Priority based Round Robin Task Scheduling Algorithm for Load Balancing in Cloud Computing

Abhijeet Malik, Prabhat Chandra
Master of Computer Application, National Institute of Technology, Kurukshetra, India

Abstract – Load Balancing is a burgeoning new field that attempts to distribute the dynamic local workload equally over all nodes in the cloud. Load Balancing is the key challenge of the cloud computing. In load balancing techniques performance of the cloud is enhanced by utilizing all resources optimally. The key objective of load balancing is to reduce resource consumption of energy and to minimize carbon emission this is the ominous need of the time. The main goal of this paper is to provide a brief overview of various Load Balancing algorithms and then to provide an efficient priority based round robin load balancing technique which prioritize various task to virtual machines on the basis of resources or processor required, number of users, time to run, job type, user type, software used, cost etc and then conveying them to various available hosts in round robin fashion. This approach improves the capability of system by enhancing various parameters such as fault tolerance, scalability and overhead etc and by minimizing resource utilization and response time. This approach is simulated and tested over CloudSim which is widely used tool to test cloud based techniques.

Index Terms – Cloud Computing, CloudSim, Load balancing, Fault Tolerance, Priority based Approach,Resource Consumption, Round Robin Allocation, Static Load Balancing .

1. INTRODUCTION

In the imminent years Cloud Computing will become a burgeoning and best ever Information Technology (IT) in the world. In Cloud Computing unlimited resources and services are provided to end users from internet at lower cost or we can say that a user has to pay for the resources they used at particular time, also known as pay per use concept [1].In cloud computing various resources, information services, software and other services are given to specific user on demand [2]. All services on user demand are distributed over internet, like operating system, Network, Storage, software, hardware and resources. These services are divided into three broad categories: -

SaaS (Software As Service): - SaaS is analogous to software provision's older thin-client model in which clients, in our scenario frequently web browsers, give the point of access to several software executing on various servers [3]. SaaS is the widely used cloud service by many consumers it transfer the task of managing and installing software to third party. It lessens the cost of licensed software and the cost of ownership as now we don't need technical staff for installation, management and up gradation of software [3]. These applications are provided on subscription model and most commonly used SaaS application are Salesforce, Google Apps, Dropbox etc[3].

PaaS (Platform As Service): - PaaS provides a platform above which software are built and set up and it operates at lower level than SaaS. It hides internal working and provides an environment to customer in which operating system and server software in addition to all underlying network infrastructure and server hardware are taken care of, therefore clients can focus on their product or service etc [3]. PaaS uses virtualization technology. Here users request as many resources as they require therefore they don't need to invest in hardware. Several PaaS providers are Red Hat's OpenShift, Heroku and Google App Engine [3].

IaaS (Infrastructure as Service): - It is the building block of cloud services and it consists of automates and scalable compute resources, cloud storage, network scalability which is metered, self provisioned and available as per demand[3]. IaaS provider gave these services via dashboard or API and clients can access their data directly and they can also outsource several technology and resources without any advance investment. IaaS clients have more control over infrastructure than PaaS and SaaS. They are mainly used for development of web scale application, PaaS and SaaS. Examples are Navisite, exoscale, and Softlayer etc [3].

Due to all these services provided by cloud load on cloud is increasing day by day due to which various load balancing techniques come into existence one of the way to balance load is scheduling the task [4], [5],[6]. A good scheduling algorithm is one which adepts according to changing environment and according to the type of user or task. Nowadays demand on cloud computing is increase so is load on the cloud therefore we need a reliable and guaranteed service all the time for this there should be a provision to allocate resources in an efficient manner to satisfy need of all clients and this can be done by allocating virtual machines preemptively and on the basis of the priority so that high priority job is executed first and this priority is decided on the basis of various factors which are discussed in section III and then after deciding priority these task are executed in round robin fashion algorithm of which is also given in section III

and at last section IV concludes the paper and throw some light on future work that can be done.

2. RELATED WORK

R. Kaur, et al [7] This paper combines the advantages of Max-Min and Min-Min algorithm and then presents a modified Max-Min algorithm which at selection time uses execution time instead of completion time unlike previous two and thus provides better resource utilization and response time. Modified Max-Min algorithm uses lower make span instead of RASA and original MAX-Min and therefore increasing the chances of execution of jobs on virtual machines.

Ali Moghtadaeipour, et al [8] in this paper to increase fault tolerance and to minimize energy consumption a task replication technique is used. Here author uses three fuzzy methods due to uncertainty of environment and other parameter determination. All these fuzzy inference increases fault tolerance for virtual prioritization, virtual machines and task prioritization by replication of tasks. In this approach ranking of tasks or virtual machines are done to increase fault tolerance and to verify results MATLAB tool is used and this approach is verified on the basis of various parameters such as success rate, workload and energy consumption.

K.Q. Yan, et al [9] This paper presents a two phase scheduling algorithm under three level cloud computing is proposed which is the combination of OLB(Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) scheduling algorithm[7]. This approach decreases execution time of each task, increase fault tolerance and provide better resource manipulation. The proposed method can be extended to manage and maintain when a node operate in three level hierarchical cloud computing network.

Jeffrey M. Galloway, et al [10] this paper introduces a load balancing approach for IaaS cloud architecture which is power based. This paper considers heterogeneous clouds for Power Aware Load Balancing (PALB) algorithm design by maintaining states of all computational nodes on the basis of utilization percentage, decides the number of nodes that should be operating. This approach provides adequate computational nodes and decrease the overall power consumption of local cloud to 70% to 97% compared to previous approaches using two sections upscale section (to power on additional node) and downscale section (powering down idle nodes) and if computational node uses less than 25% resources than it shuts down[10].

Stuti Dave, et al [11] in this paper author introduces an efficient algorithm named FAIR RR which is the improved version of existing round robin approach. This method uses a dynamic TQ (time quantum) based on algorithm execution. Cloud Analyst tool is used for verification of this algorithm against RR, throttled and active monitoring.

3. PROPOSED MODELLING

In cloud computing many clients can submit their job simultaneously thus overloading the VMs therefore in this proposed algorithm we assign some priority to each and every job and then on the basis of this priority these jobs are executed in round robin fashion. For example jobs like weather prediction, rainfall simulation, cyclone simulation etc require very large computation and require huge amount of resources therefore these should be tackled appropriately. Parameters used in this proposed algorithm for assigning priority are Number of users, servers, Time t run, Number of processor required, Memory needed, time of job submission, software used, type of job or user etc. Algorithm: For computation and assignment of priority to each job based on threshold value and allocation of resources to each job.

Step 1: [Receive user request data i.e., submission time, urgency, price, node and name of requested server]

All values are added to the linked list

Step 2: [Based on specified criterion time priority value is found for every job and its tasks]

Priority is assigned to every task for each user's job.

time_p[i] = priority value.

Step 3: [Based upon the specified criteria node priority value is found for every job and it's tasks]

priority is assigned to every task for each user's job.

node_p[i] = priority value;

Step 4: [to find whether data is in threshold or not every client's data is checked]

If (input value lies in between the threshold limit and total number of nodes <= available number of nodes)

{

[Node priority value and other metrics such as urgency are added along with respective computed time]

total [k] = time_p[i] + node_p[i] + urgency + price

Print (Ready to run available node = available number of nodes –total number of nodes)

}

Else if (input value lies in between the threshold limit)

{

total[k] = time_p[i] + node_p[i] + urgency + price

Print (lies between the limit but it is in queue)

}

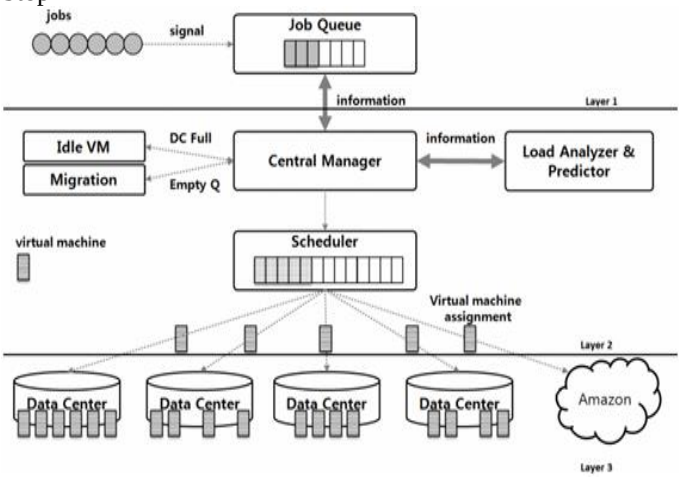
Else

Print (Exceed the condition threshold limit)

Step 5: [All sum[k] values are then sorted]

Step 6: user's request is ready to run from minimum values of total[k]

Stop



The main difficulties in the resource allocation in a cloud system are to take proper decision for job scheduling, execution of job, managing the status of job etc. Apart from traditional best fit and bin packing algorithm in this paper an algorithm is developed for the job allocation in the cloud environment to be decided by the cloud administrator. Several parameters listed above are considered for the priority based on the client and server requirements and requests by the users. After priority is given to each request then these jobs are assigned to virtual machine use round robin scheduling technique so that every virtual machine is assigned with equal amount of load on each VM algorithm for the same is explained below which is very simple so that it reduce the overhead of computer and work efficiently.

Pseudocode for Algorithm-

For (All tasks Ti)

{

If (Ti is having high priority)

{

Put the task into T1 batch

}

Else

{

Put the task into T2 batch

}

}

(End for)

Do until (all tasks in T1 mapped)

{

For (all Tasks)

{

Sorting all tasks in T1 in decreasing order of Priority and adding them to request queue then

Dequeue a task from request queue and then allocate resources with Round Robin Algorithm

}

(End for)

}

(End do)

Clear all task requests

Do until (all tasks in T2 mapped)

For (all Tasks)

{

Sorting all tasks in T2 in decreasing order of priority and adding them to request queue then

Dequeue a task from request queue and then allocate resources with Round Robin Algorithm

}

(End for)

}

(End do)

4. CONCLUSION AND FUTURE WORK

Load balancing is one of the major challenges in cloud computing. This will avoid the situation where some nodes are heavily loaded while others are idle or doing little work. It helps to achieve a high user satisfaction and resource utilization ratio. Existing algorithms have some limitations and to remove these limitations to improve the efficacy of various algorithms a new load balancing algorithm can be designed by combining the features of Round robin and min-min algorithm. In this proposed algorithm tasks are first divided according to their priorities. Here, all the incoming requests from the clients have been automatically redirected to virtual machines based on priority and also the new virtual machines are created and redirected when the requests are

overloaded. Hence, it may improve the overall performance and resource utilization of the system and helps to take further steps towards green cloud computing. It also ensures that every computing resource is distributed efficiently and fairly. This scheduling algorithm can be extended to suit other cloud platforms also.

REFERENCES

[1] NAVTEJ SINGH GHUMMAN, RAJWINDER KAUR,"Dynamic Combination of Improved Max-Min and Ant Colony Algorithm for Load Balancing in Cloud System"
[2] Anthony T.Velte, Toby J.Velte, Robert Elsenpeter, Cloud Computing A Practical Approach, TATA McGRAW-HILL Edition 2010
[3] https://www.computenext.com/blog/when-to-use-saas-paas-and-iaas/
[4] K. Al Nuaimi, N. Mohamed, and M. Al Nuaimi, and Jameela Al-Jaroodi"A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms", IEEE Second Symposium on Network Cloud Computing and Applications ,pp: 137-142 ,year 2012
[5] S. Aslam, and M. Ali Shah "Load Balancing Algorithms in Cloud Computing: A Survey of Modern Techniques", 2015 National Software Engineering Conference (NSEC 2015) IEEE, pp: 30-35, Year 2015.
[6] N. K. Mishra, and N. Mishra, "Load Balancing Techniques: Need, Objectives and Major Challenges in Cloud Computing- A Systematic Review," in International Journal of Computer Applications, Volume 131 - Number 18, Year 2015.
[7] R. Kaur, and P. Luthra, "Load Balancing in Cloud System using Max Min and Min Min Algorithm," in International Journal of Computer Applications (0975 – 8887), National Conference on Emerging Trends in Computer Technology (NCETCT-2014) ,pp:31-34 ,year 2014.
[8] Ali Moghtadaeipour, and Reza Tavoli, "A New Approach to improve load balancing for increasing Fault Tolerance and decreasing Energy in Cloud Computing", in 2nd international conference on Knowledge-based Engineering and Innovation (KBEI),pp: 982-987,year 2015
[9] K.Q. Yan, W.P. Liao, and Shun-Sheng Wang, "Towards a Load Balancing in a three-level cloud computing network ", Computer Science and Information Technology (ICCSIT),3rd IEEE International Conference on Computer,pp:108-113 ,IEEE 2010.
[10] Jeffrey M. Galloway, Karl L. Smith, Susan S. Vrbsky, "Power Aware Load Balancing for Cloud Computing", in World Congress on Engineering and Computer Science (WCECS) Vol I, ISBN: 978-988-18210-9-6,year 2011.
[11] Stuti Dave, Prashant Maheta, "Utilizing Round Robin Concept for Load Balancing Algorithm at Virtual Machine Level in Cloud Environment", International Journal of Computer Applications (0975 –8887) Volume 94 – No 4, pp 23-29,year 2014 .