www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 6 June 2015, Page No. 12333-12339

A Review on Load Balancing In Cloud Computing

[1] Peenaz Pathak,[2]Er.Kamna Mahajan
[1,2]Dept Of Computer Science
RBIEBT,Kharar,INDIA

Abstract: *In today's world every activity belongs to internet, everything is going online in such a case web applications are playing an important role providing services to the customers and when the application becomes popular, traffic is also growing. Load Balancing is required in such situations to avoid overload. This paper also introduces task scheduling as it is the most important part in cloud computing which aims at meeting users requirements and improving the resource utilization. The purpose of this paper is to review various load balancing, Task scheduling algorithms along with their merits and demerits in detail.*

Keywords: Load balancing, Cloud Computing, Task Scheduling

1. Introduction:

1.1 Load balancing:

It basically distributes the workload across multiple computing resources such as computers, a computer cluster, network links, central processing units or disk drives. Load balancing is the process by which inbound internet protocol (IP) traffic can be distributed across multiple servers. It also enhances the performance of the servers which leads to their optimal utilization. For Internet services, the load balancer is a software program that is listening on the port where external clients connect to access services. The load balancer forwards requests to one of the "backend" servers, which replies to the load balancer. This allows the load balancer to reply to the client. It also prevents clients from contacting back-end servers directly, which may have security benefits by hiding the structure of the internal network and preventing attacks on the network[1].

1.2 Load Balancing Techniques:

1.2.1 Honey Bee Behavior Inspired Load Balancing:

This algorithm is inspired by the behavior of honey bees finding the food and informing others to go and eat the food. In bee hives, there is a class of bees called the scout bees and the forager bees. First forager bees go and find their food. After coming back to their respective beehive, they dance called waggle/tremble/vibration dance. After seeing the strength of their dance, the scout bees follow the forager bees and get the food. The more energetic the dance is the more food is available. The whole process is mapped to overloaded or under loaded virtual servers[2,3]. The server processes the requests of the clients which is similar to the food of the bees. As the server gets heavy or is overloaded, the bees

search for another location i.e. client is moved to any other virtual server.

Advantages: Maximizing the throughput, minimum waiting time, minimum overhead.

1.2.2 Ant Colony Optimization Technique:

ACO is used for proper distribution of load among the nodes of a cloud. In this case ants uses the basic pheromone updating formula and node selection formula of the ACO to distribute evenly the work load of nodes in a cloud. For efficient load balancing , a tier-wise distribution of nodes is taken into consideration, here the nodes are distributed in three tier structure such that the work is properly distributed among the nodes. This system shows the proper distribution of load among nodes. The ants will traverse in such a way that they know about the under loaded and over loaded nodes in a network[4]. A pheromone table which was designed will be updated by ants as per the resource utilization and node selection[3] . Ants will move in forward direction in search of the over loaded or under loaded node. If an ant encounters an overloaded node in its movement when it has previously encountered an under loaded node then it will go backward to the under loaded node to check if the node is still under loaded or not and if it finds it still under loaded then it will redistribute the work to the under loaded node.

Advantages: High Resource Utilization , performance of the network is increased

1.2.3 Throttled Load Balancing Algorithm:

In this algorithm the client first requests the load balancer to find a suitable Virtual Machine to perform the required operation. The process first starts by maintaining a list of all the VMs each row is individually indexed to speed up the lookup process. If a match is found on the basis of size and availability of the machine, the Throttled Virtual Machine Load Balancer returns the VM id to the Data Center Controller[5]. The Data Center Controller sends the request to the VM identified by that id. Data Center Controller notifies the Throttled Virtual Machine Load Balancer of the new allocation.

Advantages: High load movement factor [3].

1.2.4 Task Scheduling Algorithm based on Load Balancing:

The scheduling algorithm is based on load balancing to meet dynamic requirements of users and obtain high resource utilization. In cloud computing, task scheduling is a multi-objective optimization problem. The main objective is to minimize job spanning i.e. the total job completion time[6]. Balanced scheduling will decrease the job spanning. A job may consist of series of tasks. This algorithm includes various techniques such as First Come First Serve in which the jobs are queued in order of which come first. Round Robin technique in which jobs are dispatched in FCFS logic and the time slice of the process will decide the allocation. Min Min technique where small jobs are executed first and large jobs are waiting for more time. In Max–Min technique they select

the largest job to be executed first, later the small jobs are executed and takes long time[7].

Advantages: Max Resource Utilization, Minimum waiting time, Minimum response time, Maximum throughput

1.3Task Scheduling:

Task Scheduling is the assignment of start and end times to a set of tasks, subject to certain constraints. The scheduling of tasks in cloud means choosing the best suitable resource available for execution of tasks or to allocate computer machines to tasks in such a manner so that the completion time is minimized[8]. The main reason behind scheduling tasks to the resources in accordance with the given time bound, which involves finding out a complete and best sequence in which various tasks can be executed to provide best results to the user. In scheduling algorithm, a list of tasks is created by giving priority to each and every tasks. The tasks are further chosen according to their priorities and will be assigned to the available processors and computer machines. We have two basic types of scheduling :

- Static scheduling which schedule tasks in a known environment i.e. it already has the information about complete structure of tasks and mapping of resources before execution, estimates of task execution time.

- Dynamic scheduling should not only be dependent on the submitted tasks to cloud environment but also on the current states of system and computer machines to make scheduling decision.

The basic scheduling criteria involves[9]:

1) CPU utilization – keep the CPU as busy as possible

2) Throughput: No of processes that complete their execution per time unit

3) Turnaround time – amount of time to execute a particular process

4) Waiting time – amount of time a process has been waiting in the ready queue

5) Response time – amount of time it takes from when a request was submitted until the first response is produced.

1.4 Task Scheduling algorithms:

1.4.1 First Come First Serve (FCFS):

In this, the process that requests the CPU first is allocated the CPU first. Its implementation is easily managed with FIFO queue. When the CPU is free, it is allocated to the process which is at the head of the queue. It is a non-pre-emptive scheduling algorithm[10]. The CPU is assigned to the processes in the order they request for it. The FCFS scheduling algorithm is non preemptive[11]. Once the CPU has been allocated to a process, that process keeps the CPU until it releases the CPU, either by terminating or by requesting I/0.

1.4.2 Shortest Job First(SJF):

In this scheduling algorithm, the CPU is allotted to the process which has the smallest next CPU

burst. The SJF uses the FCFS to break tie (a situation where two processes have the same length next CPU burst)[11]. The SJF algorithm can be pre-emptive or non-pre-emptive. In preemptive SJF scheduling, the execution of a process that is currently running is interrupted in order to give the CPU to a newly arrived process with a shorter next CPU burst. On the other hand, the non-pre-emptive SJF will allow the currently running process to finish its CPU burst before a new process is allocated to the CPU.SJF scheduling is used popularly in long-term scheduling.

1.4.3 Priority Scheduling :

With each process a priority is associated and CPU will be allocated to the process with the highest priority. Priority scheduling is not fixed it can be preemptive or non preemptive[10].The priority of the process arriving at the ready queue is compared with the priority of the currently running process. If the priority of the newly arrived process is higher than the currently running process then scheduling is preemptive. A non preemptive priority scheduling algorithm will insert the new process at the head of the ready queue[11].

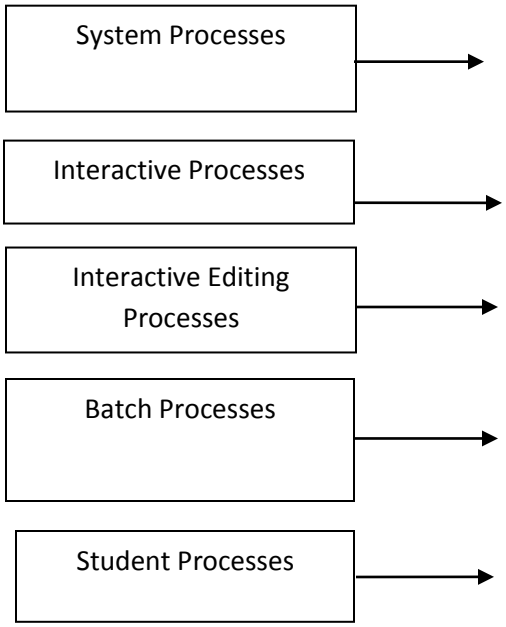
1.4.4 Round Robin Scheduling :

The round-robin(RR) scheduling algorithm is designed especially for time-sharing system. A small unit of time called a time quantum is defined. In this preemptive scheduling, the processes are transmitted in a first-in-first-out sequence but each process is allowed to run for only a limited amount of time.

1.4.5 Multilevel Queue Scheduling:

The processes can be classified into different groups depending upon their situation. These groups include foreground processes(interactive) and background (batch) processes[11].

Highest Priority



Lowest Priority

1.4.6 Multi Queue Scheduling:

The MQS method gives importance to select job dynamically in order to achieve the optimum cloud scheduling problem and hence it utilize the unused free space in an economic way. This approach enhances the scheduler to group the various burst time based jobs into a particular queue[7] which are categorized into

small, medium and long based on ascending order. The proposed scheduling algorithm achieves the optimum usage of resources for cloud computing and attains high resource utilization and provides Quality of System in cloud environment.

1.5 Cloud Computing:

Cloud Computing is a new trend emerging in IT environment with huge requirements of infrastructure and resources. Computation in cloud is done with the aim to achieve maximum resource utilization and cost minimization. Cloud computing involves virtualization, distributed computing, utility computing, networking, software and web services. The cloud architecture is mainly distributed into three main layers, namely: infrastructure, platform and software[4]. Cloud computing has an advantage of delivering a flexible, high-performance and on-demand Services. Cloud has different meaning to different stakeholders[12]. Cloud provides a variety of resources, including platforms for computation, data centers, storages, Networks, firewalls and software in form of services. At the same time it also provides the ways of managing these resources such that users of cloud can access them without facing any kind of performance related problems[8].

2. Related Work:

1) **Ruhi Gupta [3]** explained Load Balancing as one of the most important parts of the current virtual environment. In this paper a complete survey of various existing load balancing techniques along with their merits, demerits and comparison between different techniques was done based on various parameters. Different scheduling algorithms were simulated for executing user request ,each algorithm was observed and their scheduling criteria like average response time, data center service time and total cost of different data centers were found.

2) A brief introduction to different load balancing strategies, algorithms, methods was given by **Deshmukh et al.[1]**. By investigating the comparative behavior of load balancing with different parameters, dynamic load balancing proved to be more reliable. So dynamic load balancing method was applied in case where traffic was equally distributed across different servers. This load balancing technique was efficient that clearly increased the performance and overloading problem was also avoided.

3) **Aggarwal et al.[10]** In this paper, a comparative study of different scheduling algorithms based on the different parameters such as average waiting, average turnaround, average response time, average CPU utilization and throughput was done. The scheduling of tasks in cloud means choosing the best suitable resource available for execution of tasks.

4) **Karthick et al.[7]** A Multi Queue Scheduling (MQS) algorithm was described to reduce the cost of both reservation and on-demand plans using the global scheduler. This MQS was based on burst time using dynamic job selection, a queuing method was implemented which increased the satisfaction of the user and utilized the free unused space of resources in an economic way.

5) **Wang et al.[6]** In this paper, (original adaptive algorithm) AGA was used to enhance the overall performance of cloud computing environment. JLGA algorithm was also intended to achieve task scheduling with least makespan and load balancing. At the same time, greedy algorithm was adopted to initialize the population, to describe the load intensive among nodes and weights multiple fitness function.

3. Conclusion:

Load balancing task scheduling is a process of managing of different task on the basis of their

priority or job execution order. Jobs have to be derived on different processors for execution. This paper gives us a brief idea about the load balancing, task scheduling and its techniques that will help us for further study. In this paper, various algorithms, techniques have been discussed for load balancing on cloud computing environment along with their merits and demerits.

4.References:

1) Ankush P. Deshmukh ,Prof. Kumarswamy Pamu: "Applying Load Balancing: A Dynamic Approach" published by Volume 2, Issue 6 , June 2012 © 2012, IJARCSSE

2) Rajesh George, Rajan V.Jeyakrishnan: "A Survey on Load Balancing in Cloud Computing Environments" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.

3) Ruhi Gupta: "Review on Existing Load Balancing Techniques of Cloud Computing" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 2, February 2014 .

4) Suresh M ,Shafi Ullah Z ,Santhosh Kumar B: "An Analysis of Load Balancing in Cloud Computing" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 10, October – 2013

5) Subasish Mohapatra, K.Smruti Rekha,Subhadarshini Mohanty: "A Comparison of Four Popular Heuristics for Load Balancing of Virtual Machines in Cloud Computing" International Journal of Computer Applications (0975 – 8887) Volume 68– No.6, April 2013.

6) Tingting Wang,ZhaobinLiu, Yi Chen, Yujie Xu, Xiaoming Dai: "Load Balancing Task Scheduling based on Genetic Algorithm in Cloud Computing" published in 2014 IEEE 12th International Conference on Dependable, Autonomic and Secure Computing

7) AV.Karthick, Dr.E.Ramaraj, R.Ganapathy Subramanian: "An Efficient Multi Queue Job Scheduling for Cloud Computing" 2014 World Congress on Computing and Communication Technologies.

8) Raja Manish Singh,Sanchita Paul, Abhishek Kumar: "Task Scheduling in Cloud Computing: Review" International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7940-7944

9) Soumen Santra ,Hemanta Dey,Sarasij Majumdar,Gauri Shankar Jha : "New Simulation Toolkit for Comparison of Scheduling Algorithm on Cloud Computing" 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)

10) Himani Aggarwal , Er. Shakti Nagpal : "Comparative Performance Study of CPU Scheduling Algorithms" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 6, June 2014

11) Silberschatz, Galvin, Gagne: "Operating System Concepts" 7th Edition

12) Mayanka Katyal, Atul Mishra: "A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment" International Journal of Distributed and Cloud Computing Volume 1 Issue 2 December 2013