Middle-East Journal of Scientific Research 24 (6): 1995-2002, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.06.23583

Load Forecasting for Optimal Resource Allocation in Cloud Computing Using Neural Method

[1]Rathinapriya Vasu, [2]E. Iniya Nehru and [3]G. Ramakrishnan

[1]M.E. Software Engineering, Easwari Engineering College, Chennai, India
[2]Senior Technical Director, National Informatics Centre, Chennai, India
[3]Professor, Information Technology, Easwari Engineering College, Chennai, India

Abstract: Cloud computing is the fast growing technology that is widely used for distributing and acquiring the required resources through the internet. It has equipped a modular and economically profitable environment. Huge consumption of power in the cloud computing framework is mainly because of the availability of large number of servers that are featured in the cloud datacenters. Energy optimization has been a major consent for various operations in a cloud environment. This kind of power optimization is mainly carried out for economic reasons and it could be done in a most efficient manner by forecasting the future load and allocating the resources accordingly .This kind of load forecasting is done so as to disseminate the load to different efficient servers based on the past downtime history, so that the energy could be reduced in a drastic manner. The servers in the cloud zone are tracked using their reliability record and this information could be used as a benchmark for allocating the resource to the requested job. When the load forecasting and server reliability are carried out simultaneously, an efficient Virtual machine could be allocated to complete the job with a relatively less power consumption.

Key words: Load forecasting · Cloud computing · Neural prediction · Host management · Server reliability

INTRODUCTION

Cloud computing is a modernistic breakthrough in the field of InformationTechnology where the applications are provided as services to end users based on a pay as you use model. This makes it easy for the users to access the information from anywhere at any time. The traditional computer setup requires the user to be in same location where the data storage is available, the cloud makes the repository and retrieval location different from each other.

Cloud computing is one of the majestic inventions in the InformationTechnology scenario since the development of the PC. Cloud computing is becoming a part and parcel of the applications that we use today, it is getting a wider attention from every section of people. Merrill has characterized that almost 12% of the entire world software industries are completely based on the internet [2]. It gives a clearvision regarding the process of utility computing. Cloud computing is a tremendous methodology that hands over the deliveryof on-demand computing services that starts from the application requested by the user to the cloud datacenter. It allows the customers to choose any kind of resources available based on their dynamic needs. Cloud computing prepares itself for three types of service models namely: (i) Software as a Service(SaaS), here the applications requested by the end user are provided as a service(ii) Platform as a Service (PaaS), this provides the platform for the customized applications that is to be deployed in the cloud environment .(iii) Infrastructure as a Service (IaaS), it typically deals with the delivery of the hardware and the associated software so that the requested application could be deployed [3].

These days a large number of researchers focus on hosting an application with high computing efficiency, without even worrying about the fact of the huge energy consumption for the dynamically generated load. These datacenters constitute a huge number of servers along with large networking equipment which eventually leads to huge consumption of power. Based on the reports stated in the, in the United States the datacenters has approximately made 1.5% of the total power consumed in

Corresponding Author: M.E. Software Engineering, Easwari Engineering College, Chennai, India.

the year 2006, that has increased to a greater quantity in the recent years. In the upcoming era the cost that is spent for operating a server would be increased to a heft y amount rather than the amount used for buying it.

The energy costs has been awfully increased these days that makes the necessity for idealizing the cost of the servers used. It is so important that the cloud datacenters assures the QoS and optimally reduce the power as well.. Today most of the Web-based industries invest a huge capital in installing a cloud framework for their application activities.Almost every social media uses the cloud scenario for all their storage and data retrieval purposes all these facts leads to huge consumption of power [4]. In the current scenario energy is saved using Dynamic Voltage Frequency Scaling system that uses an adaptive algorithm which is mostly used in adjusting the hardware that is the operating clock of the system, which doesn't yield an accurate result .The other technique is shutting down of the idle servers, although it reduces power to a certain extent, shutting down of servers in a dynamic environment leads to huge wastage of resources.

This paper main intention is to design, implement and evaluate a neural load forecasting technique for optimal resource allocation in cloud computing that significantly consumes a lesser power in the virtualized servers. The following technique signifies a comparatively accurate prediction methodology that forecasts the future load, using the past downtime history of the servers. This technique makes sure that the requested job is allocated to an optimal server, the one that is deserved to complete the job with lesser amount of power consumption.

Neural framework uses a completely unique computational methodology for various problem solving techniques in various areas like mathematics, engineering, chemistry, biology , power utility, etc., that are really very difficult to solve for both computers and users as well.

The remaining part of the paper is coordinated as follows. In the following section, epitomize some of the related work regarding the load forecasting and server state management. The Section III illustrates the system architecture that is been used for the Load forecasting. Section IVportrays the complete information regarding the implementation methodology. Section V tabulates the experimental results and finally Section VIprovides a brief sketch about the conclusion future works.

Related Works

Power Consumption: Power consumption plays a quite important and majestic role in cloud computing, Truong Duy [6], proposed a paper on the usage of a green computing algorithm that are tremendously used for energy savings in the cloud environment. Many research scholars focus on deploying an application that works fast and solve the timely need, which eventually leads to huge wastage of resources. In order to overcome this process, scholars should focus much on reducing the power consumption. It is quite very obvious to reduce power in two different ways, one is by allocating the required the server in an efficient way using Dynamic Voltage Frequency Scaling methodology and the other is by shutting down of the servers when they are not in use. Shutting down of the servers leads to huge wastage of resources in a dynamic environment and it takes quite a lot of implementation cost and maintenance as well. So green computing could be used which forecasts the CPU utilization capacity of the available host, so that proper planning could be done so as to allocate the resource in a quite optimal manner with lesser wastage of the power.

Virtualization: The key technology that plays a major role in cloud computing is Virtualization [8], the main objective of the virtualization technology is that it could make full use of the pretty penny mainframe assets. This technology is used in such a way that it allows a host to run multiple operating system simultaneously. The virtualization technology is considered to be a boon as it has given a solution tothe problem of equipment purchase and maintenance expenses as well.

Cloud computing makes a wide use of virtualization, it would allow the user to access the resources via internet, that is used a method of service. It breaks the traditional myth of computing technology, it makes the user and the place of the available resource independent .It makes it possible that the user could access the resource from anywhere anytime. It introduces a concept of mirroring which eventually makes it possible to create identical copies of the same datacenter that could be used to reduce the site traffic.

Artificial Neural Network: A neural network based approach was introduced by Kara [10], features for classification in the feed forward neural network was identified. Here three different layers were used, namely the input layer, hidden layer and the output. The basic dataset is fed into the input layer, this works on based upon a supervised learning methodology .This kind of learning is like initially a face is recognized based on the color, texture and organs, but over a period of time the system automatically identifies the human face, in a similar way it adapts to the environment and learns by example. ANN mainly deals with this kind of classification.

Server State Management: Mohan raj [11], proposed a paper on server sleep state in order to reduce the power consumption in a virtual environment.. The day by day advancements in the cloud environment scenario lead to the tremendous construction of huge infrastructures known as datacenters. But reducing the power consumption has become the at most necessity at this stage which would eventually reduce the operational cost. One main option that could be used to reduce the power is to remove the servers that were remaining IDLE for a longer time .By removing these unused servers, it could help saving the energy that was being wasted for these servers.

Based on the above literary works we bring out a new contemporary methodology, that forecasts the future load based on the reliability record of the available servers, that eventually leads to less energy consumption. This optimizes the resource allocation So that the load could be fairly distributed among the available reliable servers that consumes less power.

System Architecture: System architecture is the visionary representation which helps in defining the entire structure and the behavior of the system completely. It helps in identifying a way so that the products could be solicited, systems could be viewed as an architectural overview of the overall system. The main aim of the proposed system is to forecast the future load in an exclusively optimal way that avoids the wastage of resources and also enables the process of allocating the job to a best reliable server that consumes less power. In order to achieve this objective of resource allocation with less energy utilization we use the following architecture:

Figure 1 represents the entire workflow of the load forecasting architecture. The input dataset is taken from the ganglia monitoring system. These real time load are fed into the Load analyzer, which translates the unstructured data into a structured format based on time series. The load predicting device is responsible for predicting the incoming load and the load analyzing device examines the current performance of all the available cloud resource nodes

Real time load is taken from the ganglia monitoring system, these are converted to the structured data based on time series. These are fed to the load balancer. Server manager ranks the available servers based on their downtime history. Based on this ranking and the predicted load, an optimal virtual machine is allocated.

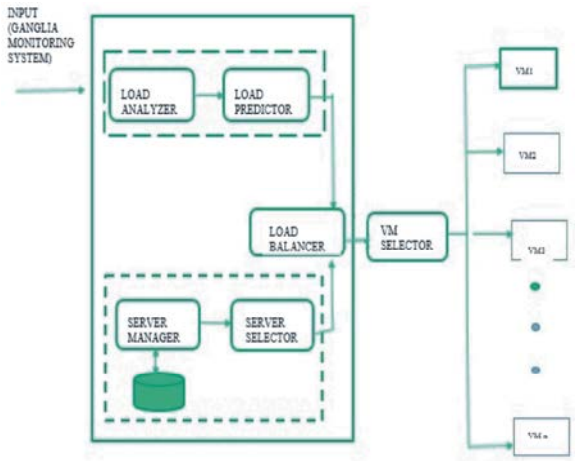

Fig. 1: System architecture

Algorithmic Flow: Figure 2 represents the algorithmic flow of the above process. Here two different algorithm are used, one is the host management and the other is the host management algorithm:

Load Analyzer and Load Predictor: Load Analyzer is one that analyzes the load, it is a major framework process which could be done only after the establishment of the cloud infrastructure. The Load analyzer what it does is, it formulates the unstructured data to a structured format based on the time series. Unstructured dada is nowhere helpful in forecasting the future load, only sequential structured data could help in forecasting the future load.

Load predictor is one which is then used to forecast the future data. Here a neural back propagation methodology is used where there are five input nodes and using which the output is calculated using the process of training the dataset that is typically a supervised learning methodology. Once after the analysis is done the data is formatted to a structured one which enables the quick method of forecasting the future load in an accurate manner.

Supervised learning is the main advantage of this process where learning is done automatically, it reduces the process of feeding the data once and once again every time that is needed. It makes sure it learns by example, leading to the process of avoidance of the wastage of resources..

Server Manager and Server Selector: Ranks all the hosts based on its reliability over the past along with other performance metrics and recommends the best VM for allotting new load from the load balancer.

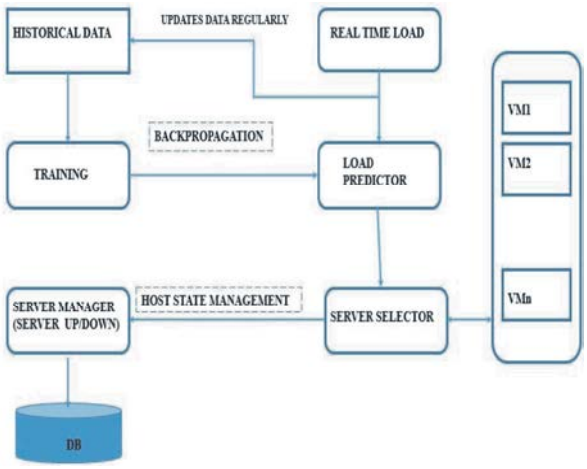

Fig. 2: Algorithmic flow of the system

Load Forecasting: Load forecasting uses Back propagation algorithm to forecast the future load based on the past downtime history. Neural Lab uses the backpropagation algorithm internally and is used as a tool to perform the prediction process.

Host State Management: The host state management algorithm leverages theresultant prediction in managing the host state's decision making process without compromising the commitment towards SLA. Host state manager continuously monitors the health of the hosts including start time, end time and uptime of each and every physical server

Implementation: Here the implementation of the process is done with Eucalyptuscloud, which is an open source software framework that is used to execute the Infrastructure as a Service (IaaS) framework, this makes it possible to run the entire virtual machine for different servers available in the cloud datacenter.

Fast Up Slow Down (FUSD): In the FUSD (Fast Up and Slow Down) kind of methodology the load that could be used in the future are forecasted using its past historic compilations. The load forecasting is carried out using the following equation:

$$E(t) = \propto * E(t-1) + (1-\propto) * O(t),\ 0 \leq \propto \leq 1,$$

Here E(t) expresses the estimated load and O(t) represents the observed load during the time t. α is portrayed as a constant that emulate the tradeoff between the constancy and communion. The pioneer of the FUSD methodology have used it to forecast the load utilized by the CPU on a Domain Name Server. They are used for the estimation of thefuture load each and every minute and forecast them immediately.

This algorithm works properly , if the prediction is in a very sequential order , i.e., if the observed load is O(t) is 20, 30, 40 and 50, then it would be more accurate to predict the next one as 60, it doesn't forecast the intermediate load values. In order to depict the negative values of both the increasing and decreasing order, the above formula is changed as follows for $-1 \leq \propto \leq 0$.

$$E(t) = - |a| * E(t-1) + (1 + |a| * O(t)$$

This depicts two different values for an increasing order and a decreasing order, leading to a confusion to choose the exact value between the two. In order to depict a more accurate forecasting of the load the formula has been slightly changed to.

$$E(t) = m * A(t - 1),$$

Here the estimated load value could be calculated using the actual load with respect to time and m is the multiplier, where the value and its value could be calculated as;

$$m = \frac{A(t-1)}{A(t-2)}$$

This also doesn't forecast the actual future load to be allocated to the VM, it forecasts a very close value nearer to the accurate. The main drawback of using this is it wastes a huge resource to forecast the accurate the one, in order to overcome the wastage of resource and allocate a perfect VM Back propagation could be used as it is based on supervised learning process, it doesn't waste much of the resource.

Neural Predictor: These Neural methodology kind of forecasting istypically used toadminister the indexing of the servers in the cloud datacenter. Based on the downtime history records of the server, the load couldbe forecasted.Back propagation is a process of training or learning the algorithm rather than the network itself. It learns by example , it uses two main functionalities a forward pass and reverse pass , which is used to depict the target in a most accurate way.

$$Y_{n+1} = h((\sum_{i=1}^{n}(X_i + b))/n)$$

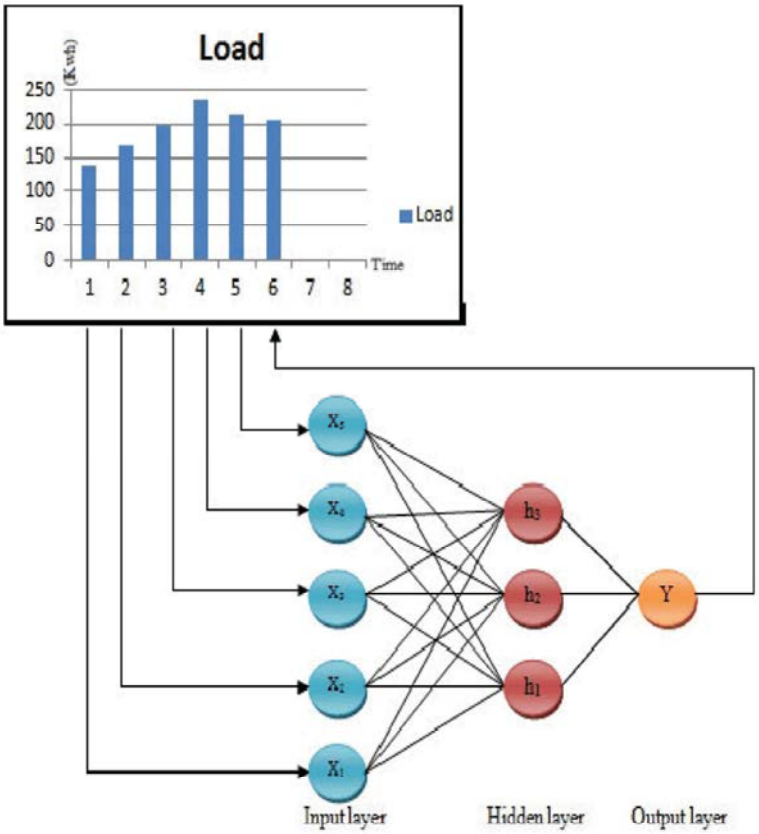

Fig. 3: Back Propagation

Figure 3 represents the back propagation flow in an accurate manner. A multi-layer perceptron archetypal is exclusively constructed withthe back propagationalgorithm for forecasting the future load. This neural frameworkmodel consists of three layers along with five input nodes which isused for executive purpose so as to find out the hidden layer and the output layer.

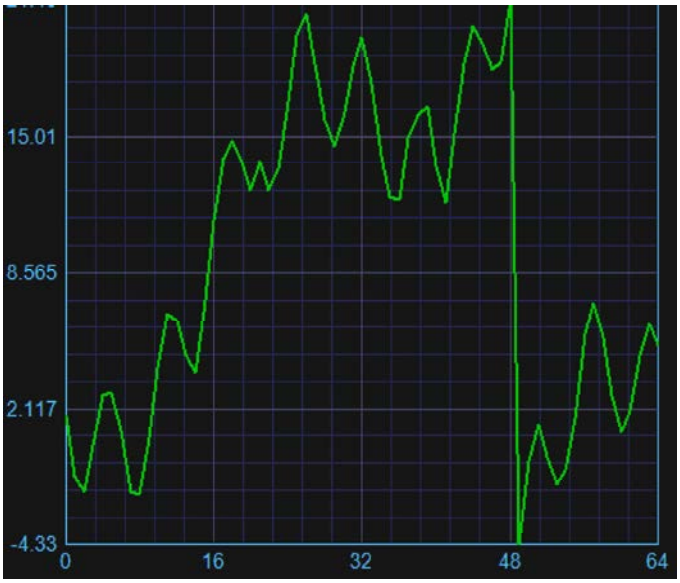

Fig. 4: Sample Load input graph

Figure 4 represents the variations between the loads based on the different time interval. The forecasting of the future load wouldbe measured based on the training undergone by the Perceptrons in a supervised manner.

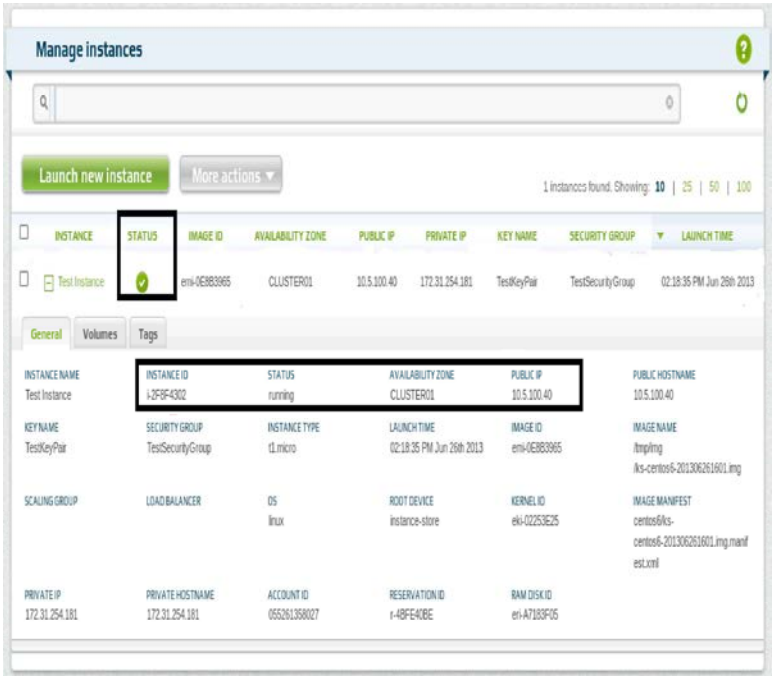

Fig. 5: Instances created using Eucalyptus cloud

Figure 5 represents the instances created using the Eucalyptus cloud. The instances are the storage, memory and CPU utilization, these depict the individual power consumption of the three. Using that we couldn't predict the overall power consumption. So inorder forecast the future load these values are coupled to the ganglia monitoring system, which gives the power consumed by the host totally.

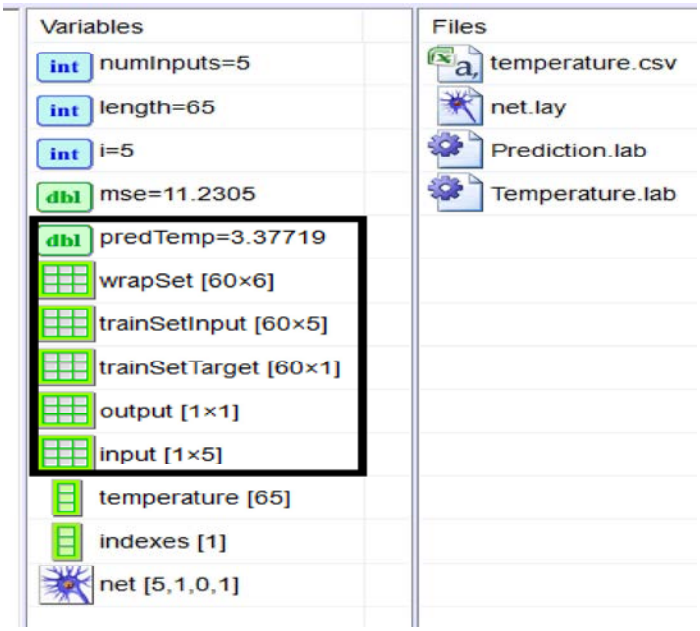

Fig. 6: Forecasted Future Load

Figure 6 represents the forecasted future load. Based on this value a reliable server is allocated to the requested job. This could be viewed using a HAProxy table. This Load forecasting helps in allocating a perfect server to the requested job with a comparatively less power consumption.

Experimental Results: An instance is created using the real time load. Eucalyptus cloud an open source cloud is used to create the real time instances. This produces a table where the power consumed by individual part is obtained. These values are coupled to the ganglia monitoring system so that the whole unit power is calculated. This unstructured power is analyzed by the load analyzer and formatted to a structured format. These values are fed to the neural lab where prediction of the future load is done. Servers are monitored frequently and then ranked based on the response time and throughput. This predicted load and server reliability are used to allot a specific VM that completes the work with a less power. Based on this process a sample dataset is obtained using the ganglia monitoring system based on the time series.

FUSD	BACK PROPAGATION	DEVIATION
10	12.4	2.4
7	8.2	1.2
6.3	7.9	1.6
10.8	11.5	0.7
11.3	12.8	1.5
8.2	10.3	2.1

Fig. 7: Table representing the comparison between the efficiency of load prediction between FUSD and Back propagation.

Figure 7 represents the table that depicts the forecasted load value based on the time series of two different algorithm FUSD and Back propagation. The graph below depicts that the FUSD forecasts a little lesser value always than the actual load whereas Back propagation predicts a more accurate value with lesser amount of resource wastage. FUSD approximately produces a deviation of approximately 1.5 % than the actual value.

Conclusion and Future Work: This paper has put forward a new methodology that advocated to deal with the load forecasting for optimal resource allocation in cloud computing using neural method. Initially the instances were created using the Eucalyptus cloud, where the individual power consumption of CPU, RAM and

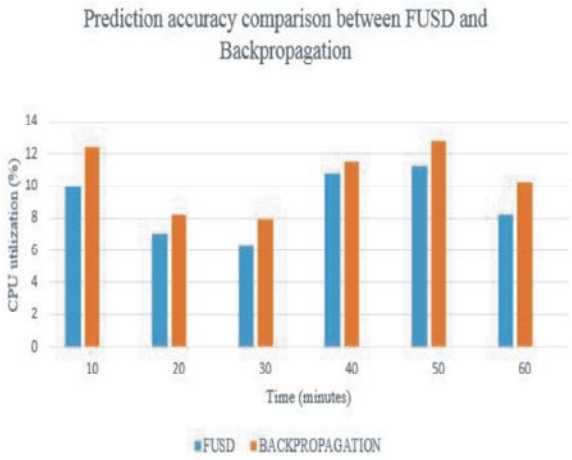

Fig. 8: Performancecomparison between FUSD and Back propagation in Load forecasting.

Storage are obtained, which is not that useful for load forecasting. So these values are coupled with the ganglia monitoring system and the power consumed by the host totally is obtained. Then using the neural lab the future load is forecasted, then based on this forecasted value a reliable server is allocated to the requested job.

The accustomed system deployment is being assessed for its pursuance. Sagacity in the neural framework architecture could be incorporated so as to improve theforecasting accuracy.

REFERENCES

1. Ali Yadavar Nikravesh, Samuel A. Ajila, Chung-Horng Lung, 2015. 'Towards an autonomic auto-scaling Prediction System for Cloud Resource Provisioning'. International Symposium on software engineering for adaptive and self- managing systems, pp: 23-31.
2. Lorido-Botran, T., J. Miguel-Alonso and J.A. Lozano, 2014. "A Review of Auto-scaling Techniques for Elastic Applications in Cloud.
3. Merrill Lynch, 2008. 'The Cloud Wars: $100+ billion at stake. Merrill Lynch Research Note', pp: 118-127.
4. Seyed Mohammad Ghoreyshi, 2013. 'Energy Efficient Resource Management of Cloud Datacentres under Fault Tolerance constraints'. IEEE, 5: 671-690. Environments," Journal of Grid Computing, 12(4).
5. Nikravesh, A.Y., S.A. Ajila and C.H. Lung, 2014. "Cloud resourceautoscaling system based on Hidden Markov Model (HMM)", Proc. of the 8th IEEE International Conference on Semantic Computing.

6. Duy, T.V.T., Y. Sato and Y. Inoguchi, 2011. ”A Prediction-Based Green Scheduler for Datacenters in Clouds.” IEICE Transactions on Information and Systems, Vol. E94-D, No. 9, pp: 1731-1741.
7. Ajila, S.A. and A.A. Bankole, 2013, “Cloud client prediction modelsusing machine learning techniques,” Proc. of the IEEE 37th Computer Software and Application Conference.
8. Von Laszewski, G., L. Wang, A.J. Younge and X. He, 2009. "Power-Aware Scheduling of Virtual Machines in DVFS-enabledClusters".Paper presented at the Proc. of IEEE InternationalConference on Cluster Computing 2009, New Orleans, LA, USA.
9. Chen, X., X. Liu, Z. Huang and H. Sun Regionknn, 2010. A scalable hybrid collaborative filtering algorithm for personalized cloud service recommendation””. In Proc. 8th Int'l Conf. cloud Services (ICWS'10), pp: 9-16.
10. Kara, M. Acar Boyacioglu and O.K. Baykan, 2011. “Predicting direction of stock price indexmovement using artificial neural networks andsupport vector machines: The sample of theIstanbul stock exchange, Expert Systems with Applications”, 38(5): 5311-5319.
11. Mohan Raj, V.K. and R. Shrimm, 2012. ”A Study on Server SleepState Transition to Reduce Power Consumption in a VirtualizedServerCluster Environment’’, Communication Systems and Networks, 4<'International Conference.
12. Abu Sharkh, M., M. Jammal, A. Shami and A. Ouda, 2012. “Resource Allocation in a Network-Based Cloud Computing Environment: Design Challenges”. IEEE Communications Magazine.
13. Taleb, T. and A. Ksentini, 2013. “Follow me cloud: interworking federated cloudsand distributed mobile networks.,” IEEE Network, 27(5): 12-19.
14. Zhen Xiao, Weijia Song and Qi Chen, 2013.” Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment”. Ieee Transactions on Parallel and Distributed Systems, 24(6): 5-7.
15. Nelson, M., B.H. Lim and G. Hutchins, 2005. “Fast Transparent Migration for Virtual Machines,” Proc. USENIX Ann. Technical Conf.
16. Caron, E., F. Desprez and A. Muresan, 2010. "Forecasting for Gri and Cloud Computing On-Demand Resources Based on Pattern Matching, " 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), IEEE, pp: 456-463.
17. Ismail, Z. and R. Efendi, 2011. "Enrollment Forecasting basedon Modified Weight Fuzzy Time Series, " Journal ofArtificial Intelligence, 4: 110-118.
18. Yong Wang, Dawu Gu, 2009. “Back Propagat ion Neural Networkfor Short -term Elect ricity Load Forecast ing with WeatherFeatures”, Internat ional Conference on Computat ional Intelligenceand Natural Computing.
19. Turchenko, V., P. Beraldi, F. De Simone and L. Grandinetti, 2011. “ Short-term stock priceprediction using mlp in moving simulationmode”, in Proceedings of the IEEE 6thInternational Conference on Intelligent DataAcquisition and Advanced Computing Systems(IDAACS), 2: 666-671.
20. Aarti Singh, Manisha Malhotra, 2012. ”Agent Based Framework for Scalability in Cloud Computing”, IJCSET, ISSN : 2229-3345 3(4): 41.