WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Load Balancing in Cloud Computing

Amol Bhalerao

Keraleeya Samajam(Regd.) Dombivli's Model College, Thakurli (East), Maharashtra, India
Email id :- amolbhalerao.model@gmail.com

ABSTRACT

Cloud Computing is a structured model that defines IT services where data and resources retrieved from the cloud service provider through internet through a well-structured web tool and application set up and delivered to users according to their needs. Sharing the resource group may initiate a problem with the availability of these resources causing a deadlock situation. One way to avoid deadlocks is to distribute the workload of all virtual machines among them. This is called load balancing. The goal of load balancing virtual machines is to reduce power consumption and provide maximum use of resources, thereby reducing the number of rejected jobs. As the number of users increases in the cloud, load balancing has become a challenge for the cloud provider. The purpose of this article is to discuss the concept of load balancing in cloud computing and how it improves and maintains the performance of cloud systems and also contains the comparison of various existing static load balancers as well as load balancers. conventional dynamic loads.

Keywords—Cloud Computing, Load Balancing, Resource Allocation, Scheduling, Deadlocks

INTRODUCTION

Cloud computing or the future of next-generation computing offers its customers virtualized network access to applications and / or services. No matter where the customer is accessing the service from, they are automatically directed to the available resources. Sometimes our system hangs or it seems it takes a few decades for the pages to come out of the printer. All of this happens because there is a queue of requests waiting for their turn to access the resources that are shared between them. But these requests cannot be served since the resources required by each of these requests are withheld by another process or requested by virtual machines. One of the causes of all these problems is blocking.

Load balancing is a new approach that helps networks and resources by providing high throughput and minimal response times

In cloud platforms, resource allocation (or load balancing) takes place majority at two levels. At first level: The load balancer assigns the requested instances to physical computers at the time of uploading an application attempting to balance the computational load of multiple applications across physical computers. At second level: When an application receives multiple incoming requests, each of these requests must be assigned to a specific application instance to balance the computational load across a set of instances of the same application The following sections discuss the concept of load balancing, its needs and purposes, the types and comparison between the traditional computing environment and the cloud computing environment, as well as the different algorithms. The conclusion of and the references are then discussed.

Load Balancing

Load balancing is the process of improving system performance by moving the workload between processors. A machine's workload is the total processing time required to complete all tasks assigned to the machine. Load balancing virtual machines means that none of the available machines are idle or partially loaded while the others are heavily loaded. Load balancing is one of the important factors to increase the working performance of the cloud service provider. The benefits of workload balancing include an increase in the rate of resource utilization, which further improves the overall performance of the, thus achieving maximum customer satisfaction.

In cloud computing, if users are increasing load will also be increased, the increase in the number of users will lead to poor performance in terms of resource usage, if the cloud provider is not configured with any good mechanism for load balancing and also the capacity of cloud servers would not be utilized properly. This will confiscate or seize the performance of heavy loaded node. If some good load balancing technique is implemented, it will equally divide the load (here term equally defines low load on heavy loaded node and more load on node with less load now) and thereby we can maximize resource utilization. One of the crucial issues of cloud computing is to divide the workload dynamically.

Load Balancing Goals

i. The load balancing goals discussed by authors of [1],[2] include:
ii. Substantial improvement in performance
iii. Maintaining system stability
iv. Increasing system flexibility to adapt to changes
v. Create a fault tolerant system by making backups.

Classification of the Load Balancing Algorithm

Depending on the orientation of the process, they are classified as:

a) Initiated by the sender: The sender starts the process; the client sends the request until a recipient is assigned to receive its workload

b) Initiated by the recipient: The recipient initiates the process; the receiver sends an acknowledgment request from a sender willing to share the workload.

c) Symmetrical: This is a combination of the type of load balancing algorithm initiated by the sender and receiver.

Based on the current state of the system, they are classified into: -

1. Static Load Balancing: - With the static load balancing algorithm, the decision to change the load does not depend on the current state of the system. It does require knowledge of applications and system resources.The performance of the virtual machines is determined at the time the order is received. performed by the slave processors and the result is returned to the master processor. Static load balancing algorithms are not preventative, and therefore eachmachine has at least one assigned task of its own. Aims to minimize the execution time of tasks and to limit communication overhead and delays. This algorithm has the disadvantage that the task is not assigned to processors or machines until it has been created, and the task cannot be modified while it is being executed to any other machine for load balancing. The four different types of static load balancing techniques are the round robin algorithm, the central manager algorithm, the threshold algorithm, and the randomized algorithm

2. Dynamic Load Balancing :- In this type of load balancing algorithm, the current state of the system is used to make load balancing decision so that the load shift depends on the current state of the system. dynamically switch from an overloaded machine to an underutilized machine for faster execution. This means that it enables process pre-emption that is not supported by the static load balancing approach.A key benefit of this approach is that your decision about load balancing is based on the current state of the system, which helps improve the overall performance of the system by dynamically migrating the load.

Traditional V / S Computing Cloud Computing Environment

There are many similarities and differences between traditional scheduling algorithmsand scheduling VM resources in a cloud computing environment.First, the main difference between the cloud computing environment and the traditional computing environment is the goal of planning. In the traditional computing environment, you mainly plan processes or tasks so that the granularity and the data transferred are small; whereas in a cloud computing environment the planning target is VM resources, so the granularity is great and the s transmitted are great too. Second, in a cloud computing environment, compared to the virtual machine deployment time, thescheduling algorithm's time can almost be overlooked.

Need of Load Balancing

We can balance the load on a machine by dynamically changing the local workload of the machine to remote nodes or machines that are less used. This maximizes user satisfaction, minimizes response time, increases resource utilization, reduces the number of order rejections and increases the throughput rate of the system. Load balancing is also necessary to achieve green computing in the clouds [9]. The factors responsible are:

I. Limited power consumption: - Load balancing can reduce power consumption by avoiding redundant nodes or virtual machines due to excessive workload.
II. Reduction of CO2 emissions: - Energy consumption and CO2 emissions are two sides of the same coin. Both are directly proportional to each other. Load balancing helps reduce power consumption, which automatically reduces CO2 emissions and thus achieves green computing.

Load Balancing Algorithms

The paper describes three load balancing algorithms, the round robin algorithm, equally spread current execution load and Throttled Load balancing [8].

a) Round Robin: - Round Robin uses the time division mechanism. The name of the algorithm suggests that it works in the form of rounds, where each node is assigned a time slot and it has to wait till its turnis assigned to each node. Each node is assigned a time slot in which it has to perform its task. The complicityof thisalgorithm is less than that of the other two algorithms.An open source simulation run by the software algorithm known as the Cloud Analyst. This algorithm is the standard algorithm used in the simulation. This algorithm simply assigns work in a round robin fashion that does not take into account the load on different machines.

b) Equally spread current execution load:- This algorithm requires a load balancer to monitor the jobs that are requested to run. The job of the load balancer is to queue the jobs and deliver them to various virtual machines. The balancer regularly checks the queue for new jobs and then assigns them to the list of free virtual server. The Balance also maintains the list of tasks assigned to virtual servers, which allows them to see which virtual machines are free and need to be assigned new jobs. The experimental work for this algorithm is done with the Cloud Analyst simulation. The name suggests that this algorithm distributes the execution load evenly across different virtual machines.

c) Throttled Load Balancing: -The Throttled Load balancing algorithm works by finding the right virtual machine to assign a particular job to. The job manager has a list of all virtual machines, based on this indexed list, it assigns the desired job to the corresponding machine to. If the job is more suitable for a particular machine than this job, assign it to the appropriate machine. If there are no virtual machines available to accept jobs, the job manager waits for the client to request and queues the job for fast processing.

d) ARA (Adaptive Resource Allocation): - In the ARA algorithm for the adaptive allocation of resources in cloud systems, which tries to counteract the harmful effects of Burrstones by allowing a certain randomness in the decision-making process and thus improving the performance and overall system availability[3]. The problem with this strategy is that it only takes into account Poisson inflows as well as the exponentially distributed service time and the fixed number of options.The following figure shows the schematic representation of the algorithm used for load balancing in the cloud computing environment The figure also shows the three algorithms contained in this document using the cloud analysis simulation tool. This tool is based on the cloud simulation. The cloud simulation provides a GUI interface that helps do the work experimentally.

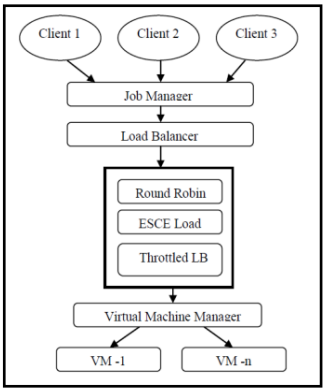

Fig. 1 Source: [7]

DYNAMIC LOAD BALANCING POLICIES AND STRATEGIES

The policies described in [4] and [5] are as follows:

1. Location Policy :- The policy used by a processor or a machine to share the activity transferred from an overloaded machine is defined as a location policy.
2. Transfer Policy:- The policy used to select a task or process from a local machineto transfer to a remote machine is called a transfer policy.
3. Selection Policy:- The policy used to identify the processors or machines participating in load balancing is called as selection policy.
4. Information Policy:- The policy responsible for collecting all the information on which the load balancing decision is based is called an information policy.
5. Load Estimation Policy:- The policy used to decide how to approximate the total workload of a processor or machine is

called the load estimation policy.

6. Process Transfer Policy:- The policy used to decide whether to perform a task that should be done locally or remotely is called a process transfer policy.
7. Priority Assignment Policy:- The policy used to prioritize the execution of local and remote processes and activities is called the prioritization policy.
8. Migration Limiting Policy:- The policy used to set a limit on the maximum number of times a task can migrate from one machine to another

COMPARISON CHART

Parameter	**Round Robin**	**Throttled**	**Active VM load Balancer**
Dynamic/Static	Static	Dynamic	Dynamic
Resource Utilization	Less	Less	More
Fault Tolerance	No	Yes	No
Overload Rejection	No	No	Yes

Fig.2 Comparison of various algorithmsSource: [6]

Parameter	**Round Robin**	**Equally Spread Current Execution**	**Throttled**
Number of request / 3 hour	35	35	35
Response Time	142.25s	142.16s	124.62s
Processing Time	35.78s	36.69s	18.26s

Fig.3 Response time of various algorithms Source [6]

We used the number of requests as shown in the table above for each load balancer policy one by one, and based on that, the result calculated for metrics like response time, processing time queries has been posted. The table shows that the overall response time of Round Robin policy and Equally Spread Current Execution policy is almost the same, while that of Throttled policy is low compared to the other two policies.

QUALITATIVE MATRIX FOR LOAD BALANCING

The different qualitative metrics or parameters that are considered important for load balancing in cloud computing are discussed as follows:

1. Throughput:- The total number of jobs that have completed execution is called throughput. High throughput is required for best system performance.
2. Associated Overhead:- The amount of overload produced by running the load balancing algorithm. There is minimal overhead for a successful implementation of the algorithm.
3. Fault Tolerant :- It is the ability of the algorithm to operate correctly and smoothly even under the failure conditions of any arbitrary node in the system.
4. Migration Time:- The time required to migrate or transfer a task from one machine to another machine in the system. This time should be minimal to improve system performance.
5. Response Time:- This is the minimum time required for a distributed system running a specific load balancing algorithm to respond.
6. Resource Utilization:- It is the degree of use of system resources. A good load balancing algorithm allows maximum use of resources.

7. Scalability:- Determines the ability of the system to perform load balancing algorithms with a limited number of processors or machines.
8. Performance:- Represents the efficiency of the system after performing load balancing. If all of the above parameters are optimally fulfilled, it will significantly improve the performance of the system.

CONCLUSION

Since such cloud computing is a large area of research and one of the main research topics is dynamic load balancing, the following research will focus on the algorithm by mainly considering two parameters on the one hand , the load on the server, and secondly, the current server performance.

The goal of load balancing is to increase customer satisfaction and maximize the use of resources and dramatically increase the performance of the cloud system and minimize response time and reduce the number of rejected jobs thus reducing the energy consumed and the carbon emission rate.

References

[1] David Escalnte and Andrew J. Korty, "Cloud Services: Policy and Assessment", EDUCAUSE Review, Vol. 46, July/August 2011

[2] Parin. V. Patel, Hitesh. D. Patel, Pinal. J. Patel, "A Survey on Load Balancing in Cloud Computing" IJERT, Vol. 1, Issue 9, November 2012

[3] JianzheTai,JueminZhang,JunLi,WaleedMeleis and NingfangMi "A R A: Adaptive Resource Allocation for Cloud Computing Environments under Bursty Workloads"978-1-4673-0012-4/11 ©2011 IEEE

[4] Ali M Alakeel, "A Guide To Dynamic Load Balancing In Distributed Computer Systems", International Journal of Computer Science and Network Security, Vol. 10 No. 6, June 2010

[5] Abhijit A Rajguru, S.S. Apte, "A Comparative Performance Analysis of Load Balancing Algorithms In Distributed Systems Using Qualitative Parameters", International Journal of Recent Technology and Engineering, Vol. 1, Issue 3, August 2012

[6] Jitendra Bhatia, Tirth Patel, Harshal Trivedi, Vishrut Majmudar," HTV Dynamic Load Balancing Algorithm for Virtual Machine Instances in Cloud",18,Dec2 012,Pages 15-20 IEEE

[7] R. Mata-Toledo, and P. Gupta, "Green data center: how green can we perform", Journal of Technology Research, Academic and Business Research Institute, Vol. 2, No. 1,May 2010, pages 1-8.

[8] Dr. Hemant S. Mahalle, Prof. Parag R. Kaveri ,Dr.Vinay Chavan," Load Balancing On Cloud Data Centres" , International Journal of Advanced Research in Computer Science and Software Engineering,Vol.3,Issue 1,January 2013.

[9] R. Shimonski, Windows 2000 And Windows Server 2003, Clustering and Load Balancing Emeryville, McGrow-Hill Professional Publishing, CA, USA, 2003.