International Journal of Advance Research in Science and Engineering
Vol. No.4, Issue 08, August 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

A REVIEW ON SLA AWARE LOAD BALANCING ALGORITHM USING JOIN-IDLE QUEUE IN CLOUD COMPUTING

Mehak Choudhary

Computer Science and Engineering, SKIET, Kurukshetra University, (India)

ABSTRACT

The emerging technology in the area of Information technology is Cloud Computing. Cloud Computing is the term associated with the virtualization, networking, software and services offered by web. The elements involved in cloud computing are clients, datacenter and distributed server. One of the main problems in cloud computing is load balancing. Balancing the load means to distribute the workload among several nodes evenly so that no single node will be overloaded. Load can be of any type that is it can be CPU load, memory capacity or network load. In this paper we presented an architecture of load balancing and algorithm which will further improve the load balancing problem by minimizing the response time while maintaining SLA. In addition to these for even distribution of load among virtual machine we use JIQ that will further minimize queue length of virtual machine.

Keywords: Cloud Computing, Load Balancing, SLA, JIQ

I. INTRODUCTION

Cloud computing is a computing paradigm that has changed the parallel and distributed computing system by widening the user's range by virtualization that is hardware and software infrastructure over internet. Cloud computing is also an economic medium to acquire and manage IT resources. Cloud computing also provides IT capabilities that are hardware, software and services from third party over network. Google Docs, Drop Box and Gmail are some examples who are using cloud computing concept.

Cloud computing are divided on the basis of two perspectives that are 1.Capable 2. Accessible. On the basis of capable perspective cloud provide three different types of services that are SaaS (Software as a Service) which is a way where applications are delivered over internet as a service instead of installing and maintaining the software we can simply access through internet. Some real time example of SaaS is Google Apps, MS office 365. PaaS (Platform as a Service) as from name it is clear that it will provide computing platforms which includes Operating System, execution environment of programming language, database, web server etc over internet. Example of PaaS is Google App Engine, Window Azure etc. IaaS (Infrastructure as a Service) it provides the infrastructure and physical resources and storage online. Example of IaaS is Amazon EC2, Google Compute Engine.

On the basis of accessible perspective cloud computing are of three types: 1. Public Cloud where resources are

provided as a service over the internet on the basis of pay-per use which ease the users because they don't need to install software or purchase hardware. Example Google App Engine 2.Private Cloud where resources are deployed and arranged within an organization.3.Hybrid Cloud which is combination of the clouds private and public. Cloud computing is very advantageous because of its scalability, ability to increase storage, throughput and it is also cost effective.

Load Balancing is the term which is very commonly used with cloud computing because it is one the issue of cloud computing. Balancing the load means to divide the load among various resources in any system evenly for effective utilization of resources and improving response time. Load balancing algorithm is basically categorized into two ways depending on current state of system. 1. Static Algorithm which is not current state dependent because it depends on the previous knowledge of system. 2. Dynamic Algorithm which is dependent on system's current state there is no need of previous knowledge of system.

Dynamic load balancing algorithms can be further divided into two ways that is distributed and non-distributed. In distributed algorithm, all the nodes execute load balancing algorithm. In non-distributed load balancing algorithm load balancing is done by either one node or group of nodes. Load Balancing are used for creating backup in case of any system failure, for performance improvement and increasing throughput.

There are many existing load balancing algorithm some of them are Round-Robin algorithm, which is a static approach for dividing the traffic equally because it has some limitations so a weighted Round-Robin algorithm was introduced in which particular weights were assigned to the server, then according to weight traffic will be assigned. Also there is Join-Idle-Queue algorithm in which initial balancing of load is done among idle processors and then jobs are assigned to the processor for minimizing the queue length at each processor. It then reduces the system's load.

SLA (Service Level Agreement) is the level of agreement between service provider and users on the basis of performance and availability. SLA metrics are created for all the services on the basis of CPU capacity, memory size, storage and boot time. There must be agreement among user and service provider in terms of their agreement or disagreement for agreement or disagreement. So, load balancing algorithms are used to maintain SLA.

II. RELATED WORK

Velagapudi,Pratap.M and Mohammad Kemal[1] sated the concept of load balancing techniques in cloud computing by which dynamic work load is distributed among multiple nodes evenly through which there will be no overloading in a single node and improvement in performance and resource utilization. This paper discusses some existing load balancing algorithms which are classified into two categories static algorithm and dynamic algorithm. Static Algorithm divides the traffic equally among servers and also not dependent on current state because it is dependent on previous knowledge of system. Dynamic Algorithm selects lightest server among whole server for traffic load balancing. Here it is dependent on current state only. Many existing algorithms are reviewed with their performance comparison. Load balancing algorithms which are discussed and compared on the basis of their performance are Round-Robin algorithm, Weighted Round Robin algorithm, Join-Idle queue etc.

Yashpalsingh Jadeja and Kirat Mali [2] discussed the concept of cloud computing and architecture of cloud computing is also explained. Earlier the concept of parallel computing and distributed computing was used

commonly, after that grid computing came into existence and now cloud computing is recent trend in IT. Cloud computing is the sharing of resources without paying for the installation, infrastructure and manpower. With cloud computing distributed resources is used properly. Cloud concept uses the concept of virtualization, interoperatability, quality of services and delivery models of cloud that is private, public and hybrid. Cloud computing uses the facility of pay-per-use of application per client. Paper also discusses the architecture that comprises of two parts frontend and backend where front end is client and backend is cloud that is internet. Cloud computing offers services that are used on the basis of pay-per-use. The services offered by it are SaaS (Software as a Service), Paas (Platform as a Service) and IaaS (Infrastructure as a Service). Cloud is of four types on the basis of location public, private, hybrid and community. Also advantages of cloud computing are discussed that it is easy to manage, manage disasters, green computing by energy conservation. Some issues still need to be discussed.

Muhammad Alhamad, Taram Dhillon and Elizabeth Chang [3] stated the design of SLA negotiation in cloud computing. Here strategies are discussed on the basis of which there will be agreement between user and service provider. Functional and non-functional requirements are scalable, available etc.SLA metrics are categorized into two categories performance metrics and business related metrics. Metrics for SLA in IaaS is CPU capacity, scale-up etc. Similarly for PaaS and SaaS are integration, scalability, reliability and usability. For negotiation between user and service provider first criteria is direct agreement that is done online. Other method is negotiation through trusted systems and also if there is more than one system. The results produced here can be used further as basic tool in load balancing systems.

Chung-Chang Li and Kuochen Wang [4] proposed tldlb (Two Level Decentralized Load Balancer) which is architecture for decentralized balancing of load and nn-dwrr (Neural Network Weighted Round Robin) algorithm for balancing the load. In this architecture there are two levels to balance the load among system thus two load balancers are used that are local load balancer and global load balancer. Local load balancer has two tasks that are initially to monitor the load in virtual machines in the particular virtual zone on the basis of metrics CPU, memory, network bandwidth and disk I/O utilization. Then the second task is to use nn-dwrr algorithm to balance the load among virtual machines. Global load balancers are used to exchange the load information of one virtual zone to another virtual zone. Both the load balancing is done while maintaining the SLA levels.

Yi Lua, Qiaomin Xiea, Gabriel Kliatb, Alan Gellerb, James R.Larusb, Albert Greenberge[5] proposed a algorithm for distributed load balancing for large systems. It basically inform dispatcher about the idle processors at the time of their idleness but if large number of dispatchers are informed then there will be queuing plus one dispatcher will waste there cycle at idle processor which will affect the response time poorly. For solving these problems JIQ uses two level load balancing scheme. At initial level, average queue length should be minimized and at second level on the basis of idle processors load is balanced at each dispatcher.JIQ helps in reducing the load on system and faster response time. In this algorithm an I-queue data structure is used which helps in communication between primary and secondary load balancing.

Akshay Jain, Anagha Yadav, Lohit Krishan and Jibi Abraham [6] proposed a model for improved load balanced environment by determining the overloaded hosts and choosing best virtual machine and destination host for migration dynamically. Here decision of migration of virtual machine is decided by calculating the mean of all the hosts which is called the threshold value. There are two steps involved in using this proposed model that is

firstly by determining overloaded host and secondly, determining the best virtual machine for migration from overloaded host and destination host for migration. Threshold value will decide the overloaded host. Determination of overloaded host is decided by the hosts which are closer to threshold band value. The destination host are decided by the hosts which are balanced that is which are inside the threshold band are balanced so overloaded host can be migrated there. This algorithm results in 31.75% less time for execution of load then the naïve algorithms.

III. PROBLEM FORMULATED

For balancing the load among Cloud computing architecture there must be load balanced among data centers and virtual machines for throughput maximization. Currently the balancing of load has some problems that are:

- Chung-Chang Li and Kuochen Wang [4] proposed a two-level decentralized load balancer (tldlb) model for balancing the load among virtual machines. Here neural-network weighted round robin (nn-wrr) algorithm is used for dispatching requests to virtual machines and minimization of SLA violation rate. Although method shows better results. But there were some limitations in using nn-dwrr because it created more overhead. For improving the performance with better results Join-Idle Queue (JIQ) is used instead of nn-dwrr.
- Now in JIQ, average queue length is maintained while dispatching requests to particular virtual machine. We use this concept of minimum queue length in our algorithm for balancing the load among virtual machines.JIQ initially checks for the idleness of virtual machines if machines are idle then it allocates the task among them by maintain the queue length.

IV. PROPOSED METHODOLOGY

The base research by Chung-Cheng Li [4] and Yi Lua [5] helps us to conclude the proposed algorithm and architecture which is basically combination of two methodologies. The steps are as follows:

- The server receives a client request for some resources (cloudlets) on the cloud.
- The broker allocates these cloudlets to the virtual machines in virtual machine list with the minimum execution time and threshold value.
- Threshold value is decided on the basis of total number of cloudlets and number of virtual machines which further helps to maintain minimum queue length.
- SLA (Service Level Agreement) is created on the basis of response time. If virtual machine that is assigned to the cloudlet on the basis of minimum queue length complete the task within the response time provided by SLA then cloudlet will be accepted otherwise not.

Table 1. Table Showing Scheme Used for Balancing Load.

Task	Category using	Issues resolved
Resource Allocation	At virtual machine level	Makespan is minimized.
Task Scheduling	Space-sharing Time-Sharing	Minimize Response time

Above table shows the presently working scheme in proposed methodology as we are working at VM level for balancing load and using both the policies space sharing and time sharing.

Earlier many load balancing algorithms were used for balancing the load among virtual machines. There were sequential load balancing algorithms which include round-robin where tasks are assigned on the basis of time slices or quantum. In this paper we have proposed a better load balancing technique using hybrid of JIQ and SLA aware load balancing which will balance load by minimizing response time and queue length. Also the concept of SLA helps to calculate better results. Our objective is to further balance the load while maintaining SLA and reducing the waiting time, response time and makespan.

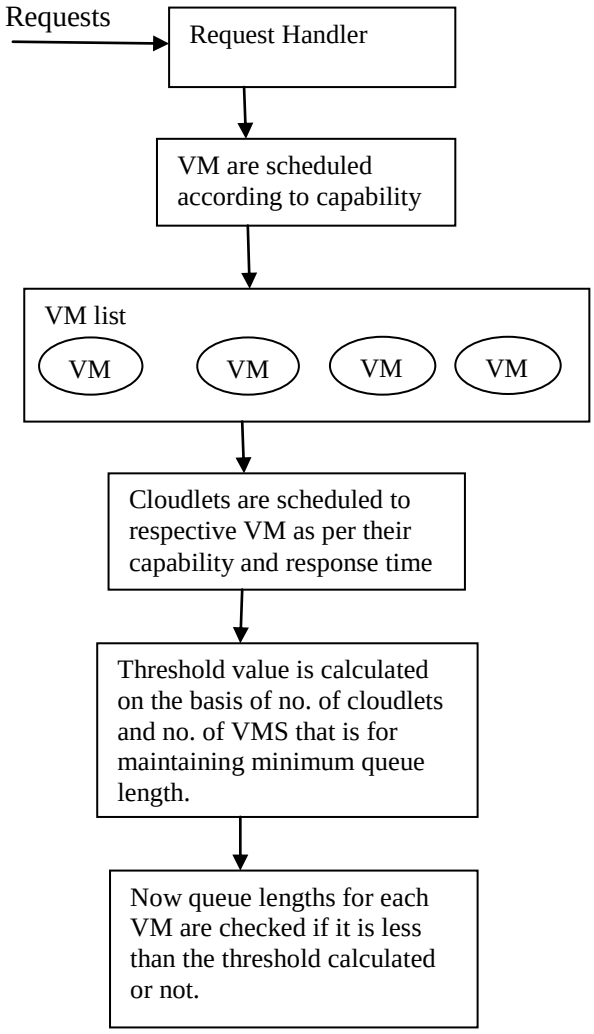

Figure 1. Architecture of Proposed Methodology

Architecture of proposed methodology explains the proper working of the system. Here, at very first stage requests are assigned to request handler. Then requests are assigned to VM present in VM list. A threshold value is required to be calculated that will further maintain the minimum queue length of each virtual machine. Requests are associated with VM on the basis of minimum response time and minimum queue length.

V. CONCLUSION

This paper presents the load balancing concept in cloud computing which further uses the combination of two technologies that are SLA aware load balancing and JIQ. Here we try to balance the load among VMs by calculating execution time and maintaining minimum queue length. Objective of the methodology is to reduce the average response time, makespan and average waiting time of tasks.

REFERENCES

[1] Velagpudi Sreenivas ,Prathap.M and Mohammad kemal "Load Balancing Techniques: Major Challenge in Cloud Computing – A Systematic Review", IEEE, 2014.

[2] Yashpalsinh Jadeja and Kirit Mali "Cloud Computing- Concepts,Architecture and Challenges",IEEE,pp:877-880,2012

[3] Mohammad Alhamad,Tharam Dhillon and Elizabeth Chang "Conceptual SLA Framework for Cloud Computing", IEEE, 2010.

[4] Chung-Cheng Li and Kuochen Wang "A SLA-awareLoad Balancing Scheme for CloudDatacenters", IEEE, pp: 58-63, 2014.

[5] Yi Lua,Qiaomin Xiea,Gabriel Kliatb,Gellerb,James R.Larusb and Albert Greenberge "Join-Idle Queue-A novel Load Balancing Algorithm for Dynamically Scalable Web Services", ELSEVIER, 2011.

[6] Akshay Jain,Anagha yadav,Lohit Krishanan and Jibi Abraham "A Threshold based Band Model for Automatic Load Balancing in Cloud Environment", IEEE, 2013.

[7] Wikipedia "Cloud Computing", http://en.wikipedia.org/wiki/Cloud_computing accessed on 4th March.

[8] Rajwinder Kaur and Pawan Luthra "Load Balancing in Cloud computing", pp: 374-381, 2014.

[9] Ram Prasad Padhy and P Gautam Rao "Load Balancing Cloud computing system", 2014.

[10] Wei-Yu Lin,Guan-Yu Lin and Hung-Yu Wei "Dynamic Auction Mechanism for Cloud Resource Allocation", IEEE, pp: 591-592, 2010.

[11] Bhaskar Prasad Rimal ,Eunmi Choic and Ian Lumb "A Taxonomy and Survey of Cloud Computing", IEEE.pp:44-51, 2009.

[12] Jayant Beliga,Robert W.A.Ayre,Kerry Hinton and Rodney S.Tucker "Green Cloud Computing:Balancing Energy in Processing,Storage and Transport",IEEE,2010.

[13] Yashpalsinh Jadeja and Kirit Mali "Cloud Computing- Concepts,Architecture and Challenges",IEEE.pp:877-880,2012.