This full-text paper was peer-reviewed and accepted to be presented at the IEEE WiSPNET 2016 conference.

Stochastic Based Load Balancing Mechanism for Non-Iterative Optimization of Traffic in Cloud

Suriya Begum
Dept. of Computer Science & Engg,
New Horizon College of Engineering,
Bangalore, India
E-Mail: suriyabegumphd@gmail.com

Prashanth C S R
Dept. of Computer Science & Engg,
New Horizon College of Engineering,
Bangalore, India

***Abstract*— A precise Quality-of-Service in cloud can be offered if the traffic management uncertainty issues are addressed properly. It is also found that developing a cost-efficient load balancing technique in cloud environment is still an unsolved problem after reviewing the work done in load balancing till date. Hence, we introduce a novel algorithm called as SBLB i.e. Stochastic Based Load balancing that applies non-iterative and state-based approach of optimization to map the appropriate resources to be allocated for virtual machines dynamically. SBLB performs resource availability and maps state-transition approach to address the complexity of resource allocation towards the virtual machines based on its states (overloaded, under-loaded, and normal). The comparative study outcome shows a minimized response time, increased number response processed per seconds, and lower instances of virtual machine in contrast with existing system.**

Keywords-component; Cloud, Load Balancing, Resource Allocation, Traffic Management, Network Monitoring

I. INTRODUCTION

The area of cloud computing has been witnessing a pace of technological advancement with its maturity level yet to be achieved. As known widely that cloud computing is all about offering the ubiquitous forms of services from multiple devices irrespective of locations [1]. This ubiquitous service furnishes various forms of potential capabilities to its consumers for catering up their dynamic requirements e.g. downloading/installing applications, accessing resources, etc. One of the potential advantages of cloud services is its minimization of operational expenditure as the corporate doesn't require either hardware or any extra software to work on cloud. The console of cloud services are directly accessed from client-specific applications e.g. browser. Another benefit of the cloud / ubiquitous computing is its capability of higher degree of distributive networking in order to support massive applications to double the number of users online, which is not possible in non-cloud services. However, traffic management in cloud is not so easy task and till date there is no full-proof traffic management service that has maintained 100% zero downtime. Hence, load balancing is one of the critical problems in cloud which still requires further closer investigation. Basically, load balancing offers cloud mainly scheduling or provisioning of the resources over distributed networking environment so that all the demands of job processing in the cloud users are met at any cost. Therefore by applying load balancing features in cloud environment, the service providers claims better availability of resources to be provisioned, careful utilization of an event, lesser drainage of energy, minimized cost of utilized resources. Hence, it can be seen that resource allocation (or provisioning / scheduling) plays a huge contributory role in load balancing in cloud environment specifically in virtual machine. Resource allocation can be stated as the process of mapping the appropriate resources to different forms of processing elements in such a way that task allocation for any should not be either overloaded or under loaded [2][3]. In this regards, virtual machine are the prime targets for applying load balancing as it is located on the physical servers. However, there is a huge difference between resource allocations as task allocation when it comes to virtual machine. Resource allocation related more to virtual machine level while task scheduling is carried out by time sharing processes. The problems associated with resource allocations are to guarantee the availability while that of task scheduling is reducing the cumulative response time.

Moreover, in order to support higher degree of elasticity in the cloud services, it is mandatory that allocation of the resources to be carried out in *dynamic* fashion. Owing to the alterations of the inbound workloads, the allocated resources must also be subjected for appropriate change in order to support the processing of job in virtual machine. The biggest research set-back is that mechanism of load balancing in cloud is completely different from other conventional networking system. Therefore, load balancing in cloud computing requires a special attention to be laid for understanding the driving force behind traffic modelling and resource allocation so that an efficient balance would be ensured. The present paper outlines a novel algorithm where the prime contribution is to formulate a simple and non-iterative optimization technique for enhancing the performance of load balancing in virtual machines. The target is also to decrease the response time with less instances of virtual machine to process the massive load of uncertain traffic. Section II discusses about the fundamental briefing of load balancing in cloud followed by recent research work being carried out in literatures in Section III. Section IV outlines the problems that have been identified from existing techniques. The proposed study is discussed in Section V followed by research methodology in Section VI. Algorithm Implementation is briefed in Section VII followed by Result discussion in Section VIII. Finally, summary of the paper is carried out in Section IX.

978-1-4673-9338-6/16/$31.00 ©2016 IEEE

II. LOAD BALANCING IN CLOUD

The mechanism of load balancing over cloud is all about an effective distribution of the traffic and necessary supporting resources (Fig.1). Essentially, the load balancing mechanism used over cloud assists to accomplish minimized cost of technological implementation in order to cater the job request by the users. The reasons of traffic congestion in cloud are mainly for noisy neighbors, inferior performance of application hosted by virtual machine, and uncertain traffic scenarios. *Noisy neighbors* is a term used in cloud computing where the tenants applies monopoly of channel capacity, CPU utilization, disk usage, as well as other subsidiary resources that can adversely affect the performance of the virtual machine. Due to these noisy neighbors, various associated virtual machine as well as relevant cloud applications sharing similar infrastructure will evenly encounter highly performance degradation [4][5].

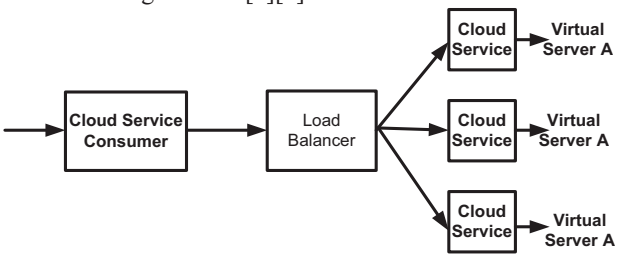

Figure 1 Load Balancing Scheme in Cloud

At present in order to mitigate the adverse effect of noisy neighbors, usage of virtual local area network is quite frequent. Hence, it is required that inbound traffic to be properly engaged among the virtual machines for better load balancing. At present, there are many service providers in cloud that offer significant load balancing techniques with elasticity as the prime selling point. Some of the present applications in load balancing are:

- Azure (Microsoft) provides an effective traffic-based tool that distributes the cloud services over numerous data centers.
- Google also offers workload distribution as a service in its infrastructure. Google also have a potential engine that performs network traffic distribution among the instances of virtual machines.

According to the study carried out by Katyal and Mishra [6], the existing load balancing algorithms are mainly classified into 5 types viz. i) static, ii) dynamic, iii) centralized, iv) distributed, and v) hierarchical mainly. The problems with static load balancing techniques are less supportability of scalability and flexibility. Dynamic load balancing algorithm suffers from computational complexity while centralized forms of load balancing algorithm suffer from less tolerance towards potential faults. Distributed load balancing algorithm is one of the most frequently seeked technique among the researchers but still it suffers from overhead in communication and algorithm complexity. Finally, hierarchical load balancing algorithms also suffers from lower complexity as well as inferior fault tolerance in virtual environment. The next section will discuss some of the recently presented research manuscript highlighting load balancing techniques in cloud.

III. RECENT STUDIES TOWARDS LOAD BALANCING

This section discusses about the recently evolved research work dedicated to load balancing techniques in cloud computing environment. We have already reviewed various existing techniques towards cloud [7][8].

Most recently, study towards job scheduling has been carried out by Kang and Choo [9] who have developed a technique of performing clustering and decision making in large-scale of cloud deployment. Nuaimi et al. [10] have presented an optimization scheme for data centers in cloud. The approach was to exploit the storage to enhance the load balancing mechanism for large traffic. Usage of resource scheduling as a media to control the load is seen in the work of Chen et al. [11]. The author uses fuzzy logic to performing load balancing mechanism. Same author have also published a manuscript highlighting the usage of load balancing in OpenFlow network [12]. A unique technique of migration for balancing the load in cloud has been witnessed in the work carried out by Song et al. [13]. The work carried out by Zhong et al. [14] has used statistical parameter considering the mobile users to ensure proper load balancing technique. Lim and Park [15] have studied about the resource sharing techniques and introduced a mechanism using middleware that can minimize the traffic intensity to some extent. Schimidt et al. [16] have used elastic charecteristics of cloud for maintain the traffic intensity of large scale cloud deployment. Shu et al. [17] have contributed a study that focuses on energy conservation as well as resource allocation technique to enhance the communication performance of the green cloud. Song et al. [18] have emphasized on the usage of game theory as the possible technique to perform load balancing in cloud computing. Usage of scheduling application for the mobile applications as an approach to minimize and manage traffic load over cloud platform was seen in the study of Wei et al. [19]. Wang and Casale [20] have presented a study emphasizing on the usage of conventional round robin technique over cloud interface for accomplishing load balancing. The study also emphasizes on enhancing the capacity of the virtual machine. Fan et al. [21] have presented a study using clustering approach over the cache to ensure a cost effective load balancing scheme. Xianfeng and Tao [22] have used genetic algorithm as well as swarm intelligence-based approach to perform load balancing in cloud environment. A completely novel approach called as *firefly* was introduced by Florence and Shanthi [23] for mechanizing the availability of the virtual machines. Tchernykh et al. [24] have implemented a unique technique of load balancing that equally emphasize on energy efficiency considering voice-over internet protocol traffic in cloud.

Our prior study have presented a mathematical modelling that combinely performs routing as well as resource scheduling to maintain a better form of load balancing in cloud environment [25]. The results has proven better rate of task completion and higher number of job completion in contrast to existing technique of load balancing. Although, all these techniques have significant contribution towards incorporating load balancing, but more exploration can be done to evolve with more better, less sophisticated, and cost efficient scheme. The next section discusses about the problem identification.

IV. Problem Identification

This section discusses about the limitations of the existing approaches towards load balancing and identifies the missing gap from the literatures. By reviewing the existing approaches, it is quite clear that majority of the approaches deploy resource scheduling approaches as the cost effective measures towards confirming load balancing in cloud deployment. However, we find that there is little consideration that has not been found to be considered by the prior researchers:

a. The existing approaches considers formulating conflicting theory of virtualization without any robust assumption that resource allocation associated with virtual machine is highly complicated task as better configuration of resources have to be dynamically formulated from the millions of configurations based on the incoming traffic which is highly uncertain. The assumptions and dependencies of resource scheduling is quite vague and non-practical in existing research work.

b. The recently developed approaches that focus on the managing and scheduling virtual machines have not emphasized on the dynamicity (or randomness) of the workloads under processed or to be processed. This approach are quite impractical over the virtual machines that hosts the services.

The above two are the major problems that have been identified from the existing studies pertaining to resource allocation as load balancing techniques. With the change of inbound traffic towards the virtual machines, if the configurations of the virtual machine is not pre-identified and selected than it leads serious violation of SLA (Service Level Agreement) and also leads to degradation of communication performance. We have observed that existing approaches normally uses a technique where the resources to be allocated is fine-tuned till it meets the demands of SLA; however, this is quite not possible and will lead to higher delay in service relays. This is because in order to cater up resource requirement in virtual machine, there is need of various fine tuning network and resource control parameters, which were never discussed in any research papers.

We have come across various studies where optimization technique is adopted for identifying the best resources to cater up the load balancing requirement of existing traffic. Although, these forms of approach can identify short termed resource allocation to meet the current workload, but there is no guarantee that same behaviour will be retained over a larger period of time. Moreover application of optimization technique may also lead to higher time and resource consumption just to execute the algorithm without any guaranteed results of the approach. Hence, cost effectiveness of the approaches is drastically missing from the literature pertaining to load balancing in cloud computing.

Such forms of problems can be only addressed if an analytical modelling is carried out considering all the above stated problems to be addressed. The next section discusses about one such approach, which we have developed specifically to overcome the ongoing problems in load balancing to offer cost effective traffic management in cloud.

V. Stochastic-based Load Balancing

The prime aim of the proposed system is to develop a unique and cost effective load balancing scheme that can autonomously undertake decision based on the situation of the traffic in cloud environment. The study essentially pertains to the resource management system that can directly affect the load balancing system to be taken place from virtual machine. The schematic architecture of the proposed system is shown in Fig.2. The mechanism of the proposed system is as follows: We monitor the amount of resource utilization for all the virtual machine and categorized into three different types i.e. i) over-utilization, ii) under-utilization, and iii) normal utilization. We map this score of resource utilization with respect to similar three different states that corresponds to the traffic and resource information for the existing virtual machines. Our algorithm SBLB (Stochastic Based Load Balancing) checks for all the inbound traffic and check for the information of the state availability of the virtual machine. The novelty in this part of implementation is our algorithm takes decision of corresponding action (β) to be undertaken to meet the demands of the incoming traffic without even using a computationally sophisticated queuing mechanism.

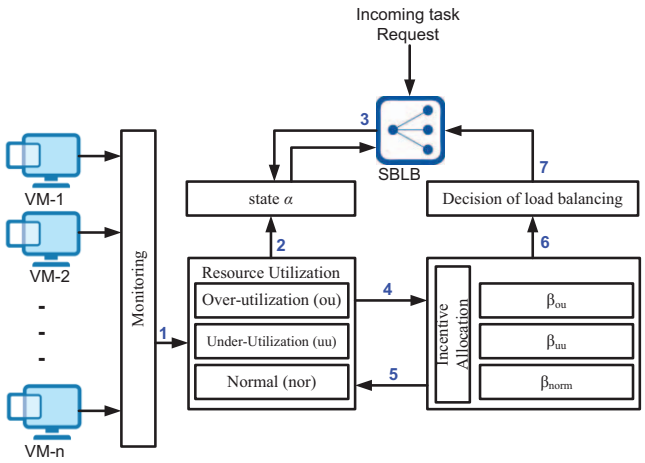

Figure 2 Schematic Architecture of SBLB

The proposed technique also implements a novel incentive-based mechanism in order to govern the action invocations for meeting the demands of load balancing. For an example, if the incoming traffic requires higher version of resources, it should opt for virtual machine with under-utilized states. Similarly, if the incoming traffic has very lower resource demands it could use only normal or sometimes over-utilized states of virtual machine. However, dynamic demands of the traffic could be only processed of such state components (*ou*, *uu*, *nor*) could have inter-operability. We apply interoperability among the state components by applying necessary incentives just to meet every demands of the traffic. For an example, we apply the incentive allocation block taking the decision about the amount of resources to be allocated if there is a possible transition of states among the state components. Hence, we apply stochastic modelling approach to perform predictive analysis about the behaviour of SBLB that is executed over the virtual machines in cloud environment. The next section discusses about the research methodology adopted to implement this system.

VI. Research Methodology

The proposed system considers analytical research methodology where a very unique and non-conventional approach has been used for developing a load balancing algorithm over cloud environment. The proposed technique is divided into following modules:

- **Network Model**: We use our prior framework [25] for developing network model that consists of virtual machines, uncertain number of clients, rack servers, different types of processes running, job status etc as the research variables. However, the emphasis is mainly laid over virtual machine where the proposed SBLB will be executed inside the virtual machine which is responsible for hosting the services.
- **Stochastic-based Model**: SBLB uses a state-based transition mechanism for performing two purposes viz. i) understanding the type of traffic and ii) making decisions to invoke certain specific action based on appropriately detected traffic state. The framework that runs SBLB will collect various inbound traffic information from the virtual machine along with existing resource being allocated to process the existing jobs by virtual machine. The collected information is maintained in a static array which is time-bound for appropriate time-based analysis. However, as the algorithm will be required to bear various heterogeneous information about the jobs (e.g. size of jobs, memory, CPU utilization, processed jobs, unprocessed jobs etc), hence, we use state-based modelling to further generalized the discrete traffic behaviour over cloud. We also apply stochastic theory for performing this modelling in order to incorporate predictive analysis for load balancing over uncertain traffic condition over cloud. The technique performs decision of routing the incoming traffic and allocates the resources to the virtual machines based on the explicit states of the traffic. We call this as *action* that corresponds to set of activity of resource management to control the load balancing based on states.
- **Incentive Model**: The prime purpose of the incentive model is to manage various virtual machine to infer about the correct or incorrect decision of processing the suitable or unsuitable (or unprioritized) jobs in the cloud environment. We define a table (Table 1) with its states of incentives to be allocated in the form of adopted states to process a current traffic and adopted action to process the anticipated traffic. Essentially, such states are the direct representations of resources being over or under-utilized so that management of load balancing can be carried out in predictive fashion. The incentive model checks for the performance of virtual machine for the current traffic and it invokes a set of action that can perform more resource scheduling in order to leverage the communication performance. Hence, proposed incentive-based modelling basically is carried out in adhoc fashion where the load balancing is only carried out when the traffic condition corresponds to any of the specified states in the proposed technique.

The next section discusses briefly about the algorithm that implements this research methodology.

VII. Algorithm Design

The development of the proposed algorithm is carried out in analytical approach. The proposed algorithm considers three different states viz. overload (*ol*), normal (*nor*), and under-load (*ul*) of the present traffic scene. Any threshold-based approach can be used to define this. For an example, traffic with more than, less than and equal to X terabytes of load over cloud observed in specific duration of observation time can be notified as *ol*, *ul*, and *nor* states.

Algorithm for Stochastic-Based Load Balancing (SBLB)

Input: α (state), β (action), $I_{\alpha,\beta}$(incentive signal), C_F(Channel Correction factor, ϕ (cloud service set parameter).

Output: load balancing

Start

1. init α, β
2. Apply local optimization

   $LO= g(I_{\alpha, \beta}, \mu)$
3. Apply global optimization

   $GO=\arg_{max}(g(I_{\alpha, \beta}, \mu))$
4. Apply load-balancing

$$\Delta\beta = \left| \left(\frac{n_{VM}}{\phi} \right)_{cur} - \left(\frac{n_{VM}}{\phi} \right)_{\exp} \right| .C_F$$

5. Design incentive function, $I_{strat}=[I_{ol}, I_{nor}, I_{ul}]$
6. Apply line-4 accordingly w.r.t $I_{\alpha,\beta}$

End

The algorithm takes the input of such states and a specific set of action (which needs to counteract with specific state of traffic). This will also mean that invoking of β_t will result transition of state α_t to α_{t+1}.The algorithm applies dual layer of optimization as shown in step-2/3. The local-level optimization is carried out by using a function g that mainly performs concatenation operation of incentive $I_{\alpha,\beta}$ and μ. The variable $I_{\alpha,\beta}$ will mean an incentive being given for a traffic channel based on its respective states. The variable μ will mean scalar multiplication of discount factor, best policy, and state transition function. We infer that a better load balancing could only be performed if local optimization gathers more and more information about the traffic states in each iteration. The global optimization is carried out only by considering the predictive values of the states using random modelling theory. Line-4 implements the formulations of load-balancing technique, which is basically a decision of action. The first component of Δβ is basically the difference between the current and expected values of number of virtual machines with respect to certain cloud service set parameter ϕ. The second component of Δβ is basically channel correction factor in order to read and consider the channel-based errors. Finally, an incentive strategical function I_{strat} is designed, which applies equation of load balancing based on three states of traffic. Incentive allocation matrix in Table 1 assists in this regards.

Table 1 Incentive allocation matrix

row=states (α)	col=action(β)			
	States	*ul*	*nor*	*ov*
	ul	$\exp^{-\|sel\_\beta\|}$	1	-1
	nor	-1	1	-1
	ov	1	1	$1-\exp^{-\|sel\_\beta\|}$

In order to have clarity, we define state to be either from size of queue, cloud response time of service, etc. The variable *sel_β* will mean the adopted set of action that should determine the amount of the resources from the virtual machine to be engaged or disengaged based on the traffic states. It should be understood that proposed system is all about predictive analysis of load balancing using probability theory and random modelling.

VIII. Result Discussion

This section discusses about the brief of the outcomes being accomplished by implementing the proposed algorithm SBLB. For better evaluating the effectiveness of the study, we compare SBLB with similar category of work done by Aslanzadeh et al. [26]. The author has used optimization technique over cloud simulator (CloudSim) to perform load balancing of virtual machine. We slightly fine tune this work to fit in our evaluation test bed by changing the iterations number from 0-10000. The graphical results are briefly discussed as below.

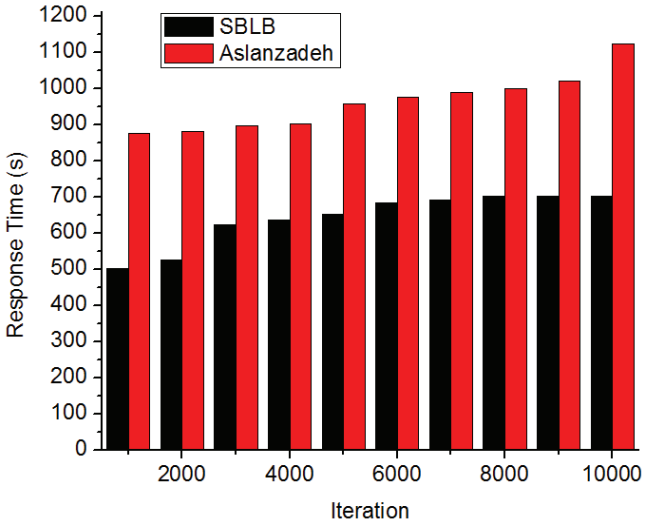

Figure 3 Comparative Analysis of Response Time

Fig.3 shows increase of response time with the increase of iteration. We randomly incorporate traffic load in increasing order of increasing rounds of iteration to see its possible impact on response time. We define response time as time duration after the job is allocated to job is relayed by the virtual machine. The technique applied by Aslanzadeh [26] uses iterative optimization technique in order to perform load balancing in cloud environment. However, we have applied a non-iterative optimization technique based on state-based transition, which cut shorts 70% of the processing time and therefore, the response time is quite superior compared to the existing system.

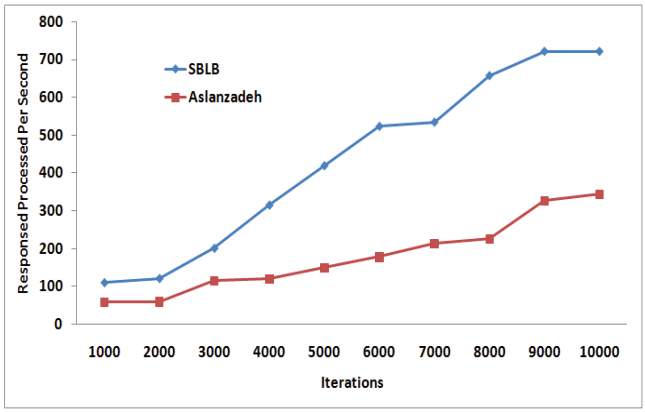

Figure 4 Comparative Analysis of Response Processed Per Second

Fig.4 showcases the number of response processed per seconds by the virtual machine. The outcome shows existing technique of Aslanzadeh et al. [26] could only process half of the jobs cumulatively as compared to the proposed SBLB with increasing number of iterations of traffic load. Hence proposed system could offer better load balancing

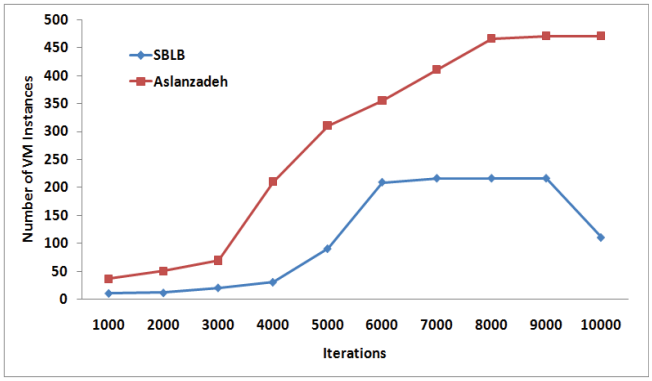

Figure 5 Comparative Analysis of VM Instances Running

Fig.5 showcases the comparative analysis of the number of instances of the virtual machines running for processing the current jobs. The outcome shows that number of VM instances for existing system increases almost linearly owing to the particle swarm optimization technique applied for accomplishing better configurations of the virtual machine. Hence the existing system is in continuous search of the virtual machine and hence it almost engages majority of the virtual machines for even a small job just to increase the response time. However, the proposed system checks the states of the resources applicable to virtual machines and with aid of interactive communication between the state-component. It supports even a sudden invocation of actions in order to meet the dynamic requirements of user. Hence, less number of virtual machine instances is quite enough to maintain the similar response time as shown in Fig.3. Therefore, the proposed system can be said to offer cost effective solution using stochastic based non-iterative optimization approach of load balancing over cloud environment.

IX. CONCLUSION

This paper presents a novel approach with higher simplicity to develop a simple algorithm for ensuring cost effective cloud computing. Our algorithm SBLB is much focused on reviewing the states of the virtual machines to understand the correct set of conditions required for allocating resources. The novelty here is existing system blindly allocates the heavy resources where probability of resource wastage is always there. However, uniqueness in our case is there is no wastage of resources as first the resources to be allocated are estimated based on the states and necessary action is taken to allocate the appropriate resources. We also develop an interaction among the state components in case of state transition so that dynamically the algorithm can take autonomous decision to make change in its existing allocate. The interesting part of the study is non-iterative optimization technique as compared to existing optimization techniques used for load balancing. Our outcomes prove the fact by exhibiting lesser value of response time, more capability to handle incoming traffic, and proper utilization of virtual machines.

REFERENCES

[1] R. Sampathkumar, Disruptive Cloud Computing and IT: Cloud Computing Simplified for every IT Professional, Xlibris Corporation, 2015

[2] Mastorakis, George, Resource Management of Mobile Cloud Computing Networks and Environments, IGI Global, 2015

[3] X. Li, J. Qiu, Cloud Computing for Data-Intensive Applications, Springer, 2014

[4] G. J. Selig, Implementing Effective IT Governance and IT Management, Van Haren, 2015

[5] B. Gregg, Systems Performance: Enterprise and the Cloud, Prentice Hall, 2013

[6] M. Katyal, A. Mishra, "A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment", *International Journal of Distributed and Cloud Computing*, Vol.1, Iss.2, 2013

[7] S. Begum, Prashanth C.S.R, "Investigational Study of 7 Effective Schemes of Load Balancing in Cloud Computing", *International Journal of Computer Science Issues*, Vol. 10, Iss.6, No 1, 2013

[8] S. Begum, Prashanth C.S.R, "Review of Load Balancing in Cloud Computing", *International Journal of Computer Science Issues*, Vol. 10, Issue 1, No 2, January 2013

[9] B. Kang and H. Choo, "A cluster-based decentralized job dispatching for the large-scale cloud", *Springer- EURASIP Journal onWireless Communications*, 2016

[10] K. A. Nuaimi, N. Mohamed, M.A. Nuaimi, and J. A. Jaroodi, "Partial Storage Optimization and Load Control Strategy of Cloud Data Centers", *Hindawi Publishing Corporation*, eScientific World Journal, 2015

[11] Z. Chen, Y. Zhu, Y. Di, and S. Feng, "A Dynamic Resource Scheduling Method Based on Fuzzy Control Theory in Cloud Environment", *Hindawi Publishing Corporation Journal of Control Science and Engineering*, 2015

[12] W. Chen, Z. Shang, X. Tian, and H. Li, "Dynamic Server Cluster Load Balancing in Virtualization Environment with OpenFlow", *Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks*, 2015

[13] X. Song, Y. Ma, and D. Teng, "A Load Balancing Scheme Using Federate Migration Based on Virtual Machines for Cloud Simulations", *Hindawi Publishing Corporation Mathematical Problems in Engineering*, 2015

[14] H. Zhong, Q. Lin, J. Cui, R. Shi, and L. Liu, "An Efficient SDN Load Balancing Scheme Based on Variance Analysis for Massive Mobile Users", *Hindawi Publishing Corporation Mobile Information Systems*, 2015

[15] Y. Lim and J. Park, "Sensor Resource Sharing Approaches in Sensor-Cloud Infrastructure", *Hindawi Publishing Corporation International Journal of Distributed Sensor Networks*, 2014

[16] D. C. Schmidt, J. White, C. D. Gill, "Elastic Infrastructure to Support Computing Clouds for Large-scale Cyber-Physical Systems", *IEEE International Symposium of Object/Component/Service-Oriented Real-Time Distributed Computing*, pp.56-63, 2014

[17] W. Shu, W. Wang, and Y. Wang, "A novel energy-efficient resource allocation algorithm based on immune clonal optimization for green cloud computing", *Springer- EURASIP Journal on Wireless Communications and Networking*, 2014

[18] S. Song, T. Lv, and X. Chen, "Load Balancing for Future Internet: An Approach Based on Game Theory", *Hindawi Publishing Corporation Journal of Applied Mathematics*, 2014

[19] X. Wei, J. Fan, Z. Lu, and K. Ding, "Application Scheduling in Mobile Cloud Computing with Load Balancing", *Hindawi Publishing Corporation Journal of Applied Mathematics*, 2013

[20] W. Wang, G. Casale, "Evaluating Weighted Round Robin Load Balancing for Cloud Web Services", *IEEE International Symposium of Symbolic and Numeric Algorithm for Scientific Computing*, pp.393-400, 2014

[21] B. Fan, H. Lim, D. G. Andersen, "Small Cache, Big Effect: Provable Load Balancing for Randomly Partitioned Cluster Services", *ACM*, 2011

[22] Y. Xianfeng and L. HongTao, "Load Balancing of Virtual Machines in Cloud Computing Environment Using Improved Ant Colony Algorithm", *International Journal of Grid Distribution Computing*, Vol. 8, No.6, pp.19-30, 2015

[23] A. P. Florence and V. Shanthi, "A Load Balancing Model Using Firefly Algorithm In Cloud Computing", *Journal of Computer Science*, vol.10 , Iss.7, pp.1156-1165, 2014

[24] A. Tchernykh, J. M.C. Mendoza, J. E. Pecero, P. Bouvry, D. Kliazovich, "Adaptive Energy Efficient Distributed VoIP Load Balancing in Federated Cloud Infrastructure" *IEEE International Conferece on Cloud Networking*, pp.27-32, 2014

[25] S. Begum, Prashanth C.S.R, "Mathematical Modelling of Joint Routing and Scheduling for an Effective Load Balancing in Cloud", *International Journal of Computer Applications*, Vol.104, No.4, October 2014

[26] S. Aslanzadeh, Z. Chaczko, "Load balancing optimization in cloud computing: Applying Endocrine-particale swarm optimization," *IEEE International Conference on Electro/Information Technology,* pp.165-169, May 2015