EFFECTIVE LOAD ALANCING AND RESOURCE ALLOCATION IN CLOUD ENVIRONMENT

Sagar Ghuge¹, Rushikesh Narhare², Himanshu Pritwani³, Divyanshu Gunjal⁴

¹Department of Computer Engineering
²Department of Computer Engineering MIT College of Engineering, Pune, India

ABSTRACT:

Cloud Computing provides different facilities such as SaaS (Software as a Service), PaaS (Platform as a Service) and IaaS (Infrastructure as a Service). Cloud Computing is fast rising area in the today's computing environment. Cloud computing is broadly used by many industries. There are different issues in cloud computing such as Virtual Machine Migration, Service Provisioning (Resource Allocation), and load balancing. The main issue in the cloud computing is load balancing and allocating resources dynamically. Load balancing is the mechanism that distributes the load among the different nodes in the whole cloud and achieves the high user satisfaction and appropriate resource sharing. Appropriate load balancing realizes in minimizing resource consumption, over provisioning, task failure. Availability of cloud system is one of the main concern of the cloud computing. Load balancing methods are practical across the cloud service to confirm the availability of the resources by minimizing use of resources and proper load balance. Resource consumption can be done with proper load balancing which helps not only in reducing the cost but making enterprise greener. Another feature of the resource provisioning is to guarantee that the requirements of the all applications are meeting appropriately. Load balancing is essential in distributed processes, Parallel processes, grid computing and many more applications. Resources are allocated on user’s request and priority bases. In this paper, we discuss the available load balancing techniques and allocation of the resources effectively.

Keywords: Load Balancing, Resource Allocation, Virtual Machine, Cloud Computing.

[1] INTRODUCTION

Cloud Computing is a new technology which is having more demand now days. Computing is fast rising area which is used in IT industry. Internet is often represented as a cloud and term is defined as cloud Computing. Cloud Computing provides the different computing facilities like hardware, software, platform as a computing resource to other. Cloud Computing is a growing concern in the software industry [2]. Cloud system terms down due to management, power outage or refutation service attacks. Fundamentally, it determines the time required for the system is more than the system consecutively running the jobs, the interval time between failures. It is widely being adopted by the industry and they are facing many issue in the cloud computing like load balancing, resource management, energy
management etc. [1]. Main theme of cloud is nothing but that the user data is not only stored locally but also stored in center data of internet which is managed by the cloud service provider. When user requires the data, cloud provides the data to use through any terminals connected to the internet. Cloud Computing represents a major change in how we store information and run application [5]. Load Balancing is a core challenging issue in cloud computing. In cloud computing environment, load balancing has a significant impact on the performance. Load balancing is one of the basic to consume all the resources in the parallel and distributed system. Unbalancing workload means that some of the resources have been used out, while others are still wasted and this type of workload is handled by Virtual machine migration [5]. There is need for load balancing in large scale system such as a cloud computing domains. Here, the workload is divided among all nodes in the whole cloud to achieve good resource allocation and great user satisfaction. Load balancing is a new technique that provides networks and resources by providing maximum throughput with minimum response time [10].

[2] LITERATURE SURVEY

The literature survey has been done in three parts- Load balancing for cloud computing, Reliability and Resource Allocation Methods. In computations, resources associated with different system in different dwellings are dynamically obtainable and scattered geologically. Cloud computing is very complex environment so issue of load balancing arises [5]. User’s resource requirements vary depending upon their goals, time constraints, priorities and budgets [14]. Divisions of workloads or jobs are done by load balancing to simplify the complex environment. Using multiple components with load balancing, instead of a single component, it may increase reliability through redundancy. Cloud load balancer offers a cost effective and reliable cloud based fail over solution. Balancing technology manages traffic and makes decision about which node allocates the resources and depends upon the priority of workloads [12]. Load balancing helps in improving the performance, stability and accommodate future modification. It can be implemented with hardware, software or combination of both. Workloads arrive at the main controller and then queuing that job to any particular node. After that node choose a cloud partition and check whether the cloud state is ideal normal or heavy. If heavy states occur, then again it requires the cloud partition which means reschedule the jobs if load is not balanced. When job is arrived at the cloud partition balancer then assign workload to particular processor when it is free according to strategy and then release the processor after completion of whole job. [1]

Round Robin Load Balancer is a static algorithm which does not consider the prior load state of the node at the time of allocation of jobs. The round robin scheduling algorithm is used for allocating jobs. It selects the first node randomly and then, allocates jobs to all other nodes in a round robin manner. This algorithm is not adopted for cloud computing so recommended Weighted Round Robin algorithm is used in which each node is assigned some specific weight. Node receives appropriate number of request depending upon the weight assigned to each particular node. Throttled Load Balancer is Dynamic load balancing algorithm to perform the required task. Request is send by client to load balancer, based on the request, suitable virtual machine is allocated to perform the particular operation. In Cloud Computing virtual machines are grouped together based on the request they are handling [1]. Honeybee Foraging Algorithm is derivative from behavior of honeybee [5]. Detector and gleaners are two types of honeybees in which detector goes outside and find the honey source. After finding the source they return and wiggle in front of available honeybee. Then gleaner goes outside and gains the honey from those sources and return to beehive and start wiggle dance, this dance indicates how much food is left. M.Randles proposed the decentralized honeybee based algorithm for
self-organization. [2][4] Cartron R. Stanojevic et al. proposed mechanism for cloud control, which fuses the use of load balancing (LB) and distributed rate limiting (DRL). Load balancing distribute the load on servers in a way so that the servers have equal load. This minimizes the associated cost to complete the jobs, and DRL make sure that the resources are spread in way to possess fair minded resource allocation and also adopt the capacities of dynamic workload so performance levels at all servers are equal. This algorithm is easy to implement with very low computation and communication overhead [6].

To find the reliability of the system which handles the load, consider the factors such as throughput which is used to calculate number of task whose execution has been completed in unit time [12]. In given scale of time throughput should be high to improve the performance of the system [5]. Response Time is defined as amount of time taken in distributed cloud environment to return with a load balancing methodology [12]. The response time should be minimized for effective system performance [5]. Resource Utilization is used to check how many resources are utilized [5]. It should be optimized for an efficient load balancing. In Resource allocation techniques Zhen Xiao, Weijia Song, Qi Chen discuss the virtualization concept for the resource allocation in the cloud computing. Also, they introduced the concept of skewness in order to measure the unevenness in the multidimensional resource utilization of a server. Mainly focus on the skewness because of that they utilize the resource nicely. Virtual machine and migration of virtual machines are used for mapping. Develop a resource allocation method that can avoid overloading in the cloud system.

[3] PROBLEM STATEMENT

Propose a model to establish cloud set up and balance the load among the Cloud system by considering some parameters such as throughput, efficiency to predict resource allocation. Resource Utilization techniques in which optimum selection can be done on the basis of cost of service and Quality of Service offered by the service.

[4] PROPOSED SYSTEM

In this project load Balancing algorithms are designed for all objectives to equally manage the load on processors and maximize their utilization while minimizing the total task execution time. There is inconsistent distribution of the workload which effect in server overburdening and server may crash. Because of the large amount of the load it is difficult to control workloads to improve system performance and maintain stability. To overcome the problem of the server crash and the inequality of the load, the model is basically developed for balancing the load or the jobs rather using already available hardware which only handle the virtualization. The proposed system is basically divided into three model such as server, submitters (cloud users) and the processors. The proposed system mainly focuses on the Qos factor such as the cost and time. Optimum resource utilization is also one of the main parameter in the project. In this project essentially concentration on software deployment using load balancing and resource allocation is done. It encompasses the concept of resource provisioning dynamically rather than using the hardware component such as hyper visor or framework support to the cloud. In the proposed work mainly consider the balancing of load
without prediction and dynamically allocate the resources to the task which give an optimal solution.

Firstly, when the jobs are requested by the user they are sent to the main controller. The main controller then selects the best partition and assigns the job to that balancer. The job is assigned to the balancer if it is in ideal or normal state. When the balancer is in the heavy state, then the job is again sent back to the main controller to assign it to another balancer. [1] When the job arrives at the balancer, it is necessary to assign this job to one of the nodes connected to the balancer. This is done by using Improved Round Robin algorithm. In Improved Round Robin algorithm, load status table is maintained in which the nodes are arranged from lowest to highest based on their load degree [1].

[4.1] METRICS FOR LOAD BALANCING IN CLOUDS

4.1.1 Throughput
Throughput is used to calculate the no. of tasks whose execution has been completed. It should be high to improve the performance of the system.

4.1.2 Overhead Associated
Overhead Associated determines the amount of overhead involved while implementing a load-balancing algorithm. It is composed of overhead due to movement of Tasks, inter-processor and inter-process communication. This should be minimized so that a load balancing technique can work efficiently.

4.1.3 Fault Tolerance
Fault Tolerance is the ability of an algorithm to perform uniform load balancing in spite of arbitrary node or link failure. The load balancing should be a good fault tolerant technique.

4.1.4 Migration Time
Migration time is the time to migrate the jobs or resources from one node to other. It should be minimized in order to enhance the performance of the system.

4.1.5 Response Time
Response Time is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.

4.1.6 Resource Utilization
Resource Utilization is used to check the utilization of resources. It should be optimized for an efficient load balancing.

4.1.7 Scalability
Scalability is the ability of an algorithm to perform load balancing for a system with any infinite number of nodes. This metric should be improved.

4.1.8 Performance
Performance is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays.
[4.2] FLOW OF SYSTEM

In the system, number of users can request the job to the server. Request is received by the central cloud server. Processors extract their own properties such as RAM, memory, cores etc., and then send these characteristics to the central server. Here, the central server is main controller. Central server responds to each request. It further processes the request to the load balancer. Different load balancers are connected to the main controller. The main controller passes the request to the balancer only if it is in ideal or normal state i.e. nodes connected to the balancer are available. [1] If the balancer is in the heavy state then the job again goes back to the main controller. Then it is forwarded to the balancer which is ideal or normal. When the request is processed to the balancer, it forwards the request to one of the nodes connected to the balancer. Thus we need to select the node to which the job is to be forwarded. For this, Improved Round Robin Algorithm is used. In Improved RR, nodes are arranged in the ascending order in the load status table. The order of the load status table depends on the load degree. Load degree can be calculated by using the methods given below. Accordingly, the job is assigned to the node with lowest node degree. Then the job is executed by the node and the result is given back to the cloud user. Server submits result to cloud user with the less response time.

Figure: 1. Flow of the System
[4.3] MATHEMATICAL MODEL

Let S=system

- $S=\{I, P, F, O, Su, Fa\}$
- $I=\{I_1, I_2, I_3, \ldots, I_n\}, i \leq n$
  Where, $I_i =$ ith request given by user
  $n =$ total number of requests.

- $P=\{P_1, P_2, \ldots, P_m\}, 1 \leq i \leq m$
  Where, $P_i =$ static or dynamic parameter such as CPU processing speed,
  Memory Size, Memory Utilization

$m =$ total number of parameters

F: Load degree (N)

- Load degree($N$) = $\sum_{i=1}^{m} \alpha_i P_i$
  Where $N =$ current node
  $\alpha_i =$ weights that may differ for different kinds of job.

- $O=\{O_1, O_2, \ldots, O_n\}$
  Where,
  $O =$ Successful completion of the $i^{th}$ request.

$Su =$ Success state where $Su \in O$
$Fa =$ Failure state $c < n |O = \emptyset$

[5] CONCLUSION

The adoption of cloud computing is being driven by trends towards globalization, consumer acceptance of the technology and organization needs to focus on core competences. In Cloud Computing Load Balancing techniques are used to obtain assessable improvement in resource utilization and availability of cloud computing environment and user satisfaction. Need of load balancing in cloud is effectually considered for improving network performance and resource utilization. Load Balancing can be combined with resource allocation system that can balance the load effectively. Hence, this will improve the overall performance and resource utility of the system. It also ensures that every computing resource is distributed efficiently and fairly.

In future, we can use the artificial intelligence and data mining techniques for resource utilization. Use of different algorithm which may be improve the accuracy of the project.
REFERENCES


