RESEARCH ARTICLE OPEN ACCESS

# Ensuring Reliability and High Availability in Cloud by Employing a Fault Tolerance Enabled Load Balancing Algorithm

G.Gayathri [1], N.Prabakaran [2]
Department of Computer Applications
St.Peter's University, Chennai, TamilNadu, India
Department of Computer Applications
St. Joseph College of Information Technology
Songea, Tanzania

**ABSTRACT**
Cloud computing enables a wide range of users to access distributed, scalable, virtualized hardware and/or software infrastructure over the Internet. Load balancing and fault tolerance are the two important issues which guarantees high availability and the reliability of the cloud. Usually availability of the service is ensure by efficient load balancing algorithms. But a system or a service is reliable only when it is able to perform correctly even in a faulty condition. A fault may be a link failure, server crash, etc., This paper proposes a fault tolerance enabled load balancing algorithm which balances the incoming load among various host in the resource pool as well as preserves the fault tolerance property by maintaining redundant copies of the service in various hosts.
***Keywords:-*** cloud computing, Load balancing, fault tolerance, performance.

## I. INTRODUCTION

Cloud computing is a distributed computing paradigm that focuses on providing a wide range of users with distributed access to scalable, virtualized hardware and/or software infrastructure over the internet. Despite this technical definition cloud computing is in essence an economic model for a different way to acquire and manage IT resources. An organization needs to weigh cost, benefits and risks of cloud computing in determining whether to adopt it as an IT strategy. The availability of advance processors and communication technology has resulted the use of interconnected, multiple hosts instead of single high-speed processor which incurs cloud computing.

In cloud computing environment, the random arrival of tasks with random utilization of CPU service time requirements can load a specific resources heavily, while the other resources are idle or are less loaded [2]. Hence, resource control or load balancing is major challenging issue in cloud computing. Load balancing is a methodology to distribute workload across multiple computers, or other resources over the network links to achieve optimal resource utilization, maximize throughput, minimum response time, and avoid overload.

A cloud computing platform dynamically provisions, configures, reconfigures, and de-provisions servers as needed. Servers in the cloud can be physical machines or virtual machines spanned across the network. Thus it utilizes the computing resources (service nodes) on the network to facilitate the execution of complicated tasks that require large-scale computation. Selecting nodes (load balancing) for executing a task in the cloud computing must be considered, and to exploit the effectiveness of the resources, they have to be properly selected according to the properties of the task.

## II. DATA CENTER

In any enterprise data center, the foremost concerns are data integrity and the continued availability of services. Any interruption or compromise in this area can lead to dire consequences, which are often times beyond recovery. This situation can be aggravated in data centers deploying virtual infrastructures. Because multiple virtual servers reside on a single physical server, when the lone physical server goes down, all of the virtual application servers are then put out of service. In data centers hosting a virtual server farm, whenever FT is enabled for a virtual server, a copy of that virtual server is automatically created and run on another physical server. [13]

In today's highly demanding data center environment, a wide range of threats to the enterprise could adversely affect its operation and availability in an instant.

Data centers aggregate all kinds of resources (e.g. data, software, hardware) to provide various services with virtualization technology. The data center, which contains many physical hosts, can be viewed as a resource pool by virtualizing all the physical resources as a whole system. With virtualization technology, many virtual machines (VMs) are created running on the physical hosts of the data center, and theVMs provide various secure and reliable services.

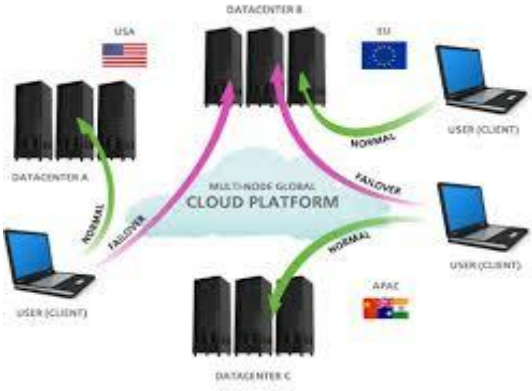


Fig. 1 Data Centers Operating through cloud.

Generally, a physical host holds several VMs, and eachVM can provide a special service for the customer .There are a different number of VMs running on different physical hosts, which result in load unbalancing in the data center and may introduce congestion on individual servers . In order to improve the performance of the data center by maximizing the throughput and minimizing the response time of the system, it is necessary to balance loads among the physical hosts for the cloud environment [1]. There are many physical hosts in a data center, and thus load balancing
withVM migration among the physical hosts plays an important role in providing stable and high-performance services [2–4].

Sometimes several physical hosts may have heavy loads; in other words, there are more than the average number of VMs running on the physical hosts with the result that the services running on it cannot guarantee their requirements. By migrating VMs from the heavy load hosts to the light ones to balance the load among the hosts, the service performance of the data center can be improved [1, 5].

## III. RELATED WORK

There have been many studies on load balancing based on virtualization technology. Based on VM migration, Wilcox [1] proposed a load balancing scheme named modified central scheduler load balancing (MCSLB), which migrates VMs from the heaviest load host to the lightest host. Wang et al. [3] proposed a scheduling algorithm to utilize better executing efficiency and maintain the load balancing of the system by combining opportunistic load balancing and load balance min– min scheduling algorithms. Zhang and Zhang [4] presented a load balancing mechanism based on ant colony and complex network theory in the open cloud computing federation to realize load balancing in the distributed system.

In order to achieve the best load balancing and reduce or avoid dynamic migration, Hu et al. [5] proposed a scheduling strategy on load balancing of VM resources based on a genetic algorithm with the consideration of system variation and historical data. Randles et al. [6] investigated three possible distributed load balancing algorithms inspired by Honeybee Foraging Behavior, Biased Random Sampling and Active Clustering for cloud computing scenarios. By integrating live OS migration into the Xen VM monitor, Clark et al. [7] discussed the procedure of live VM migration, and presented the design, implementation and evaluation of high-performance OS migration built on top of the Xen VMM.

Owing to hardware failure, communication link errors and malicious attack, fault tolerance is one of the most critical issues in distributed systems [8–10]. However, the researches of load balancing discussed above do not take the fault tolerance of the services into account. Different VM migration schemes result in various influences on the fault-tolerant level of the services.

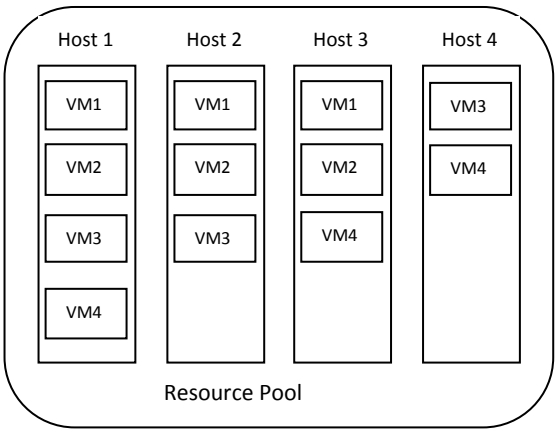


Fig .2 Several Virtual Machines running on different hosts

As shown in Fig. 2, in the cloud computing environment, the data center has many VMs to run the service, and the fault tolerant level of the service is guaranteed by distributing the VMs of the service onto various physical hosts. Fault-tolerant level can be described as: if service i can work normally when ki hosts break down, the fault-tolerant level of service i is defined as ki.[11]. To provide reliable services, the fault-tolerant level should be ensured while migratingVMs to balance loads.

To guarantee the fault-tolerant level of all services provided by the data center while balancing the load based on VM migration among the hosts a Fault Tolerance enabled load balancing algorithm based on VM migration, is proposed in this paper.

## IV. LOAD BALANCING WITH FAULT TOLERANCE

Normally in data centers, to ensure fault tolerance the virtual machine instances are duplicated and made available in different hosts. In cloud computing environment users request for services. As the request gets in the load balancer has to assign the request to the virtual machines running inside the host. As this scenario continues, the hosts are gradually loaded with the jobs. At some point of time the hosts will have different loads. Some maybe heavily loaded others may be lightly loaded.

When a host is heavily loaded, that is to its maximum capacity, due to overload the host itself may crash. If the host crashes all the VMs running in the host becomes unavailable for the users.

To avoid this situation, multiple copies are maintained in various hosts to ensure availability and reliability of the services. But when a particular host is loaded heavily, it is the duty of the load balancing algorithm to move some VMs to lightly loaded hosts. But while moving the VMs the algorithm should ensure the fault tolerant levels of the entire system i.e., if particular host is down for some reason, the redundant VM should take care of the request.

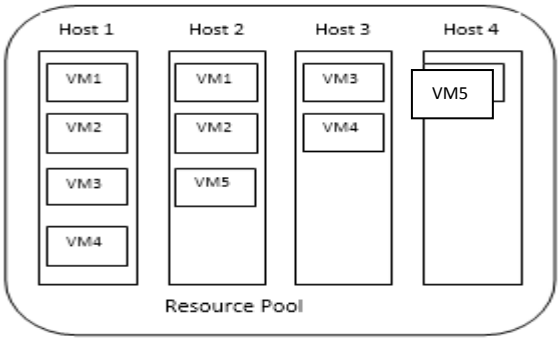


Fig. 3 Scenario in which a host is heavily loaded

In Fig 3, Host 1 is loaded heavily, the load balancer has to choose any one of the lightly loaded host available in the resource pool. The destination host is chosen so that the load is distributed evenly and also the fault tolerance level is also maintained. For example in Fig 2, two copies of a VMs are maintained because if one fails the other will take care of the service.

Now to distribute the load evenly any one the VM must be moved from the source host to destination host. But the destination host is the one which should not run the same kind of VM, because if two copies of the same VM exist in a single host, if that host fails then that particular VM will become unavailable and the fault tolerance of the system is not maintained.

Now the proposed algorithm will find the lightly loaded host which does not run the identical copy of the VM already. This way the fault tolerance level of the system is maintained. In the Fig 3, if VM4 is moved, it should be moved to Host 4, because it is the lightly loaded and also it does not run the VM4 already. VM4 cannot be moved to Host2 since it is already loaded with three VMs and VM4 cannot be moved to Host 3 because it is already running another copy of VM4. Similarly VM2 should be moved to either Host 3 or Host 4 but not to Host 2.

Fig. 4 illustrate the load balanced state and preserving the fault tolerance level of the system also.

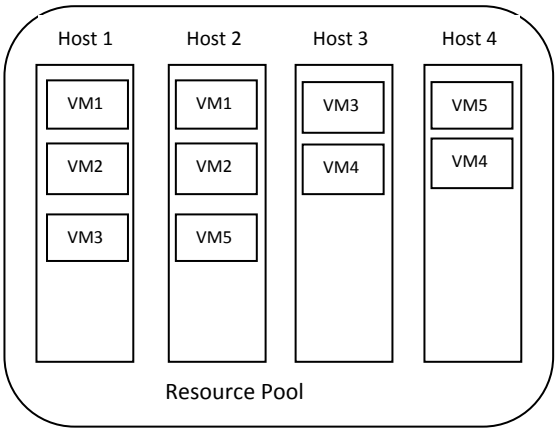


Fig. 4 Load balanced with fault tolerance

## V. FAULT TOLERANCE ENABLED LOAD BALANCING ALGORITHM.

Step1:Maintain the status of all the host ,like number of VMs and its service number.
Step 2: Find the heavily loaded host.
Step 3: if a particular host is heavily loaded
choose the VM that has to be moved.
Step 4: Find the host which is lightly loaded
Step 5: Check whether the particular host is not hosting the VM of same type.
Step 6: if a suitable host is found migrate the VM to the newly found destination host.
Else
do not move the VM from source host.
Step7: Record the changes happened in the status table.

Step 8:If any other VM has to be moved then
Go to Step 3.
Else
Go to Step 2.
Step 9: Monitor the resource pool.

Thus the proposed algorithm helps to balance the load on the servers and also preserves the fault tolerance property of the system.

## VI. FUTURE WORK

The fault tolerance enabled load balancing algorithm can be tested in any simulation tools and its efficiency and performance may be estimated.

## VII. CONCLUSION

In this paper the authors have devised an algorithm for fault tolerance enable load balancing in the cloud environment. This algorithm is devised to enhance the availability and reliability of the cloud to the next level. The working model and the performance of the algorithm will be studied in the future.

## REFERENCES


[1] Wilcox Jr, T.C. (2009) Dynamic load balancing of Virtual machines hosted on Xen. Master's Thesis, Brigham Young University, USA.

[2] Meng, X., Pappas, V. and Zhang, L. (2010) Improving the Scalability of Data Center Networks with Traffic-Aware Virtual Machine Placement. 2010 Proc. IEEE INFOCOM, San Diego,CA, March 15– 19, pp. 1–9. IEEE, NewYork.

[3] Wang, S.,Yan, K., Liao,W. andWang, S. (2010) Towards a Load Balancing in a Three-level Cloud Computing Network. 2010 3rdIEEE Int. Conf. Computer Science and Information Technology(ICCSIT), Chengdu, China, July 9–11, pp. 108–113. IEEE,NewYork.

[4] Zhang, Z. and Zhang, X. (2010) A Load Balancing MechanismBased on Ant Colony and Complex Network Theory in OpenCloud Computing Federation. 2010 2nd Int. Conf. Industrial Mechatronics and Automation (ICIMA), Wuhan, China, May 30–31, pp. 240–243. IEEE, NewYork.

[5] Hu, J., Gu, J., Sun, G. and Zhao, T. (2010) A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment. 2010 3rd Int. Symp. Parallel Architectures, Algorithms and Programming (PAAP), Dalian, China, December 18–20, pp. 89–96. IEEE, NewYork.

[6] Randles, M., Lamb, D. and Taleb-Bendiab, A. (2010) A Comparative Study into Distributed Load Balancing Algorithmsfor Cloud Computing. 2010 IEEE 24th Int. Conf. Advanced Information Networking and Applications Workshops (WAINA),Perth, Australia, April 20–23, pp. 551–556. IEEE, NewYork.

[7] Clark, C., Fraser, K., Hand, S., Hansen, J., Jul, E., Limpach C., Pratt, I. and Warfield, A. (2005) Live Migration of Virtual Machines. Proc. 2nd Conf. Symp. Networked Systems Design & Implementation-Volume 2 (NSDI'05), Berkeley, CA, pp. 273– 286. USENIX Association, Berkeley, CA.

[8] Deng, J., Qiu, M. and Wu, G. (2010) Fault Tolerant Data Collection in Heterogeneous Intelligent Monitoring Networks. 2010 IEEE 5th Int. Conf. Networking, Architecture and Storage (NAS), Macau, China, July 15–17, pp. 13–18. IEEE, NewYork.

[9] Zhu, X., Qin, X. and Qiu, M. (2011) Qos-aware fault-tolerant scheduling for real-time tasks on heterogeneous clusters. IEEE Trans. Comput., 60, 800–812.

[10] Qiu, M., Liu, J., Li, J., Fei, Z., Ming, Z. and Sha, E. (2011) A Novel Energy-Aware Fault Tolerance Mechanism for Wireless Sensor Networks. 2011 IEEE/ACM Int. Conf. Green Computing and Communications

(GreenCom), Chengdu, China,August 4–5, pp. 56–61. IEEE, NewYork.

[11] Lin Yao, Guowei Wu, Jiankang Ren, Yanwei Zhu and Ying Li(2013) Guaranteeing Fault-Tolerant Requirement Load Balancing Scheme Based on VM Migration\

[12] Soumya Ray and Ajanta De Sarkar(2012) Execution analysis of load balancing algorithms in cloud computing environment http://computer.howstuffworks.com/data-centers.htm