Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Cloud Load Balancing Services Survey and Research Challenges

Himanshi
Student of CSE Dept.
DIET Karnal, India

Sunil Ahuja
Lect. In CSE Dept.
DIET Karnal, India

Abstract— Cloud Computing is a kind of dispensed computing where massively scalable IT-related capabilities are provided to multiple external customers "as a service" using internet technologies. The cloud providers have to achieve a large, general-purpose processing infrastructure; and virtualization of infrastructure for different consumers and services to provide the multiple application services. This paper explores autonomic approaches for optimizing provisioning for heterogeneous workloads on enterprise Grids and clouds. Central to these problems lies the establishment of a successful load balancing algorithm. The load can be Central Processing Unit load, memory capacity, delay or network load. Load balancing involves distributing the load among numerous nodes of a distributed system to enhance both resource usage and job response time while also avoiding a predicament where some of the nodes are heavily filled while other nodes are idle or doing very little work. Load balancing helps to ensure that all the processor in the system or every node into the system does roughly the equal level of work at any immediate of time. This technique can be sender initiated, receiver started or symmetric type (mix of transmitter initiated and receiver initiated types). This paper reviews Load balancing strategies for Cloud infrastructures.

Index Terms—Cloud Computing, Load Balancing, Infrastructures, Public Clouds.

I. INTRODUCTION

Cloud computing [1] has recently received considerable attention both in academic and commercial community as a new computing paradigm to provide dynamically scalable and virtualized resource as a service online. By this implies, users are able to get access to the resources, such as applications and data, through the cloud everywhere so when on demand. Currently, a few large businesses, such Amazon, Bing, Yahoo!, Microsoft, IBM, and Sun are building their own cloud platforms for customers and companies to access the cloud resources through services. Recently, with all the rapid growth of virtualization technology [2], more and more data centers employ this technology to create new generation data center to aid cloud computing [3] as a result of advantages such host combination, live migration, and resource isolation [4]. Live migration of virtual machines [5] means the virtual machine is apparently responsive everyday during the migration process from the customers' perspective. Compared with traditional suspend/resume migration, live migration holds many benefits such as for example energy saving, load balancing, and online maintenance. Many live migration methods are proposed to enhance the migration efficiency [6]. Because the live migration technology widely found in modern cloud computing information center, real time migration of multiple virtual machines becomes more and more regular. Distinct from the single virtual machine migration, the live migration of multiple virtual machines faces many completely new problems, such as migration failures as a result of inadequate sources in target machine, migration conflicts due to the concurrent migrations, additionally the migration thrashing as a consequence of the dynamic changes of virtual machine workloads. Every one of the preceding dilemmas must certainly be overcome to maximize the migration efficiency in virtualized cloud data center environments. In this paper, we learn the live migration efficiency of multiple virtual machines from experimental perspective and investigate different resource reservation methods and migration strategies in the live migration process. We first describe the live migration framework of several virtual machines with resource reservation technology. Then we perform a few experiments to research the impacts of various resource reservation methods in the performance of real time migration both in source machine and target machine.

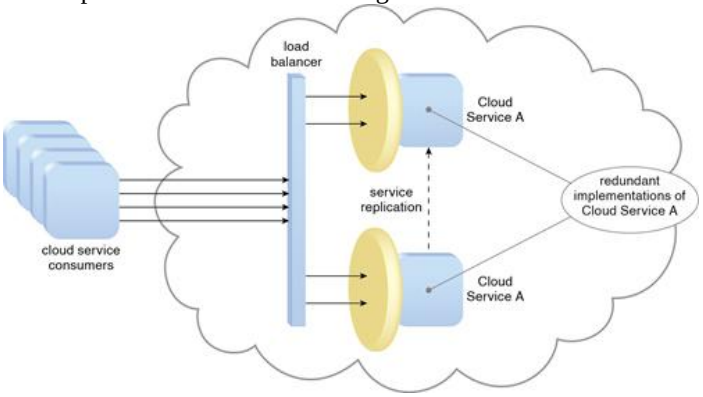

Figure 1 Load Balancing Service in Cloud systems

Additionally, we also review the efficiency of parallel migration strategy and workload-aware migration strategy. The metrics such downtime, total migration time, and workload performance overheads are measured. Experiments reveal some brand-new discovery of live migration of multiple virtual devices. On the basis of the noticed results, we present corresponding optimization methods to boost the migration efficiency.

II. CLOUD COMPUTING CHALLENGES

There are various challenges of Cloud Load Balancing which are explained below.

1. Spatial Allocation of the Cloud Nodes

Some algorithms are projected to be effectual merely for an intranet or closely placed nodes whereas contact delays are negligible. Though, it is a trial to design a load balancing algorithm that can work for spatially distributed nodes. This is because supplementary factors have to be seized into report such as the speed of the web links amid the nodes, the distance amid the client and the task processing nodes, and the distances amid the nodes encompassed in bestowing the service. There is a demand to develop a method to manipulation load balancing mechanism amid all the spatial distributed nodes as being able to efficiently tolerate elevated delays.

2. Storage/ Replication

A maximum replication algorithm does not seize effectual storage utilization into account. This is because the alike data will be stored in all replication nodes. Maximum replication algorithms impose higher prices as extra storage is needed. Though, partial replication algorithms might save portions of the data sets in every single node established on every single node's skills such as processing manipulation and capacity. This might lead to larger utilization, yet it increases the intricacy of the load balancing algorithms as they endeavor to seize into report the potential of the data set's portions across the disparate Cloud nodes.

3. Algorithm Intricacy

Load balancing algorithms are favored to be less convoluted in words of implementation and operations. The higher implementation intricacy should lead to a extra convoluted procedure that might cause a little negative presentation issues. Furthermore, after the algorithms need extra data and higher contact for monitoring and domination, delays should cause extra setbacks and the efficiency will drop. Therefore, load balancing algorithms have to be projected in the simplest probable forms.

4. Point of Failure

Controlling the load balancing and accumulating data concerning the disparate nodes have to be projected in a method that avoids possessing a solitary point of wreck in the algorithm. A little algorithm can furnish effectual and competent mechanisms for resolving the load balancing in a precise pattern. Though, they have the subject of one controller for the finished system. In such cases, if the controller fails, next the finished arrangement should fail. Each Load balancing algorithm has to be projected in order to vanquish this challenge. Distributed load balancing algorithms seem to furnish a larger way, yet they are far extra convoluted and need extra coordination and manipulation to purpose correctly.

5. Throughput:

It is the finished number of tasks that have finished killing for a given scale of time. It is needed to have elevated across locale for larger presentation of the arrangement.

6. Associated Overhead:

It describes the number of overhead across the implementation of the load balancing algorithm. It is a constitution of movement of tasks, inter procedure contact and inter processor. For load balancing method to work properly, minimum overhead ought to be there.

7. Fault tolerant:

We can delineate it as the skill to present load balancing by the appropriate algorithm lacking arbitrary link or node failure. Every single load balancing algorithm ought to have good obligation agreement approach.

8. Migration time:

It is the number of period for a procedure to be transferred from one arrangement node to one more node for execution. For larger presentation of the arrangement this period ought to be always less.

9. Response time:

In Distributed arrangement, it is the period seized by a particular load balancing method to respond. This period ought to be minimized for larger performance.

10. Resource Utilization:

It is the parameter that gives the data inside that extant the resource is utilized. For effectual load balancing in arrangement, optimum resource ought to be utilized.

11. Scalability:

It is the skill of load balancing algorithm for a arrangement alongside each finite number of processor and machines. This parameter can be enhanced for larger arrangement performance.

12. Performance:

It is the finished efficiency of the system. If all the parameters are enhanced next the finished arrangement presentation can be improved.

III. LOAD BALANCING

Load Balancing in Distributed Networks

In distributed webs like mesh webs, P2P grid, mobile ad-hoc, wireless and overlay webs, balancing the load in an adaptive manner enhances the web presentation considerably. Further, the ultimate aim of load balancing is as follows :

- Even allocation of load to every single resource
- Minimization of processing period for every single job
- Maximum utilization of every single resource

Grid serves as a comprehensive arrangement for association across that the maximum resource utilization is achieved. Load balancing in grid is attained by employing the deduction of onward and retrograde ants as a competency locations and manipulation word to recognize the appropriate resources as well. The authors gave resource manipulation processing in grid established on tree ideal and ant dominion model. The jobs are devoted to the resources established on its priority, node's workload and resource availability. There asserted four definitions for competent job allocation to the resource.

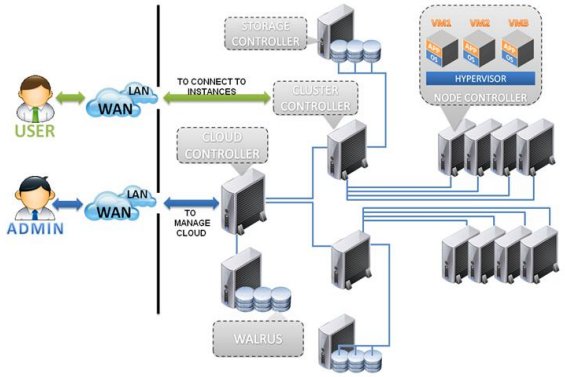

Figure 2 Structural Overview of Cloud Controller

Load Balancing Methods in Cloud Computing:

As is well-known, cloud computing exploits a collection of computing resources to accelerate the killing of huge scale tasks. The algorithm is completely established on the average completion of every single task allocated to the node, node's capacity and utilization of computing resources. As considering concerning the effectual load balancing in cloud [7], vibrant resource allocation plays a momentous role. It is uttered that, usually, cloud computing firms need parallel data processing to make the placement of services extra facile in their product portfolio.

The design follows the master-slave outline as illustrated in Figure 2. It is asserted that beforehand tolerating a task to the system, the adjacent contraption in cloud has to be commenced, that runs the Job Manager (JM). The JM is accountable for consenting client's job and design them. The cloud controller is an interface, accountable for contact by that the JM can allocate or de-allocate the job to VMs established [8] on present job killing phase. The actual killing of tasks is grasped out by the set of instances shouted Task Manager (TM). The procedure aids to enhance the finished resource utilization and considerably cutting the processing cost

Metrics for Load Balancing In Cloud

Distinctive parameters examined in the continuing load balancing strategies in cloud computing are debated below.

- **Throughput:**

It is utilized to assess the number of tasks gave in a given constituent time. Throughput ought to be elevated to enhance the system's performance.

- **Fault Tolerance:**

It is described as the skill of an algorithm to finish uniform load balancing above the network. An competent load balancing method ought to be a good obligation tolerant technique [9].

- **Overhead Associated:**

It denotes the number of overheads enforced across the implantation of load balancing algorithm. Overheads are encompassed because of the task migration, inter-processor contact and inter-process communication. In finish, the attendance of overheads ought to be minimized to enhance the presentation of load balancing technique.

- **Response Time:**

It is described as the number of period seized to retort alongside a load balancing methodology in a distributed cloud environment. The metric ought to be minimized for competent system's performance.

- **Migration Time:**

It is the period seized to move the resources or jobs from one node to another. It ought to additionally be decreased for enhancing the presentation of the system.

- **Performance**:

It is the parameter utilized to check the efficiency of the system. The presentation of the arrangement has to be enhanced in a substantial manner.

- **Resource Utilization:**

Resource utilization ought to be optimized for effectual load balancing schema.

- **Scalability**:

The Scalability parameter in cloud computing ought to be improved. It is described as the skill of an algorithm to present balancing loads for a arrangement alongside each large number of nodes[10]

IV. RELATED WORK

Rahman, M. et al, in "Load Balancer as a Service in Cloud Computing" 2014 [11], the authors describe The explosive growth of cloud computing in recent years has led to a massive increase in both the amount of traffic and the number of service requests to cloud servers. This growth trend of load poses serious challenges to the cloud load balancer in efficient balancing of the load, already a daunting job. The cloud load balancing is a highly researched field where numerous solutions to balance load have been proposed. Unfortunately, no research papers provided a comprehensive review focusing Load Balancer as a Service (LBaaS) model. In this paper, they first understand the concepts of load balancing, its importance and desired characteristics in cloud. Then they provide complete review on the existing load balancing strategies, their strength, shortcomings and a comparative study. Finally, they presented load balancer as a service model adopted by the major market players, and their observation, future needs and challenges.

Ajit, M. et al, in "VM level load balancing in cloud environment" 2013 [12], the authors describe As Cloud Computing is spreading globally and number of users demanding more cloud services and better results are growing rapidly, cloud load balancing become a very interesting and important research area. Generally, cloud is based on powerful datacenters that handle large number of users, so it must be featured with load balancer to achieve reliability which depends on the way it handles the load. Cloud load balancing helps to enhance the overall cloud performance. Many algorithms were suggested for assigning the users requests to Cloud resources to provide services efficiently. This paper presents the analysis of three contemporary algorithms in cloud analyst tool to resolve the issue of cloud load balancing as a preparation phase for new load balancing technique. A Weighted Signature based load balancing (WSLB) algorithm is proposed to minimize users response time. Further, this paper also provides the anticipated results with the implementation of the proposed algorithm.

Al-Rayis, E. et al, in "Performance Analysis of Load Balancing Architectures in Cloud Computing" 2013 [13], the authors describe The Cloud computing is a rapidly emerging distributed system paradigm that offers a huge amount of IT resources as utility services at a reduced cost and flexible schemes. The key of such flexibility is an efficient load balancer that offers better management and utilization of virtualized underlying cloud infrastructures. However, most of the existing load balancers in cloud computing are based on either centralized or fully distributed architectures while the idea of harnessing multiple load balancers in a hierarchical structure to improve the sever load and job response time is still under studied. Therefore, this paper, aims at bridging this gap by providing a comparative study between the three load balancing architectures in cloud computing: centralized, decentralized and hierarchical load balancers. The experimental results suggest that the hierarchical architecture for load balancers best suits the public cloud environment and call for further research to test whether these results can be generalized for other types of clouds.

Sahu, Y. et al, in "Cloud Server Optimization with Load Balancing and Green Computing Techniques Using Dynamic Compare and Balance Algorithm" 2013 [14], the authors describe Cloud computing is a business oriented concept to provide online IT resources and IT services on demand using pay per use model where main goal of cloud service provider is to use cloud computing resources efficiently and gain profits marginally. One of the challenging areas in cloud computing is frequent optimization of cloud server. It mainly concerns with the load balancing of cloud data centers to improve efficiency of the host machine and minimize number of active host machine to support green computing concept. To balance the load of entire data center, they need to transfer the virtual machines of the overloaded host to the light weighted host using migration techniques. In this paper, they introduce a threshold based Dynamic compare and balance algorithm (DCABA) for cloud server optimization. Unlike the traditional server optimization strategies which consider only load balancing and scheduling of resources based on the usage of CPU, RAM and BW in physical servers, DCABA also minimizes the number of host machines to be powered on, for reducing the cost of cloud services. Our approach can serve the purpose of service cost reduction in cloud industry with effective utilization of available resources.

Zehua Zhang et al, in "A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation" 2010 [15], the authors describe Although cloud computing is generally recognized as a technology which will has a significant impact on IT in the future. However, Cloud computing is still in its infancy, many crucial problems need to be solved for the realization of the fine scenery which theoretically depicted by cloud computing. Load balancing is one of these problems; it plays a very important role in the realization of Open Cloud Computing Federation. They proposal a load balancing mechanism based on ant colony and complex network theory in open cloud computing federation in this paper, it improves many aspects of the related Ant Colony algorithms which

proposed to realize load balancing in distributed system, Furthermore, this mechanism take the characteristic of Complex Network into consideration. Finally, the performance of this mechanism is qualitatively analyzed, and a prototype is developed to enable the quantitative analysis, simulation results manifest the analysis.

Al Nuaimi, K. et al, in "A partial replication load balancing algorithm for distributed Data as a Service (DaaS)" 2013 [16], the authors describe This paper presents an easy and direct algorithm to solve the issue of load balancing in providing Data as a Service (DaaS) in the Cloud. The algorithm is based on some earlier approach for efficient dual direction data downloading. Our contribution is in solving the issue of the high storage demand when storing replicated data in multiple Cloud nodes. Rather than storing full replicas, their approach devises a model to store partial replicas of the data on multiple distributed Cloud servers. Moreover, they provide a direct method to download this data from multiple Cloud servers by coordinating the download process among the different nodes and different parts of the replicas on the Cloud. They implemented and evaluated their algorithm on the top of the CloudSim toolkit.

Guo Fen et al, in "Performance Weighted Deploying and Scheduling Strategy Research for Virtual Machine on Clouds" 2013 [17], the authors describe A performance weighted deploying and scheduling strategy for virtual machine on clouds (PWDSS) is introduced in this paper concerning users' requests of virtual resources and cloud load balancing. This approach follows three stages: first to use a monitor toolkits to collect the cloud performance data from the virtual machines and physical machines of cloud, and to standardize them, Second, to propose a cloud platform load balancing measurement model, in which the weighted vectors and matrix are set according to the customer requirements, Third, to give an algorithm to select the best appropriate physical machine in the measuring model obtained in stage 2, and then to deploy the new virtual machine, forecasting the load balancing value of every physical machine when the new virtual machine is deployed on it. The experimental results demonstrate that the proposed PWDSS can achieve better effects of system load balancing. At the same time, it can also meet the user requirements better.

Jihe Wang et al, in "Design and Optimization of Traffic Balance Broker for Cloud-Based Telehealth Platform" 2013 [18], the authors describe The use of cloud computing for the better health care is more and more important. Patient's real-time physiological signals, such as electrocardiogram (ECG) and blood pressure, should be transmitted to hospital servers for remote monitoring, and stored in Data Centers (DCs) so that the authorized doctors are able to access the patient's disease history. This implies challenges in brokering between the cloud consumers and providers when a huge number of people gets the real-time services from the distributed medical organizations. This paper proposes a probability-based bandwidth model in a telehealth cloud system, which helps cloud broker to allocate the most efficient computing nodes and links. This brokering mechanism considers the location of Personal Health Record (PHR) in cloud and schedules the real- time signal with a low information transfer between different hosts. Furthermore, their broker uses a bandwidth evaluation for the model, and they also compare various predicting methods to obtain the best bandwidth allocating algorithm. They simulate an inter-host environment for measuring the performance of their bandwidth allocating method with various data coherence protocols, which controls the domain of PHR in cloud, and the results show that their model is effective at determining the best performing service, and the inserted service validates the utility of their approach.

Ardagna, D. et al, in "MODAClouds: A model-driven approach for the design and execution of applications on multiple Clouds" 2012 [19], the authors describe Cloud computing is emerging as a major trend in the ICT industry. While most of the attention of the research community is focused on considering the perspective of the Cloud providers, offering mechanisms to support scaling of resources and interoperability and federation between Clouds, the perspective of developers and operators willing to choose the Cloud without being strictly bound to a specific solution is mostly neglected. They argue that Model-Driven Development can be helpful in this context as it would allow developers to design software systems in a cloud-agnostic way and to be supported by model transformation techniques into the process of instantiating the system into specific, possibly, multiple Clouds. The MODAClouds (MOdel-Driven Approach for the design and execution of applications on multiple Clouds) approach they present here is based on these principles and aims at supporting system developers and operators in exploiting multiple Clouds for the same system and in migrating (part of) their systems from Cloud to Cloud as needed. MODAClouds offers a quality-driven design, development and operation method and features a Decision Support System to enable risk analysis for the selection of Cloud providers and for the evaluation of the Cloud adoption impact on internal business processes. Furthermore, MODAClouds offers a run-time environment for observing the system under execution and for enabling a feedback loop with the design environment. This allows system developers to react to performance fluctuations and to re-deploy applications on different Clouds on the long term.

Xian Wang et al, in "A Global Optimal Service Selection Approach Based on QoS and Load-Aware in Cloud Environment" 2013 [20], the authors describe The global optimal Web service selection based on quality of service (QoS) in cloud environment has become a research focus when there are lots of the same or similar services. In this environment, it is possible that many service users request the same or similar services at the same time, which will result in users' unsatisfied requirement and services' load imbalance. The existing service selection approaches usually suppose that service's load capacity is infinite and user always select the service with the best expected QoS despite the amount of user requests. Therefore, it is very important problem how to get a tradeoff between the users' satisfied requirement and service load balance. To solve this problem, this paper presents a global optimal service selection approach based on QoS and load-aware in cloud environment. In this approach, they first build a user QoS utility model which describes the relationship between user's request and services' QoS, and design a service's load capacity model to achieve the load capacity of a service. Then, they use 0-1 integer programming to build a global optimal model based on QoS utility of users and services' load capacity, and provide the optimal service selection algorithm for users. Finally, by conducting

large-scale experiments based on a Web service dataset, they show that their approach can effectively help users to select high qualified services while keeping load balance of services in cloud environment.

Qin Liu et al, in "Dynamic Grouping Strategy in Cloud Computing" 2012 [21], the authors describe Cloud computing has emerged as a new type of commercial paradigm. As a typical cloud service, each file stored in the cloud is described with several keywords. By querying the cloud with certain keywords, a user can retrieve files whose keywords match his query. An organization that has thousands of users querying the cloud can set multiple proxy servers inside itself to reduce the querying cost. All users can be classified into different groups, and the users in a group will send their queries to the same proxy server, which will query the cloud with a combined query, i.e., the union of keywords in a group of queries. In such an environment, an important problem is cost efficiency, i.e., how to classify users into different groups so that the total number of returned files is minimized. Observing that this is mainly affected by the number of keywords in the combined queries, their problem is translated to classifying n users into k groups in the case of k proxy servers, so that the number of keywords in k combined queries is minimized. Since more common keywords in a group of queries will generate less keywords in the combined queries, they should group users with the most common keywords together. Two additional aspects needed to be addressed are load balancing and robustness, i.e., the workloads among proxy servers are balanced and each user obtains search results even if some proxy servers fail. To solve above problems simultaneously, they propose mathematic grouping and heuristic grouping strategies, where mathematic grouping solves the relaxed problem by using a local optimization method, and heuristic grouping is based on the classical heuristic clustering algorithm, K-means. Extensive evaluations have been conducted on the analytical model to verify the effectiveness of their strategies.

Zohar, E. et al, in "PACK: Prediction-Based Cloud Bandwidth and Cost Reduction System" 2014 [22], the authors describe In this paper, they present PACK (Predictive ACKs), a novel end-to-end traffic redundancy elimination (TRE) system, designed for cloud computing customers. Cloud-based TRE needs to apply a judicious use of cloud resources so that the bandwidth cost reduction combined with the additional cost of TRE computation and storage would be optimized. PACK's main advantage is its capability of offloading the cloud-server TRE effort to end-clients, thus minimizing the processing costs induced by the TRE algorithm. Unlike previous solutions, PACK does not require the server to continuously maintain clients' status. This makes PACK very suitable for pervasive computation environments that combine client mobility and server migration to maintain cloud elasticity. PACK is based on a novel TRE technique, which allows the client to use newly received chunks to identify previously received chunk chains, which in turn can be used as reliable predictors to future transmitted chunks. They present a fully functional PACK implementation, transparent to all TCP-based applications and network devices. Finally, they analyze PACK benefits for cloud users, using traffic traces from various sources.

Yi Zhao et al, in "Adaptive Distributed Load Balancing Algorithm Based on Live Migration of Virtual Machines in Cloud" 2009 [23], the authors describe EUCALYPTUS, an open source cloud-computing framework, is still lack of load balancing. In the paper, they provide a kind of implementation by adaptive live migration of virtual machines. They design and implement a simple model which decreases the migration time of virtual machines by shared storage and fulfills the zero-downtime relocation of virtual machines by transforming them as Red Hat cluster services. During the migration process, they also keep the inclusion relationship between VLAN and virtual machines. They propose a distributed load balancing algorithm COMPARE_AND_BALANCE based on sampling to reach an equilibrium solution. The experimental results show that it converges quickly.

Wei Yuan et al, in "Towards Efficient Deployment of Cloud Applications through Dynamic Reverse Proxy Optimization" 2013 [24], the authors describe With the increase of users and the deployment requests, the issue of dynamic deployment in PaaS becomes prominent. Different approaches of application deployment have been deeply discussed, but the issues like fast response to a large number of concurrent deployment requests are rarely focused. In this work, they extend Nginx as a dynamic reverse proxy to support dynamically remote configuration for better elasticity of cloud applications in PaaS, and then further optimize it for improving performance under a large number of concurrent configuration requests. Three optimization approaches are proposed: Batch Request Committing (BRC), Batch File Processing (BFP) and In Memory Configuration (IMC). They give a detailed implementation of each method, and a qualitative analysis on three optimization approaches has been made. Finally, a series of experiments are presented to validate the optimization effect. The experiment results show that the maximum throughput per second has increased significantly, and the average response time of each request has decreased dramatically.

Chandramohan, D. et al, in "Hybrid authentication technique to preserve user privacy and protection as an end point lock for the cloud service digital information" 2013 [25], the authors describe This paper presents a new approach for privacy preserving of user data and publishing in cloud storage area. Hybrid authentication technique overcomes the limitations of attackers and general intruders and preserves better utilization of user's confidential data by providing access only to authorized persons. In this proposed technique it is illustrated how this technique used to prevent user secret data attributes kept in digital cloud storage environment. Moreover the proposed system shows a better data privacy preserving utility by users and handling sensitive data during vulnerable attempts and attacks to the cloud storage area.

V. CONCLUSION AND FUTURE SCOPE

Load balancing in cloud computing systems is a big challenge. A distributed solution is required always in need. Because it is not always practical feasible or cost efficient to maintain one or more idle services just as to fulfill the required demands. Jobs cannot be assigned to appropriate servers and clients individually for efficient load balancing as

cloud is a very complex structure and components are present throughout a wide spread area. 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. In future, we would like to improve load balancing in cloud systems using Ant colony Optimizations.

REFERENCES

[1] Zissis, Dimitrios, and Dimitrios Lekkas. "Addressing cloud computing security issues." Future Generation computer systems 28, no. 3 (2012): 583-592.

[2] Khiyaita, A., M. Zbakh, H. El Bakkali, and Dafir El Kettani. "Load balancing cloud computing: state of art." In Network Security and Systems (JNS2), 2012 National Days of, pp. 106-109. IEEE, 2012.

[3] Mao, Ming, and Marty Humphrey. "A performance study on the vm startup time in the cloud." In Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on, pp. 423-430. IEEE, 2012.

[4] Jhawar, Ravi, Vincenzo Piuri, and Marco Santambrogio. "Fault tolerance management in cloud computing: A system-level perspective." Systems Journal, IEEE 7, no. 2 (2013): 288-297.

[5] Barrett, Enda, Enda Howley, and Jim Duggan. "Applying reinforcement learning towards automating resource allocation and application scalability in the cloud." Concurrency and Computation: Practice and Experience 25, no. 12 (2013): 1656-1674.

[6] Caton, S.; Haas, C.; Chard, K.; Bubendorfer, K.; Rana, O.F.,"A Social Compute Cloud: Allocating and Sharing Infrastructure Resources via Social Networks",IEEE,Services Computing, IEEE Transactions on, 2014

[7] Chun-Wei Tsai; Rodrigues, J.J.P.C.,"Metaheuristic Scheduling for Cloud: A Survey",IEEE,Systems Journal, IEEE, 2014

[8] Chun-Wei Tsai; Wei-Cheng Huang; Meng-Hsiu Chiang; Ming-Chao Chiang; Chu-Sing Yang,"A Hyper-Heuristic Scheduling Algorithm for Cloud",IEEE,Cloud Computing, IEEE Transactions on, 2014

[9] Dastjerdi, A.V.; Buyya, R.,"Compatibility-Aware Cloud Service Composition under Fuzzy Preferences of Users",IEEE,Cloud Computing, IEEE Transactions on, 2014

[10] Hung, P.P.; Mui Van Nguyen; Aazam, M.; Eui-Nam Huh,"Task scheduling for optimizing recovery time in cloud computing",IEEE,Computing, Management and Telecommunications (ComManTel), 2014 International Conference on, 2014

[11] Rahman, M.; Iqbal, S.; Gao, J.,"Load Balancer as a Service in Cloud Computing",IEEE,Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on,2014

[12] Ajit, M.; Vidya, G.,"VM level load balancing in cloud environment",IEEE,Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference on,2013

[13] Al-Rayis, E.; Kurdi, H.,"Performance Analysis of Load Balancing Architectures in Cloud Computing",IEEE,Modelling Symposium (EMS), 2013 European,2013

[14] Sahu, Y.; Pateriya, R.K.; Gupta, R.K.,"Cloud Server Optimization with Load Balancing and Green Computing Techniques Using Dynamic Compare and Balance Algorithm",IEEE,Computational Intelligence and Communication Networks (CICN), 2013 5th International Conference on,2013

[15] Zehua Zhang; Xuejie Zhang,"A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation",IEEE,Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on,2010

[16] Al Nuaimi, K.; Mohamed, N.; Al Nuaimi, M.; Al-Jaroodi, J.,"A partial replication load balancing algorithm for distributed Data as a Service (DaaS)",IEEE,High Performance Computing and Simulation (HPCS), 2013 International Conference on,2013

[17] Guo Fen; Min Hua-Qing; Yang Jie,"Performance Weighted Deploying and Scheduling Strategy Research for Virtual Machine on Clouds",IEEE,Emerging Intelligent Data and Web Technologies (EIDWT), 2013 Fourth International Conference on,2013

[18] Jihe Wang; Bing Guo; Meikang Qiu; Zhong Ming,"Design and Optimization of Traffic Balance Broker for Cloud-Based Telehealth Platform",IEEE,Utility and Cloud Computing (UCC), 2013 IEEE/ACM 6th International Conference on,2013

[19] Ardagna, D.; di Nitto, E.; Mohagheghi, P.; Mosser, S.; Ballagny, C.; D'Andria, F.; Casale, G.; Matthews, P.; Nechifor, C.-S.; Petcu, D.; Gericke, A.; Sheridan, C.,"MODAClouds: A model-driven approach for the design and execution of applications on multiple Clouds",IEEE,Modeling in Software Engineering (MISE), 2012 ICSE Workshop on,2012

[20] Xian Wang; Jianxun Liu; Buqing Cao; Mingdong Tang,"A Global Optimal Service Selection Approach Based on QoS and Load-Aware in Cloud Environment",IEEE,High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on,2013

[21] Qin Liu; Yuhong Guo; Jie Wu; Guojun Wang,"Dynamic Grouping Strategy in Cloud Computing",IEEE,Cloud and Green Computing (CGC), 2012 Second International Conference on,2012

[22] Zohar, E.; Cidon, I.; Mokryn, O.,"PACK: Prediction-Based Cloud Bandwidth and Cost Reduction System",IEEE,Networking, IEEE/ACM Transactions on,2014

[23] Yi Zhao; Wenlong Huang,"Adaptive Distributed Load Balancing Algorithm Based on Live Migration of Virtual Machines in Cloud",IEEE,INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on,2009

[24] Wei Yuan; Hailong Sun; Xu Wang; Xudong Liu,"Towards Efficient Deployment of Cloud Applications through Dynamic Reverse Proxy Optimization",IEEE,High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on,2013

[25] Chandramohan, D.; Vengattaraman, T.; Rajaguru, D.; Baskaran, R.; Dhavachelvan, P.,"Hybrid authentication technique to preserve user privacy and protection as an end point lock for the cloud service digital information",IEEE,Green High Performance Computing (ICGHPC), 2013 IEEE International Conference on,2013

