A Survey on and Performance Analysis of Load Balancing Algorithms using Meta Heuristics approach in Public Cloud-Service Provider's Perspective

R. Ramya
Department of Computer Science and Engineering
PSNA College of Engineering and Technology,
Dindigul, India

S. Puspalatha
Department of Computer Science and Engineering
PSNA College of Engineering and Technology,
Dindigul, India

T. Hemalatha
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology,
Dindigul, India

M. Bhuvana
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology,
Dindigul, India

***Abstract*— Cloud Computing provides higher level services and shared pools of configurable resources with minimal management effort over the internet. Cloud Computing achieves coherence and economic of scale by means of shared resources. The growth of cloud computing depends on the high availability of low cost computers, storage devices and networks as well as hardware virtualization and utility computing. Service Level Agreement (SLA) is a service commitment tool between a cloud service provider and a cloud customer. Using SLA in cloud computing, reduce the chances of disappointing the cloud customer and manage the expectation between the service provider and the customer. The challenge is, it is very difficult to handle all request by the cloud providers at a time during peak hours and to keep up SLA. So when there is an uneven request arrival pattern, the cloud resources may either be underutilized or over-utilized. In order to balance the load, the load balancer plays the major role in cloud computing. The load balancing algorithm equalizes the workload and computing resources in a cloud computing environment. It allows an organization to manage their workload demands by allocating resources among multiple servers and thus it minimizes the response time, minimize the waiting time, maximize the throughput time, maximize the resource utilization and minimize the communication delays of the server by meeting the SLA. The major goal of this study is to review both the existing static and dynamic load balancing algorithms proposed till now and to design and implement a Load balancer that uses a Meta Heuristics approach – Ant Colony Optimization technique to perform balancing the load so that the SLA is met evenly without any issues. Each of the load balancing algorithms is compared with other algorithms theoretically and experimentally one of it is tested with the proposed system using AWS Cloud PaaS (Platform as a Service).**

Keywords*— *AWS PaaS, Cloud Computing, Load Balancing, Service Level Agreement, Static and Dynamic Load Balancing.

I. INTRODUCTION

The cloud computing is the practice of accessing, storing, managing and processing the data from the remote server hosted on the internet, instead of using a local server or a personal computer. Depend on the sharing of resources in order to achieve consistency and economies of scale. Many organization expands more on the resource in computer infrastructure and maintenance, and so the cloud computing aims are to focus on the organization core business. Cloud providers often use the "pay-as-you-go" model.

A. Essential Characteristics

The five essential cloud characteristics are,

- On-demand self-service is a service provided by the cloud provider. Whenever the user demands the resources, the cloud service provides the resources to the cloud user. The cloud computing provides computing power, networks and software in a flexible way to the cloud user.
- In broad network access**,** all the cloud resources are accessed via the internet through a standard mechanism like mobile phone, tablets, computers, workstation etc.
- In resource pooling, using the multi-tenant model, the cloud providers resources are pooled into a cloud. When the user demands the resource, different virtual resources are assigned and reassigned dynamically. The resources may be a network, processing memory, storage, network bandwidth.
- Measured service is defined as the resource usage are controlled and optimized by the cloud system automatically by leveraging the metering capability to the type of services. Measures and monitors the - provision of resources for effective service. It provides transparency to both the cloud user and also a cloud provider.
- Rapid elasticity is the ability to provide scalable service. The cloud resources can be elastically provisioned and released as per the demands of the user. In the cloud user point of view, the resources in the cloud are appeared to be unlimited and it can be appropriate at any time.

B. Cloud services models

There are three types of the cloud computing service model. They are

- Software as a Service (SaaS) is defined as the ability of the cloud user to use only the cloud providers application running on the cloud infrastructure. The application can be accessed through a web browser or any program interface. The cloud users are not allowed to manage or control the cloud infrastructure or even application capabilities. Example of SaaS is Google spreadsheet.
- Platform as a Service (PaaS) is defined as the ability of the cloud user to deploy the application on the cloud infrastructure using cloud providers programming language, tools, services, and libraries. The cloud users are allowed to control only the deployed application and configuration setting of the application. Example of PaaS is Google App Engine.
- Infrastructure as a Service (IaaS) is defined as the ability of the cloud user to use the computing resource like processing, storage, network etc. the cloud user is allowed to deploy and run the software. Example of IaaS is Google Compute Engine.

C. Cloud Deployment Model

There are four types of deployment model, they are,

- Private cloud is defined as the cloud infrastructure is solely operated for a single organization. Example of private cloud is Windows serve "Hyper V"
- Public cloud is defined as the cloud resources are available to anyone who wants to use or purchase them. Example of a public cloud is Google doc, spreadsheet.
- Community cloud provides the cloud resources for a limited number of clients or organization for a specific community. Example of a community cloud is Salesforce.com.
- Hybrid cloud is defined as the cloud infrastructure is the combination of two or more clouds. Example of hybrid cloud is Cloud Bursting for load balancing between clouds.

D. Cloud Computing in Load Balancing

Load balancing is the process of distributing multiple loads and resources to multiple servers in a cloud environment. Cloud load balancing achieves high performance in the lower cost than the traditional load balancing technology. And also load balancing achieves high scalability and high availability of resources. The workload is divided into many servers or computing resource to enable a better cloud resource utilization in which no single node is overloaded. In order to distribute traffic across servers, load balancing is used in data center networks. Using load balancing, the cloud users can efficiently use the network bandwidth and also reduce the provision costs.

Advantages of Load Balancing

There are few advantages of load balancing and some of them are

- Increased scalability
- Redundancy
- Reduced downtime and increased performance
- Increased flexibility.

Disadvantage of Load Balancing

Some of the cons of load balancing are,

- Require additional configurations to maintain the connection between the client and the server.
- Hardware-based load balancer costs more.

II. LITERATURE SURVEY

To achieve better load balancing, researchers developed many new load balancing algorithms to solve the complexity of cloud computing and there is a number of load balancing algorithm. Thus, this section describes the related work of load balancing algorithms and it is obtained in two ways: Static load balancing algorithms and dynamic load balancing algorithm. Fig. 1. represents the types of load balancing algorithms.

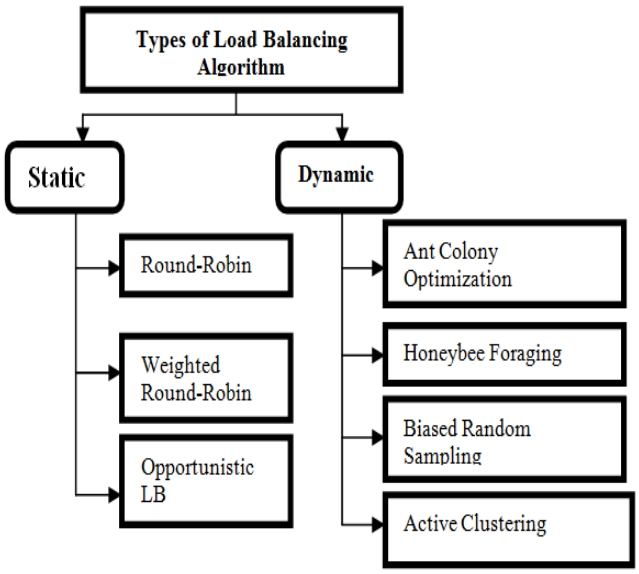

Fig. 1. Types of Load Balancing Algorithms

A. Static Load Balancing

The static load balancing algorithm distributes the load to server equivalently by getting the prior knowledge of the application[9]. These algorithm does not consider the current state of the node. The algorithm work based on the previous information[8]. The system which has less variation in load suits with static load balancing algorithm. One of the disadvantages of static load balancing algorithms is that the task is not supposed to shift to another machine during execution.

B. *Round Robin Algorithm*

In this algorithm[6], [3] time slice mechanism is used to process the data. Generally, the round-robin algorithm randomly selects the first node and then allocates jobs to another node in a round robin fashion.

The tasks are assigned to the processor in a circular order. This algorithm equally distributes the load to all nodes. This type of scheduling in a cloud is similar to the process scheduling in round robin fashion. The algorithm starts with a node and moves to the next node, then a VM is assigned to that node. This is repeated until the scheduler returns to the first node again.

C. *Weighted Round Robin Algorithm*

The weighted round robin[2] is developed to overcome the crucial problems in the round robin algorithm. In this algorithm [1], the jobs are distributed to the server on the basis of weight assigned to each server. A larger value is assigned to the processor with greater capacities. It receives the appropriate number of request on the basis of a value assigned to the node.

D. *Opportunistic Load Balancing Algorithm(OLB)*

OLB[6] provides better execution efficiency to balance the load. The load balancing Min-Min algorithm minimize the execution time of each task and OLB scheduling balance the load by making every node in working state. OLB attempts to make each node in a busy state. This scheduling algorithm first allocates the task and divides the task into subtasks of three- level cloud computing network and solves the workload in the least time.

III. DYNAMIC LOAD BALANCING ALGORITHM

The dynamic load balancing algorithm searches for the lightest server in the whole network and then it balances the tasks. In here, the workloads are distributed among different computing server by searching lightest server and process in a run time[8]. The dynamic algorithm[3] is considered as a complex algorithm but it results in better fault tolerance and overall performance level is high. Each job is inserted into the queue through the queue manager. At any time if a node is having a high value, then it transfers to it the lightest server.

A. *Ant colony optimization(ACO)*

ACO[7] is the meta-heuristics algorithm to solve the combinatorial problems by mimicking the behavior of ants. The aim is to find the optimal path and colonies of ants based on their behavior. When the request is sent, this algorithm records the data by checking whether the data is over or underutilized by visiting the node one by one in a forwarding direction. The solution is continuously updated rather than updating their own result. ACO[3] algorithm produces the optimal solution for any number of jobs and machine that are used in it. All nodes in the graph are connected to one another and ants are randomly moved to find the optimal solution.

B. *Honey Bee Foraging*

The main idea of this algorithm[7] is followed by the behavior honeybees. It balances the load among heterogeneous nodes in the cloud using a decentralized load balancing method. The two types of honeybees are finders and reapers. The finder honeybees find the honey source from the outside of the honeycomb.

Then the quality and quantity of honey available are determined in the honeycomb. The reapers then go outside of the honeycomb and reap the honey and then it returns to the beehive to indicate the remaining food left in there. Similar to that case, heavily loaded node task is removed and assigned it to the lightly loaded node by considering its priority.

C. *Biased Random Sampling*

In this algorithm[7], all the network are represented in the form of the virtual graph. So in a graph, each server is considered as the vertex and the in degree represents the available free resources of the node in the graph.

The load balancer gets prior information from the in-degree and then it allocates the jobs to the node. The in-degree is decreased when the job is allocated and increased when the job is executed. To add and delete a process, the random sampling technique is used.

All the processes have a certain threshold value, which indicates the maximum traversal of the node. The traversal length is known as walk length. When the load balancer receives the request, it then compares the current walk length with the threshold value and then it allocates the request to the appropriate node.

D. *Active Clustering*

Active clustering [1] is the extended version of random sampling. it works on the basis of clustering concept which group the similar nodes and based on the grouped node it works. This processing of grouping is based on the concept of matchmaker node.

A matchmaker node is a node which initiates the process and selects another process to work. The matchmaker node connects to the immediate neighbor node as an initial node and then gets disconnected. This process is done repetitively to balance the load. Table I. compares the types of load balancing algorithms discussed.

TABLE I. COMPARISON OF LOAD BALANICNG ALGORITHMS

S.No	Type.	Load Balancing Algorithms	Merits	Demerits
1	STATIC	Round robin algorithm[6]	Utilize all resources in a balanced manner. Ensure fairness to all allocated Nodes	A non-uniform of workload distribution is not suitable Migration time is high.
2		Weighted Round Robin Algorithm[1],[2]	Proper load balancing is done	Not preferred since the prediction of execution time is not possible.
3		Opportunistic load balancing algorithm[6]	Better efficiency Proper load balancing is done	Slow processing Poor make span time
4	DYNAMIC	Ant Colony Optimization Algorithm[7]	Fault tolerance Good scalability Achieve load balancing for Complex network.	Less throughput High power consumption.
5		Honeybee Foraging Algorithm[7]	Less response time Less waiting time Increase system diversity	Less throughput High response time
6		Biased Random Sampling Algorithm[7]	Achieve load balancing across all system node. Perform better with High and Similar population.	Performance overhead Less throughput.
7		Active Clustering Algorithm[1]	High availability of resources High throughput	Degrades as system diversity increases Less performance.

IV. IMPLEMENTATION

The cloud virtual machine datasets are collected from the AWS EC2 instances. CPU speed, CPU utilization, memory utilization, number of cores in the system , RAM size, Virtual memory size, response time are the attributes which are the fetched in the AWS. The fetched datasets scales or smoothens the raw values and converted into to the attribute of smoothened datasheets. The output is given as a input to the Ant Colony optimization algorithm and all the information is updated into the load monitor database. The Ant Colony Optimization algorithm used here is a meta heuristics algorithm which is used to predict the virtual machine that are further allocated to the users on demand there by Service Level Agreement(SLA) violation may not occur for load balancing system.

The heuristics algorithm has guaranteed for the optimal solution with any number of jobs that are used in it. Thus, ACO is an unsupervised machine learning task of inferring a function that describes the structure of unlabeled load. When the cloud user submits the request or job to the request handler, the request handler routes the request to the load balancer. Load balancer plays the vital role to allocate the instances to the user with the use of load estimator. The load estimator estimates the load factor using a machine learning algorithm for load balancing like ACO algorithm ant then updates the status to the load monitor database. Meta virtual machine monitor is used to collect the information about the local virtual machine and updates into the database. On the basis of meta VMM and database, the load balancer allocates the workload to the request handler. Then it responds to the user. Fig. 2. depicts the architecture of the proposed work

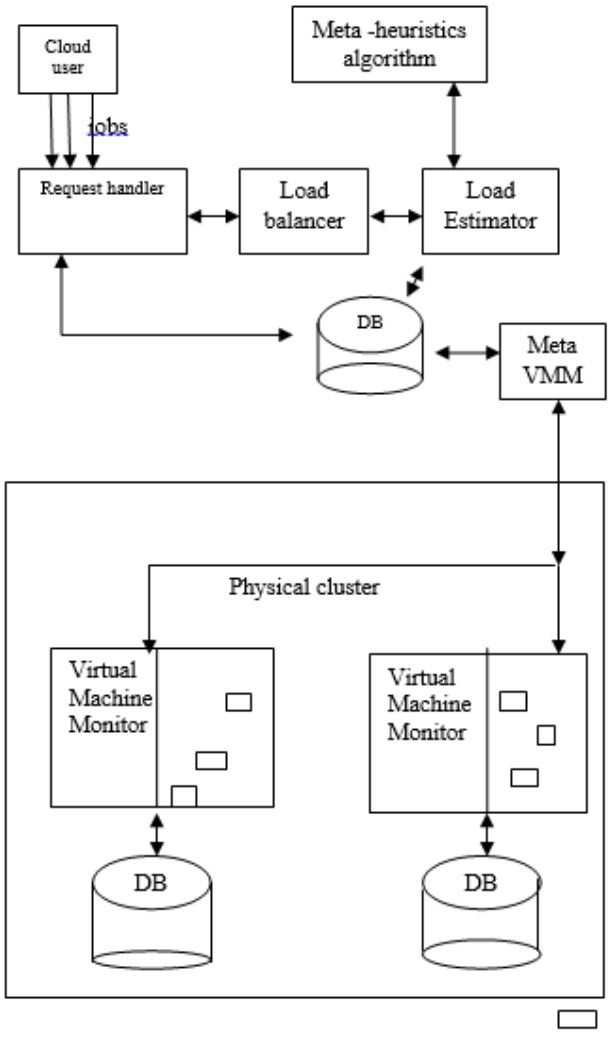

Fig. 2. Architecture of the proposed work

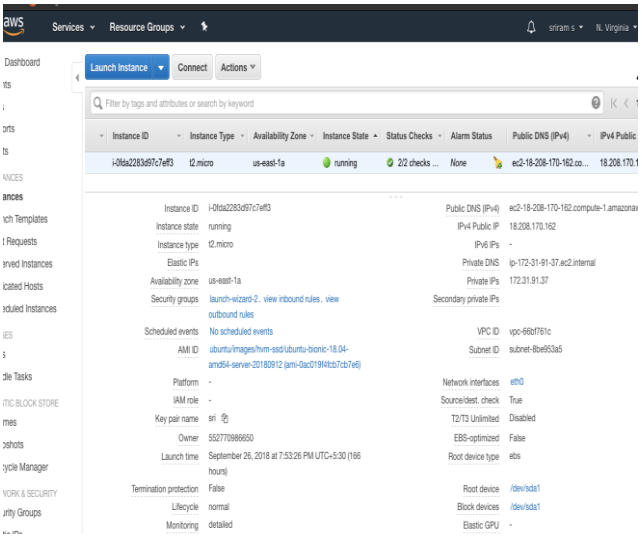

Fig. 3. High Level Design of PaaS

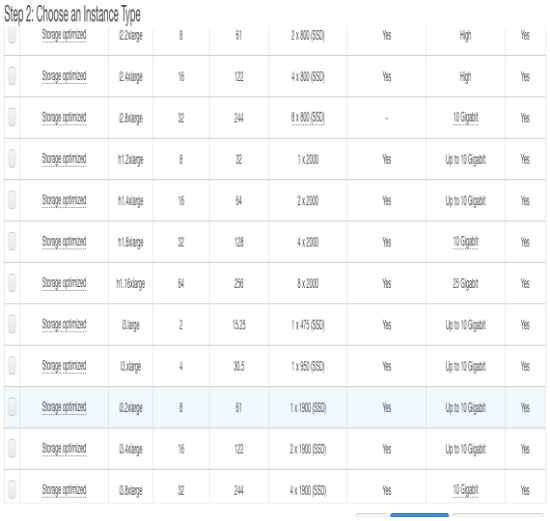

Fig.4. Selection of instances

V. EXPERIMENTATION

The proposed system is experimented using Platform as a Service available in Amazon Web Services. In AWS cloud, various types of instance are provisioned on demand with various configurations which are suited for different type of applications. Few of the different types of EC2 instances which are available with the different combinations of CPU, Memory, disk and Networking in the form of Platform as a Service are in the following TABLE II The instances are broadly categorized in to Accelerated Computing, Storage Optimized, memory optimized, compute optimized and General Purpose instance[9]. For our experimentation only T2 and T3 Instances are used which are General Purpose Instance.

The Fig. 2. explains the AWS cloud platform contains the launch instance information like instance ID, Instance status, instance type etc. The Fig 4. selects the instance type from the AWS platform for storage optimization and memory optimization. The Fig 5. explain about the configuration details about the instances like number of instances to be created, network used, reservation capacity etc. the Fig 5. clearly explains about the additional storage setting devices like attaching additional Elastic Load Balancing(ELB) volumes and instances.

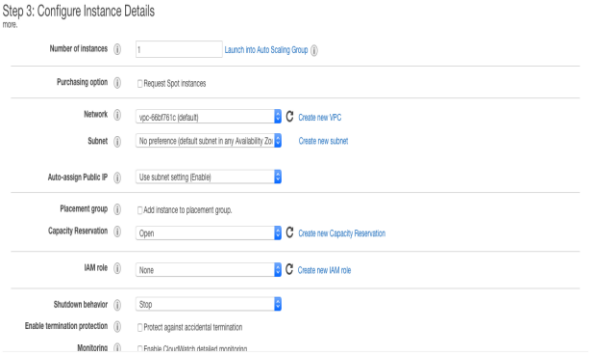

Fig. 5. Instance Configuration

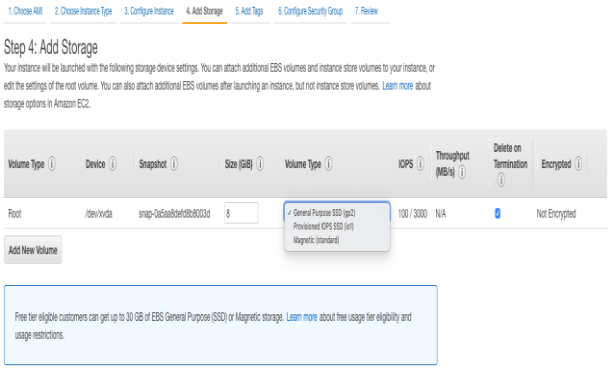

Fig. 6. Storage Setting in AWS

The instances are broadly categorized in to Accelerated Computing, Storage Optimized, memory optimized, compute optimized and General Purpose instance. For our experimentation only T2 and T3 Instances are used which are General Purpose Instance. .

For our experimentation only T2 and T3 Instances are used which are General Purpose Instance. In this experimentation several instances of type T2 and T3 with varying memory and network capacities are used to collect the response time and make span time for various set of tasks. The tasks are chosen to simple web services to web applications with database connectivity and MySQL. Then scaling is performed to smooth the values collected at different intervals for a period of time and several durations.

The Ant Colony Optimization algorithm is executed to classify collected data set along with the instance type and other features to segregate the instances so that such instances may be used for further allocation of task. Detailed experimentation for balancing the load among the Virtual Machines with different VCPU and Memory is going on which will be presented in the next paper.

TABLE II. DIFFERENT TYPES OF INSTANCES AVAILABLE IN AWS

Sl. No.	Instance Type	Specification	Details
1.	T2	3.3 GHz & 3.0 GHz Intel Scalable Processor	High Frequency Intel Xeon Processor with VCPUs of 1/2/4/8 with memory of ranging from 0.5 to 32 GB with Elastic Block store (EBS)
2.	T3	2.5 GHz Intel Scalable Processor	Max. 8 VCPU with 32 GiB (Gigabyte) Memory with EBS
3.	M4	2.5 GHz Intel Xeon	Max. 96 VCPU with 384 GiB Memory with (EBS)
4.	M5	2.5 GHz Intel Xeon with Advanced Vector Instruction set	Max. 96 VCPU with 384 GiB Memory with (EBS)

TABLE III. COMPARISON OF STATIC AND DYNAMIC LOAD BALANCING

S.No	**Static Load Balancing**	**Dynamic Load Balancing**
1	Distribute the work among process prior to the execution of the algorithm	Distribute the work among process during the execution of the algorithm
2	Behaviour is predictable	Behaviour is unpredictable
3	Lesser resource utilization	Better resource utilization
4	Less reliable	More reliable
5	Less efficient	More efficient
6	Minimal communication delay	More communication delay

VI. CONCLUSION

This study reviewed the existing load balancing algorithms available in cloud computing and each has their different aspects to balance the load. Overutilization or underutilization of resources leads to poor performances. To achieve efficient and maximum resource utilization, the load balancing is necessary for cloud computing. Based on the literature survey, Table I presents several load balancing algorithm and the advantage and disadvantage of the load balancing algorithm. This study Table II shows the different types of instances available in AWS. Table III compares static and dynamic load balancing algorithm Different load balancing algorithm has been discussed and each of them varies in certain criteria. Further, we discussed the challenges of the load balancing algorithm so that efficient algorithms are developed in the future. Then experimentation is done using AWS PaaS instances of type T2 & T3 with varying VCPU, Memory and Network capacities. In the future, this work will be extended to test the proposed load balancing algorithm which will be compared with the existing round robin algorithm used in AWS Cloud.

REFERENCES

[1] Manpreet kaur, Verma BK, " A Review on Varoius load Balancing Algorithms with merits-Demerits in Cloud Compuitng" IJAERD, vol 5, P-ISSN(P) 2348-6406, 2018.

[2] Deepa T, Sharon Amulya Joshi S, " A Survey on Load Balancing Algorithms in Cloud", IJCERT, Vol 3, ISSN(O) 2349-7084, 2016.

[3] Asha, Bharath Kumar, Girish V, " Load Balancing in cloud Computing", IJRTER, vol 4, ISSN: 2455-1457,2018.

[4] Sanjay Mani Tripathi, Sarvpal Singh, " A Literature Review on Algorithms for the Load Balancing iin Cloud Computing Environments and their Future Trends", Mathematical and Computer Modelling, Computer modeling & New Technology 21(1) 64-73, 2017.

[5] Zahra Mohammed Elngomi, Khalid Khanfar, " A Comparative Study of Load Balancing Algorithm s: a Review paper", IJCSMC, vol 5, ISSN 2320-088X, 2016.

[6] Abhijit Aditya, Uddalak Chatterjee and Snehasis, " A Comparative Study of Different Static and Dynamic Load Balancing Algorithms in Cloud Computing with Special Emphasis on time Factor, "INPRESSCO, vol 5, ISSN 2277-4106,2015.

[7] Deepa T, Dhanaraj cheelu, " A comparative study of static and dynamic load balancing in cloud computing", IEEE, ISSN 978-1-5386,2017.

[8] Ramesh prajapati, Dushyantsinh, Rathod, " Comparative of Static and dynamic load balancing in grid computing", IJTRE, vol 2, ISSN 2347-4718, 2015.

[9] online resource: www.aws.amazon.com .