Review on Efficient Approach for Load Balancing on the cloud computing enviroment

Gurpreet Kaur[1], Shivani Ahuja[2]
[1]*IET Bhaddal, Ropar*
[2] *IET Bhaddal, Ropar*
(E-mail: gurpreetkaurkang2@gmail.com)

***Abstract*—** Cloud computing, a framework for enabling convenient, and on-demand network access to a shared pool of computing resources, is emerging as a new paradigm of large-scale distributed computing. In this paper we have given the characteristics of cloud computing and its application architecture. The next section covers three types of cloud. It has widely been adopted by the industry, though there are many existing issues like Load Balancing, Virtual Machine Migration, Server Consolidation, Energy Management, etc. that are not fully addressed. Central of these issues is load balancing. We have also given review of existing work on load balancing issue in cloud computing through which a problem is formulated. In the end of paper we have given a flow of work that can be carried out to reduce the problem for which there is need to propose a new hybrid optimization algorithm.

Keywords*—cloud computing, load balancing, hybrid optimization algorithm, types of cloud, architecture*

I. INTRODUCTION

Cloud computing is a new technology and it is becoming popular because of its great features. In this technology almost everything like hardware, software and platform are provided as a service. A cloud provider provides services on the basis of client's requests. An important issue in cloud is, scheduling of users requests, means how to allocate resources to these requests, so that the requested tasks can be completed in a minimum time and the cost incurred in the task should also be minimum. In case of Cloud computing services can be used from diverse and wide spread resources, rather than remote servers or local machines. There is no standard definition of Cloud computing. Generally it consists of a bunc*h of distributed servers known as masters, providing demanded services and resources to different clients known as clients in a network with scalability and reliability of datacenter. The distributed computers provide on-demand services. Services may be of software resources (e.g. Software as a Service, SaaS) or physical resources (e.g. Platform as a Service, PaaS) or hardware/infrastructure (e.g. Hardware as a Service, HaaS or Infrastructure as a Service, IaaS). AmazonEC2 (Amazon Elastic Compute Cloud) is an example of cloud computing services [2].

Cloud computing has recently become popular due to the maturity of related technologies such as network devices, software applications and hardware capacities. Resources in these systems can be widely distributed and the scale of resources involved can range from several servers to an entire data center. To integrate and make good use of resources at various scales, cloud computing needs efficient methods to manage them [4]. Consequently, the focus of much research in recent years has been on how to utilize resources and how to reduce power consumption. One of the key technologies in cloud computing is virtualization. The ability to create virtual machines (VMs) [14] dynamically on demand is a popular solution for managing resources on physical machines. Therefore, many methods [17,18] have been developed that enhance resource utilization such as memory compression, request discrimination, defining threshold for resource usage and task allocation among VMs. Improvements in power consumption, and the relationship between resource usage and energy consumption has also been widely studied [6,10]. Some research aims to improve resource utilization while others aim to reduce energy consumption. The goals of both are to reduce costs for data centers. Due to the large size of many data centers, the financial savings are substantial.

II. CHARACTERISTICS OF CLOUD COMPUTING

A. Self Healing

Any application or any service running in a cloud computing environment has the property of self healing. In case of failure of the application, there is always a hot backup of the application ready to take over without disruption. There are multiple copies of the same application - each copy updating itself regularly so that at times of failure there is at least one copy of the application which can take over without even the slightest change in its running state.

B. Multi-tenancy

With cloud computing, any application supports multi-tenancy - that is multiple tenants at the same instant of time. The system allows several customers to share the infrastructure allotted to them without any of them being aware of the sharing. This is done by virtualizing the servers on the available machine pool and then allotting the servers to

multiple users. This is done in such a way that the privacy of the users or the security of their data is not compromised.

C. Linearly Scalable

Cloud computing services are linearly scalable. The system is able to break down the workloads into pieces and service it across the infrastructure. An exact idea of linear scalability can be obtained from the fact that if one server is able to process say 1000 transactions per second, then two servers can process 2000 transactions per second.

D. Service-oriented

Cloud computing systems are all service oriented - i.e. the systems are such that they are created out of other discrete services. Many such discrete services which are independent of each other are combined together to form this service. This allows re-use of the different services that are available and that are being created. Using the services that were just created, other such services can be created.

E. SLA Driven

Usually businesses have agreements on the amount of services. Scalability and availability issues cause clients to break these agreements. But cloud computing services are SLA driven such that
when the system experiences peaks of load, it will automatically adjust itself so as to comply with the service-level agreements. The services will create additional instances of the applications on more servers so that the load can be easily managed.

F. Virtualized

The applications in cloud computing are fully decoupled from the underlying hardware. The cloud computing environment is a fully virtualized environment.

G. Flexible

Another feature of the cloud computing services is that they are flexible. They can be used to serve a large variety of workload types - varying from small loads of a small consumer application to very heavy loads of a commercial application.

III. CLOUD COMPUTING APPLICATION ARCHITECTURE

We know that cloud computing is the shift of computing to a host of hardware infrastructure that is distributed in the cloud. The commodity hardware infrastructure consists of the various low cost data servers that are connected to the system and provide their storage and processing and other computing resources to the application. Cloud computing involves running applications on virtual servers that are allocated on this distributed hardware infrastructure available in the cloud. These virtual servers are made in such a way that the different service level agreements and reliability issues are met. There may be multiple instances of the same virtual server accessing the different parts of the hardware infrastructure available. This is to make sure that there are multiple copies of the applications which are ready to take over on another one's failure. The virtual server distributes the processing between the infrastructure and the computing is done and the result returned.

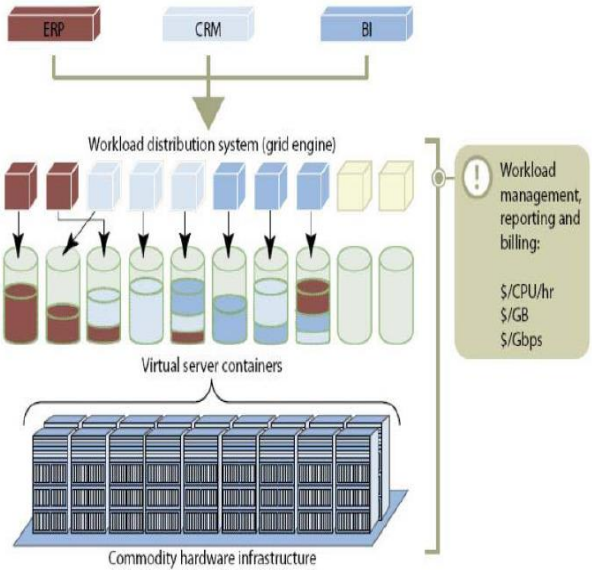

Fig. 1 The basic Cloud computing application architecture

There will be a workload distribution management system, also known as the grid engine, for managing the different requests coming to the virtual servers. This engine will take care of the creation of multiple copies and also the preservation of integrity of the data that is stored in the infrastructure. This will also adjust itself such that even on heavier load, the processing is completed as per the requirements. The different workload management systems are hidden from the users. For the user, the processing is done and the result is obtained. There is no question of where it was done and how it was done. The users are billed based on the usage of the system - as said before - the commodity is now cycles and bytes. The billing is usually on the basis of usage per CPU per hour or GB data transfer per hour.

IV. CLOUD TYPES

Together with virtualization, clouds can be defined as computers that are networked anywhere in the world with the availability of paying the used clouds in a pay-per-use way, meaning that just the resources that are being used will be paid. In the following the types of clouds will be introduced.

A. Public Clouds

A public cloud encompasses the traditional concept of cloud computing, having the opportunity to use computing resources from anywhere in the world. The clouds can be used in a so-called pay-per-use manner, meaning that just the resources that are being used will be paid by transaction fees.

B. Private Clouds

Private clouds are normally datacenters that are used in a private network and can therefore restrict the unwanted public to access the data that is used by the company. It is obvious that this way has a more secure background than the traditional public clouds. However, managers still have to worry about the purchase, building and maintenance of the system.

C. Hybrid Clouds

As the name already reveals, a hybrid cloud is a mixture of both a private and public cloud. This can involve work load being processed by an enterprise data center while other activities are provided by the public cloud. Below an overview of all three cloud computing types is illustrated.

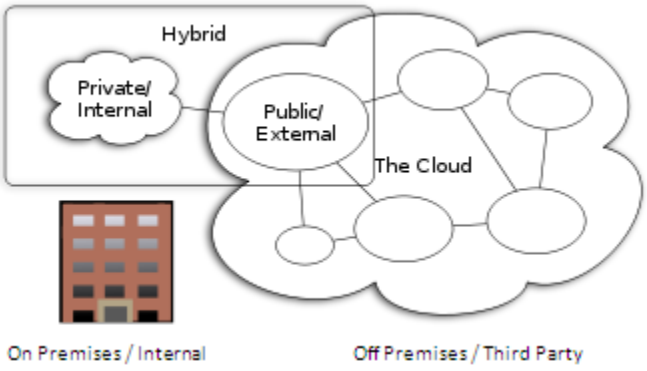

Fig. 2 Cloud Computing Types

V. LOAD BALANCING ON CLOUD COMPUTING

With the increasing popularity of cloud computing, the amount of processing that is being done in the clouds is surging drastically. A cloud is constituted by various nodes which perform computation according to the requests of the clients. As the requests of the clients can be random to the nodes they can vary in quantity and thus the load on each node can also vary. Therefore, every node in a cloud can be unevenly loaded of tasks according to the amount of work requested by the clients. This phenomenon can drastically reduce the working efficiency of the cloud as some nodes which are overloaded will have a higher task completion time compared to the corresponding time taken on an under loaded node in the same cloud. This problem is not only confined only to cloud but is related with every large network like a grid, etc.

Load balancing in large distributed server systems is a complex optimization problem of critical importance in cloud systems and data centers. Load balancing algorithms are classified as static and dynamic algorithms. Static algorithms are mostly suitable for homogeneous and stable environments and can produce very good results in these environments. However, they are usually not flexible and cannot match the dynamic changes to the attributes during the execution time. Dynamic algorithms are more flexible and take into consideration different types of attributes in the system both prior to and during run-time [2]. These algorithms can adapt to changes and provide better results in heterogeneous and dynamic environments. However, as the distribution attributes become more complex and dynamic. As a result some of these algorithms could become inefficient and cause more overhead than necessary resulting in an overall degradation of the services performance.

VI. PREVIOUS WORK

Atyaf Dhari et al. 2017) proposed Load Balancing Decision Algorithm (LBDA) to manage and balance the load between the virtual machines in a datacenter along with reducing the completion time (Makespan) and Response time. Findings: The mechanism of LBDA is based on three stages, first calculates the VM capacity and VM load to categorize the VMs' states (Under loaded VM, Balanced VM, High Balance VM, Overloaded). Second, calculate the time required to execute the task in each VM. Finally, makes a decision to distribute the tasks among the VMs based on VM state and task time required. Improvements: We compared the result of our proposed LBDA with Max- Min, Shortest Job Firstand Round Robin.

Mohammad Goudarzi et al. (2017) evaluated the efficiency of the proposed solution using both simulation and testbed experiments. The evaluation study demonstrated that proposal can outperform existing optimal and near-optimal counterparts in terms of weighted execution cost, energy consumption and execution time. Due to nowadays advances of mobile technologies in both hardware and software, mobile devices have become an inseparable part of human life. Along with this progress, mobile devices are expected to perform various types of applications. **Muhammad Baqer Mollah et al. (2017)** presented the main security and privacy challenges in this field which have grown much interest among the academia and research community. Although, there are many challenges, corresponding security solutions have been proposed and identified in literature by many researchers to counter the challenges. We also present these recent works in short. Furthermore, we compare these works based on different security and privacy requirements, and finally present open issues. The rapid growth of mobile computing is seriously challenged by the resource constrained mobile devices. However, the growth of mobile computing can be enhanced by integrating mobile computing into cloud computing, and hence a new paradigm of computing called mobile cloud computing emerges.

M. Vanitha et al. (2017) proposed, in- volving a well-organized use of resources, which is known as the dynamic well-organized load balancing (DWOLB) algorithm. This is a powerful algorithm for reducing the energy that is consumed in cloud computing. Cloud computing is used in almost all domains today. Through the use of cloud-based applications, it has become easier for an internet user to make use of the services and re- sources that are widely available. The cloud service provider undertakes to deliver all the

subscribers' requirements as per the service level agreement (SLA). These resources must be well-protected since they are used by many subscribers.

R.R. Kotkondawar et al. (2014) described a Cloud computing is the most recent technology in today's world of computing and it overcomes deficiencies of traditional ways of computing. Cloud computing is a new way of providing the essential services to cloud users on "Pay As You Go" basis. Cloud computing provides different features like on demand access, flexibility, instant response, pay per use etc. to customers. In order to provide all these features to cloud users, cloud computing systems must be structured and managed efficiently to provide the Quality of Services (QOS) to users. Various technological concepts such as abstraction and virtualization are used that hides the implementation details from an average cloud user. Cloud load balancing plays a very important role in providing all the cloud features to users which is the main topic of interest in our research. Different archictures apply altogether different load balancing algorithms. The research includes the Study of different approaches of effective management of cloud systems. The study includes load balancing approaches in different system architectures like Centralized, Distributed and Cluster based architecture. Finally various algorithms have been compared based on the different parameters like response time, efficiency and throughput etc. **Kumar Nishan et al. (2012)** proposed an algorithm for load Distribution of workloads among nodes of a cloud by the use of Ant Colony Optimization (ACO). This is a modified Approach of ant colony optimization that has been applied from the perspective of cloud or grid network systems with the Main aim of load balancing of nodes. This modified algorithm has an edge over the original approach in which each ant build their own individual result set and it is later on built into a complete solution. However, in this approach the ants continuously update a single result set rather than updating their own result set. Further, as they know that a cloud is the collection of many nodes, which can support various types of application that is used by the clients on a basis of pay per use. Therefore, the system, which is incurring a cost for the user should function smoothly and should have algorithms that can continue the proper system functioning even at peak usage hours.

Klaithem Al Nuaimi et al. (2012) have investigated the different algorithms proposed to resolve the issue of load balancing and task scheduling in Cloud Computing. They discussed and compared these algorithms to provide an overview of the latest approaches in the field. Load Balancing is essential for efficient operations in distributed environments. As Cloud Computing is growing rapidly and clients are demanding more services and better results, load balancing for the Cloud has become a very interesting and important research area. Many algorithms were suggested to provide efficient mechanisms and algorithms for assigning the \client's requests to available Cloud nodes. These approaches aim to enhance the overall performance of the Cloud and provide the user more satisfying and efficient services. **Zheng Hu et al. (2012)** introduced the failure and recovery scenario in the current Cloud computing entities and propose a Reinforcement Learning (RL) based algorithm to make job scheduling in the current computing Cloud fault tolerant. We carry out experimental comparison with Resource-constrained Utility Accrual algorithm (RUA), Utility Accrual Packet scheduling algorithm (UPA) and LBESA to demonstrate the feasibility of proposed approach. **Andre Martin et al. (2011)** presented a new fault tolerance approach based on active replication for Stream Map Reduce systems. Presented approach is cost effective for cloud consumers as well as Cloud providers. Cost effectiveness is achieved by fully utilizing the acquired computational resources without performance degradation and by reducing the need for additional nodes dedicated to fault tolerance.

VII. Problem Formulation

In a cloud environment, there may be any number of host machines and each host machine has different-different load due to virtual machines as per the client's demand. The load of a host machine may be of various types such as CPU load, Memory load, Storage load and Network related load etc. If the load of any host machine exceeds its capacity then it affects its efficiency. In runtime, any client application service may change their resource (CPU, RAM, Storage and Bandwidth etc.) demand and this causes the host system to be imbalanced. If this imbalanced situation occurs due to overloading then system is balanced using load balancing techniques by distributing the extra workload to the whole clouds host heaving light loads. This helps to improve the overall performance of the cloud system.

Load Balancing is defined as a process of making effective resource utilization by reassigning the total load to the individual nodes of the collective system and thereby minimizing the response time of the job. Load Balancing algorithms are classified as Static and Dynamic algorithms. Static algorithms are most suitable for homogenous and stable environments. However, they cannot match the dynamic changes to the attributes during execution time. Dynamic algorithms take into consideration different types of attributes in the system both prior to and during run time. These algorithms can adapt to changes and provide better results in heterogeneous and dynamic environments.

In the past, a number of load balancing algorithms have been developed specifically to suit the dynamic cloud computing environments such as INS (Index Name Server) algorithm[A], WLC (Weighted Least Connection) algorithm[B], DDFTP (Dual direction Downloading algorithm from FTP servers)[C], LBMM (Load Balancing Min-Min) algorithm[D], ACO(Ant Colony Optimization) algorithm[E] and Bee-MMT(Artificial Bee Colony algorithm- Minimal Migration time)[F]. We are going to use the PSO (Particle Swarm Optimization) algorithm for load balancing in dynamic cloud environments as particle swarm has already get better results than genetic and ACO in grid computing[G]. Performance of Particle Swarm Optimization has also been

approved better in distributed system [12]. In the proposed research, the bat optimization algorithm for task scheduling and load balancing on cloud computing will be implemented. The proposed Hybrid algorithm (Cuckoo search optimization and Bee Colony optimization) will also compare with the load balancing decision algorithm for evaluation purpose. The results of the proposed work will be analyzed on the basis of makespan, execution time and response time.

VIII. OBJECTIVES

The key objective of this research work is to optimize the performance of the cloud architecture. Overloaded nodes across the server and storage side often lead to performance degradation and are more vulnerable to various failures. To remove this limitation the load must be migrated from the overloaded resource to an underutilized one without causing harm and disruption to the application workload. Objectives for this research work are:

1. To study and understand the task shcduling and load balancng approach on cloud.
2. To implement exising load balancing decision algorithm and proposed bat optimization on cloud environment.
3. To analyze the behavior of the proposed Hybrid on the basis of following parameters:

(a) Execution Time
(b) Response Time
(c) Makespan

IX. RESEARCH METHODLOGY

In cloud computing, load balancing is required to distribute the dynamic local workload evenly across all the nodes. It helps to achieve a high user satisfaction and resource utilization ratio by ensuring an efficient and fair allocation of every computing resource. Proper load balancing aids in minimizing resource consumption, implementing fail-over, enabling scalability, avoiding bottlenecks and over-provisioning etc.

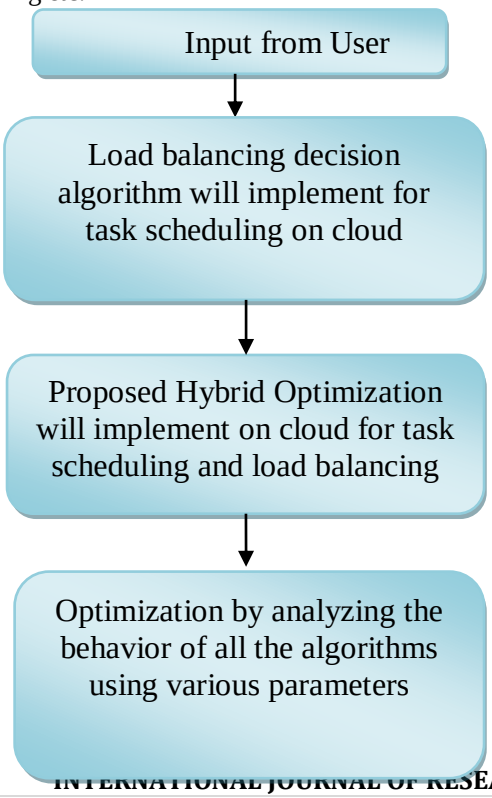

Fig. 3: Flowchart of Proposed work

Input: - Required parameters for cloudlets and vm's are taken from user.
Output: - Improves load balancing at cloud with better response time, data processing time and throughput.

X. Conclusion

In this paper we have given basic introduction about cloud computing and its application architecture. The next section of this paper covers three basic types of cloud. In cloud Computing, Load Balancing is essential for efficient operations in distributed environments. As Cloud Computing is growing rapidly and clients are demanding more services and better results, load balancing for the Cloud has become a very interesting and important research area. Load balancing ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. Our objective is to implement an effective load balancing algorithm for balancing the load on cloud. In this paper we have given the proposed idea that will help in dealing with load balancing approach and improve the results in terms of various parameters.

References

[1] A. Dhari, K. I. Arif , "An Efficient Load Balancing Scheme for Cloud Computing," in the Indian Journal of Science and Technology, Vol 10(11), , March 2011

[2] A. Nahir, A. Orda, D. Raz "Schedule First, Manage Later: Network-Aware Load Balancing," In the Proceedings IEEE INFOCOM, 2013

[3] H. Jamal, A. Nasir, K.Ruhana, K.Mahamud and A.M. Din, "Load Balancing Using Enhanced Ant Algorithm in Grid Computing", Proceedings of the Second International Conference on Computational Intelligence, Modelling and Simulation, pp. 160-165, 2010.

[4] I. Azawi Mohialdeen, "Comparative Study Of Scheduling Algorithms In Cloud Computing Environment," Journal of Computer Science, 9 (2): 252-263, 2013 ISSN 1549-3636 © 2013 Science Publications.

[5] J. T. Juemin, Z. Jun Li, W. Meleis, N. Mi "ARA: Adaptive Resource Allocation for Cloud," in the proceeding IEEE 2011

[7] K. Nishant, P. Sharma, V. Krishna, Nitin, R. Rastogi "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," in the 14th International Conference on Modeling and Simulation, IEEE 2012

[8] K. Al Nuaimi, N. Mohamed, M. Al Nuaimi, J. Al-Jaroodi "A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms," in the proceeding IEEE 2012.

[9] M. Goudarzi, M. Zamani, A. Toroghi Haghighat, "A fast hybrid multi-site computation offloading for mobile cloud computing," in the Journal of Network and Computer Applications, Elsevier 2017.

[10] M. Baqer Mollah, Md. Abul Kalam Azad, A. Vasilakos, "Security and privacy challenges in mobile cloud computing: Survey and way Ahead," in the Journal of Network and Computer Applications, Elsevier 2017.

[11] M. A. Sharkh, A. Ouda, A. Shami, "AResource Scheduling Model for Cloud Computing Data centers," in the proceeding IEEE 2013.

[12] M. sudha, M. Monica, "Investigation on Efficient Management of Workflows in Cloud Computing Environment", International Journal of Computer Science and Engineering (IJCSE), Volume 02, Number 05, August 2010, pp. 1841- 1845.

[13] P. Patel, D. Bansal, L. Yuan, A. Murthy, A. Greenberg Ananta: Cloud Scale Load Balancing," in SIGCOMM, August12–16, 2013, HongKong, China.

[14] Prof. Dr. Jayant. S. Umale, Miss. Priyanka, A. Chaudhari, "Survey on Job Scheduling Algorithms of Cloud Computing," International Journal of Computer Science and Management Research , 2013.

[15] S. Razzaghzadeh, A. Habibizad Navin, A. M. Rahmani, M. Hosseinzadeh, "Probabilistic modeling to achieve load balancing in Expert Clouds," Elsevier B.V 2017.

[16] S. Maguluri, R. Srikant, L. Ying, "Stochastic Models of Load Balancing and Scheduling in Cloud Computing Clusters," IEEE INFOCOM 2012 Proceedings, pp.702- 710, March 2012

[17] Vanitha, P. Marikkannu, "Effective resource utilization in cloud environment through a dynamic well-organized load balancing algorithm for virtual machines", Elsevier Ltd, 2017.