A QoS Load Balancing Scheduling Algorithm in Cloud Environment

Sana J. Shaikh[*1], Prof. S.B.Rathod[#2]

[*]Master in Computer Engineering, Computer Department, SAE, Pune University, Pune, India
[#]Master in Computer Engineering, Computer Department, SAE, Pune University, Pune, India

Abstract *- As we all know that the rapid development in Internet especially in cloud computing, the scheduling algorithm plays very important and vital role in day-today life. For implementing the process and handling the resources, the proper load balancing technique is required in cloud environment. In distributed environment, it is very difficult to achieve the resources with having different configuration and capacity. To optimize a particular outcome, the load balancer can map the task to resource that based on some particular objectives and utilize a task that takes necessary objectives the most commonly used load balancing objectives are tasks completion time and resource utilization. The cloud workflow background that completely generalizes and describes the workflow scheduling optimization problems based on QoS (Quality of Service) under the architecture of cloud. In the first stage , Service Level Agreement (SLA) based scheduling algorithm determines the priority of the tasks and assign the task to the respective cluster. In the second stage, the Idle-server monitoring algorithm balanced the load among the server within the each cluster. Our main goal is to understand the existing load balancing scheduling techniques and develop an optimized load balancing scheduling algorithm which gives maximum benefit to cloud environment. This paper outlines a comparative study that has been done to assess these Scheduling algorithms on the cloud computing environment.*

Key Words *- Cloud computing, Quality of Service, Load balancing scheduling techniques, Load balancing algorithm.*

I .INTRODUCTION

The cloud load balancing is one type of load balancing method that is performed in cloud computing environment. Load balancing is process of distributing or dividing workloads across multiple computing system or resources. A load balancing reduces cost and maximizes availability of resources which is associated with document management systems. In order to suit user requirements, it uses a precise method to map the tasks to appropriate cloud resources, though by default maximum strategies are static in nature.

As we all know that the load balancer holds the current state of system. We call it good scheduler when it does not changes report of resource availability and the existing status of cloud resources and able to generate resourceful schedules so the overall performance of the system is improved. An important issue is that when operating with load-balanced services it shows how to handle information that must be kept across the multiple requests as per user in a user's session. If this information is stored on one backend server locally then subsequent requests are going to different backend servers so that it is unable to find that previous information. To introduce this performance issue, the cached information should be recomputed in which the request of load balancing requests to different backend servers.

Ideally cluster of servers behind the load balancer should be session-aware, so that if a client connects to any backend server at any time, the user gets unpredicted experience. This is usually achieved with in-memory database or shared database. In distributed resources, scheduling problem is process that maps and manages the implementation of independent tasks. In order to meet the user's specific need, process can provide appropriate resources to ensure that the workflow can be successfully completed. Cloud Computing is state which gives proper and on-demand network access to shared pool of computing resources like network, storage, servers and services that are to be rapidly released with the efficient way in minimum management.[7]

Primarily cloud computing provides following types of service models:

A. Software as a Service Model –

In Software as a Service model, where customers can request for desired software, use it and pay only for the duration of time it was used, instead of purchasing, installing and maintaining on their local machine. An Example for SaaS is Google Docs.

B. Platform as a Service –

In Platform as a Service model, where complete resources are needed to design develop, testing , deploy and hosting an application are provided as services without spending money for

purchasing and maintaining the servers, storage and software. Example for PaaS is Google App Engine.

C. *Infrastructure as a Service –*

In Infrastructure as a Service model, where, infrastructure like a virtualized serer, memory, and storage are provided as services. An Example of IaaS is Amazon Elastic Compute (EC2) and Simple Storage Service (S3).

Cloud computing has following assets:

- Customers can scale up and scale down the resources dynamically as needed.
- Customers can pay only for how much the resources were used.
- Service Provider fully manages the service. Customers no longer need to concern about purchase, installation and maintenance of server and software updates.
- No investment on server, software and licensing.
- Users can access cloud from anywhere with an internet connection.

At present, cloud computing is suffering from some challenges like security, QoS, Power Consumption and Load Balancing etc. Currently, as there is an increase in technology and consumer demands, there is excessive workload which calls for the need of the load balancer. The concept of balancing the load among the server in cloud has an important effect on the performance. The uneven distribution of load among the servers results in server overloading and may lead to the crashing of servers. This degrades the performance. Load balancing is the technique that distributes the load equally among the servers which avoids the overloading of servers, server crashes and performance degrades. Load Balancing is an important factor that good response time, effective resource utilization. Thus the effective load balancing is needed.

The cloud computing model for delivering services through which the resources are retrieved from a centralized pool of resources. The cloud management software has to be managing resources at large scale. The main challenge is providing perform isolation and making efficient use of underlying hardware. This is a basic approach that involves a user accessing a resource when it is idle on random basis. The cloud has become an alternative Cloud computing is rapidly gaining popularity and number of cloud user is increasingly day by day.

II. RELETED WORK

This section describes the related work of QoS scheduling algorithm in cloud environment. The main challenge of cloud computing is distribution of workload in well balanced manner. So the distribution should be done among the different nodes so that resources should be properly utilized. To optimize this problem, good load balancer should be used [1]. In distributed workflow, the process that can provide the appropriate resources to ensure that the workflow can be successfully completed in order to meet user's need. In other words, the workflow scheduling algorithms are workflow instances of system instances by relevant rules and relational allocation of idle system resources so that the workflow can be easily implemented. The scheduling algorithms mainly have two types as: Market driven algorithm and Performance driven algorithm.

The Performance Driven algorithm can optimize the performance of system without considering the cost and map the workflow tasks to resources according to policies. There are two representative algorithms of Performance driven algorithm as: Heterogeneous Earliest Finish Time algorithm and throughput maximizing strategy. The Market Driven scheduling algorithms manage resource allocation of any task and it considers the cost. The representative algorithms are Backtracking [5], Generic Algorithm [2], LOSS and GAIN algorithm [3], Deadline allocation algorithm (Deadline Distribution Algorithm) [4] and QoS based deadline allocation scheduling algorithm [6]. As we know the cloud has greatly simplified capacity provisioning process, it poses several challenges in the area of Quality-of Service (QoS) management. Quality of Service demoted the performance level, reliability and availability offered by infrastructure and application.

The cloud computing is technique where group of servers are distributed in data center that allows centralized data storage and online access to computing resources or services. As the request enters, it has to be distributed equally among the servers otherwise results in server overloading, performance degrades and not effective utilization of resources. Effective load balancing technique improves response time of the task as well as utilizes the resources effectively.

A. *Backtracking*

Backtracking is general algorithm that finds all the solution to some computational problem, notably constraints satisfaction problems, which

incrementally builds candidates (backtracks) to the solutions and it determines that candidate cannot possibly be completed to valid solutions. Backtracking can be applied for different problems that admit the concept of partial candidate solution and relatively quick test of whether it can possibly be completed to valid solutions. Backtracking is important method for solving problems such that crosswords, Sudoku and many other puzzles. It is most popular and convenient technique for parsing. But when the problem is large then it is very difficult to backtrack each problem to find solution and sometimes it becomes very time consuming job so the backtracking is not efficient for large problems.

B. Generic Algorithms

By applying the principle of evolution, Genetic algorithms provide robust search techniques that allow a high-quality solution to be derived from a large search space in given polynomial time. The Generic Algorithm always combines the exploitation of the best solutions from the past searches with the explorations of new regions of the solution space. And the solution of any problem in search space can be represented by individuals. So this algorithm is very popular. The fitness function in population determines a quality of individuals.

The fitness value shows the comparison between the individuals that means how good the individual is. A genetic algorithm has the following steps:

i. Create population consisting of randomly generated solutions.

ii. Generate new offspring by applying genetic operators, crossover, namely selection, and mutation, one after the other.

iii. Evaluate the fitness value of each individual in population.

iv. Repeat steps ii and iii until the solution.

Disadvantage of this scheduling is complex and time consuming so it is not reliable.

C. QDA Scheduling Algorithm

A QoS- based Deadline Allocation Algorithm; QDA in short, considers cloud computing environment and the characteristics of workflow. The QDA algorithm refers the main sub-deadline allocation criteria of Comprised Time Cost Scheduling Algorithm. The CTC algorithm uses QoS utility function value as a service resource selection condition and it takes user performance into account. Here are some specific steps of QDA Algorithm:

Input: A set of workflow instances with the same type of task.

Output: A set of scheduling results, which gives the corresponding service resource for each task within the instance set

QDA Algorithm works as follows:

i. First, check if there is any uncompleted task in last round, if found then preferentially deal with this task.

ii. Calculate average execution time of different tasks in the last instance during this scheduling round. First, Select the last instance and determine of which tasks it consists. Second, Use the existing performance evaluation techniques to predict the Expected Execution Time of various instances executing in each resource node.

iii. After calculating this, each task's sub-deadline in the last instance, the same work should be done for other instances. In QDA Algorithm, Assume the scheduling task of instances follow the Stream-Pipe mode, which means staggering the deadline of a longer number of concurrent instances with the same type to avoid the competition of cheap resources. A longer term task has an opportunity to get the cheaper resources released by the shorter term task. These tasks can be executed in parallel and will not exceed their assigned deadline.

iv. The QoS parameters from the service providers to calculate the corresponding QoS utility function and divide the four candidate sub-sets according to four different kind of user preferences. After that, rank all the cloud services within each candidate sub-sets in accordance with QoS utility function values in ascending order.

v. For each sub-task, select its candidate service resource in the corresponding user preference candidate service subset, which meets its sub-deadline and has the minimum QoS utility function.

vi. After the allocation of all the sub-tasks to its corresponding service resource, the one round scheduling is executed.

vii. Repeat Step (i) and Step (6) for next round till scheduling ends.

D. Compromised Time Cost Algorithm

According to characteristics of cloud computing ,the workflow have large number of instances so taking consideration of execution time and execution cost as key factors, the algorithm that can get minimum total execution cost is Compromised Time Cost (CTC) scheduling[8] algorithm. This Compromised Time Cost scheduling algorithm is based on QoS constraints under cloud environment which can give better performance on SwinDeW-C [8] platform. This algorithm mainly based on two factors as execution time and execution cost. This uses the round robin scheduling policies to control the execution cost and execution time of any system or application.

User can change the execution time and execution cost of workflow instances during the running process in order to find more satisfactory balance between execution cost and execution time. So the overview of this workflow scheduling algorithms in cloud environment, it is easy to find that these algorithms are only concern with certain aspect of QoS, mainly in terms of cost and time parameters.

III. PROPOSED METHOD

Based on above discussion, both the algorithms are having some disadvantages. The backtracking algorithm is not efficient and the generic algorithm is not reliable means it is complex and time consuming scheduling algorithm and the QoS scheduling algorithm proposed in this paper to overcome them. Currently, due to the increased usage of cloud, there is a tremendous increase in workload. The uneven distribution of load among the servers results in server overloading and may lead to the server crash. This affects the performance.

Cloud computing service providers can attract the customers and maximize their profit by providing Quality of Service (QoS). Providing both QoS and load balancing among the servers are the most challenging research issues. Hence, in this paper, a framework is designed to offer both QoS and balancing the load among the servers in cloud. This paper proposes a two stage scheduling algorithm. The servers with different processing power are grouped into different clusters. In the first stage, Service Level Agreement (SLA) based scheduling algorithm determines the priority of the tasks and assigns the tasks to the respective cluster. In the second stage, the Idle-Server Monitoring algorithm balances the load among the servers within each cluster.

The proposed algorithm has used the response time as a QoS parameter and is implemented using CloudSim simulator. So we can say that the algorithm provides better response time, waiting time, effective resource utilization and balancing load among the servers as compared to other existing algorithms.

Two stages load balancing scheduling algorithm can be done as:

A. SLA Based Scheduling Algorithm

In the first stage of this algorithm, the Service level agreement is done based on the scheduling technique which determines the input is accepted from user and executes them in priority manner. As per the priority the task scheduling will be done. The highest priority will be get first chance. Whenever the first stage is completed the second stage is come to front commonly known as Load Balancing. SLA based scheduling algorithm compute the priority of the task by considering task length, cost and deadlines.

B. Idle-Server Monitoring Algorithm

In this second phase, the idle-server monitoring algorithm runs within each cluster to monitor the set of servers in its cluster. The algorithm checks for any idle server in its cluster. If found, it assign the task to the identified server. If the task cannot be assigned to server, the task is put into the queue. Within each cluster, Idle-Server Monitoring Algorithm maintains all the server status in the table.

Thus, the Idle-Server Monitoring algorithm within the medium processing power server's cluster checks for any idle high processing power cluster. If a free server is found, it assigns its task to the identified high processing server. Similarly, Idle-Server Monitoring Algorithm within low processing power server's cluster checks for any free medium processing power server within medium processing power server's cluster. If free server is found, it assigns its task to the identified medium processing power server. By doing this, the resources are utilized effectively and the tasks are completed within the scheduled time. The system experiences no overloads and reduces request rejection.

IV. CONCLUSION

In cloud computing environment, load balancing and scheduling are very wide concepts.

In this paper we are specifically focused on load balancing. In cloud computing, designing an algorithm with an aim to perform load balancing to resource in an optimized way has been a complicated task. During load balancing there are various techniques and constraints are applies but as cloud computing is too vast all aspects are not being able to capture at a same time. This paper explores various methods of load balancing in cloud computing environment. The management and balancing the resources is complex and therefore the demand for two stage load balancing scheduling algorithm is applying. The proposed load balancing method is based on Idle-Server Monitoring algorithm and Service Level Agreement algorithm. The method also optimizes the number of physical resources while satisfying conflicts. This method has better make span than all previous load balancing scheduling algorithm methods. And hence improves the output of the system. By combining these different parameters an efficient load balancing scheduling algorithm can be obtained which can improve the overall performance of the cloud services.

The main aim of proposed algorithm is to satisfy both SLA and Load balancing among the servers. In this paper, the two stage scheduling algorithm SLA based scheduling algorithm and Idle-Server Monitoring Algorithm is elaborated. SLA based scheduling algorithm schedules the tasks to the respective cluster based on SLA and Idle-Server Monitoring Algorithm balances the load among the clusters as well as cluster.

REFERENCES

[1] Madhurima Rana, Saurabh Bilgaiyan, Utsav Kar - A Study on Load Balancing in Cloud Computing Environment Using Evolutionary and Swarm Based Algorithms, 2014

[2] Yu J, Buyya R. Scheduling scientific workflow applications with deadline and budget constraints using genetic algorithms. Scientific Programming Journal, 2006, 14(3 /4): 217-230.

[3] Sakellriou R, Zhao H, Tsiakkour E, et al. Scheduling workflows with budget constraints, 05-22 Pisa, Italy:University of Pisa, Dipartimento di Informatica, 2005:347-357.

[4] Yu J, Buyya R, Than CK.A cost-based scheduling of scientific workflow applications on utility grids. The 1st International Conference on E-Science and Grid Computing. Washington, DC: IEEE Computer Society, 2005: 140-147.

[5] Menasc D A, Casalicchio E. A framework for resource allocation in grid computing.Proceedings of 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems. Washington, DC: IEEE Computer Society, 2004:259-267.

[6] Huifang Li, Siyuan Ge, Lu Zhang. A QoS- based Scheduling algorithm for Instance-intensive Workflow in Cloud Environment. 26th Chinese Control and Decision Conference (CCDC), 2014:4094-4099.

[7] Mark D. Ryan, ―Cloud computing for Enterprise Architectures: Concepts, Principles and Approaches‖ , 2013

[8] Liu Ke, Jin Hai, Chen Jinjun, Liu Xiao, Yuan Dong, Yang Yun.A compromised-time-cost scheduling algorithm in SwinDeW-C for instance-intensive cost-constrained workflows on a cloud computing platform. International Journal of High Performance Computing Applications, 2010, 24(4): 445-456.