Review on Performance Evaluation for Cloud Computing

Sumit Katiyar [1], Raj Kumar Paul [2],
[1,2]Department of Engineering Computer Science & Engineering,
RKDF University, Bhopal, India

Abstract: Cloud Computing Technology serves computing resources as a service. Cloud mainly focused on optimum utilization of computing resources. Major cloud services are SaaS, PaaS, SaaS and IaaS. Cloud user uses cloud services based on "Pay and use" concept. Now a days, cloud users and services are getting increased. So it quit challenging for cloud service providers to provide efficient services in optimum cost to cloud users. In cloud computing load balancing plays a vital role to improve cloud performance. A load balancing method transfer or migrates a task for over loaded machine to under loaded machine without effecting current user running tasks. Various load balancing methods are suggested by cloud researchers. Existing load balancing methods encounters with several issues such a poor throughput, higher waiting time, improper balancing, poor make span time. In this survey we are presenting review of various performance improvement parameters and load balancing methods for cloud computing.

Keywords: Cloud Computing, Load Balancing, Performance Improvement.

I Introduction

Cloud computing technique is shifting from hypothesis to observe. As a recent large scale level computing with distributed paradigm, which supported virtualized resource pool, cloud computing resultant from the ever-increasing interaction and profound development of parallel, distributed, grid and utility computing with various internet services, etc. If solely with accesses to cloud knowledge centers, cloud computing may facilitate users everywhere the globe to on-demand leverage a variety of IT related computing services supported a strong concept of "pay per use" model means user only needs to pay for services which he is utilizing, such as infrastructure services, platform services and various software services.

In adding, cloud computing too corresponds with the essential plan of inexperienced computing. By elastic able, highly scalable and dynamic management of the computing resource pool, cloud computing techniques provides efficient resources utilization in less cost [1–4]. Service request planning is one among the key necessary ways to realize those. The main goal of designing and using a job planning is to realize a high performance computing and therefore the best system outturn. Standard job planning method does not be able to provide for planning within the cloud computing environments. As per a straight forward classification job planning method in cloud computing environment will be classify into two major categories, one is "Batch mode heuristic planning algorithms or BMHA" and another is "on-line mode heuristic algorithms" [5].

In BMHA, Jobs are queued and picked up into a collection after they take place within the system. The planning algorithmic program can begin when a hard and fast amount of your time. The key samples of batch mode heuristic planning algorithms are mostly based on algorithmic programs are initial come back initial Served planning algorithm (FCFS), spherical Robin planning algorithmic program (RR), Min–Min algorithmic program and Max Min algorithmic program. By using "on line mode heuristic planning", algorithmic program, Jobs is scheduled after they arrive within the system. The cloud computing atmosphere could be a heterogeneous system and therefore the speed of every processor varies quickly, the "on line mode heuristic", planning algorithms are so much useful for a cloud computing atmosphere. Most match task planning algorithmic program (MFTF) is applicable example of On-line mode heuristic planning.

II Cloud Computing and Load Balancing

Cloud computing could be a model for enabling appropriate, on-demand network access to a joint pool of configurable computing resources that may be quickly provisioned and discharged with least management effort. The underlying plan of cloud computing is that the separation of applications from the in operation systems and therefore the hardware on that they run [6].

Cloud computing distribute applications via the web, that are accessible from Internet browsers, desktop laptop and mobile apps whereas the code and knowledge are hold on servers at a distant location. In the past, loads of people upset regarding losing our documents and files if one thing unhealthy happened to our laptop, sort of a virus or a hardware malfunction. Nowadays, our knowledge is migrating on the far side the boundaries of our personal computers and everyone our knowledge would still safely reside on the online and which are accessible by any one form any locations by using Internet-connected pc, with in the earth thanks to cloud computing.

2.1 Performance Evolution

Issue in Cloud Computing - Cloud computing resources must be compatible, high performance and powerful. High performance is one of the cloud advantages which must be satisfactory for each service. Higher performance of services and anything related to cloud have influence on users and service providers. Hence, performance evaluation for cloud providers and users is important. There are many methods for performance prediction and evaluation; we use the following methods in our evaluation:

- Evaluation based on criteria
- Evaluation based on characteristics
- Evaluation based on simulation

Following factors affect the performance of cloud computing.

2.2 Load Balancing in Cloud Computing

It is a method of re-assigning the entire load to the individual nodes of the shared system to build resource utilization efficient and to get better response time of the job, concurrently removing a state in which a few of the nodes are over loaded where as a number of other node are under loaded [7].

A dynamic load balancing algorithm does not consider the earlier performance of the system, that is, it depends on the current behavior of the system. The main things to think about while mounting such algorithm are: evaluation and comparison of load, stability of different system, performance of system, communication between the nodes, nature of job to be transferred and selection of nodes. This load can be considered in terms of Network loads, CPU load, amount of memory utilize [8].

2.3 Significance of Load Balancing in Cloud Computing

Load balancing is the pre necessities for raising the cloud performance and for entirely utilizing the resources. Load balancing is main issues associated to cloud computing. The load perhaps memory, CPU capacity, network load or delay loads. It is always required that work load must be shared among the various nodes of the distributed system so as to improve the resource utilization and also for better performance of the computing organization [9].

This can aid to keep away from the situation where a few of the nodes are either overloaded or under loaded in the network. Load balancing able to be either centralized or decentralized. Load Balancing algorithms are used for implementing. Nowadays cloud computing is a set of numerous data centers which are sliced into virtual servers and located at different geographical location for providing services to clients [10].

The arrival of load can affect some server to be overloaded while other servers possibly idle or under loaded. Uniformly distributing the load improves the performance of the cloud by transferring load from the overloaded server. Well-organized scheduling and efficient resource allotment is a characteristic of cloud model based on which the system's performance is considered. These characteristics have resulted on cost optimization, which can be then achieved by improving the response time and processing time.

2.4 Why Load Balancing?

-The major objectives of load balancing are-

1. To make improvement in performance.
2. To contain a backup plan in case the system fails even moderately.
3. To keep the system stability.
4. To accommodate future amendment in the system.

III Literature Survey

Worked on Load Balancing in Cloud Computing Environment Using Improved Weighted Round Robin Algorithm for No preemptive Dependent Tasks. Cloud computing uses the concepts of scheduling and load balancing to migrate tasks to underutilized VMs for effectively sharing the resources.

The scheduling of the no preemptive tasks in the cloud computing environment is an irrecoverable restraint and hence it has to be assigned to the most appropriate VMs at the initial placement itself. Practically, they arrived jobs consist of multiple interdependent tasks and they may execute the independent tasks in multiple VMs or in the same VM's multiple cores. Also, the jobs arrive during the run time of the server in varying random intervals under various load conditions.

The participating heterogeneous resources are managed by allocating the tasks to appropriate resources by static or dynamic scheduling to make the cloud computing more efficient and thus it improves the user satisfaction. Objective of this work was to introduce and evaluate the proposed scheduling and load balancing algorithm by considering the capabilities of each virtual machine (VM), the task length of each requested job, and the interdependency of multiple tasks. Performances of the proposed algorithm were studied by comparing with the existing methods.

Kumar *et al.* [2] worked on Load Balancing in Cloud Data Center Using Modified Active Monitoring Load Balancer. Cloud Computing is a hot topic of research for the researchers these days. With the rapid growth of Internet technology cloud computing have become main source of computing for small as well big IT companies. In the cloud computing milieu the cloud data centers and the users of the cloud-computing are globally situated, therefore it is a

big challenge for cloud data centers to efficiently handle the requests which are coming from millions of users and service them in an efficient manner.

Load balancing in this environment means equal distribution of workload across all the nodes. Load balancing provides a way of achieving the proper utilization of resources and better user satisfaction. Hence, use of an appropriate load balancing algorithm is necessary for selecting the virtual machines or servers. This paper [4] focused on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers.

The proposed algorithm [4] in this paper had been implemented using Cloud Analyst simulator and the performance of the proposed algorithm is compared with the three algorithms which are preexists on the basis of response time. The experiment carried out in the paper clearly shows that the proposed algorithm performs better than the existing algorithms.[5] In this research paper authors presented an optimal Load Balancing method for cloud computing method. This method is based on efficient utilization of Virtual Machines. Load balancing method have an important and major concerns in the cloud computing environment. Cloud computing technology comprises of various hardware, software based resources, and managing these resources will play an important role in executing a remote location based request of cloud user. Now these days clients from various parts of the world are requesting or demanding for the various services in a rapid rate.

Author described in this present scenario the load balancing methods should be very efficient in allocating the user request and also ensuring the usage of the cloud resources in an efficient and intelligent manner so that underutilization of the resources will not occur in the cloud environment. In this research work, author [6] presented a novel Virtual machine assign load balance method, which allocates all the incoming user requests to all the available virtual machines in an efficient manner.

After that the performance of the method were analyzed by using simulator cloud sim and various results are calculated and compared with existing active Virtual Machine based load balance algorithm. Simulation results clearly show that the proposed algorithm of author distributes the cloud load on all available virtual machines without under or over utilization. In this research paper author represented a novel method for load balancing in cloud data center. In a large, scale cloud computing environment the end users and cloud data centers both are distributed geographically based across the globe. The most challenging task for a cloud data centers is to, how to handle and manages various services of the millions of requests, which are arriving very frequently from cloud users efficiently and correctly.

In this research paper [8] describe that in cloud computing load balancing method is requires to distribute the cloud workload dynamically equally in between all the cloud nodes. Efficient Load balancing techniques helps cloud service provider and user to achieve a high-level user satisfaction and optimal resource utilization and ensuring an efficient and fair allocation of every computing resource.

Proper load balancing [9] in cloud environment aids in minimizing resource consumption, enabling scalability, implementing fail over, avoiding bottlenecks and provisioning. In this research work researcher proposed a new concepts of, "Central Load Balancer", based load-balancing algorithm, for balancing the load among virtual machines in cloud data center. Experimental results clearly shows that authors proposed algorithm achieved much better load balancing in to a wide scale area cloud, computing environment as compared to previous existing load balancing algorithms.

Shaw *et al.* [11] presented a survey on scheduling and load balancing methods in Cloud Computing environment. Author describes, now these days' cloud computing technology is the most innovative and emerging method due to its some unique features and various qualities such as elasticity of resource provisioning and the pay per use based pricing model, this model enables cloud users to pay only as per their need. Cloud services user can able to access anywhere and anytime through required commodity hardware only, cloud service demands are increasing day by day. Cloud services must be providing a higher performance output to the cloud user and beneficial for the Cloud Service Provider (CSP).

To achieving this important goal, many challenges are to be faced. Efficient Load balancing is one of them, which helps the Cloud Service Provider to meet the various qualities of service requirements of the cloud users and same times maximize his profit by efficient optimum use of the various cloud resources. For balancing the total load in to the cloud environments, the workloads and resources should be properly arrange and schedule in to an efficient manner [11].

Load balancers are used to identify which back end servers are overloaded, for various balancing and scheduling algorithms. The selected cloud server allocates various IT resources, and arranges or schedules all these applications dynamically on various free virtual machines, which are located on the same physical machine [12]. It is also the responsibility of the cloud service provider dynamically allocate the various virtual machines across physical machines, for equal and efficient work load distribution and to avoid situation for all types of over utilization or underutilization of any cloud resources.

In this research paper [12], authors described a new method to solve the problem of load balancing and task scheduling in Cloud Computing environments and described some of their short comings for further future development. Authors also described Virtual Machines migration issues involved in load balancing.

Nuaimi *et al.* [13] presented, a survey on load balancing techniques and various challenges are addressed for cloud computing. Cloud, computing is a new emerging technology, which has a new standard for large scale parallel computing and distributed computing. Cloud computing provides shar-

ing of resources, information, various software packages and other computing resources as per client requirements at particular time. Cloud computing is growing continuously and rapidly. More users that are new attracted towards various utility computing, better, and fast service.

For better management and utilization of available cloud resources, more efficient methods for load balancing are still required. Therefore, that study of loads balancing methods in cloud computing environments are interesting area of research for researchers. In addition, by efficient and better load balancing in cloud, increased performance of cloud environments and user gets better services. In this research paper, [14] presented and discussed different loads, balancing techniques used to solve the issue in cloud computing environment.

IV Existing Load Balancing Methods in Cloud Computing

Round-Robin Algorithm- This is static load balancing algorithm which uses the round robin method for allocating work. It chooses the first node at random and then, assigned job to all other nodes in a round robin manner. Without any kind of priority the jobs are assigned to the processors in round order. Due to the non-uniform allocation of workload, this algorithm is not suitable for cloud computing. Some nodes get deeply loaded and some nodes get calmly loaded since the running time of any procedure is not known in advance [4]. Opportunistic Load Balancing Algorithm- It is static load balancing algorithm so it does not believe the current workload of the VM.

This efforts to keep every node busy. This algorithm deals rapidly with the unexecuted jobs in random order to the present available node. Every task is assigned to the node randomly. It present load balance schedule without fine results. The task will process in slow in way because it does not calculate the existing execution time of the node [8].

Weighted Round Robin Algorithm-The weighted round robin considers the resource capabilities of the VMs and assigns higher number of tasks to the higher capacity VMs based on the weightage given to each of the VMs. But it failed to consider the length of the tasks to select the appropriate VM [9]. Min-Min Load Balancing Algorithm- This is also a static load balancing algorithm so the parameters related to the task are known in advance. In this algorithm the cloud manager initially deals with the jobs having least execution time by assigning them to the processors according to the ability of complete the job in particular completion time. The jobs having highest execution time has to wait for the unspecific phase of time [2].

4.1 Honeybee Foraging Behavior

It is a nature inspired Algorithm for self-organization. Honey bee achieves global load balancing through local server actions. The performance of the system is enhanced with increased system diversity. The main problem is that throughput is not increase dwith an increase in system size. When the diverse population of service types is required then this algorithm is best suited [15].

4.2 Active Clustering

In this algorithm same type nodes of the system are grouped together and they work together in groups. It works like as self-aggregation load balancing technique where a network is rewired to balance the load of the system. Systems optimize using similar job assignments by connecting similar services. System Performance improved with improved resources. The throughput is improved by using all these resources effectively.

4.3 Compare and Balance

This algorithm is uses to reach an equilibrium condition and manage unbalanced systems load. In this algorithm on the basis of probability (no. of virtual machine running on the current host and whole cloud system), current host • randomly select a host and comparetheir load. If load of current host is more than the selected host, it transfers extra load to that particular node. Then each host of the system performs the same procedure. This loadbalancing algorithm is also designed and implemented to reduce virtual machines migration time. Shared storage memory is used to reduce virtual machines migration time [14].

4.4 Lock-free Multiprocessing Solution for LB

It proposed a lock-free multiprocessing load balancing solution that avoids the use of shared memory in contrast to other multiprocessing load balancing solutions which use shared memory and lock to maintain a user session. It is achieved by modifying kernel. This solution helps in improving the overall performance of load balancer in a multi-core environment by running multiple load-balancing processes in one load balancer [11].

4.5 Ant Colony Optimization

Ant algorithms is a multi-agent approach to difficult combinatorial optimization problems. Example of this approach is traveling salesman problem (TSP) and the quadratic assignment problem (QAP). These algorithms were inspired by the observation of realant colonies. Ant's behavior is directed more to the survival of the colonies.They not think for individual.

4.6 Shortest Response Time First

The idea of this algorithm is straight forward. In this each process is assigned a priority which is allowed to run. In this

equal priority processes are scheduled in FCFS order. The (SJF) algorithm is a special case of general priority Scheduling algorithm.In SJF algorithm is priority is the inverse of the next CPU burst. It means, if longer the CPU burst then lower the priority. The SJF policy selects the job with the shortest (expected) processing time first. In this algorithm shorter jobs are executed before long jobs. In SJF, it is very important to know or estimate the processing time of each job which is major problemof SJF [7].

4.7 Based Random Sampling

This algorithm is based on the construction of the virtual graph having connectivity between the all nodes of the system where each node of the graph is corresponding to the node computer of the cloud system. Edges between nodes are two types as Incoming edge and outgoing edge that is used to consider the load of particular system and also allotment the resources of the node. It is very good technique to balance the load [9].

4.8 Max-Min Load Balancing Algorithm

Max Min algorithm works similar as the Min-Min algorithm except for the following: after finding the minimum implementation time, the cloud manager deals with jobs having maximum implementation time. The assigned task is detached from the list of the jobs that are to be assigned to the processor and the implementation time for all other jobs is updated on that processor. Due to its static approach the necessities are known in advance then the algorithm performed fine [4].

V Challenges in Load Balancing in Cloud Computing

In cloud environment, cloud service providers and cloud users are often two different parties that have their own interests; they do not share their detailed resource states and workload characteristics. However, we have a tendency to tend to argue that many cloud services and cloud-oriented applications are not economical and it have several challenges.

5.1 Existing Load Balancing Methods for Cloud Computing have Following Challenges

- Slower Response Time-Slower response time shows poor performance for the system.
- Higher Execution Time-Higher execution time shows poor performance.
- Selection of Load balancing-Dynamic Load is balancing shows better performance.
- Selections of partitioning method-Many load balancing methods are based on static partitioning, which are less efficient in large environment.
- Prediction of Task arrival patterns-Jobs are arrive from various nodes in cloud atmosphere, so it is quite difficult to identified exact arrival pattern.
- Priority of Task- During load balancing it is also challenging to execute jobs priority wise.

VI Conclusion and Future Work

Cloud Computing has a glorious and great future, but still various crucial problems are still need to be solved for the actual use of various resources of Inter Cloud computing. One of the very important problem , these problems set, are load distribution and load balancing with better server utilization and less complexity. It plays a crucial role in the development of user request in Inter Cloud environment with a minimum response time.

In a cloud-computing environment, resources are assigned to a user on request basis for a cloud application. When a cloud user or a user application makes a request to cloud service provider, as per the availability of the resource CSP assign resource. In future we will develop an efficient load balancing method for cloud computing and will compared the performance of proposed method with various existing.

REFERENCES

[1] D. Chitra Devi and V. Uthariaraj, "Load balancing in cloud computing environment using improved weighted round robin algorithm for nonpreemptive dependent tasks," *The Scientific World Journal*, vol. 2016, pp. 1–14, Feb 2016. [Online]. Available: http://dx.doi.org/10.1155/2016/3896065

[2] A. Kumar and M. Kalra, "Load balancing in cloud data center using modified active monitoring load balancer," in *2016 International Conference on Advances in Computing, Communication, Automation (ICACCA) (Spring)*, April 2016, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICACCA.2016.7578903

[3] K. Qureshi, B. Majeed, J. H. Kazmi, and S. A. Madani, "Task partitioning, scheduling and load balancing strategy for mixed nature of tasks," *The Journal of Supercomputing*, vol. 59, no. 3, pp. 1348–1359, Mar 2012. [Online]. Available: https://doi.org/10.1007/s11227-010-0539-3

[4] M. Rusek, G. Dwornicki, and A. Orłowski, "A decentralized system for load balancing of containerized microservices in the cloud," in *Advances in Systems Science*. Cham: Springer International Publishing, 2017, pp. 142–152.

[5] G. Soni and M. Kalra, "A novel approach for load balancing in cloud data center," in *2014 IEEE International Advance Computing Conference (IACC)*, Feb 2014, pp. 807–812.

[Online]. Available: https://doi.org/10.1109/IAdCC.2014.6779427

[6] S. K. Mishra, B. Sahoo, and P. P. Parida, "Load balancing in cloud computing: A big picture," *Journal of King Saud University - Computer and Information Sciences*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1319157817303361

[7] S. Joshi and U. Kumari, "Load balancing in cloud computing: Challenges amp;amp; issues," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, Dec 2016, pp. 120–125. [Online]. Available: https://doi.org/10.1109/IC3I.2016.7917945

[8] M. Rana, S. Bilgaiyan, and U. Kar, "A study on load balancing in cloud computing environment using evolutionary and swarm based algorithms," in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, July 2014, pp. 245–250. [Online]. Available: https://doi.org/10.1109/ICCICCT.2014.6992964

[9] G. Shao and J. Chen, "A load balancing strategy based on data correlation in cloud computing," in *2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC)*, Dec 2016, pp. 364–368.

[10] H. Shoja, H. Nahid, and R. Azizi, "A comparative survey on load balancing algorithms in cloud computing," in *Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, July 2014, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICCCNT.2014.6963138

[11] S. B. Shaw and A. K. Singh, "A survey on scheduling and load balancing techniques in cloud computing environment," in *2014 International Conference on Computer and Communication Technology (ICCCT)*, Sept 2014, pp. 87–95. [Online]. Available: https://doi.org/10.1109/ICCCT.2014.7001474

[12] M. K. A. M. Alnazir, A. B. A. N. Mustafa, H. A. Ali, and A. A. O. Yousif, "Performance analysis of cloud computing for distributed data center using cloud-sim," in *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, Jan 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICCCCEE.2017.7867662

[13] K. A. Nuaimi, N. Mohamed, M. A. Nuaimi, and J. Al-Jaroodi, "A survey of load balancing in cloud computing: Challenges and algorithms," in *2012 Second Symposium on Network Cloud Computing and Applications*, Dec 2012, pp. 137–142. [Online]. Available: https://doi.org/10.1109/NCCA.2012.29

[14] V. N. Volkova, L. V. Chemenkaya, E. N. Desyatirikova, M. Hajali, A. Khodar, and A. Osama, "Load balancing in cloud computing," in *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, Jan 2018, pp. 387–390. [Online]. Available: https://doi.org/10.1109/EIConRus.2018.8317113

[15] D. B. L.D. and P. V. Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments," *Applied Soft Computing*, vol. 13, no. 5, pp. 2292–2303, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494613000446