SURVEY ON VARIOUS PERFORMANCE IMPROVEMENT POLICIES FOR CLOUD COMPUTING

Ms. Nikita Gupta¹, Prof. Angad Singh² and Prof. Amit Kanskar³

¹M. Tech Scholar, Department of IT NIIST Bhopal
²Assistant Professor, Department of IT NIIST Bhopal
³Professor & Head, Department of IT NIIST Bhopal

Abstract-Cloud computing technology serves “computing resources as a service” to cloud users. Cloud users can “Pay and use” computing service from cloud service provides. Cloud computing mainly based on virtualization technology and main idea of the cloud is “sharing of computing resources”, which attract users to utilizes cloud services. Most of the IT companies are working on cloud, which increases cloud users. Higher growth of cloud services and cloud consumers results in a delay in services. So it is necessary for cloud service provider to get optimum utilization of computing resources for better performance. In cloud computing system performance totally depends on efficient and optimistic utilization of computing resources. That can be achieved by efficient load distribution policies. Load balancing is the process of apportioning the load among various working nodes of a distributed system to improve both resource utilization and job response time while avoiding a situation where some of the nodes are heavily loaded while others are underloaded. Load balancing ensures that every node in the system does the approximately equal amount of work as per their capacity at any instant of time. Various cloud researchers are working on cloud load balancing methods and suggest different algorithms. This survey paper presents a detailed study of different load balancing policies available for cloud computing and also performs a comparative analysis based on various parameters.

Keywords-Cloud computing, Performance, Load balancing, Cloud service, Virtualization, Virtual machines

I. INTRODUCTION

Cloud computing is a computing provided over the internet. Cloud computing is a rapidly growing technology in today’s Internet world. With the advent of cloud computing, a new age of the Internet has really started. Along with the use of cloud computing in excess of any applications or services running on the Internet so far [1,2]. Although the number of users on the Internet is huge compared to those who used the devices on the cloud, the growth rate of users in cloud computing is growing. When computing is done by distant data centers rather than local it is called as network-centric computing and network-centric content. With this idea and with advancement in Internet technologies two new computing models are widely accepted, Grid computing and Cloud or Utility computing, in which cloud computing is latest. Cloud computing is a new paradigm of internet technology for the provisioning of computing infrastructure. It refers to on-demand applications and hardware delivered as a service through virtualization of hardware and systems software in the datacenters [7]. The term “Elastic Computing” means the ability to change the number of resources used with the dynamic changes in workload in real time as per requirement. These attractive features of Cloud are making huge industry moment toward it and making it as the next dominant computing paradigm. As cloud computing is in its initial phase, it is suffering from various issues like security, virtualization, capacity allocation, load balancing, and energy optimization.
Cloud inherits some of these challenges from parallel and distributed computing but it faces major challenges of its own. This paper mainly focuses on performance improvement policies for cloud computing such as efficient load balancing (LB), which is the major problem in cloud computing. The principle aspect of cloud computing is virtualization that deals with the construction and management of virtual machines efficiently. As the number of consumers and requests for the services are increasing day by day in cloud computing, therefore load balancing is an important research area for handling the users’ requests efficiently. The motivation of this work is to help those researchers who are new to this problem area of Cloud. This paper will help researchers to understand why LB is required in Cloud and take them to the end of work happened for this problem.

II. CLOUD COMPUTING

Cloud computing term is referred as “The Cloud” which used as a “Metaphor” for the “internet” so cloud computing is a “type of Internet-based computing”. It is a new and an innovative idea of 21st Century for IT industries 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 on cloud computing is that the separation of applications from the in operating systems and therefore the hardware on that they run.

![Figure 1 Why Cloud Computing?](image)

Cloud computing distribute applications via the web, that is accessible from internet browsers, desktop laptop, and mobile apps whereas the code and knowledge are held 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 anyone forms any locations by using Internet-connected pc, within the earth thanks to cloud computing.
2.1 CLOUD COMPUTING PERFORMANCE ISSUES

In a 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.

![Figure 2 Cloud Challenges [CC Survey report 16]](image)

However, we have a tendency to tend to argue that many cloud services and cloud-oriented applications are not economical and it has several performance issues.

- **a)** Optimum utilization of computing resources
- **b)** Security and privacy of user data
- **c)** Load balancing or Task Scheduling policies
- **d)** Government policies
- **e)** Data Storage and Retrieval

III. LOAD BALANCING IN CLOUD COMPUTING

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 operating systems and therefore the hardware on that they run. Load balancing is the process of apportioning the load among various working nodes of a distributed system to improve both resource utilization and job response time while avoiding a situation where some of the nodes are heavily loaded while others are underloaded. Load balancing ensures that every node in the system does the approximately equal amount of work as per their capacity at any instant of time.

3.1 Why Load balancing?

Load Balancing is the major concern in any network or system because it affects three main aspects of the system, i.e., performance, functionality and so the cost in the cloud. It is a resource management technique to utilize all resources at a minimum. It is a technique to evenly distribute the workload among slug nodes in the network.

Load balancing methods are important in cloud because it provides-

- **Scalability** - The load balancing algorithm must provide scalability in terms of addition of new resources to address the everyday increasing demand of services with a greatly increased number of users. This also demands flexibility in accommodating the change.
- **Better response time** - The load balancing must be implemented and executed well enough to provide the best possible response to a user.
- **Cost-effectiveness** - A good load balancing algorithm must aim for better overall system performance with the cost being quite reasonable.
Prioritization - The tasks must be prioritized so that the critical tasks do not have to face the problem of starvation, or if addressed, the problem of a late response.

Fair node utilization - The serving nodes must be utilized efficiently so that no single node is overwhelmed, leaving certain others totally free, or lightly loaded.

IV. EXISTING LOAD BALANCING METHODS

In cloud computing load balancing methods have mainly two categories static and dynamic. In literature surveys following load balancing methods are used for cloud performance.

- **Round Robin Algorithm** - Round Robin is a very famous load balancing algorithm, in which the processes are divided between all processors [4]. The process allocation order is maintained locally independent of the allocations from remote processors. In Round Robin, it sends the requests to the node with the least number of connections, so at any point of time some node may be heavily loaded and other remain idle this problem is reduced by CLBDM.

- **Central Load Balancing Decision Model (CLBDM)** - CLBDM is a central load balancing decision model, which is suggested by [2]. This method is mainly based on session switching at the application layer. The improvement is that in the cloud, it calculated the connection time between the client and the node, and if that connection time exceeds a threshold then the connection will be terminated and the task will be forwarded to another node using the regular Round Robin rules.

- **Map Reduce-based Entity Resolution** - Map Reduce is a computing model and an associated implementation for processing and generating large datasets. Map task and reduce task two main task in this model which written by the user, Map takes an input pair and produces a set of intermediate value pair and Reduce task accepts an intermediate key and a set of values for that key and merges these values to form a smaller set of value [11]. Map task read entities in parallel and process them, this will cause the Reduce task to be overloaded.

- **Ant colony optimization (ACO)** - This method is suggested in [5] of ant colony optimization. In ACO algorithm when the request is initiated the ant starts its movement. Movement of ant is of two ways
  - **Forward Movement** - Forward Movement means the ant in continuously moving from one overloaded node to another node and check it is overloaded or underloaded if ant finds an overloaded node it will continue moving in the forward direction and check each node.
  - **Backward Movement** - If an ant finds an overloaded node the ant will use the backward movement to get to the previous node, in the algorithm [11] if ant finds the target node then ant will commit suicide, this algorithm reduced the unnecessary backward movement, overcome heterogeneity, is excellent in fault tolerance.

- **Load balancing of virtual machine resources** - In paper [3] author proposed a scheduling strategy on load balancing of VM resources that uses historical data and the current state of the system. This strategy achieves the best load balancing and reduced dynamic migration by using a genetic algorithm. It helps in resolving the issue of load-imbalance and the high cost of migration thus achieving better resource utilization.

- **Index Name Server Algorithm (INS)** - The INS algorithm proposed in [12] the goal is to find an algorithm to minimize the data duplication and redundancy. INS is able to handle the load balancing dynamically. INS have some parameters which help in calculating the optimum selection point like that HashCode of the block of data to be downloaded, the position of the server, the transition quality, the maximum bandwidth. Another calculation point whether the connection can handle additional nodes or not.

- **Opportunistic Load Balancing (OLB)** - Sang proposed OLB is a static load balancing algorithm that has the goal of keeping each node in cloud busy [7]. However, OLB does not calculate the execution
time of the node, due to this the tasks to be processed in a slower manner and will cause bottlenecks since requests might be pending waiting for nodes to be free.

- **Honeybee Foraging Behavior**-[1] investigated a decentralized honeybee based load balancing technique that is a nature-inspired algorithm for self-organization. It achieves global load balancing through local server actions. Performance of the system is enhanced with increased system diversity but throughput is not increased with an increase in system size. It is best suited for the conditions where the diverse population of service types is required.

- **Biased Random Sampling**- In paper [13] authors investigated a distributed and scalable load balancing approach that uses random sampling of the system domain to achieve self-organization thus balancing the load across all nodes of the system. The performance of the system is improved with a high and similar population of resources thus resulting in an increased throughput by effectively utilizing the increased system resources. It is degraded with an increase in population diversity.

- **Active Clustering**- In paper author [11] investigated a self-aggregation load balancing technique that is a self-aggregation algorithm to optimize job assignments by connecting similar services using local re-wiring. The performance of the system is enhanced with high resources thereby increasing the throughput by using these resources effectively. It is degraded with an increase in system diversity.

![Cloud Load Balancing](image)

**Figure 3 Cloud Load Balancing [8]**

4.1 **Comparisons of Existing Works**- Following method are used for comparison

| Author Ref | Algorithm | Description | Parameters Used | Benefits | Challenges |
|------------|-----------|-------------|-----------------|----------|------------|
| [1] Anureet kaur | Honeybee foraging algorithm | Decentralized honeybee-based nature-inspired load balancing technique for self-organization. | Makespan, Waiting Time, Computation cost | Works well under distributed load | Increase in resources does not improve throughput equally |
| [2] Entisar S. Alkayal | PSO( Particle Swarm Optimization) | More Sophisticated in terms of execution time and throughput. | Path length, Capacity, Makespan | Best in finding path | Slower in large load |
| [5] D | Weighted | Works as the Min- | Makespan, | Reduces the | Smaller |
VI. CONCLUSIONS & FUTURE WORKS

This paper presents a comparative study of various load balancing methods for cloud computing. Also, covers the detailed study of cloud load balancing and challenges in load balancing. Cloud computing is a promising technology, is a model for providing the resources needed for customer service oriented. Platform virtualization is one of the essential characteristics of cloud computing, virtualization refers to factors affecting the organization and business performance, such as information technology resources, hardware, software, operating system, and service. Cloud computing is characterized by distributed computing and this is also the risk of bottlenecks that may occur during the allocation of resources for load balancing. In future work, we will develop an efficient load balancing method for cloud computing to improve cloud performance. The proposed method will be compared with a various existing method to the analysis of performance.

REFERENCES

[1] Anureet Kaur, Bikrampal Kaur,” Load Balancing in tasks using Honeybee Behavior Algorithm in Cloud Computing”, IEEE 2016, PP 978-983
[2] Entisar S. Alkayal, Nicholas R. Jennings, “Efficient Task Scheduling Multi-Objective Particle Swarm Optimization in Cloud Computing”, 2016 IEEE 41st Conference on Local Computer Networks Workshops, PP 17-23

[3] Ekta Rani, Harpreet Kaur, “Efficient Load Balancing Task Scheduling in Cloud Computing using Raven Roosting Optimization Algorithm”, International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017, PP 2419-2425.

[4] Ankit Kumar, Mala Kalra, “Load Balancing in Cloud Data Center Using Modified Active Monitoring Load Balancer”, IEEE 2016 978-983.

[5] D. Chitra Devi and V. Rhymend Uthariara, Load Balancing in Cloud Computing Environment Using Improved Weighted Round Robin Algorithm for Non-preemptive Dependent Tasks, Hindawi Publishing Corporation e Scientific World Journal Volume 2016, PP 1-14.

[6] Z. Xiao, W. Song, and Q. Chen, “Dynamic resource allocation using virtual machines for cloud computing environment,” IEEE Transactions on Parallel and Distributed Systems, vol.24, no. 6, pp. 1107–1117, 2013.

[7] L.D.DhineshBabuaandP.Venkata Krishna, “Honeybeebehavior inspired load balancing of tasks in cloud computing environments,” Applied Soft Computing Journal, vol.13, no.5, pp.2292–2303, 2013.

[8] J. Cao, K. Li, and I. Stojmenovic, “Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centers,” IEEE Transactions on Computers, vol.63, no.1, 2014, pp 45–58.

[9] R. N. Calheiros and R. Buyya, “Meeting deadlines of scientific workflows in public clouds with tasks replication,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 7, pp. 1787–1796, 2014.

[10] R. Basker, V. Rhymend Uthariaraj, and D. Chitra Devi, “An enhanced scheduling in weighted round robin for the cloud infrastructure services,” International Journal of Recent Advances in Engineering & Technology, vol.2, no.3, pp.81–86, 2014.

[11] Adamuthe, A. C., Pandharpatte, R. M., & Thampi, G. T. (2013, November). Multiobjective Virtual Machine Placement in Cloud Environment. In Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), 2013 International Conference on (pp. 8-13). IEEE.

[12] Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., & Warfield, A. (2003). Xen and the art of virtualization. ACM SIGOPS Operating Systems Review, 37(5), 164-177.

[13] Beegom, A. A., & Rajasree, M. S. (2014). A Particle Swarm Optimization Based Pareto Optimal Task Scheduling in Cloud Computing. In Advances in Swarm Intelligence (pp. 79-86). Springer International Publishing.

[14] Sakellari, G., & Loukas, G. (2013). A survey of mathematical models, simulation approaches and test beds used for research in cloud computing. Simulation Modelling Practice and Theory, 39, 92-103.

[15] Suresh, M., Santhosh, K. B., & Karthik, S. (2014, March). A Load Balancing Model in Public Cloud Using ANFIS and GSO. In Intelligent Computing Applications (ICICA), 2014 International Conference on (pp. 85-89). IEEE.

[16] Tripathi, P. K., Bandyopadhyay, S., & Pal, S. K. (2007). Multi-objective particle swarm optimization with time variant inertia and acceleration coefficients. Information Sciences, 177(22), 5033-5049.

[17] Vinothina, V., Sridaran, R. and Ganapathi, P. (2012) "A survey on resource allocation strategies in cloud computing." International Journal of Advanced Computer Science and Applications 3.6: 97 - 104.