graph shows the comparative performance with the response time associated with existing as well as overloaded load balancing model, while using the Y axis showing the consequence of improved response time with increased number of jobs throughout X axis. This specific graph demonstrates in which overloaded load balancing model performs properly as number of jobs increases. Because the number of jobs improves the proposed overloaded load balancing model gives minimum response time.

Table 1: comparative performance of response time

No of jobs	response time (existing model)	response time (Proposed Model)
25	7	6
50	12	10
75	22	18
100	29	20
125	42	29
150	55	35
175	68	40
200	75	48
225	80	51
250	85	55
275	90	59
300	95	64



Figuer 3: Comparative performance of the response time of existing and overloaded load balancing model

5. Conclusion and Future Work

Load Balancing is a vital task within Cloud Computing environment to attain maximum usage of resources. Various works continues to be completed in these studies area to improve the performance as well as efficiency from the cloud system. Our proposed system provides a simple approach associated with load balancing within the cloud while using the thought of cloud partitioning. However, more work has to be carried out this area. In the future study, some other load balancing algorithms can be located out, simply because other algorithms can provide much better load balancing results. Numerous load balancing algorithms should be compared. Extremely effective algorithms needs to be designed in order that the distribution from the load on the list of nodes within the cloud partition will probably be simpler and minimize enough time complexity. Cloud division just isn't an easy task. Several things have to be thought to be while making cloud partition, including the nodes in the cluster could possibly be still apart.

References

[1] Michael Vrable_, Stefan Savage, and Geoffrey M. Voelker,BlueSky: A Cloud-Backed File System for the Enterprise

[2] Tejinder Sharma, Vijay Kumar Banga, Efficient and Enhanced Algorithm in Cloud Computing, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3,Issue-1, March 2013

[3] Jasmin James, Dr. Bhupendra Verma, EFFICIENT VM LOAD BALANCING ALGORITHM FOR A CLOUD COMPUTING ENVIRONMENT, Jasmin James et al. / International Journal on Computer Science and Engineering (IJCSE)

[4] Nidhi Jain Kansal1, Inderveer Chana, "Cloud Load Balancing Techniques: A Step Towards Green

Computing", IJCSI International Journal of Computer Science, 2012.

[5] Shantanu Dutt, "New Faster Kernighan-Lin-Type Graph Partitioning Algorithms", IEEE ,1993

[6] Tarun Kumar Ghosh, Rajmohan Goswami, "Load Balanced Static Grid Scheduling Using Max-Min Heuristic" , 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, 2012.

[7] Zehua Zhang, Xuejie Zhang, "A Load balancing mechanism based on Ant colony and complex network theory in Open Cloud Computing Federation", 2nd International Conference on Industrial Mechanism and Automation,2010.

Authors

P. Vijay Kumar Completed my B.Tech in Shree institute of technical education, Pursuing M.Tech in Chadalawada Ramanamma Engineering college. His areas of interests include Image Processing, software Engineering, Object Oriented systems and Data Mining & Image Mining.

Prof. R. Suresh received B.E.degree from SVNIT(REC) in 1996 and M.Tech., degree from JNTU Hyderabad in 2001 and pursuing the Ph.D degree from JNTUA, Anantapuramu. From 1998 to Till date he is with the JNTUA, Anantapur working at different levels. His areas of interests includes Image Processing, software ingineering , Object Oriented systems and Data Mining & Image Mining. His Current interests include Texture model approach to Mammograms' classification and Detection.