Server Consolidation Based Dynamic Load Balancing Approach In Cloud Computing Environment

Shailee M. Majmudar[1], Krunal J. Panchal[2]
[1]PG Student, Department of Computer Engineering, L. J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India
[2]Assistant Professor, Department of Computer Engineering, L. J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India

ABSTRACT

Power efficiency is one of the main issues that will drive the design of data centers, especially of those devoted to provide Cloud computing services. In virtualized data centers, consolidation of Virtual Machines (VMs) on the minimum number of physical servers has recognized as a very efficient approach, as this allows unloaded servers to be switched off or used to accommodate more load, which is clearly a cheaper alternative to buy more resources. The consolidation problem must be solved on multiple dimensions, since in modern data centers CPU is not the only critical resource: depending on the characteristics of the workload other resources. The problem is so complex that centralized and deterministic solutions are practically useless in large data centers with hundreds or thousands of servers.

Key words: *Virtual machine, Data center, Resource management, Cloud services*

1. INTRODUCTION

All main trends in information technology, for example, Cloud Computing and Big Data, are based on large and powerful computing infrastructures. The ever increasing demand for computing resources in companies and resource providers to build large warehouse-sized data centers, which require a significant amount of power to be operated and hence consume a lot of energy. The virtualization paradigm can be exploited to alleviate the problem, as many Virtual Machine (VM) instances can be executed on the same physical server. This enables the consolidation of the workload, which consists in allocating the maximum number of VMs in the minimum number of physical machines. Consolidation allows unneeded servers to be put into a low-power state or switched off (leading to energy saving), or devoted to the execution of incremental workload (leading to savings, thanks to the reduced need for additional servers). Unfortunately, efficient VM consolidation is hindered by the inherent complexity of the problem. Virtualization is an important and core technology for cloud computing. It allows the abstraction of fundamental elements of computing such as hardware, storage and networking. Virtualization technology has helped the cloud data centers to effectively increase resource utilization, reduce electricity costs and ease management complexities. But there are many challenges in providing services with reliability and performance guarantee in such a complex virtualized environment involving server consolidation.

2. BACKGROUND

Load balancing is the major concern in the cloud computing environment. Cloud comprises of many hardware and software resources and managing these will play an important role in executing a client's request. Now a day's clients from different parts of the world are demanding for the various services in a rapid rate. In this present situation the load balancing algorithms built should be very efficient in allocating the request and also ensuring the usage of the resources in an intelligent way so that underutilization of the resources will not occur in the cloud environment. In the present work, a novel VM-assign load balance algorithm is proposed which allocates the incoming requests to the all available virtual machines in an efficient manner.

Load balancing is the process of improving the performance of the system by shifting of workload among the processors. Workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine. Load balancing is done so that every virtual machine in the cloud system does the same amount of work throughout therefore increasing the throughput and minimizing the response time. Load balancing is one of the important factors to heighten the working performance of the cloud service provider. Balancing the load of virtual machines uniformly means that anyone of the available machine is not idle or partially loaded while others are heavily loaded. One of the crucial issue of cloud computing is to the workload dynamically. The benefits of distributing the workload includes increased resource utilization ratio which further leads to enhancing the overall performance thereby achieving maximum client satisfaction.

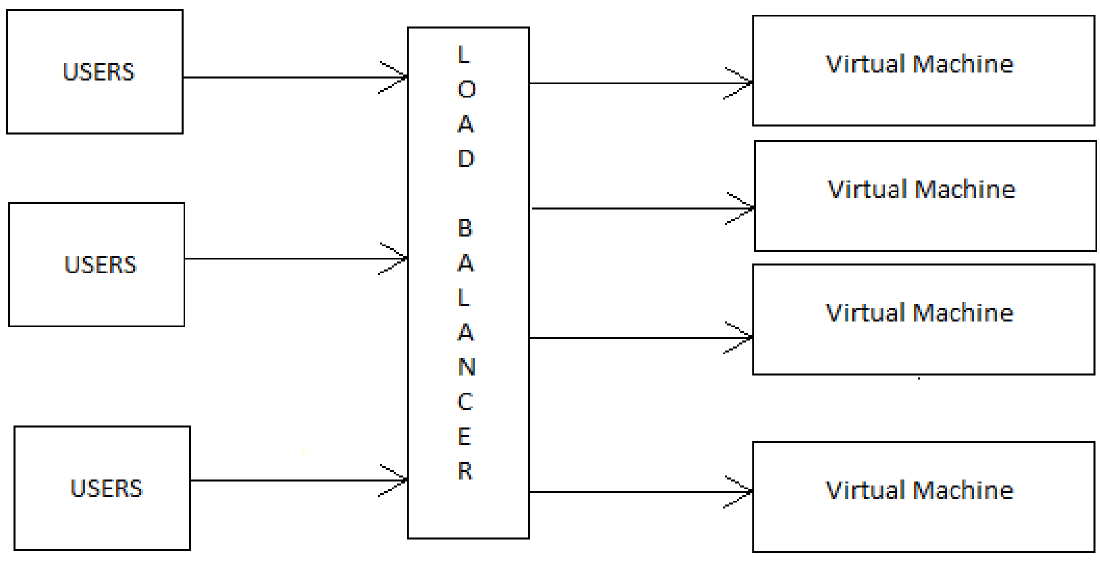


Fig 1.1 general load balancer

3. PROPOSED SYSTEM

There will be designed a novel Auto Load-aware Scale scheme We will describe scale in and scale out strategy based on prediction algorithm. A new proactive technique for auto-scaling of resources that changes the number of resources for the private cloud dynamically based on system load is proposed. The technique that supports both on-demand and advance reservation requests uses machine learning to predict future workload based on past workload.

Work flow :

- Periodically Probing of all current Vm's for checking their load is to be deployed on a central cloud controller.
- Machine learning based Api's will be used for future prediction based on Heuristics i.e past statistics.

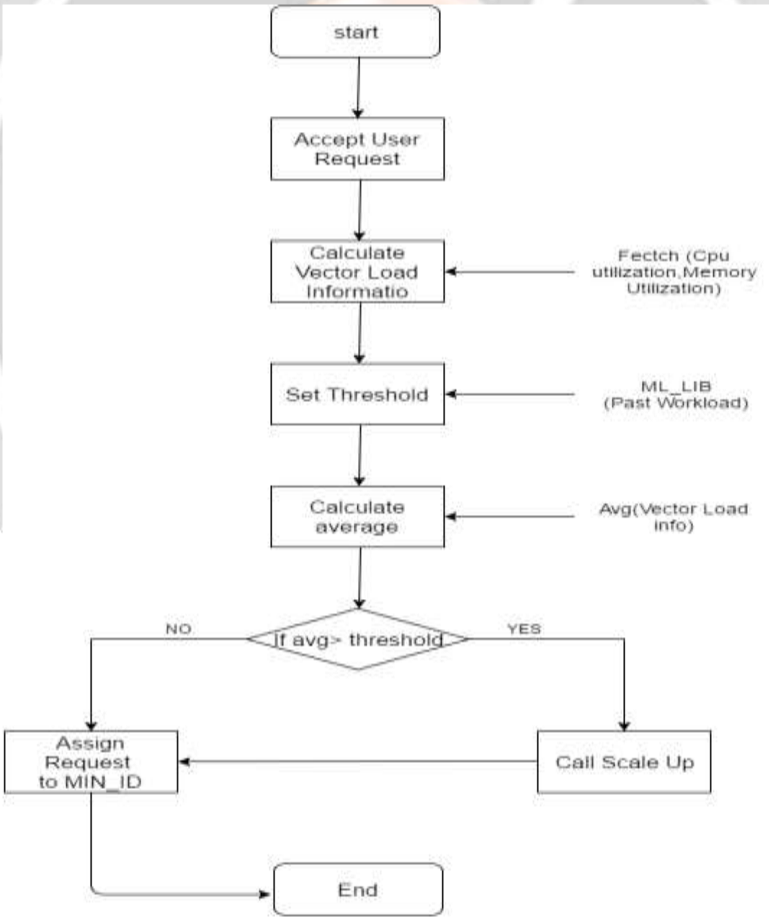
- If statistic indicator is below threshold then auto scaling will be called for launching new Vms.
- Load balancing results into an optimal resource scheduling.

Proposed System Algorithm :

Input[CPU Util , Memory Util]

- Start Procedure
- Repeat For All VM Instances
- Vector_ load _info <> = fetch < CPU_ utill , Mem_ util >
- End For
- Threshold = ML_LIB(Past_ Work_ load)
- Avg <- Average(Vector _load _info <>)
- if Avg (Vector _load_ info) > Threshold
- call Scale _up()
- else
- Map Request to Vm with Minimum loaded
- End Procedure

Flowchart :



4. RESULT ANALYSIS

Here we have shown the result analysis by implementing the proposed system and comparision of existing and proposed system.

```
<terminated> Sirinfo [Java Application] C:\Program Files (x86)\Java\jdk1.8.0_65\bin\javaw.exe (Mar 14, 2016, 9:47:05 AM)
GeoIP Database loaded: GEO-533LITE 20160105 Build 1 Copyright (c) 2016 MaxMind Inc All Rights Reserved
2015-05-03...15:32:36...66.249.64.205...United States...CA
2015-05-03...15:56:04...176.108.188.1...Ukraine...11
2015-05-03...15:32:34...63.111.67.72...United States...null
2015-05-03...15:56:05...101.190.80.116...Australia...02
2015-05-03...15:32:35...198.245.60.28...Canada...QC
2015-05-03...15:32:32...110.55.4.190...Philippines...07
2015-05-03...15:32:33...208.88.224.203...United States...FL
2015-05-03...15:56:00...124.170.132.39...Australia...04
2015-05-03...03:36:54...5.31.151.27...United Arab Emirates...null
2015-05-03...15:32:31...122.170.112.131...India...09
2015-05-03...15:56:02...173.252.88.91...United States...CA
2015-05-03...15:56:07...103.240.32.22...India...09
2015-05-03...15:56:08...203.106.155.175...Malaysia...09
2015-05-03...15:32:38...179.111.136.145...Brazil...null
2015-05-03...15:56:09...90.191.110.104...Estonia...null
```

Fig 4.1 output of data displayed for prediction

```
RunUtilizationFatch (1) [Java Application] C:\Program Files (x86)\Java\jdk1.8.0_65\bin\javaw.exe (Mar 11, 2016, 8:48:30 PM)
i-9778fa4f CPU: 2%
i-9f78fa47 CPU: 2%

i-9778fa4f Memory: 85.98393%
i-9f78fa47 Memory: 74.83913%

i-9778fa4f Response Time: 1826.0 ms
i-9f78fa47 Response Time: 861.0 ms

Calculating utilization_factor
i-9f78fa47 Utilization: 38.419563
i-9778fa4f Utilization: 43.981964


Printing queue....
i-9f78fa47 is in queue
i-9778fa4f is in queue

i-9778fa4f CPU: 2%
i-9f78fa47 CPU: 2%

i-9778fa4f Memory: 85.952515%
i-9f78fa47 Memory: 74.847%

i-9778fa4f Response Time: 2977.0 ms
i-9f78fa47 Response Time: 1092.0 ms

Calculating utilization_factor
i-9f78fa47 Utilization: 38.4235
i-9778fa4f Utilization: 43.976257

Printing queue....
i-9f78fa47 is in queue
i-9778fa4f is in queue
```

Fig 4.2 Output of parameters value fetching repeat after fix time span

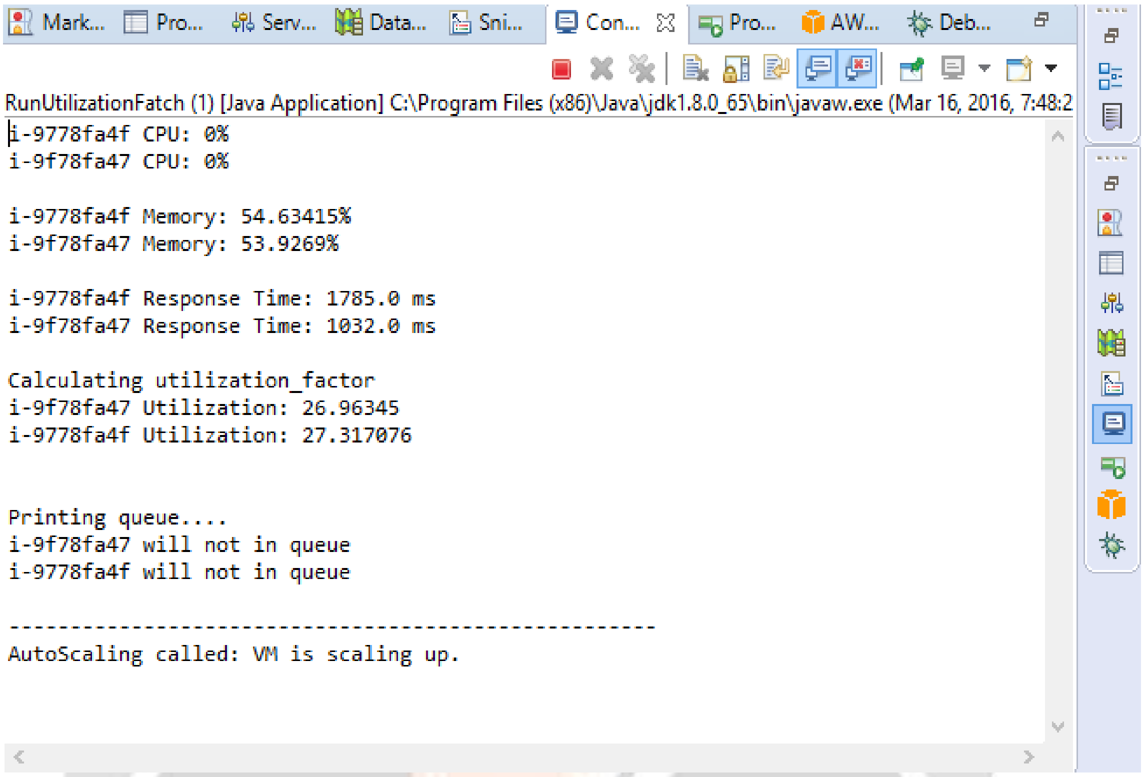


Fig 4.3 Scaling up procedure

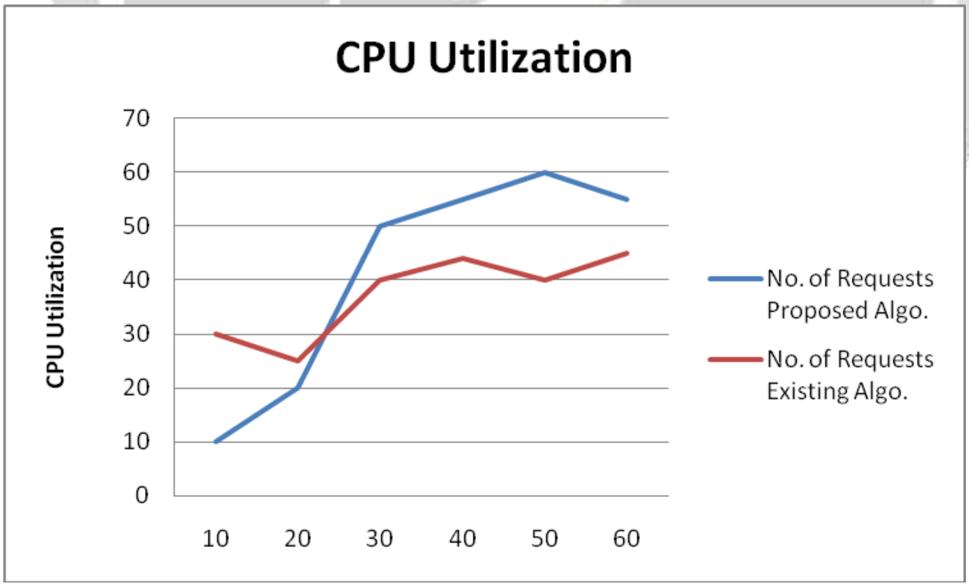


Fig 4.4 CPU Utilization Graph

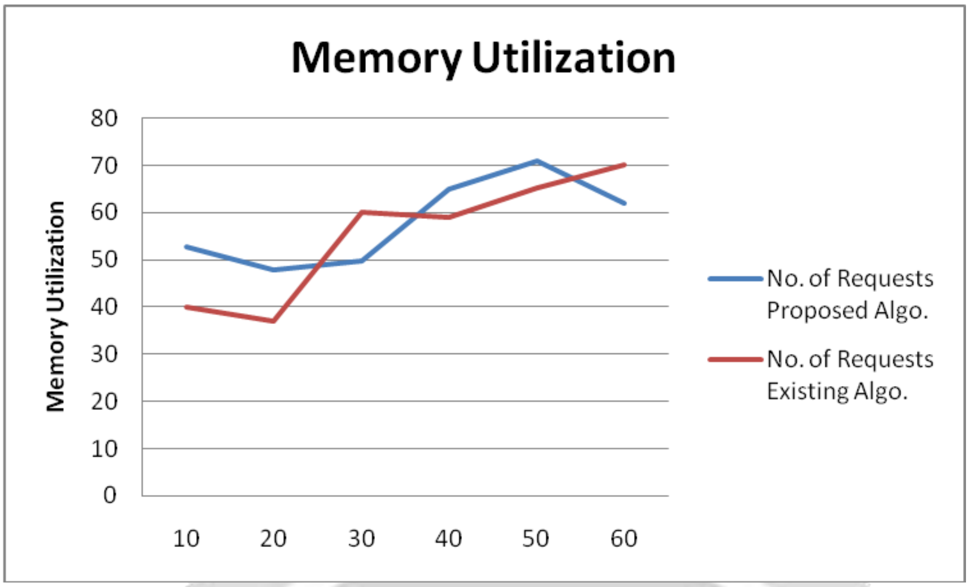


Fig 4.5 Memory Utilization Graph

5. CONCLUSION AND FUTURE WORK

The approach we applied will improve the efficiency of resource utilization by using machine learning technique in cloud computing environment. By following this approach using dynamic threshold policy it can work efficiently for different dynamic conditions in cloud computing. In future we can focus on the other parameters like temperature of the cpu , As it can provide the more dynamic and efficient approach . The placement of the VM is also needed to be considered for dynamicity in the system. More dynamic threshold policies can be applied to existing algorithm.

6. REFERENCES

[1] Shridhar G.Damanal , G. Ram Mahana Reddy, “Optimal Load Balancing in Cloud Computing By Efficient Utilization of Virtual Machines” IEEE Cloud Computing, 2014, pp 1-4,DOI: 10.1109/COMSNETS.2014.6734930

[2] Agraj Sharma, Sateesh K. Peddoju, Response Time Based Load Balancing in Cloud Computing”, IEEE Control, Instrumentation, Communication and Computational Technologies , 2014, pp 1-4, DOI: 10.1109/NOMS.2014.6838340

[3] Jie Bao, Zhihui Lu, Jie Wu, Shiyong Zhang, Yiping Zhong, “Implementing a Novel Load-aware Auto Scale Scheme for Private Cloud Resource Management Platform”, IEEE Cloud Computing,2014,pp 1287-1293,DOI: 10.1109/ICCICCT.2014.6993159

[4] Mohammadreza Mesbahi, Amir Masoud Rahmani, “Cloud Light Weight: a New Solution for Load Balancing in Cloud Computing”, IEEE Data Science and Engineering , 2014, pp 44-53,978-1-4799-5461

[5] Velagapudi Sreenivas , Prathap.M , Mohammed Kemae, “load balancing techniques: major challenge in cloud computing –a Systematic review” IEEE Electronics and Communication System,2014, pp1-6DOI:10.1109/ESC.2014.6892523

[6]M. Randles, D. Lamb, and A. Bendiab, “A comparative Study into distributed load balancing algorithms for cloud computing”, IEEE 24thInternational Conference on Advanced Information Networking andApplications Workshops (WAINA), pp.551-556, April 2011, DOI : 10.1109/WAINA.2010.85

[7] W. Bhathiya, "CloudAnalyst a CloudSim-based tool for modelling and analysis of large scale cloud computing environments", MEDC Project, Cloud Computing and Distributed Systems Laboratory, University of Melbourne, Australia, pp. 1-44, June 2012, DOI :10.1109/ESNA.2012.34

[8] Z. Zhang and Xu. Zhang "A Load balancing mechanism based on ant colony and complex network theory in open cloud computing federation", 2nd International Conference on Industrial Mechatronics and Automation (ICIMA), Wuhan, China, vol. 2, pp.240-243, May 2013, DOI :10.1109/NATCSE.2012.6745

[9]M. Hines and K. Gopalan, "Post-copy based live virtual machine migration using adaptive pre-paging and dynamic self-ballooning", inProceedings of the ACM/Usenix International Conference on Virtual Execution Environments (VEE'09), pp.51-60, March 2012, DOI :10.1209/ESSD.2013.564

[10]http://www.webopedia.com/TERM/C/cloud_computing.html, 24 Nov 2015 22:20.

[11] Cloud Computing Principles and Paradigms , Rajkumar Buyya, James Broberg , Andrrzej Goscinski , New delhi 20111.pp 135