Load Balancing Method for Infrastructure as a Service (IaaS) In Cloud Computing: Survey

Nikhil S. Dharane
UG student, Dept of Computer Engineering, SKNCOE, Pune, India

Abhijeet S. Kulkarni
UG student, Dept of Computer Engineering, SKNCOE, Pune, India

Akshay U. Bhawthankar
UG student, Dept of Computer Engineering, SKNCOE, Pune, India

A.A. Deshmukh
Assistant Professor,
Dept of Computer Engineering,
SKNCOE, Pune, India

ABSTRACT
The cloud is becoming an important service in internet computing. Infrastructure as a Service provides on-demand virtual machines to users. Load balancing plays an important role in the deployment of virtual machines onto physical hosts. Resource requirement of virtual machine is hard to predict. Thus to deal with this issue, lots of load balancing methods are present. So in this paper we discuss these load balancing methods to provide overview of latest approaches in IaaS. Different load balancing methods have different parameters. So, we compared them on the basis of parameters used for their method. We have proposed our own load balancing method after comparing all available methods. In our proposed method, we first use load forecasting to know load on hosts. Then on the basis of threshold value we have to migrate VMs to another host. If load is below or higher then we have to shut down or migrate VM respectively. So, by using load forecasting we are trying to overcome drawbacks of existing load balancing methods.

KEYWORDS
Load forecasting, Exponential smoothing, Threshold, Migration

1. INTRODUCTION
Cloud computing is recent technology that deals with distribution of resources and services. Services are divided as software services (SaaS), physical resources (PaaS), infrastructure services (IaaS).

IaaS (Infrastructure as a Service), where infrastructure or actual hardware is provided to customers who are responsible to install operating systems and necessary software as per their usage. IaaS cloud is usually provided to users in the form of Virtual Machines (VMs), such as Amazon EC2. In an IaaS cloud, users can apply VMs on-demand to deploy and run their applications also provide services to their clients which will be helpful to clients. From the perspective of users, this way of applying and using resources does not only save the cost of providing services, but also improves the reliability.[1]

IaaS has major issues like resource management, network infrastructure, virtualization, data management etc. IaaS provides benefits like: scalability, QoS, reduction in overheads, cost effectiveness. Different types of resources like physical and logical are provided in IaaS. Physical resources includes CPU, memory, storage. Logical resources includes operating system, energy.[2]

In IaaS cloud, there are physical servers with a large number of virtual machines. These virtual machines are hosted with many heterogeneous applications. In order to optimize the utilization of computing resources and also saving energy consumption of cloud data centers, the applications running on the virtual machines will be migrated either to the same server or to another physical or virtual server. Identifying when it is best to migrate an application in a virtual machine has a direct impact on resource optimization. Performance optimization can be best achieved by an efficiently monitoring the utilization of computing resources. So, we need a comprehensive intelligent monitoring agent to analyze the performances of virtual machines.[3]

Nowadays, cloud computing is focusing on how efficiently infrastructure is made available and available resources are used. Load balancing is main factor in cloud computing. Load balancing means distribute available workload across multiple nodes to ensure that no resource is underutilized or overwhelmed. So good load balancer is required to adapt changing environment strategies and types of tasks.[4]

Live VM migration is used for load balancing. It is used to transfer active VM from one physical host to another without disrupting the VM. VM migration achieves load balancing. [5]

2. LITERATURE SURVEY

2.1 A Model Based Load-Balancing Method in IaaS Cloud [1]:
Zhenzhong Zhang, Limin Xiao, Yuan Tao, Ji Tian, Shouxin Wang, Hua Liu et al algorithm for load balancing. They proposed an algorithm which first forecasts load and estimates resource requirements of virtual machines. A scalable framework for load-balancing which uses their resource requirement forecasting model. Experiments show that this method can accurately estimate the resource requirements of virtual machines, and work well in load-balancing framework. CPU utilization and memory usage are taken as parameters. But it does not provide any experiments which will be done on other parameters.

2.2 Agent based resource monitoring system in IaaS [3]:
A. Meera, S.Swamynathan et al one algorithm. It is agent based resource monitoring algorithm. Agents are used to perform some specific tasks. Agents collect information

about VMs. This information is domain specific. Individual packets of every VM are checked by agent instead of whole cloud. It uses CPU performance, memory usage as checking parameters for VMs. But this algorithm has some SLA issues.

2.3 Genetic algorithm based load balancing strategy for cloud computing [4]:

Kousik Dasgupta, Brototi Mandal, Paramartha Dutta, Jyotsna Kumar Mondal, Santanu Dam et al genetic algorithm for load balancing. In this crossover, mutation and chromosome concepts are used. Chromosome is optimal solution. Mutation value is used to decide which chromosome is optimal. No of instructions and processing unit are used as parameters. This algorithm considers these values and solution for load balancing. But this algorithm uses very simple approach. Only same values of crossover and mutation are taken.

2.4 Task based system load balancing in cloud computing using particle swarm optimization [5]:

Fahimeh Ramezani, Jie Lu and Farookh Hussain developed task based system for load balancing. It uses particle swarm optimization technique. Instead of migrating whole VM this system will migrate extra tasks. VM does not loss information. CPU, memory, bandwidth, hard disk usage are taken as parameters to find different results. This algorithm is applicable to simple task scheduling optimization. So more detailed algorithm is required for larger systems where load balancing is required efficiently.

2.5 Trust and Reliability based Load balancing Algorithm for Cloud IaaS [6]:

Punit Gupta, Mayank Kumar Goyal, Prakash Kumar et al a suitable trust model based on the existing model that is suitable for trust value management for the cloud IaaS parameters. Based on the achieved trust values, a suitable load balancing algorithm is proposed for better distribution of load which further enhance the QOS of services being provided to the users. Other algorithms do not consider the property of VMM but it has not taken into consideration the properties of a VMM in a datacenter. So a trust management model is developed to overcome this problem, by taking into consideration VMM characteristics which vary from datacenter to datacenter. Then these trust value are been used by load balancing algorithm proposed to improve the QOS provided to the user and better utilization of resources. Trust based model uses time parameters only. According to that trust values are calculated. So that VM can be categorized into trusted and un trusted VMs. It shows high efficiency when compared with other non trust algorithms.

2.6 Load Balancing in Public Cloud [7]:

Shrikant M. Lanjewar, Susmit S. Surwade, Sachin P. Patil, Pratik S. Ghumatkar, Prof Y.B. GURAV et al algorithm for load balancing in public cloud. This model divides the public cloud into several cloud partitions. When the environment is very large and complex, these divisions simplify the load balancing. The cloud has a main controller that chooses the suitable partitions for arriving jobs while the balancer for each cloud partition chooses the best load balancing strategy. The load balancing strategy is based on the cloud partitioning concept. After creating the cloud partitions, the load balancing then starts. When a job arrives at the system, with the main controller deciding which cloud partition should receive the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, this partitioning can be accomplished locally. If the cloud partition load status is not normal, this job should be transferred to another partition. Load balancing model uses static and dynamic parameters i.e. CPU, Memory respectively. It initialize parameters and checks load degree.

2.7 Load Balancing in Cloud Computing Using Modified Throttled Algorithm [8]:

Shridhar G.Domanal and G.Ram Mohana Reddy et al an efficient approach to handle the load at servers by considering both availability of VMs for a given request and uniform load sharing among the VMs for the number of requests served. The work aimed at efficient method for load balancing, depicted from its two different objectives. One being the response time required to serve the requests and other being the distribution of load among the existing VMs. This algorithm is efficient than round-robin and throttled algorithm. Distribution of load among the virtual machines in Round-Robin algorithm was nearly uniform, but was found less efficient considering response time. Throttled algorithm with better response time than Round Robin failed to distribute load uniformly, overloading initial VMs and leaving others underutilized. Proposed algorithm distributes load nearly uniform among VMs, with improved response time compared to existing algorithms. Simulation results have demonstrated that the proposed algorithm has distributed the load uniformly among virtual machines.

2.8 Analysis of Issues with Load Balancing Algorithms in Hosted (Cloud) Environments [9]:

Branko Radojevic, Mario Žagar presented analysis of detected issues in available load balancing algorithms and introduced new algorithm. The new algorithm communicates with parts of our computer for end user experience in order to be able to influence load balancing decisions or reactively change decision in handling critical situations. Central Load Balancing Decision Module (or CLBDM for short) interacts (monitor) with all parts of our computer system, including load balancers and application servers. Then, based on the collected information and internal computation, CLBDM will impact forwarding decisions on the load balancers.

2.9 Improving Resource Utilization Using QoS Based Load Balancing Algorithm For Multiple Workflows In IaaS Cloud Computing Environment [10]:

L. Shakkeera, Latha Tamilselvan and Mohamed Imran et al algorithm for load balancing in IaaS cloud. They have discussed QoS based load balancing mechanism. Optimized load balancing algorithm aims to utilize virtual cloud resources efficiently. In this model, they have created web

application with many modules. These modules are tasks which are submitted to load balancing server. The load balancer redirect tasks to corresponding virtual machine. If the size of database inside the machine exceeds then load balancing uses other virtual machine. Parameters used for this model are cost, average execution times, throughput, CPU usage, disk space, memory usage, network transmission and reception rate, resource utilization rate and scheduling success rate for the number of virtual machines. It improves scalability of resources using load balancing techniques. They have considered only independent tasks of virtual cloud.

2.10 RIAL: Resource Intensity Aware Load Balancing in Clouds [11]:

Liuhua Chen, Haiying Shen, Karan Sapra et al a new method of load balancing. Resource Intensity Aware load balancing system is used for each physical machine. For each physical machine, RIAL dynamically assigns weights to different resources according to their usage in physical machine instead of assigning it statically. It reduces cost and future load imbalance. Frequently communicating VMs are kept in same PM. This algorithm migrates VMs from overloaded PMs to lightly loaded PMs. It is distinguished by its resource weight determination based on resource intensity. RIAL takes into account the communication dependencies between VMs in order to reduce the communication between VMs after migration, and also tries to minimize the VM performance degradation when selecting destination PMs. RIAL is found more superior than other load balancing algorithms.

2.11 L3B: Low Level Load Balancer in the Cloud [12]:

Monika Simjanoska, Sasko Ristov, Goran Velkoski, and Marjan Gusev et al new low level load balancing in IaaS cloud. This method preserves the cloud's elasticity by dynamic activation of cloud resources and load balancing the traffic over the resources on low network level. In addition, L3B tends to load the instances in the region where they provide maximum performance. L3B improves the overall cloud performance, reduces power consumption and customers cost for renting cloud resources etc. L3B generates additional latency in delivering the request and response packets in the direction from client to server and vice versa, it provides several benefits especially if the server is hosted in VM instances. But if the central node fails, it will result in L3B decline. L3B is the fact that all L3B modules, agents and repositories require additional hardware resources, i.e. computing, memory and storage capacity. L3B in the cloud will improve the performance of the client-server model, but we also determine the condition how it can be achieved

3. COMPARISON OF VARIOUS LOAD BALANCING METHODS

Table 1. Comparison of various load balancing methods

	Model based..... [1]	Agent based..... [3]	Genetic algorithm. ..[4]	Task based system..... [5]	Trust and reality..... [6]	Load balancing..... [7]	Load balncing in cloud ...[8]	Analysis of issues. ..... [9]	Improving resource..... [10]	RIAL [11]	L3B [12]
CPU usage	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Memory usage	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Network I/O	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes
Disk I/O	Yes	No	No	No	No	No	Yes	Yes	Yes	No	Yes
MIPS	Yes	No	No	No	Yes	No	No	No	No	No	No

By comparing all available methods of load balancing we found that model based load balancing method is most suitable for balancing load efficiently [1]. Model based method [1] uses all required parameters and has load forecasting feature.

4. PROPOSED WORK

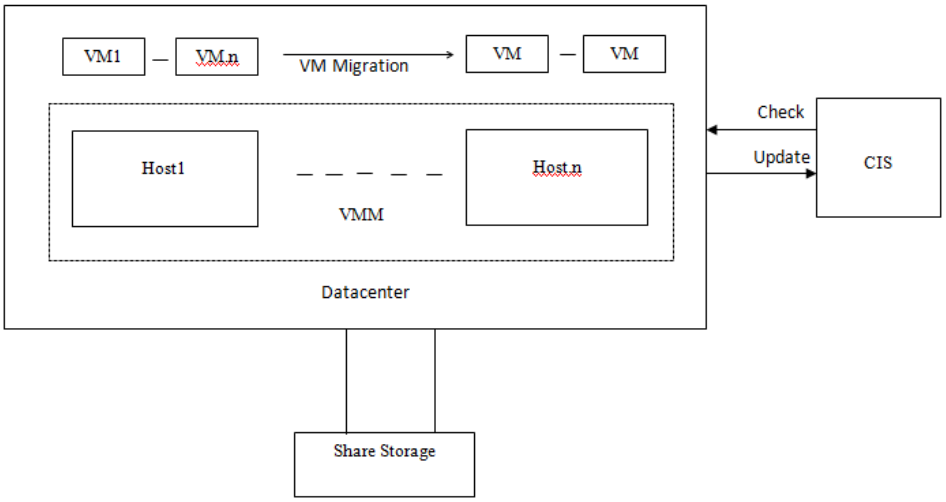


Fig 1. System architecture

Our model will have two cases one for VM starting and other for load is increasing or falling below threshold. Fig 1 shows system architecture. First we get the next several hours prediction load of the starting VM. Then, we select n hosts that have lower load. Then, one suitable host will be select from these n hosts for the VM running on. The principle for choose this host is that, if this VM running on the host, the load-balancing factor will be the minimum in next several hours. Prediction of load can be done by using triple exponential smoothing [13] formula

$$R_t = \alpha * (y_t - S_{t-L}) + (1-\alpha) * (R_{t-1} + G_{t-1}) \quad (1)$$

$0 < \alpha < 1$

$$G_t = \beta * (S_t - S_{t-1}) + (1-\beta) * G_{t-1} \quad (2)$$

$0 < \beta < 1$

$$S_t = \gamma * (y_t - S_t) + (1-\gamma) * S_{t-L} \quad (3)$$

$0 < \gamma < 1$

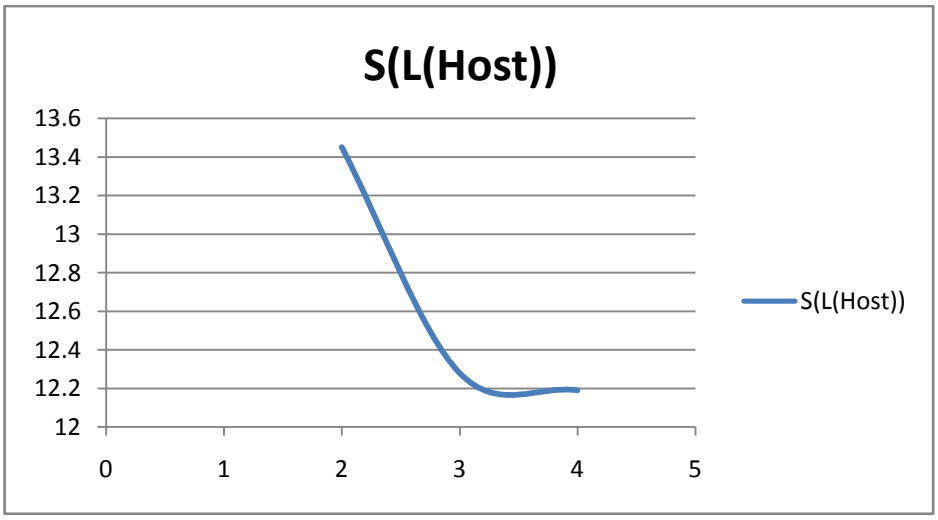
α, β, γ are constants

R_t be the estimate of the deseasonalized level.
G_t be the estimate of the trend
S_t be the estimate of seasonal component (seasonal index)
L be period

For the second case which the load of some hosts exceeds or falls below the threshold, several hosts with lowest load and highest load will be selected, and some suitable VMs on high load hosts would be chosen to migration to the hosts with lower load. In the extreme case, the load of every host is below the threshold, we need migrate the VMs of some hosts and shutdown these hosts. Conversely, if the load of every host is higher than the threshold, new hosts would start. Load balancing factor is as follows:

$$S(L(Host_x)) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(L_i(Host_x) - L(Host_x))^{2\backslash}}$$

n = no of hosts that VM could running on
L = load
S = Load balancing factor



Graph 1: Number of hosts Vs Load balancing factor

x axis = No of hosts y axis = load balancing factor (CPU utilization)

This graph which is simulated on Matlab 7.0 shows that when number of hosts increases then load balancing factor decreases. So load is perfectly balanced between all available hosts.

5. CONCLUSION

This paper is based on cloud computing technology which has a very vast potential. One of the major issues of cloud computing is load balancing because overloading of a system may lead to poor performance which can make the technology unsuccessful. So there is always a requirement of efficient load balancing algorithm for efficient utilization of resources. Our paper focuses on the various load balancing algorithms, their applicability and some limitations in cloud computing environment. Also we have mentioned our proposed work by comparing all available load balancing mechanisms. In future, we will like to add more efficient load forecasting technique which will predict load of our nodes. We will use day time and night time loads as reference to predict loads of node. Also we will add loads of disk I/O, memory utilization.

6. REFERENCES

[1] Zhenzhong Zhang, Limin Xiao, Yuan Tao, Ji Tian, Shouxin Wang, Hua Liu ”A Model Based Load Balancing Method In IaaS Cloud “,2013 42nd international conference on parallel processing

[2] Sunilkumar S. Manvi , Gopal Krishna Shyam "Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey" Elsevier Journal of Network and Computer Applications

[3] A.Meera, S.Swamynathan "Agent based Resource Monitoring system in IaaS Cloud Environment", 2013 International Conference On Computational Intelligence: Modeling Techniques and Applications

[4] Kousik Dasgupta, Brototi Mandal, Paramartha Dutta, Jyotsna Kumar Mondal, Santanu Dam "Genetic algorithm based load balancing strategy for cloud computing" 2013 International Conference On Computational Intelligence: Modeling Techniques and Applications

[5] Fahimeh Ramezani, Jie Lu and Farookh Hussain "Task Based System Load Balancing Approach in Cloud Environments " 2014, Springer-Verlag Berlin Heidelberg

[6] Punit Gupta, Mayank Kumar Goyal, Prakash Kumar " Trust and Reliability based Load Balancing Algorithm for Cloud IaaS " 2013 3rd IEEE International Advance Computing Conference (IACC)

[7] Shrikant M. Lanjewar, Susmit S. Surwade, Sachin P. Patil, Pratik S. Ghumatkar, Prof Y.B. Gurav "Load Balancing In Public Cloud" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI (Feb. 2014)

[8] Shridhar G.Domanal and G.Ram Mohana Reddy " Load Balancing in Cloud Computing Using Modified Throttled Algorithm".

[9] Branko Radojević, Mario Žagar " Analysis of Issues with Load Balancing Algorithms in Hosted (Cloud) Environments " 2011 MIPRO

[10] L. Shakkeera, Latha Tamilselvan and Mohamed Imran " Improving Resource Utilization Using QoS Based Load Balancing Algorithm For Multiple Workflows In IaaS Cloud Computing Environment" June 2013, ICTACT, Volume 04, Issue 02

[11] Liuhua Chen, Haiying Shen, Karan Spra "RIAL: Resource Intensity Aware Load Balancing in Clouds " ,2014, IEEE Conference on Computer Communications

[12] Monika Simjanoska, Sasko Ristov, Goran Velkoski, and Marjan Gusev " L3B: Low Level Load Balancer in the Cloud " 2013, IEEE, EuroCon 2013

[13] Prajakta S. Kalekar , Kanwal Rekhi School Of Information Technology "Time Series Forecasting using Holt Winters Exponential Smoothing"