International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Forecasting Cloud Resource Provisioning System using Supervised Machine Learning

Frishta Mirzad, Muhammad Rukunuddin Ghalib

Abstract: ***One of the biggest challenges cloud computing faces is forecasting correctly the resource use for future demands. Consumption of cloud resources is consistently changing, making it difficult for algorithms to forecast to make precise predictions. Using of the machine learning in cloud computing leads to many benefits. Such as chances of the enhancement in the quality of the service via forecasting future burden of works and responding automatically with dynamic scaling.***

This motivates the work presented in this paper to predict CPU use of host machines for a single time and multiple times. This paper uses three supervised machine-learning algorithms to classify and predict CPU utilization because of their capability to keep data and predict accurate time series issues. It is tried to forecast CPU usage with better accuracy while comparing to traditional methods.

Keywords: Cloud computing, Virtualizations, Resource-provisioning policies, Machine learning.

I. INTRODUCTION

Inspiration: Cloud computing is being extending at a rapid speed. Specifically as companies persist to turn their business to huge cloud vendors suchlike Microsoft Azure, AWS, google cloud platform. As a result of warm marketplace competition, suppliers have been below compel to generate delightful components and facilities, whereas managing their network expense. These elements merge to uncover suppliers to a broad diversity of volume of works (from two of the exterior consumers and their special inner businesses) that should exchange an ordinary data center structure. Supplying best work, accessibility and dependability below these circumstances can be costly unless complicated (though realistic and extendable) resource control. Unluckily, investigation on cloud resource administration till date has insufficient a comprehensive knowledge of the vital features of the volume of works of huge advertisement suppliers. For instance, previous works have done in this area but they are not giving more accuracy since the demands for resources are increasing day by day so there should be fast and accurate systems to manage and provide resources on time.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
*Correspondence Author
Frishta Mirzad*, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.
E-mail: frishta.mirzad2018@vitstudent.ac.in,
Muhammad Rukunuddin Ghalib, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.
E-mail: ghalib.it@gmail.com

In this paper, we make Predicting methods to build accurate time series forecasting of host machines CPU usage since one of the highest demanded resource is CPU in the cloud-computing environment hence it is a reason of the lack of the host machines. For measuring the performance of the host machine CPU is the important metric, so it is a hot topic for researchers to examine while predicting host efficiency.

II. LITERATURE REVIEW

Cloud computing provides trustworthy and pay per go model facilities with secure allotment of resources. Customers in cloud computing environment can simultaneously request for several resources. Afterwards, a well-organized way is needed to complete the needs of the customers to provide them all the resources. [8]. Somwya koneru et.al in [7] emphasis on growing the proficiency of the scheduling algorithm for the actual time cloud computing facilities. The resource-scheduling algorithm uses the turnaround time devices effectively by contrasting it into obtain operate and a miss operate for individual job and as well consumes to optimize the proficiency advantage. A general progress in the resource utilization and decrease in the treatment expense is displayed. The resource allocation strategy explained in the. [6] as a mainstreaming cloud supplier performance for consuming and allotting uncommon resources across the bound of cloud surroundings so to fulfill the demands of the cloud applications. Furthermore, an overview of the categorization of RAS and it is effect in the cloud system is provided. Allotting the resources for IaaS rely on predetermined allotting strategies. There the emphasis is on due date touchy strategy to assign the resources in a prosperous way by decreasing the demand denials by Haizea Haizea is an open source resource rent manager, and be able to deed as a scheduled to open source cloud toolset Nebula. [5]

III. IMPORTANCE OF RESOURCE ALLOCATION

In the cloud, computing resource allotment is the procedure of allotting accessible resources to the demanded cloud applications through the internet. Resource allotment famish facilities if the allotment is not controlled accurately. Resource provisioning resolve that issue permitting the facility providers to control the resources for every individual module.

Resource allotment policy is all about merging cloud suppliers operations for using and allotting rare resources across the extent of cloud surroundings to meet the demands of the cloud applications. It needs the kind and quantity of resources are requested by every application so that to fil out a user task. [2]

The sequence and time of allotment of resources are as well an entry for an ideal RAP. An ideal RAP must prevent the bellow standards as following.

A: **resource controversy**: situation occurs when two or more applicants attempt at the same time to access the alike resources.

B: **Poverty of the resource** occurs when there is no enough resources.

C: **Resource fragmentation**: there are condition arises when the resource are knock down. Despite that, enough resources are available but we will not be able to assign them to users.

D: **over-provisioning** of assets occurs when the applicants gets excess resources than the require one.

E**: Under provisioning** of assets happens when the customer is allocated with less number of the assets than require one.

Before the estimated time to complete a job resource users demands for the resources this may lead to under provisioning of resources.

To prevail the aforementioned disparity we have to get input from both side cloud user and cloud service provider as well.

User's angle inputs are the application requirements and service level agreements and cloud service provider inputs are status of the resources, amount of the resources and availability of the resources service provider inputs to administer resource allotment to host applications by RAP.

The consequence of RAP should fulfill the parameters such as throughput, latency, and response time. Despite the fact that cloud provides authentic resources. It also causes a critical issue in allotting and managing assets automatically among the applications.

From the vision of a cloud provider, forecasting the automatic nature of users, user appeal and application appeal are inoperative. For the cloud applicants in a minimal cost the job should completed. We need an impressive allotment system to suit cloud environment due to limited resources, resource dissimilarity, area limitation, environment requisiteness and automatic nature of resource request.

Cloud assets are contains both hardware and virtual resources. Such that physical resources are shared among, several compute demands thru virtualization and provisioning. The demand for virtualized resources is demonstrated thru a collection of parameters narrating the processing, memory and disk needs that is explained below. Provisioning fulfils the request topography virtual resources to physical ones. On demand basis, the hardware and software resources allotted to users. Virtual machines are rented for scalable computing.[3]

In big systems like clusters, data centers or grids the complexity of finding an optimize resource allocation is exponential. As resource appeal and accumulation can be automatic and unknown, several strategies for resource assigning are suggested. This paper explains four several resource assigning strategies located in cloud perimeter. [8]

IV. PROPOSED ARCHITECTURE

The model accepts demands from users then this demand goes to manager. Manager manages the requests and then the machine learning processor sees which user or customer requested for how many resources based on earlier demand and then asset provisioning operates and scaling decision is carried out. The right way to determine if a VM is provisioned with the suitable number of virtual CPUs for workload is to revise it is functionality metrics. This give us perception into features like;

- How much time a CPU spend on an application.
- How much time is took by other virtual CPUs on physical host.
- How much time t is utilized by the physical host itself
- How much time the virtual CPU consume idle because they have no work to carry out.
- The time which CPU waits for a physical host when it is serving another virtual CPU is called "CPU steal time".[4]

However, every virtual CPU must be scheduled it is fair distribution of the physical host's central processing unit cycle; the time is not inevitably distributed evenly. In majority cases, CPU steal time is "lost" and CPU steal time "gained" would parity out, though if the percentage is stays high , it lead to problems with web application The amount of time used by physical machine depends on how appropriately each VM is provisioned. When there is, too much over provisioned VMs the host CPU spend more time on scheduling and starting / terminating time slots. In this case, it have many hardware and software interruptions to deal with, while operations are being completed waiting time will be more.

In this paper, I used three machine learning algorithms, random forest, linear regression and XGBoost. The two algorithms random forest and XGBoost are used for classification and accuracy to show how much the model is accurate. Moreover, linear regression is used for predicting the time portion of CPU, which predict and shows in how much time a task will complete by CPU.

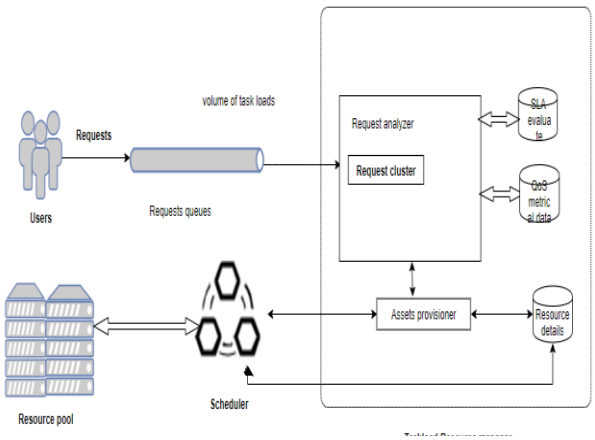

Figure 1: Proposed architecture diagram

As the below figure shows how the system works, it shows every step that how the process is being done in the systems. How the request goes to request analyzer, after checking of the available resources it is assigned to the customers.

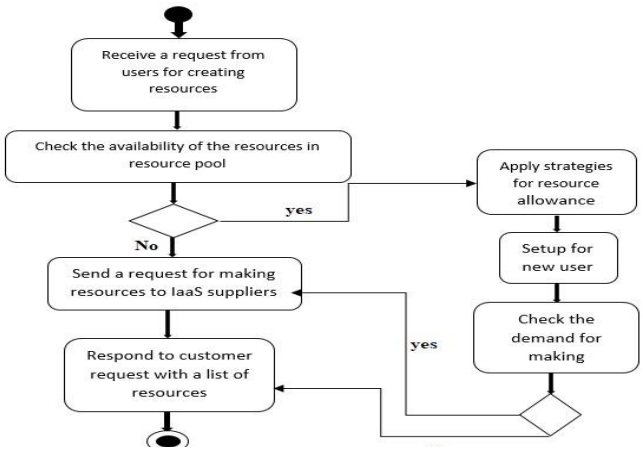

Figure 2 :Proposed architecture flow chart

V. METHODOLOGY

In this part, we discuss our approach for forecasting resource provisioning for web applications using machine learning to improve the proficiency-forecasting model. Many authors in making forecasting models have investigated random forest and linear regression. Lately XGBoost a powerful classification method has been getting considerably fame.

5.1.XGBoost

XGBoost stands for extreme Gradient Boosting. XGBoost is the accomplishment of gradient boosted decision trees layout for speed and proficiency.

It is an accomplishment of gradient boosting machines constructed by Tianqi Chen, with allotments from several developers. It is related to a wide set of tools under the distributed machine learning association or DMLC [3].

XGBoost is a library you can configure it and download it on your machine. Then you can access this library through different interfaces. The XGBoost library is laser centralized on computational speed and model efficiency. XGBoost suggest many number of advanced features or characteristics.

5.2 Three main shape of XGBoost are:

Slope boosting: this algorithm is also called gradient or slope boosting machine containing the acquisition rapidity.

Accidental slope boosting: it takes the sample of the row and column and then column per levels.

Arrange slope boosting: with both L1 and L2 regularization**.**

5.3. System features:

- During the training parallelization of tree structure by using all your CPU cores.
- By using a cluster of machines and distributed computing to train huge models.
- The large data set that do not fit into memory use out of core computing.

•To make best use of hardware optimize the cache of the data structure and algorithm.

Algorithm features: the accomplishment of the algorithm was engineered for performance of computing and memory assets. The design aim was to build the best use of available assets to train the model .there are some key algorithms accomplishment characteristics are mentioned below:

•Sparse Aware accomplishment to handle missing values automatically.

•To support parallelization of tree structure use block structure.

•To boost an already fitted model on new data do continue the training.

5.4. What algorithm does XGBoost use?

The XGBoost library accomplishes slope boosting decision tree algorithm. This algorithm has different names, gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines.

To correct the existence model errors we add new models and boosting is an ensemble technique. These models are added in a sequential manner until there no more improvement can be made.

In gradient boosting by creating new models to predict the errors of previous models then we put all these predictions together to make the last prediction this is called gradient boosting when adding new models it minimizes the loss by using gradient decent algorithm. this algorithm supports both classification predictive modeling problem and regression.[4]

VI. RANDOM FOREST CLASSIFIER

Random forest algorithm works as a large collection of de-correlated decision trees. It creates many decision trees and use them to make a classification. That is why it is technique based on the bagging technique.

Random forest or random forest decision tree is a method operates by constructing multiple decision trees during training phase.

The decision of the majority tree is chosen by the random forest as the final decision. [9]

VII. LINEAR REGRESSION

"The road to machine learning starts with Regression"

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. It is parametric in nature because it makes certain assumptions (discussed next) based on the data set. If the data set follows those assumptions, regression gives incredible results. Otherwise, it struggles to provide convincing accuracy the linear regression is a supervised learning algorithm in machine learning. It operates regression tasks. Based on the independent variable regression models an objective forecasting. This algorithm is used for searching out the correlation between parameters and predicting. There are various regression models for different purpose based-on sort of the relationship among independent and dependent variables, they viewing it and how many independent variables are used. Linear regression is used for task forecasting a dependent variable (y) based upon unrelated variable (x). Therefore, regression method figure out a linear relation among x (input), and y (output). Therefore it is name is linear regression. [9]

$Y = W_0 + W1 * X1 \ldots + W_n * X_n$

While,

n= number of input variables

x_i is the ith input

$W_{i\,is}$ the coefficient of the ith input

W0 is called interception and add more degree of freedom to the linear regression.

VIII. CONDUCTED EXPERIMENT

In this paper the following experiments are performed. The first test implicates comparison of all the algorithms and techniques operation of the training data. The second test will measure the efficiency of the every algorithm on the testing data. The aim of this test is to check if the algorithms are able of giving a better overall efficiency on the data it has not seen. The third test will measure how far into future the algorithms (random forest, XGBoost, linear regression) can predict. The outcome of these test will be exciting as it will be useful to data centers management systems to prior how much CPU is utilized on every host until situation like migration can happen. This test will measure the accuracy of the network for forecasting CPU usage in advance to one-step into future.

8.1. Details of the conducted Experiment

a. Dataset

The com- active db is a set of a computer systems activity scales. This data was gathered from a sun SPARC station 20/712 with having 128 Mbytes of memory operating in a several user university department. Consumers shall be doing a huge diversity of works like from reaching the internet, altering files or rushing very CPU-limit programs. The last data was captured consistently on two distinct reasons. On both reasons, system activity was collected each 5 seconds. The last dataset is adopted from both reasons with same numbers of observations came from every set cycle in accidental arrangement.

b. Trained Models

Firstly the model is trained with CPU utilization data set by using three machine learning method. Afterwards using of the same dataset, we instruct a new model for the two-response time and throughput. I have used the following metric RMSE- root mean square error parentage of observations whose prediction accuracy is 90.36% of authentic value to measure both training and testing output of the model. I have two-model first one is the response time the throughput and second one is CPU usage. I used python language with pycharm ide for making models using Random forest, linear regression and XGBOOST algorithms. [3]

c. Testing of the Models

The training to testing ration of the 90 to 91 % this gave ideal forecasting output for the model. This is on the basis of the records from the older works in [7, 8] related virtual machine start up time and encourage from the work of [10].

IX. RESULTS

The aim of this survey work are to check out the accuracy of the selected machine methods in predicting future resource utilization to combine SLA into the forecasting model. CPU usage forecasting model and SLA acknowledgment time forecasting are used to obtain and fulfil the aims. The first model which I have used random forest algorithm has the accuracy ratio of 90.6

The Second model using XGBoost algorithm has the accuracy ratio of 90.326. These result shows that how accurate are these algorithms for predicting the future resource usage.

X. COMPARISON OF THE APPLIED ALGORITHMS WITH OTHER ALGORITHMS

The below table shows different algorithms accuracy and root mean square error. This is how we can choose among these algorithms the one with highest accuracy and less root mean square error for our system. The accuracy of algorithms used in [12], neural network (NN) 77%, support vector regression (SVR) 60 %, linear regression (LR) 74%. With comparison to these algorithms, the algorithms, which are used in this paper used, have the high accuracy of more than 90%. Random forest accuracy 90.3 %, XGBoost accuracy 90.1%. The linear regression a supervised machine algorithm is used to show the relationship between two variables, independent variable (x) and dependent variable (y) uses the root mean square error, this function shows the error between predicted values and actual values the lesser the error the accurate the model will be. [12]

Table 1. Different algorithms RMSE comparison

Model	Accuracy	RMSE	PRED
NN	77	8.67	0.77
SVR	60	6.75	0.60
LR	74	6.72	0.74

The above table shows the different metrics comparison of different algorithms

Table 2. Used algorithms accuracy comparison

Model	Accuracy	RMSE	PRED
Random forest	90.1	nil	0.906
XGBoost	90.3	nil	0.903
LR	74	9.0	0.9

This table shows the accuracy of the algorithms which I have used in this experiment. The accuracy of the random forest algorithm is 90.1 and XGBoost (extreme gradient boosting) is 90.3. and the linear regression which an essential statistic method which is used for prediction of the numerical applications while both output (target class, feature) and attribute and features are numeric. The output is shown in a linear manner that is the combination of the attributes with Predefined weights $Y= W_0 + W1 * X1\ldots + W_n * X_n$

While,

n= number of input variables

x_i is the ith input

$W_{i\ is}$ the coefficient of the ith input

W0 is called interception and add more degree of freedom to the linear regression.

XI. CPU USAGE PREDICTION

We take into account two load-test runs, in which the precise system was subjected to two various loads. Below tables are the statistics of the two run systems, which was measured during the run time. Table 1.

Table 3. System one statistic

Run 1 ----- BUSY_TIME 1,159,299 # CPU times are in hundredths of a second IDLE_TIME 286,806 OS_CPU_WAIT_TIME 1,621,100 %busy 80.17 response time 0.2 # seconds

Table 4. System two statistic

Run 2 --------- BUSY_TIME 1,305,666 IDLE_TIME 125,475 OS_CPU_WAIT_TIME 5,071,200 %busy 91.23 response time 1.5

Even though CPU utilization was pretty high in both situation, but their operation could not have been more variant. Response time of the Run 1 was 200 ms while the response time of the Run 2 was 1500 ms. It is estimated that the single transaction takes 200 ms time. Therefore, run 1's reply time is similar as the base line number though the CPU utilization was as upper as 80%.

in a manner in case a system is operating at 20 % CPU , it simple means you are spoiling 80% CPU. In trading terms, this means you are spoiling costly money. Actually, it was stated that CPU time is not a thing, which you can save it in a bank for later usage.

In the load test which we have done in run 2 OS-CPU-WAIT –TIME of 5071200 vs the BUSY-TIME of 1305666 says that the system consumed 4 times the time waiting for the CPU to become accessible better than exactly running on it. It is cause of there were more processes ready to execute that there were CPUs accessible. In an individual CPU system when the load is smaller then 1 it indicates an average, each process which required the CPU can use it instantly free of being closed. Inversely while the load is bigger then 1, it indicates on average, right there have been processes prepared to execute, yet may not because of un availability of the CPUs.[11]

So there is a big difference between 100 using of CPU and workload = 1 and workload = 10 and CPU usage. A system consist of multiprocessor, number of the load with number of CPU may be interpreted together. For instance while the system has 4 CPU, later here could be conflict in case the workload is bigger then 4. (the test load which we have done in run 2 system had 4 CPUs while the work load was 19.22) no matter the efficiency was so awful.

The below plot with 1000 points of measured value and predicted value for CPU usage the diagonal shows a perfect fit between predicted and actual values.

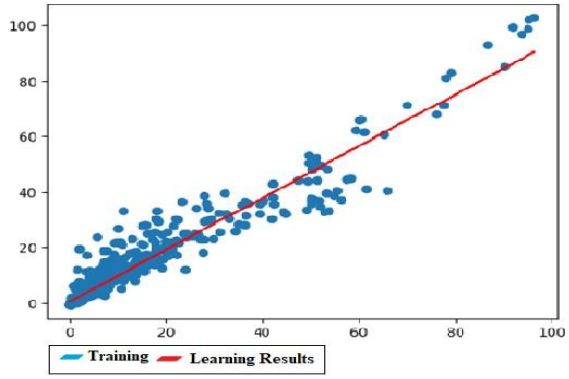

Figure 1: actual and predicted Vcpu utilization prediction

In the below figure we are predicting the portion of time that CPUs run in user mode (our target variable).

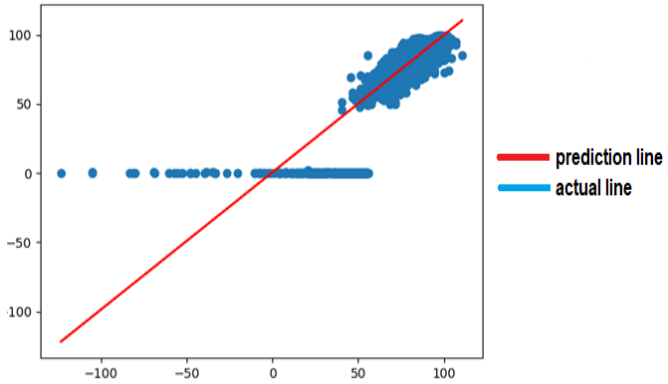

Figure 2: Actual and predicted Vcpu user time series-Linear Regression

We have taken the reading CPU feature to show how much CPU is used by customers in the cloud and based on this we are predicting the future usage of the CPU by using linear regression algorithm, we use the MRSE metric to show how the data points are fit and how much dependent are these variables.

We distributed this data into attributes (independent variable x) and labels (dependent variable y) to predict them. Next the dataset is splatted in to two parts, one part of the dataset is used for training of the models and the second part of the data we will use for the testing of the models to check the accuracy of the models. To evaluate the performance of the algorithms we use the RMSE metric to show how the algorithms

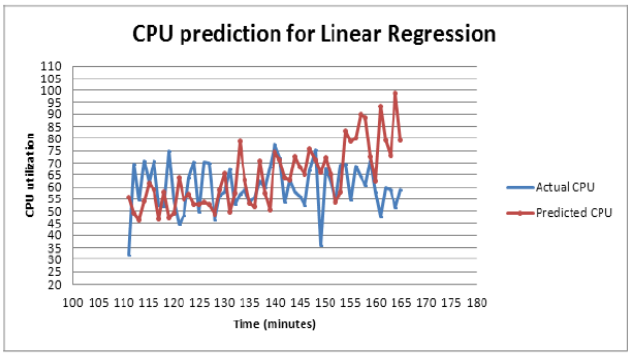

Figure 3: CPU usage prediction

The usage of the CPU in above figure shows the random usage level.

We wanted to do a realistic scenario simulation while clients sends demands in a random manner. Therefore a client demand is ending, next clients starts an application. This was within the casual start and closing circle which we noticed the fall and go up CPU usage. [13]
Figure 3 shows the actual and predicted CPU usage of linear regression algorithm.

XII. CONCLUSION AND FUTURE DIRECTIONS

Cloud computing is being extending at a rapid speed. Specifically as companies persist to turn their business to huge cloud vendors suchlike Microsoft Azure, AWS, google cloud platform. As a result of warm marketplace competition, suppliers have been below compel to generate delightful components and facilities, whereas managing their network expense.
Cloud computing is extensively used for sharing of data resources and information. Resource provisioning is used to provide assets in cloud computing environments. There are different methods to use and provision resources or assets.
This survey paper indicates to enhance allowance of the cloud assets; correct predictions of the time of loads in advance are needed. The issue this paper discuss is the time portion of the CPU and CPU usage predictions. Utilization forecasting is very significant for automatic extending of the cloud assets to attain proficiency in conditions of expense and energy usage to ensure the quality of services. The system suggested in this report is that pull out physical assets usage, saves the usage of the templates, and test regularity. For future directions, it is planned to implement this system in public cloud computing structure infrastructure, to provide a quick an on time services for clients.

REFERENCES

1. Andreas Kapsalis, Panagiotis Kasnesis, Iakovos S. Venieris, and Dimitra, "A Cooperative Fog Approach for Effective Workload Balancing", "IEEE CLOUD COMPUTING",ISSN: 2325-6095, **CD:** 2372-2568, INSPEC Accession Number: 16838311, 2017.
2. Yan Kyaw Tun, Nguyen H. Tran, Shashi Raj Pandey, Zhu Han," Wireless Network Slicing: Generalized Kelly Mechanism Based Resource Allocation", "IEEE Journalon Selected Areas in Communications", Volume: 37 , Issue: 8 , Aug. 2019.
3. Faiza Samreen, Yehia Elkhatib, Matthew Rowe, Gordon S. Blair," Daleel: Simplifying Cloud Instance Selection Using Machine Learning", "IEEE/IFIP Network Operations and Management Symposium (NOMS 2016): Mini-Conference", ISBN: 978-1-5090-0223-8, ISSN: 2374-9709, 2016
4. Matthias Lermer, Stefan Frey, Christoph Reich, "Machin Learning in cloud Environments considering External information", "IMMM 2016: advanced institute convergence information technology research center", Furtwangen University, Furtwangen, Germany.
5. Haid Goudazi , Masood pedram, "Multi- Dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing System", "IEEE 4th International Conference on Cloud Computing", Book e-ISBN : 2159-6190, 2011.
6. Pankaj Sareen, Parveen Kumar. "Resource Allocation strategies in cloud computing", , "International Journal of Computer Science & Communication Networks", Vol: 5(6),358-365, SSN:2249-5789, 2015.
7. Somwya koneru, Dario Bruneo, Audric Lhoas, Francesco Longo, Antonio Puliafito, "Analytical Evaluation of Resource Allocation Policies in Green IaaS Clouds", "IEEE Third International Conference on Cloud and Green Computing" Electronic ISBN: 978-0-7695-5114-2, 2013.
8. Aaqib Rashid, Amit Chaturvedi, "Virtualization and its Role in Cloud Computing Environment", "(international association computer science information technology", Mewar University , Rajasthan , India , 2019.
9. Pal, M. Random forest classifier for remote sensing classification."(National authority remote sensing space sciences." 2015.
10. Akindele A, Bankole, Samuel A. Ajila, "Predicting Cloud Resource Provisioning using Machine Learning Techniques", "(2013 26th IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)). Carleton University, Ottawa, Canada.", vol.6, Issue 54 , 2013.
11. Linlin Wu, Saurabh Kumar Garg and Raj kumarBuyya: SLA –based Resource Allocation for SaaS Provides in Cloud Computing Environments (IEEE, 2011), pp.195-204.
12. Abdul khalek Mosa, Norman W. paton, "Optimizing virtual machine placement for energy and SLA in clouds using utility functions", "advanced research journals,", DOI : 10.1.1186/S13677-016-7, University of Manchester Oxford Road, United Kingdom, 2016.
13. Martin Duggan, Karl Mason, Jim Duggan, Enda Hawley, Enda Barrett, "Predicting Host CPU Utilization in Cloud Computing using Recurrent Neural Networks", "IEEE The 12th International Conference for Internet Technology and Secured Transactions" ISBN: 978-1-908320-93-3, 2017.

AUTHORS PROFILE

Frishta Mirzad, Currently student of Vellore Institute of technology. I have developed deep interest in area of Cloud Computing from my schooling. This interest in Cloud Computing made me choose computer science and engineering in my graduation. Post completion of two years in graduation, I have explored lot of areas in computers science and engineering and I have completely engrossed myself in deep diving into cloud computing. With the help of my professors from my college, I am working towards publishing a paper in cloud computing.

Dr. Muhammad Rukunuddin Ghalib, He is an Associate Professor (Senior) at School of Computer science and Engineering, VIT University, Vellore, India. His research activities are carried out in Data Mining, Bioinformatics, Microarray Gene Expression Analysis, Neural Network, and Soft Computing.