**Energy Efficient Optimization**

**Of**

**Virtual Machine Placement**

A PROJECT REPORT

**(15CSE495 Project Phase I)**

*Submitted b*y

|  |  |
| --- | --- |
| ***Name*** | ***Roll number*** |
| P N DHINESH KUMAR | CB.EN.U4CSE15016 |
| MOHANA PRABHA | CB.EN.U4CSE15027 |
| NETHRA NARAYANAN | CB.EN.U4CSE15031 |
| D R DEEPAK VENGATESH | CB.EN.U4CSE15213 |

***In partial fulfillment for the award of the degree***

###### *Of*

###### BACHELOR OF TECHNOLOGY

IN

**COMPUTER SCIENCE AND ENGINEERING**



**AMRITA SCHOOL OF ENGINEERING, COIMBATORE**

**AMRITA VISHWA VIDYAPEETHAM**

**COIMBATORE 641 112**

**October 2018**

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112



BONAFIDE CERTIFICATE

This is to certify that the project report entitled **“ENERGY EFFICIENT OPTIMIZATION OF VIRTUAL MACHINE PLACEMENT”** submitted by P N DHINESH KUMAR (CB.EN.U4CSE15016), MOHANA PRABHA (CB.EN.U4CSE15027), NETHRA NARAYANAN (CB.EN.U4CSE15031), DEEPAK VENGATESH (CB.EN.U4CSE15213) in partial fulfillment of the requirements for the award of the **Degree of Bachelo**r **of Technology** in **Computer Science and Engineering** is a bonafide record of the work carried out under our guidance and supervision at Amrita School of Engineering.

PROJECT GUIDE PANEL INCHARGE

Dr. Prakash P Malathi P

Asst. Professor, CSE Asst.Professor, CSE

Department of Computer Science and Engineering

CHAIRPERSON

Dr.Latha Parameswaran

Dept. of Computer Science and Engg.

This project report was evaluated by us on :……………………………..

INTERNAL EXAMINER EXTERNAL EXAMINER

**ACKNOWLEDGEMENT**

We would like to express our deep gratitude to our beloved **Satguru Sri Mata Amritanandamayi Devi** for providing the bright academic climate at this university, which has made this entire task appreciable. This acknowledgement is intended to be a thanksgiving measure to all those people involved directly or indirectly with our project. We would like to thank our Vice Chancellor **Dr. Venkat Rangan. P** and Dr **Sasangan Ramanathan** Dean Engineering of Amrita Vishwa Vidyapeetham for providing us the necessary infrastructure required for the completion of the project.

    We express our thanks to **Dr.Latha Parameswaran**, Chairperson of Department of Computer Science Engineering, Dr. P Bagavathy Sivakumar and Prof. Prashant R Nair, Vice Chairpersons of the Department of Computer Science and Engineering for their valuable help and support during our study. We express our gratitude to our guides, Prof. Prashant R Nairfor his/her guidance, support and supervision.

    We feel extremely grateful to Dr. Vidhya Balasubramanian, Ms. Malathi P , Mr. Sabareesh B A , Ms. Kamalaveni V for their feedback and encouragement which helped us to complete the project. We would like to thank Ms. Malathi P for scheduling and organizing periodic reviews during the course of our project. We also thank the entire staff of the Department of Computer Science Engineering.

    We would like to extend our sincere thanks to our family and friends for helping and motivating us during the course of the project. Finally, we would like to thank all those who have helped, guided and encouraged us directly or indirectly during the project work. Last but not the least , we thank **God** for His blessings which made our project a success.

**TABLE OF CONTENTS**

**CHAPTER TITLE PAGE NO**

**ABSTRACT 5**

**LIST OF FIGURES 6**

**LIST OF ABBREVIATIONS 7**

**1. INTRODUCTION** 8

1.1 Background 8

1.2 Problem statement 8

* 1. Specific Objective 8

1.4 Limitations 9

**2. LITERATURE SURVEY** 10

**3. REQUIREMENT SPECIFICATIONS** 18

**4**. **ARCHITECTURE** 19

**5. PRELIMINARY RESULTS** 21

**6. REFERENCES** 22

**ABSTRACT**

A major challenge in Cloud computing is resource allocation for computational tasks. Many previous work have established a number of solutions to provide Cloud resources in an eﬃcient manner. However, in order to create a holistic resource allocation model, a prediction of the future resource consumption of upcoming computational tasks is necessary. We aim to present an approach for predicting Cloud resource utilization on a per-Virtual Machine and per-resource level. We apply machine learning-based prediction models to show how much prediction of workload can improve Virtual Machine Placement Efficiency.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Fig 4.1 | Virtual Machine to Physical Machine (Server) mapping | **15** |
| Fig 4.2 | Architecture of Proposed Solution | **16** |
| Fig 5.1 | Root Mean Square Deviation Formula | **17** |
|  | |  |

**LIST OF ABBREVIATIONS**

**Machine Learning:** Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

**Neural Networks:** A neuralnetwork is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.

**Artificial Neural Network:**Artificial neural networks or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains.

**Back propagation:** The most widely used training algorithm for multilayer and feed forward network is Back propagation. The name given is back propagation because, it calculates the difference between actual and predicted values is propagated from output nodes backwards to nodes in previous layer. This is done to improve weights during processing.

**RMSD:** The root-mean-square deviation or root-mean-square error is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.

**Virtual Machine:** In computing, a virtual machine is an emulation of a computer system. Virtual machines are based on computer architectures and provide functionality of a physical computer. Their implementations may involve specialized hardware, software, or a combination.

**CHAPTER 1**

**INTRODUCTION**

**1.1 Background**

Cloud computing can provide any type of service/resource, depending upon the requirement of the user. It is a Pay-as-you-use service. As the world is rapidly moving towards the age of cloud computing it is important that these services are quickly accessible and energy efficient. These resources are a pool of physical and virtual resources which are available anywhere, anytime on cloud. To access cloud services we need a virtual machine and to map a VM to a physical host, there must be a proper VM placement algorithm to ensure that the available resources are not wasted and are allocated and distributed across the hosts in an optimized manner. In cloud computing, Virtual Machine (VM) placement is a critical operation which aims to find the best Physical Machine to host a Virtual Machine. This has a direct effect on the performance, resource utilization and power consumption of the data centers and reduces the maintenance costs of cloud service providers. In order to provide fast data retrieval and energy efficiency we have planned to predict the workload of a Virtual Machine proactively and also come up with a efficient VM placement strategy by mapping the different Virtual Machines to Physical Machines.

There are 4 types of resources:

* CPU
* Disk
* Memory
* Network Bandwidth

**1.2 Problem Statement**

To design and develop a Virtual Machine Placement Algorithm by predicting the workload of a Virtual Machine to reduce the data retrieval time and improve the energy efficiency.

**1.3 Specific Objective**

* Preventing congestion in data centre's network
* Distributing traffic evenly in data centre's network
* Reducing energy consumption
* Maximizing resource utilization: The effective and efficient utilization of each server reduces the need for more PMs to host VMs.
* Improving Load balancing

**1.4 Limitations**

Our project focuses only on various scenarios generated using a simulator (CloudSim). The same can be extended to a large data like Google Cloud Trace data which is a huge dataset monitored over a long period of time. We would like to address this issue by working on Google Cloud Trace data as a future enhancement.

Moreover the machine learning model used needs a large amount of training data is needed to have good prediction accuracy. We also need to consider the fact that the machine learning approach could potentially outweigh the cost savings coming from cost and time reduction of optimized resource allocation and scheduling algorithms if it needs a lot of computation power.

**CHAPTER 2**

**LITERATURE SURVEY**

The following section provides a review of the literature related to the development of a Virtual Machine Placement Algorithm that makes use of usage statistics of a virtual Machine to predict the resource need and provide VM allocation and VM migration to other Physical Machines.

Resource prediction and Allocation has been discussed in **“Dynamic Resource Prediction and Allocation in Clouds using Pattern Matching”** by Harshpreet Singh1\* and Rajneesh Randhawa2 [1] where they primarily discuss about regular monitoring of resource usage such that the resources are sufficiently provided. Prerequisites created by the client are mapped with the quantity of VM's required which are then made accessible and are relegated on interest. A VM is an imitated machine which carries on much the same as the physical machine and is straightforward to the client. One physical PC machine can make the same number of VMs as the equipment accessible. For example, if there is a machine with 16 GB RAM, 3.4 GHz octa-center processor,4 GB VRAM, a RAID setup of 4 1-TB HDDs, it can undoubtedly handle 5low end VMs. A solitary client can be doled out various VMs dependent on the client's necessity. The issue emerges when another VM should be designated in a split second. On the off chance that the client needs a new VM and solicitations one, the VM can take anything from one to four minutes to begin and turn out to be completely operational. To defeat the deferral in giving the asset, a method dependent on patters as utilized in23 is streamlined and proposed. It begins when the client registers at the CP. The client gives the underlying prerequisite and a maximum necessity of the application. In view of these parameters, the introductory assets are allocated and the use is checked by the CP. After certain time unit, the observed utilization is kept as the Initial Resource Provision Array (RPA). This fills in as a brief parameter to foresee the up and coming utilization of client. The client is offered assets as bundles. A bundle has a one letter name to distinguish it. There are unique bundles that a client can ask for. New bundles can be made relying upon the interest of the clients. A bundle has different PC segments. It works based on hit and miss parameters. If the pattern matches, the parameter "Hit" gets increased. This implies the framework was fruitful in satisfying the client current solicitation relying upon the past expectations. The parameter "Miss" would have been augmented. This implies the framework bombed in accurately foreseeing the assets in the past. The Ratio of Hit and Miss tells the execution of the framework.

Virtual Machine monitoring has been discussed in **“Virtual machine monitoring in cloud computing”** byNikhil Saswade1\* Vinayak Ashok Bharadi2 and Yogesh Zanzane3[5] which discusses about creating basic cloud stage for numerous cloud related services and Automation with Cloud asset observing. Principle goal of Cloud robotization is that it ought to legitimately recognize and screen the utilization of cloud and its assets. The clients ought to acknowledge whether the framework under their utilization is completely practical or not. If, beyond the realm of imagination answers for the issue ought to be given. The security of the clients and the specialist co-ops cannot be undermined in any ways amid this procedure. Checking will be attempted to watch the assets under use, virtual machine, and time it takes to finish handling of errand.An individual stage under checking has neither rhyme nor reason since provider may not give cautioning system. In any case, they positively give utilization tally and measure of time the administration was under use. Be that as it may, imagine a scenario where client is utilizing more than one specialist co-op. By what method will the administrations be overseen at that point? For this very reason, the application needs to address not just administration from just a solitary specialist organization yet additionally two or three administrations given by other specialist organizations. Numerous mists are secured under this application. With the goal that various cloud benefits that client uses can be put under observing and alarming instruments will set for administrations from various cloud to improve openness and to remain safe of any extra use just as underutilization of cloud assets. Decrease of time and increment in usage is considered as an essential favorable position of cloud. Be that as it may, there are no accessible apparatuses appropriate for cost count and examination in Cloud condition. This paper centers around filling this hole. This gives an establishment to assessing financial effectiveness of Cloud and gives signs to cost improvement of Cloud. They have built up an estimation and examination approach into a web apparatus which is utilized in the inner Cloud condition and exhibit at first its investigation capacity on the cost dissemination and use awkwardness factor. To make one regular stage, with the goal that diverse cloud administration customers don't need to proceed to sign into various suppliers however they will have one normal SaaS to oversee and get to their everything administration related information. To build up this regular SaaS an investigation of all API's that are accessible by various cloud suppliers is finished. That API's will give yield and that will be brought and given to customer just as framework to screen and get result. This application will enable associations to proactively screen the wellbeing and execution of their basic applications conveyed on numerous mists. This application not simply screens basic application conveyed on cloud it even alarms clients while framework in procedure about use of assets and virtual machines made on cloud. This component isn't just restricted to any one cloud based specialist co-op however it is equipped for giving checking just as alarming system for different cloud based specialist organization.

Also in **“Virtual machine resource allocation algorithm in cloud environment”** Lei Zheng1\* [6] examines on the best way to determine the issue that virtual machine arrangement reservation plot squander a ton of assets and single-target organization calculation isn't far reaching, a virtual machine asset portion calculation dependent on virtual machines bunch multi-objective hereditary calculation is proposed. The calculation is separated into gathering coding and assets coding. Assets coding coordinated coding as indicated by the history asset need of virtual machines to physical machine and incorporate number of physical machine and asset need of physical machine involved by virtual machine through improved hybrid and transformation tasks. The exploratory outcomes demonstrate that the calculation is viable to diminish the quantity of physical machine and asset usage of physical machine, sparing vitality however much as could be expected.

The meaning of receptacle pressing issue is that a set S of M in size and a set P of N in size given, how all components of S are stuffed in components of P with least components of P utilized. BPP issue, a troublesome NP issue, is impossible by a known ideal calculation in polynomial time. The issue of virtual machines organization is really receptacle pressing issue. In distributed computing, how to sensibly convey virtual machines to pertinent hubs will be considered, acknowledging ideal use of assets while meeting administration targets of various applications. The virtual machines situation might be viewed as vector receptacle pressing issue. The merchandise being stuffed are the virtual machine under task, and the assets of virtual machine are the alterable size of products. The case is the physical hub, and the limit of the case is the utilization edge of hub asset. The quantity of sorts of assets is the quantity of measurements of vector receptacle pressing issue. Accepting that the quantity of physical hubs is M and the quantity of virtual machines is N, the arrangement space from the virtual machines to the physical hubs is NM . It is a NP issue comparative with canister pressing issue that requires a rough ideal arrangement.

So as to confirm the proposed calculation, the creator conducts recreation explore in CloudSim .With the reason for checking the adequacy and arrangement plot, the creator chooses the accompanying two traditional virtual machine organization algorithms（Multi-object virtual machines asset designation Algorithm, MOA）to contrast and the multi-objective virtual machine assets conveyance calculation. Best Fit Algorithm (BFA) intends to choose the physical machine that meets the asset need of virtual machine with least outstanding asset amid the virtual machine organization process, making the physical machine least residual asset. First Fit Algorithm (FFA) intends to look physical machines all together amid the virtual machine sending process, letting virtual machine legitimately sent in the physical machine that meets the asset need of virtual machine.

In the paper, subsequent to investigating the exploration circumstance of virtual machine organization conspire in distributed computing, the creator set forward improved hereditary calculation, lead bunch coding and asset need coding of virtual machine, and improve the hybrid and change task to determine the issue of vitality squander in distributed computing. Under trial condition, it demonstrates that the calculation can lessen the quantity of physical machines as well as improve the asset usage rate.

The machine learning concepts are discussed in **“Modeling Virtualized Applications using Machine Learning Techniques”** bySajib Kundu1, Raju Rangaswami2, Ajay Gulati3, Ming Zhao4 and Kaushik Dutta5 [7]. They present strategies to demonstrate the execution of a VM-facilitated application as a component of the assets assigned to the VM and the asset dispute it encounters. To address this multi-dimensional demonstrating issue, they propose and reﬁne the utilization of two AI systems: artiﬁcial neural system (ANN) and bolster vector machine (SVM). They assess these displaying systems utilizing ﬁve virtualized applications from the RUBiS and Filebench suite of benchmarks and show that their middle and 90th percentile expectation mistakes are inside 4.36% and 29.17% separately. These outcomes are considerably superior to anything relapse based methodologies just as immediate utilizations of AI procedures without our reﬁnements. They additionally present a straightforward and compelling way to deal with VM estimating and experimentally exhibit that it can convey ideal outcomes for 65% of the measuring issues that was examined and delivers near ideal sizes for the staying 35%.

They distinguish and think about the effect of key VM asset distribution and dispute parameters that influence the execution of virtualized applications. In doing as such, they ﬁnd that the I/O idleness seen by the virtual machine is a decent pointer of I/O dispute in a mutual stockpiling condition. They observationally exhibit that it is conceivable to demonstrate the execution of virtualized applications precisely utilizing only three straightforward parameters.

They apply and broadly assess two AI strategies, Artiﬁcial Neural Network (ANNs) and Support Vector Machine (SVM), to anticipate application execution dependent on these parameters. They create sub-displaying, a bunching based methodology that beats key constraints when straightforwardly applying these AI systems.

They execute and assess their displaying methods for ﬁve complex benchmark remaining burdens from the RUBiS and Filebench suites for the VMware ESX hypervisor . The middle and 90th percentile expectation mistakes are 4.36% and 29.17% separately (found the middle value of over these applications). The higher forecast blunders show up for the most part inside zones of low asset allotment and therefore low execution.

They present a basic and compelling way to deal with estimating the asset necessities for VMs within the sight of capacity I/O dispute dependent on the execution models that we create. Their VM measuring tests convey ideal outcomes for 65% of the estimating issues that was considered and delivers sizes that are near ideal for the staying 35%.

They propose to utilize propelled AI techniques to show the connection between the asset portion to a virtualized application and its execution utilizing a constrained measure of preparing information. Such a model would then be able to be utilized to anticipate the asset need of an application to meet its execution target.

In view of their examination with the ﬁve virtualized server benchmarks, they reason that precisely structuring and conﬁguring ANN and SVM based models are basic to viably address the VM execution displaying issue. They ﬁnd that making various sub models is an important advance in accomplishing high exactness expectations. For example, even with as few as 68 preparing information focuses, both the ANN and SVM based sub-models improve the expectation exactness for the ﬁlebench-ﬁleserver benchmark significantly.

The energy optimization concepts as per  **“Energy-Aware Application-Centric VM Allocation for HPC Workloads”** byH.Viswanathan1,E K Lee2,I Rodero3,D Pompili4,M Parashar5,M Gamell6 [8] propose an inventive application-driven vitality mindful methodology for Virtual Machine (VM) distribution is displayed. The proposed system guarantees high asset use and vitality efﬁciency through VM combination while fulfilling application QoS. While existing VM designation arrangements are gone for fulfilling just the asset usage necessities of uses along just a single measurement (CPU use), the proposed methodology is progressively conventional as it utilizes learning acquired through application proﬁling along various measurements. The aftereffects of our assessment demonstrate that the proposed VM designation procedure empowers signiﬁcant decrease either in vitality utilization or in execution time, contingent upon the streamlining objectives.

This paper centers around making a VM assignment demonstrate experimentally from information acquired by running HPC outstanding burdens broadly on a framework with a universally useful rack server conﬁguration, build up a proactive application-driven VM allotment calculation that utilizes this observational model, approve the methodology through broad recreations and evaluate execution gains as far as execution time and vitality utilization.

They pick an exhaustive arrangement of HPC benchmark remaining burdens. Every outstanding task at hand burdens at least one of the accompanying subsystems - CPU, memory, circle (stockpiling), and system interface.

The outcomes got from recreations utilizing genuine creation HPC remaining task at hand follows demonstrate that proactive VM distribution can signiﬁcantly add to vitality efﬁciency and advancement of the application execution relying upon enhancement objectives.

**CHAPTER 3**

**REQUIREMENT SPECIFICATIONS**

The proposed system needs access to a Virtual Machine usage statistics so that it can predict the VM workload. We will be using CloudSim simulator to generate a small network of Physical Machines and Virtual Machines and apply the proposed solution on it.

As an extension of this work the same solution will be applied on Google Cloud Trace which is a usage statistics dataset.

No other hardware requirement is needed.

**CHAPTER 4**

**ARCHITECTURE**

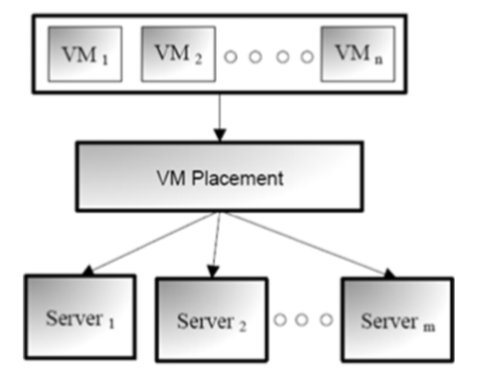


Fig 4.1: Virtual Machine to Physical Machine (Server) mapping

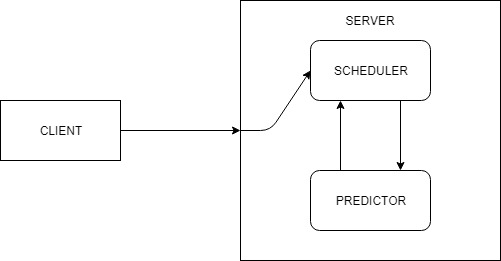


Fig 4.2: Architecture of Proposed Solution

In a cloud environment a task scheduler is used to perform the acquiring or releasing of the resources. In our architecture, a client, i.e., a consumer of Cloud services posts a request for the execution of a task to the cloud server. The server is composed of the scheduler, who is responsible for performing task scheduling and allocation, and the predictor, which supports the scheduler by predicting resource utilization. This enables the scheduler to optimize its decisions.

Upon receiving a request from the client/VM, the server must decide on how to allocate cloud resources. In existing architectures, this means ﬁnding a duration (time) and allocating resources for the VM request. Such a decision requires knowledge about the duration and resource consumption of the VM. For this, the scheduler queries the predictor to achieve a prediction of those metrics. Finally, after the execution, the cloud infrastructure reports the actual resource usage back to the predictor. Recorded traces of this data, that is, a history of predicted and actual resource utilization values, are then used by the predictor component to train its model.

We use an Artificial Neural Network (ANN) for training the predictor. The predictor takes an input vector I= (R1, R2… Rn) which denotes the VM request made to server for each type of resource. There might be ‘n’ number of resources but we will mainly concentrate on 4 types:

* CPU
* Memory
* Disk
* Network Bandwidth

So out input vector to predictor will be I= (R1, R2, R3, R4) and the output of the predictor will be a vector of resource utilization (duration of usage) U= (U1, U2, U3, U4) for each of those resources. The dataset consists of both resource needed for the VM at an instant and corresponding duration for it needs. We are planning to train this dataset to predict the duration of utilization for a given resource request. The training needs a huge amount of data and by the end of training initially the predictor will be able to predict to a certain extent. Later after every request the actual resource needed for that request along with their actual usage is added to the predictor to improve the prediction model and improving accuracy. Normalization of dataset values helps reduce the time taken for training of ANN and improve prediction performance.

**CHAPTER 5**

**PRELIMINARY RESULTS**

To determine the extent to which machine learning improves the performance of predicting resource usage we use the metric Root Mean Square Deviation (RMSD).

RMSD is calculated using the difference between predicted and actual values of the dataset and calculate root mean square deviation for it.

We calculate RMSD of Machine Learning Technique of prediction (RMSDML) and also the RMSD of a normal VM allocation algorithm without prediction (RMSDNML).

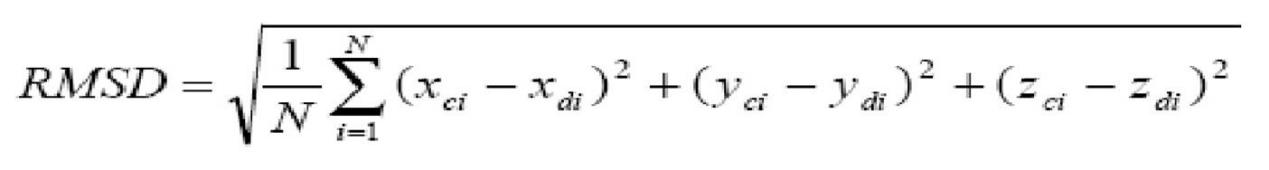


Fig 5.1 Root Mean Square Deviation Formula

We calculate the ratio between these two RMSD values – Model Relation.

Model Relation = RMSDML / RMSDNML

If Model Relation = 1 then both the techniques achieved the same results.

If Model Relation > 1 then it means Machine learning technique performs worse than Non- ML technique.

If Model Relation < 1 then it means Machine learning technique performs better than Non- ML technique.

**CHAPTER 6**

**REFERENCES**

[1] Zoha Usmania, Shailendra Singh, “A Survey of Virtual Machine Placement Techniques in a Cloud Data Center” , ICISP 2015.

[2] T.P. Shabeera, S.D. Madhu Kumar, Sameera M. Salam, K. Murali Krishnan, “Optimizing VM allocation and data placement for data-intensive applications in cloud using ACO metaheuristic algorithm” , Engineering Science and Technology, an International Journal 2017.

[3] Dabiah Ahmed , Huaglory Tianﬁeld and Yan Zhang “Glow worm Swarm Optimisation Algorithm for Virtual Machine Placement in Cloud Computing” , IEEE 2016.

[4] N. Janani, R.D. Shiva Jegan and P. Prakash, “Optimization of Virtual Machine Placement in Cloud Environment Using Genetic Algorithm”, Research Journal of Applied Sciences, Engineering and Technology 2015.

[5] Nikhil Saswade, Dr. Vinayak Bharadi, Yogesh Zanzane ,“Virtual machine monitoring in cloud computing” , 7th International Conference on Communication, Computing and Virtualization 2016.

[6] Lei Zheng ,“Virtual machine resource allocation algorithm in cloud environment”, Computer Modelling & New Technologies 2014.

[7] Sajib Kundu, Ajay Gulati, Ming Zhao, Kaushik Dutta , “Modeling Virtualized Applications using Machine Learning Techniques” , ACM 2012.

[8] H. Viswanathan, E.K. Lee, I. Rodero, D. Pompili, M. Parashar ,“Energy-Aware Application-Centric VM Allocation for HPC Workloads” , IEEE 2011.