A Method of Optimized Load Balancing For Cloud Computing

Khushboo Urmaliya[1], Ankur Mudgal[2]
[1] Dept of CSE
[2]Asst. Professor, Dept of CSE
[1, 2] Shri Ram Institute of Science & Technology, Jabalpur, Madhya Pradesh, India.

Abstract- *In a cloud environment, the aim of using optimal resources can be achieved using a load-balancing technique. The load-balancing technique assigns a set of requests into a set of resources for distributing the load. It is one of the significant issues in cloud computing and known as an NP-hard problem. Therefore, many nature-inspired meta-heuristic techniques are proposed to provide high efficiency. However, despite the importance of the nature-inspired meta-heuristic techniques for solving the problem of the load-balancing in the cloud environment, there is not a complete and detailed paper about reviewing and studying the main important issues in this domain. Therefore, this thesis presents comprehensive coverage of the nature-inspired meta-heuristic techniques applied in the area of the cloud load-balancing. The main goal of this thesis is to highlight the emphasis on optimization algorithms and the benefits that they provide to overcome the cloud load-balancing challenges. In addition, to solve the load-balancing problem in the cloud environments, the advantages and disadvantage of the nature-inspired meta-heuristic honey-Bee algorithms have been proposed and their significant challenges are considered for proposing the techniques that are more effective in the cloud computing environment.*

Keywords- Load balancing, Cloud computing, Scheduling, Honey Bee, Nature Inspired, Response Time.

I. INTRODUCTION

Cloud computing is a computing paradigm that provision the infinite number of computing resource to end user based upon their demands, anywhere and anytime in pay-as-you-go fashion where users pay only for the services they use. Users can access various types of services from cloud like resource pooling, elastic and flexible [1], scalability (horizontal as well as vertical) [2], utility services, throughput, performance, high availability, managed services etc. due to centralized management of cloud infrastructure.

Cloud service provider and users can leverage the benefit of virtualization in technology as well as dynamic resource scheduling techniques. Effective resource scheduling not only execute the tasks in minimum time but also increase the resource utilization ratio, i.e. reduces the resource consumption. Scheduling of tasks becomes a matter of great concern due to increase in workload, continuously at the cloud datacenters that may lead to the scarcity of cloud resources. Hence cloud computing is still in its infancy and more research is required to map the tasks with cloud resources efficiently and fulfill the objective of scheduling (improve the quality of service parameters). Objective of scheduling is to specify best resource for execution of tasks so that scheduling algorithm can improve various quality of services (QoS) parameters like resource utilization, task rejection ratio, reliability, energy consumption, execution cost etc. without affecting service level agreement (SLA), considering constraint (deadline, priority etc.) and avoid the load imbalance (over utilized and underutilized) problem. All the terms resource provisioning and resource scheduling comes under the resource management in cloud environment as shown in Figure 1 [3]. Further, resource scheduling is broadly categorized into several terms, resource mapping, resource brokering, load balancing and resource allocation etc. Initially, we will discuss the basic concepts of resource provisioning techniques that comes prior to scheduling techniques in cloud computing.

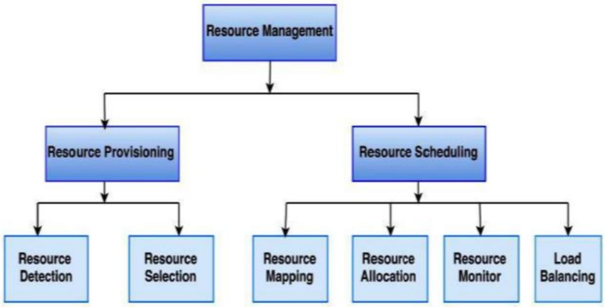

Fig 1: Resource management in cloud environment

Favours provided by cloud resource provisioning are mentioned below [4]:

- Response time and makespan time of upcoming workload is reduced by efficient resource provisioning techniques.

- Better resource utilization can reduce the problem of Overprovisioning and under-provisioning.
- If virtual machine start-up delay is less, then it provides better resource provisioning in cloud environment.
- Effective cloud resource provisioning algorithm increases the robustness as well as fault tolerance capability.
- Resource provisioning algorithm reduces power consumption without affecting SLA violation.

Aim of scheduling is to find the best cloud resources for upcoming end users' applications (tasks) to improve the key performance parameters (QoS parameters) and improve the resource utilization ratio [5]. There are various performance indicator parameters like response time, execution cost, makespan time, reliability, energy consumption etc. in cloud computing. We must analyze and improve them using efficient resource scheduling algorithm to fulfill the requirement of end users as well as service provider without affecting the service level agreement (SLA) violation. Resource scheduling becomes a prominent issue in the field of cloud computing due to heterogeneity, dynamism and dispersion of resources that are not resolved by existing scheduling algorithms. Therefore, we need a scheduling algorithm that distribute the heterogeneous workload among the cloud resources (VMs) based upon the capacity of the resources and overcomes the problem of overload and underloaded.

On the current years, the load balancing of the servers on the cloud data center is a hot topic and to balance the load of the servers, it is necessary to deploy the tasks to the optimal server with minimum load. Most of the existing works deploy the tasks to the randomly selected servers on the cloud data center. This may affect the performance and efficiency of the resources on the cloud data center and unnecessary increase the completion time of the tasks. In cloud environment, the proper load balancing mechanism provides the satisfactory performance and increase the resource availability on the cloud data center. Due to the increasing demands of the tasks, selection of the optimal servers dynamically is an important research topic in cloud.

To meet the above-mentioned challenge we propose a dynamic optimized load balancing mechanism based on honey bee optimization in cloud based model. The algorithm finds an optimal loaded server with a minimum load for each task dynamically and balances the load of the cloud data center with efficient resource utilization.

II. RELATED WORK

The main focus of load balancing is the efficient utilization of the virtual machines and balancing the incoming requests to various virtual machines [6]. A large public cloud consists of many nodes which may spread over different geographical locations. Hence partitioning of cloud is done to manage a large cloud. In Cloud Analyst [7], three different algorithms are there for load balancing:

1) Round Robin algorithm is a very simple load balancing algorithm that allocates the new cloudlets on the available virtual machines in a circular order. This algorithm is very simple and can be implemented easily. It is static in nature i.e., prior information of user tasks and system resources is required.

2) Equally Spread Current Execution (ESCE) algorithm is active VM Load balancing algorithm. It distributes the load equally on each virtual machine in a cloud environment. ESCE VM Load Balancer maintains a list of virtual machines. It continuously checks the VM list and the task queue. If a VM is found free, then cloudlet request is allotted to that VM. Meanwhile, VM Load balancer also checks for the overloaded VM so as to reduce its load by moving some load to an idle or under loaded virtual machine.

3) Throttled load balancing algorithm determines the appropriate virtual machine which can handle the assigned load with greater ease. It is dynamic in nature as it maintains the present state of the all VMs in a cloud environment. If an appropriate VM is found, then throttled VM Load balancers accept the cloudlet request and allocate it to that virtual machine.

Ant Colony Optimization algorithm proposed by Linan Zhu et al., [8] is based on the behavior of ants to solve travelling sales man problem. Ants drop pheromone liquid substance while following a path for the search of food source. Other ants follow the path based on high pheromone strength. This concept is used by Ant colony algorithm to choose optimal path from source to destination that results into reduction of response time and distribute the work load of network. Karaboga, proposed a foraging behavior of honey bee swarm in 2005 [9]. Honey bee search for their food and informs other bees in beehive about the quantity and quality of food by performing waggle dance. There are three types of bees in the algorithm:

1) Scout bees: Arbitrarily Search for food source. Perform waggle dance to show the quality of food.

2) Employed bees: Collect all the information about food source and exchange the information obtained with onlooker bees.
3) Onlooker bees: Calculate the fitness value to find the best food source.

In respect of load balancing of incoming requests, tasks from overloaded machines are referred as honey bees, these tasks are transferred from overloaded machines to under loaded machines. Dynamic nature of the algorithm makes the changes in the status of the load to be reflected. Updated load on that particular machine is taken into account for other waiting tasks [9].

Zeng Zeng and Bharadwaj proposed a request balancing strategy, also known as optimal metadata replication and request balancing strategy [10], [11]. This paper focuses on choosing appropriate metadata server when data retrieval request is initiated. Chosen metadata server gives command to raw data server for actual data retrieval. There may be replicated copies of the object on multiple metadata servers. For distributing the data to respective metadata servers, Zipf law of distribution is applied and aim at achieving minimum mean response time in a highly loaded cloud environment.
Nader Mohamed et al., designed a load balancing technique to handle download of large files called as dual direction technique [12]. This approach makes use of two Dual Direction FTP servers (DDFTP) to download a large single file. One of the server download half portion of a file from left to right while another server download another half from right to left direction. When both the servers find a common middle point of file, download operation terminates. A reliable, ordered delivery feature of TCP allows file blocks to be downloaded in a sequential manner hence reducing the overhead of coordination and data loss. This unique method of parallelizing the download enhances effective bandwidth utilization and reduces response time to give better performance.

Klaithem Al Nuaimi et al., [13] presented a simple algorithm to tackle the issue of balancing the load in giving Data as a Service (DaaS) in the Cloud. The algorithm has a basis of some prior approach for efficient data download in dual direction. Main objective of this paper is to solve the problem of the requirement of high volumes of storage when data is replicated in multiple cloud servers. Instead of storing full replicas of file, this algorithm gives a technique to store partial replicas of the data file [14]on Cloud servers. An efficient method is provided, to download the information from various distributed servers and organise them in proper order.

There is lot of research in growing phase and some research challenges are still over looked like load balancing, resource provisioning, scheduling of applications, energy consumption etc. in cloud environment. Research related to resource scheduling is still in infancy phase and needs the improvement. Here, we will discuss some review papers based upon resource provisioning and scheduling techniques that are related to our research and useful for the present survey. M. Amiriand L. Khanli [15] presented a comprehensive survey regarding to the prediction of future demand of applications in different aspects. J. Zhang et al. [16] discussed about the resource provision techniques and algorithm design. Further this is observed that results of one phase can be used for another phase, finally it focuses at virtual machine migration, availability etc. parameters but both surveys do not discuss about the over provisioning and under-provisioning problem profoundly. S. Smanchat and K. Viriyapant [17] have divided the taxonomy of workflow scheduling into two parts: scheduling criteria and scheduling generation in cloud computing. Several resource scheduling algorithms have been discussed in this paper to improve the research as well as development of scheduling algorithms. P. Dave et al. [18] have presented a comparative study based upon different scheduling techniques to measure the QoS parameters reliability, scalability, resource utilization, throughput, execution cost etc. K. Radha et al. [19] have discussed problem of resource allocation (mapping upcoming requests with available resources) and management in the field of cloud. To solve the mentioned issues, capacity allocation algorithm is proposed for multi-tier system. C. Nandakumar and K. Ranjithprabhu [20] analyzed and compared the performance of various heuristic and meta-meta-heuristic algorithm based upon QoS parameters in cloud environment. Metrics satisfied by the existing algorithms are depicted in tabulated form but Most of the algorithms did not considered energy efficiency parameter that is important for green computing and this survey is limited only for some heuristic and meta-heuristic algorithm. To overcome the limitations of min-min algorithm [21], proposed improved load balancing algorithm that reduce the time and increase the resource utilization ratio. Priority of the user is also considered in this algorithm to assure guarantee of the service. Y. Mao et al. proposed min-max scheduling algorithm that assigned the larger tasks to best resources at the starting to improve the makespan time, response time as well as resource utilization ratio of user request but algorithm face the problem of overutilization and underutilization of resources and failed to improve required parameters. To cover the limitation of max-min algorithm, modified max-min algorithm has been proposed by O. Elzeki et al. [22] that was based upon the concept of excepted execution time instead of complete processing time and improves makespan time but unable to improve other parameters like cost, energy etc. There are several improved versions of max-min algorithm that have been proposed [23] to optimize the QoS parameters in cloud computing though proposed algorithms failed to improve the key performance indicator parameters. These are static algorithms and don't work well in dynamic environment (cloud computing).

III. PROPOSED SYSTEM

Dynamic algorithms provide better results in heterogeneous environment. These algorithms are more stretchable. Dynamic algorithms can take in charge of the dynamic changes to the attributes. Main advantage of this algorithm is that, the selection of task is based on present state and this will help to improve the performance of the system. Proposed system performs dynamic load balancing using little modification in honey bee algorithm. System architecture is shown below:

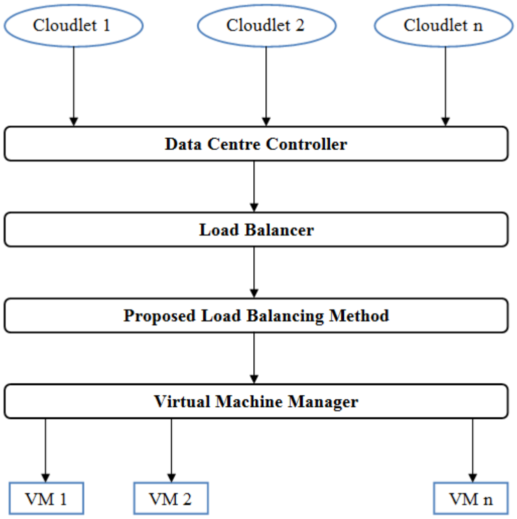

Fig 2: System Architecture

Cloud user base generates requests in the form of cloudlet and then cloudlets are received by data center controller. Load balancer component uses load balancing algorithm for balancing load among various virtual machines by instructing virtual machine manager. System architecture uses nature inspired honey bee based algorithm for efficient load balancing of resources to maximize utilization of resources.Flowchart of proposed algorithm is as below:

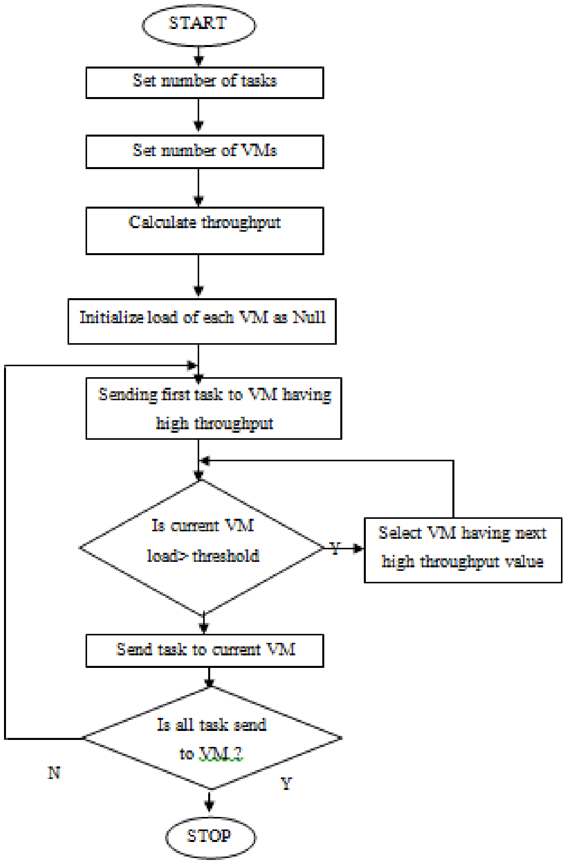
START
Set number of tasks
Set number of VMs
Calculate throughput
Initialize load of each VM as Null
Sending first task to VM having high throughput
Is current VM load> threshold
Y
Select VM having next high throughput value
Send task to current VM
Is all task send to VM ?
N
Y
STOP

Fig 3: Flowchart of algorithm

IV. RESULT

There are different sizes virtual machines are used for this evaluation. Table 1 shows the performance comparison between the proposed algorithm, Existing algorithm and Round Robin algorithm with respect to makespan using different CSP. Figure below shows the graphical representation of the results where CU = 50 and CSP=5.

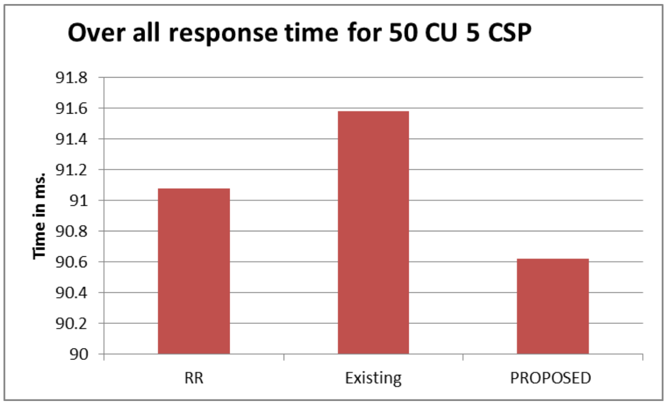

Fig 4:Performance chart for ORT

Result Chart for 20 Cloud users and 3 Cloud Service providers.

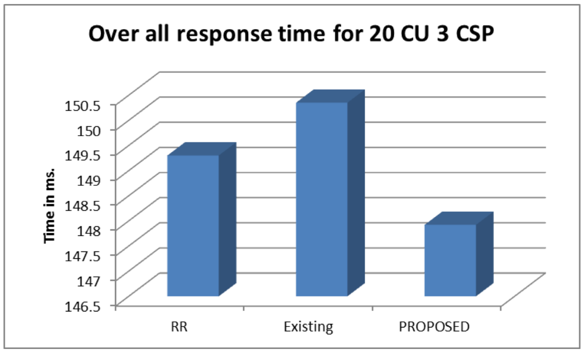

Fig 5:Performance chart for ORT

Results of above evaluations show that proposed algorithm completes user allocation with lower response time and higher performance as compared to existing Load Balancing algorithms. Performance of proposed algorithm is better than existing algorithms. Results shows that proposed algorithm behaves better in terms of response time after testing it with Cloud Analyst Simulator.

IV. CONCLUSION

This thesis focuses on the problem of load balancing in cloud computing environment. In this work, a better approach has been designed for the load balancing of data requests received at the large scale cloud data centers. In such data centers thousands of servers are connected by interconnection network. For high priority tasks, servers that are currently having less number of high priority tasks are selected so that such task gets executed faster. Proposed algorithm is compared with existing load balancing algorithms and it is observed that both response time and processing time are improved in the proposed strategy.

REFERENCES

[1] Kumar, M., Sharma, S.C., PSO-COGENT, 2018. Cost and energy efficient scheduling in cloud environment with deadline constraint" in Sustainable Computing. Informatics and Systems 19, 147–164.

[2] Kumar, Mohit, Sharma, S.C., 2017. Dynamic load balancing algorithm for balancing the workload among virtual machine in cloud computing. in Procedia Computer Science 115, 322–329.

[3] Madni, S., et al., 2016. Resource scheduling for infrastructure as a service (IaaS) in cloud computing: challenges and opportunities. J. Netw. Comput. Appl. 68, 173–200.

[4] Singh, S., Chana, I., 2016. QoS-aware autonomic resource management in cloud computing: a systematic review. ACM Comput. Surv. 48 (3) article 42.

[5] Singh, S., Chana, I., 2016. Cloud resource provisioning: survey, status and future research directions. Knowl. Inf. Syst. 49 (3), 1005–1069.

[6] Kavitha, K V and Suthan, Vinza V, "Dynamic Load Balancing in Cloud Based Multimedia System with Genetic Algorithm", International Conference on Inventive Computation Technologies (ICICT), pp. 1–4, 2016.

[7] Wickremasinghe, Bhathiya and Calheiros, Rodrigo N and Buyya, Rajkumar "Cloud Analyst: A CloudSim-Based Visual Modeller for Analysing Cloud Computing Environments and Applications", 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 446-452, 2010.

[8] Zhu, Linan and Li, Qingshui and He, Lingna, "Study on Cloud Computing Resource Scheduling Strategy Based on the Ant Colony Optimization Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 9, No. 5, pp. 54–58, 2012.

[9] Karaboga, Dervis "An Idea Based on Honey Bee Swarm for Numerical Optimization", Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department, October 2005.

[10] Sheeja, YS and Jayalekshmi, S, "Cost Effective Load Balancing Based on Honey Bee Behavior in Cloud Environment", First International Conference on Computational Systems and Communications (ICCSC), pp 214–219, 2014.

[11] Zeng, Zeng and Veeravalli, Bharadwaj, "On the Design of Distributed Object Placement and Load Balancing Strategies in Large-Scale Networked Multimedia Storage Systems", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 10, pp. 369–382, 2008.

[12] Zeng, Zeng and Bharadwaj, Veeravalli, "Optimal Metadata Replications and Request Balancing Strategy on Cloud Data Centers", Journal of Parallel and Distributed Computing, Elsevier, Vol. 74, No. 10, pp. 2934–2940, 2014.

[13] Al-Jaroodi, Jameela and Mohamed, Nader, "DDFTP: Dual-direction FTP", Proceedings of the 2011, 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, IEEE Computer Society, pp. 504–513, 2011.

[14] Darji, Vinay and Shah, Jayna and Mehta, Rutvik, "Survey Paper on Various Load Balancing Algorithms in Cloud Computing", International Journal of Scientific & Engineering Research, Volume 5, Issue 5, pp.583–588, May 2014.

[15] Amiri, M., Khanli, L., 2017. Survey on prediction models of applications for resource provisioning in cloud. J. Netw. Comput. Appl. 82, 93–113.

[16] Zhang, J., et al., 2016. Resource provision algorithms in cloud computing. A survey" Journal of Network and Computer Applications 64, 23–42.

[17] Smanchat, S., Viriyapant, K., 2015. Taxonomies of workflow scheduling problem and techniques in the cloud. J. Netw. Comput. Appl. 52, 1–12.

[18] Dave, P., et al., 2014. Various job scheduling algorithms in cloud computing: a survey. In: International Conference on Information Communication and Embedded Systems, pp. 1–5.

[19] Radha, K., et al., 2014. Allocation of resources and scheduling in cloud computing with cloud migration. Int. J. Appl. Eng. Res. 9 (19), 5827–5837.

[20] Nandhakumar, C., Ranjithprabhu, K., 2015. Heuristic and meta-heuristic workflow scheduling algorithms in multi-cloud environments—a survey. In: International Conference on Advanced Computing and Communication Systems, pp. 1–5.

[21] Chen, H., Wang, F., Helian, N., Akanmu, G., Oct. 2013. "User Priority Guided Min-Min Scheduling Algorithm for Load Balancing in Cloud Computing," in National Conference on Parallel Computing Technologies. PARCOMPTECH), Bangalore, India.

[22] Elzeki, O.M., et al., 2012. Improved max-min algorithm in cloud computing. Int. J. Comput. Appl. 50 (12).

[23] Kanani, B., Maniyar, B., 2015. Review on max-min task scheduling algorithm for cloud computing. Journal of Emerging Technologies and Innovative Research 2 (3).