A COMPENDIUM ON LOAD FORECASTING APPROACHES AND MODELS

Sudha Pelluri[1], Puranam Srinivas[2], Soma Sneha[3]

[1]*Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, India*
[2]*Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, India*
[3]*Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, India*

Abstract

Internet Application Developers are provided with great opportunities by the advancements in the field of Cloud Computing. To the users it is a necessity to get all resources without any time delay and within specific cost constraints. Now the challenge of providing resources to accommodate the large demand of ever-growing Cloud users is to be met by the Cloud Providers. Now, the competition among providers has progressed to such a stage where the load forecasting and dynamic resource allocation are among the major concerns to them. This paper illustrates the need for proper prediction mechanisms and the types of load forecasting mechanisms. A number of dominant working models are discussed in this paper. The analysis of every method belonging to a particular model is explained in detail.

Keywords—*Cloud, load forecast, time series, models.*

----------***----------

1. INTRODUCTION

The current challenges in technology have made users choose cloud computing. Cloud computing relies on two unique notions, pay per use and dynamic resource provisioning. The increase in demand has thrown new challenges to the cloud service providers like security, load forecasting, dynamic resource allocation and cost reduction. This paper discusses the approaches in Load Forecasting. It refers to predicting the future load on a machine. For example, for a web application, load forecasting implies predicting the number of requests.

Planning capacity for drastically changing workloads is a very difficult task. However capacity could be planned for an average workload or for peak load which is shown in the Fig. 1 Nonetheless each method has its own disadvantages. There is less cost incurred because of the less hardware used when planned for the average load. But the performance will suffer when peak load occurs. [1]

Auto Scaling is opted by most of the cloud providers as it enables overcome the difficulties of traditional approach of the provider having to respond to users request for resources explicitly. Auto-scaling techniques can be classified as either reactive (the system reacts to changes but does not anticipate them) or predictive (the system tries to predict future resource requirements in order to ensure sufficient resource availability ahead of time). Reactive Rule-based methods define scaling conditions based on a target metric reaching some threshold and are offered by several cloud providers such as Amazon or third party tools like RightScale and AzureWatch. Predictive auto scaling occurs as a two step procedure. One is the prediction and the second step is scaling. The scaling is done based on the prediction which is done using time series analysis, control theory, reinforcement learning or Queuing theory approaches. It has well been proved that without auto scaling, the overhead in maintaining resources is a major challenge. As the frequency of change in workload is high in a cloud, resources must be continuously allocated and relieved, which obviously leads to thrashing. So an appropriate solution would be to predict the load in future and allocate resources based on the prediction. For illustration, consider that N machines are serving M customers. Now the number of customers has been increased to M+1000 and processor utilization also increases in systems. Then number of machines must be increased. This requires Auto scaling in minimum time.

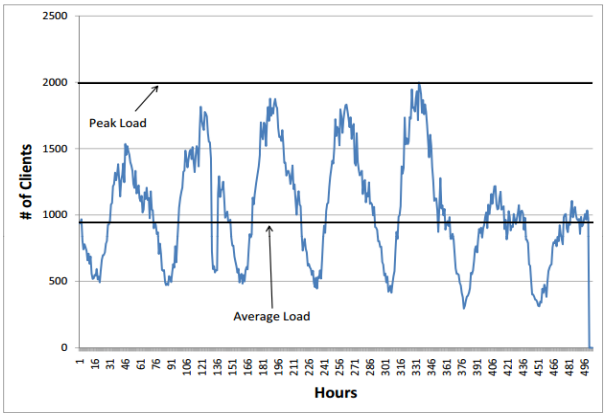

Fig 1 Peak and Average Workload [1]

2. CHALLENGES

The challenges in elastic provisioning of resources are as follows:

2.1 Workload Forecasting

Because of the high degree of heterogeneity and dynamism in the workloads, one would expect that it would be difficult to achieve accurate visibility into future resource demands. Hence, there is a demand on resource schedulers to refer to more sophisticated time based models of resource usage. The resource allocation must be changed continuously based on the workloads. At a given time, the number of resources which are allocated must be optimal even for workloads with some burstiness.

2.2 Identify Resource Requirement for Incoming Load

The resource requirement largely depends on number of customers, nature of application being used and also on type of interaction with the application the customer has. Thus the resource estimation should aptly satisfy the demand and must be provisioned within the stated Service Level Agreements (SLAs). Users lack tools using which they identify accurate requests and users have less incentives to make accurate requests [2]. So it is important to closely identify the requests which users actually need based on the Job/ task requirements.

2.3 Resource Allocation while Optimizing Multiple SLA Factors like Deadlines and Cost

There is a need to optimize the resource allocation efficiency by introducing metrics in order to satisfy the Service Level Agreements. Based on the metrics, system has to make decisions that allow users to release resources or acquire them. [3] Creating and removing resources requires programming. Acquiring resources must commensurate with time. To acquire resource, a call to API must be made, machine must then boot with specific image, application hosted must then run and also if required there must be status update. All these must be done within a less amount of time in order not to have performance decline. Further, customers tend to overestimate the resources to be used to ensure good performance. This causes overall underutilization and reduces the capacity.[4] This is called over-provisioning. Under-provisioning causes SLO violations which lead to financial penalties. [5]Over and under provisioning is discouraged.

Deadlines are important criteria in enabling task completion for various users. Dynamic specification of deadlines is a special feature in cloud, (users enter the cloud system dynamically and can give specification only then) which makes it a challenging opportunity for researchers. Static algorithms like shortest job, Round robin, priority etc., are not useful in resource allocation to meet the SLAs.

3. TYPES OF LOAD FORECASTING MECHANISMS

Load forecasting can be categorized based on time horizon. The classification is as follows: Short Term Load Forecasting, Medium Term Load Forecasting and Long Term Load Forecasting. Short Term Load Forecasting is used to predict a load up to a week ahead. Medium term Load Forecasting ranges from a week to a year. Long Term Load Forecasting predicts load more than a year ahead. To demonstrate different types of forecasting, consider electric power system operations. In this field, Long Term Load Forecasting is needed for the determination of prices and regulatory policies. Further, Medium Term Load forecasting is used for maintenance scheduling, coordination of sharing arrangements and fixing prices. In addition to that, the Short Term Load Forecasting is used for economic scheduling of generating capacity, scheduling of fuel purchases, security analysis and short term maintenance scheduling. [6]

Short Term Load Forecasting can be done using approaches like time series analysis, regression, exponential smoothing, state space analysis, pattern matching, neural networks, spectral networks and many more. This paper summarizes some of the above approaches.

To find the patterns in the data for modeling, there are several proposed methods.

- Pattern Matching: It refers to finding a similar pattern which is available using past time series which are similar to present pattern.
- Signal Processing: Using a Fast Fourier Transform, higher frequency pattern is determined which represents most similar pattern.
- Auto-correlation: The input series is shifted repeatedly and correlation is calculated between original and shifted series. If correlation crosses a threshold after x shifts then a repeated pattern is identified with duration x steps.

4. APPROACHES AND MODELS

Statistical approaches are based on a vivid mathematical model which expresses the relationship between load and several input parameters. These models are very common in determining the load forecast and are also used to depict the relationship between the load and external factors. Thus predictions are done using any of the below approaches

A. Time Series Analysis
B. Linear Regression
C. Queuing models
D. Exponential/ Double exponential smoothing
E. Neural Networks/ Artificial NN
F. Support vector machines
G. QSRM
H. Markov Chain Prediction (MC)

4.1 Time Series Analysis

In time series analysis, future values are predicted based on previously recorded values with an assumption that future will resemble the past. Time Series data consists of certain components such as trend, seasonal effect, cyclical and irregular effect. This is done mathematically with the help of certain models. They are univariate or multivariate. Univariate is used for short term forecast while multivariate can be used for any forecasts. In order to carry out time series analysis, the input required is a list of previous observations which is called as input or history window. The accuracy of any prediction depends on minimum error during modeling. Time Series modeling consists of two steps namely building a model and using model for prediction. The models used are Auto Regressive (AR) model, Moving Average (MA) model, ARMA model, ARIMA model, ARMAX model.[7]

Some important patterns that may be identified in time series analysis are

- Trend: Data with steady growth or fall over time.
- Seasonality: Data with upward or downward swing in short to intermediate time.
- Cycle: Data with upward or downward swing in long time.
- Random variations: Unpredicted changes in data over time with no appropriate pattern.[8]

In time series analysis, if the variable is chosen to be y, then y_t is its current value and y_{t-1} is the value of previous observation. The value of observation except the current one are called lags. Time series function will be similar to follows:

$$y_t = f(y_{t-1}\ldots.y_{t-n}) + e_t \qquad (1)$$

where e_t is the random shock or noise.

1) *Averaging Methods:* This method can be used for smoothing time series in order to remove noise. The future value is predicted as weighted average of previous consecutive observations. Based on the way of calculating weights, several methods are determined.

a) Moving Average MA(q): MA(q) is arithmetic mean of last q observed values. In this model, the current value is expressed as linear combination of q previous shocks.

$$y_t = b_t - lb_{t-1} - \ldots - l_q b_{t-q} \qquad (2)$$

$$l(B) = 1 - l_1 B - l_2 B^2 - \ldots.. - l_p B^p \qquad (3)$$

For an MA(1) model,

- Acf: cuts off abruptly after lag 1.
- Pacf: declines in geometric progression from its highest value at lag 1.

b) Weighted Moving Average WMA(q): In WMA(q), weights are not uniform. A recent observation has more weight.

c) Exponential Smoothing: The weights are given exponentially which decrease over time to make recent data more influential for forecast [9]. A parameter called smoothing factor is also introduced to lessen the effect of past data. Exponential smoothing does not totally exclude past data but gives them a weaker influence. [10]

$$S_t = \alpha \cdot y_t + (1-\alpha) S_{t-1} \qquad (4)$$

d) Double Exponential Smoothing: Double Exponential Smoothing is exponential smoothing of singly exponential smoothing. It is suitable for time series with linear trend. The prediction formula is as follows.[10]

$$y_{t+m} = (2y_t' - y_t) + m(y_t' - y_t)\alpha/(1-\alpha) \qquad (5)$$

2) *Auto Regression Methods AR(p):* The weights are determined using autocorrelation coefficients and solving linear equation. In Autoregressive model, the current value is expressed as linear combination of p previous values and a shock. It is denoted by AR (p).

$$y_t = a_1 y_{t-1} + \ldots + a_t y_{t-p} + b_t \qquad (6)$$

where b_t is the shock.

Introducing another term to simplifying the above, we get

$$B\, y_t = y_{t-1} \qquad (7)$$

$$\text{Similarly, } B^x y_t = y_{t-m} \qquad (8)$$

$$a(B) = 1 - a_1 B - a_2 B^2 - \ldots.. - a_p B^p \qquad (9)$$

For an AR(1) model,

- Acf: declines in geometric progression from its highest value at lag 1.
- Pacf: cuts off abruptly after lag 1. [6]

The above behavior is shown in Fig. 2.

3) *Auto Regression Moving Average ARMA(p,q):* Auto Regressive Moving Average (ARMA) is a combination of both AR and MA models. It is denoted by ARMA (p, q).

$$y_t - a_1 y_{t-1} - \ldots - a_t y_{t-p} = b_t - lb_{t-1} - \ldots - l_q b_{t-q} \qquad (10)$$

ARMA models are used to model stationary time series. A stationary time series is the one which has a constant mean and variance over time period. Their limitation is that they

can be used to model only data in a linear fashion. ARMA modeling can be carried out in a sequence of steps.

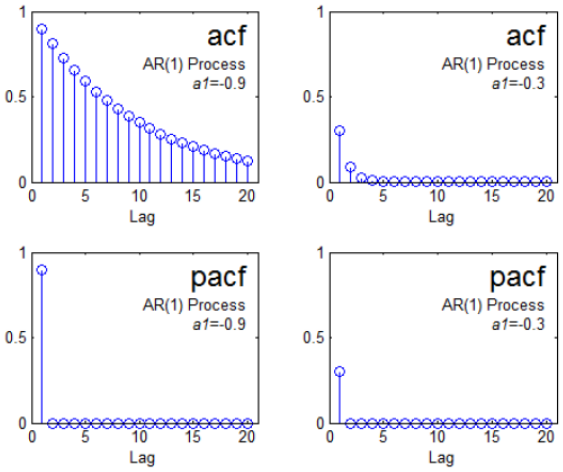

Fig 2 Acf and pacf of AR(1) [6]

- Identification: Determine the model and its order.
- Estimation: Coefficients of model must be estimated.
- Check: Confirm that parameters are appropriate and the noise is random.

In identification phase, the structure of model is determined using acf and pacf graphs. acf is plotted with auto-correlations versus lag i.e., acf at lag x measures correlation of the time series with itself lagged x years. pacf is plotted with partial correlations versus lag. At lag x, it is autocorrelation at lag x after first removing autocorrelation with AR(x-1) model. After determining the structure of the model, the parameters are approximated using maximum likelihood. The residuals are then calculated. The sequence of residuals should be white. Else, check acf to recommend some inclusions.

Table 1 [11]

Process	ACF	PACF
AR(p)	Tails off exponential decay or damped sine wave	Cuts off after lag p
MA(q)	Cuts off after lag q	Tails off exponential decay or damped sine wave
ARMA(p,q)	Tails off after lag (q-p)	Tails off after lag (p-q)

The aim of check phase is to check whether the chosen model is closer to the reality and serves the purpose of prediction. For a model to satisfy above condition firstly the residuals should be chosen random and acf of residuals should always be zero except at lag zero. Secondly, estimated coefficients must be different from zero. The model must be stable and reversible. [12]

Pre-whitening refers to removal of autocorrelation from a time series before using time series in any application. An unexpected or exceptional occurrence of an event during time series is called intervention. Examples are recession in IT market, labor strikes. [13]

4) *Auto Regression Integrated Moving Average ARIMA (p,d,q):* ARMA model is not suitable for the time series which have trend. So in order to remove trend, differencing is done. The differencing is integrated into ARMA forming ARIMA, Auto Regressive Integrated Moving Average model. It is denoted by ARIMA (p, d, q) where p is AR order, q is MA order and d is differencing order. Univariate ARIMA models are used as benchmarks for Short Term Load Forecasting. There are certain steps to build ARIMA (p, q, d) models.

Step 1: Plot the available data.
Step 2: Check if the data is stationary that is they are scattered about a constant mean level. Also check acf and pacf plots for stationarity (if acf, pacf drop quickly to zero, then it is stationary).
Step 3: If there is no stationarity that is if trend is observed, then difference the data until it becomes stationary.
Step 4: Once stationarity is obtained, look at acf and pacf plots. Check against theoretical behavior of MA and AR to see if they fit. This gives ARIMA either with no MA or with no AR that is either ARIMA (p, d, 0) or ARIMA (0, d, q) respectively.
Step 5: If there is no MA or AR behavior, then ARIMA is built. Appropriate order is chosen for building the model.

5) *Auto Regression Moving Average with exogenous input (ARMAX):* ARMAX model is Auto Regressive Moving Average with extra/external input. It is denoted by ARMAX ([x, y, z], p, q, r) where x is input signal, y is output signal, z is extra input signal, p and q are order of AR and MA respectively and r is order of external process. In this approach, the load series is patterned as the output which is generated from a linear filter that has a white-noise input and exogenous input series. [14] The former being inaccessible and the latter being accessible. By this approach, the characteristics of the linear filter are recognized based on the output and the input series. Also, the model order and its parameters are identified. Usually with the above approach the model based on the rule of mean square forecasting error is searched. But technically, the gradient based method used by the STS method is used to generate suboptimum models for forecasting error function which has many local optimum points [15] . For example, Network Control System (NCS) are the feedback control systems which use

network as a real time medium between controlled object and controller signal transmission. NCS are used in industrial processes, because of the advantages like wide open, low cost, high flexibility, etc. In the network control system, the communication bandwidth is shared by multiple control loops, which control circuit competing communication resources [16] as shown in *Fig. 3*.

4.2 Neural Networks

Neural Networks consist of set of neurons which are distributed across different layers. The input layer has one neuron for one value of past observation and one neuron for predicted value. The other layers are output layer and in between input and output layers are the hidden layers. At the training phase, input and the weights are fed into neural network and desired output is drawn at a learning rate of *p.* For time series analysis, NN captures temporal patterns in data in the form of past memory associated with the model and then uses this to know the future behavior.[17] NN are applied majorly in modeling non linear relationship among data without any prior assumption of relationship. This is an advantage of NN over ARIMA. [18]

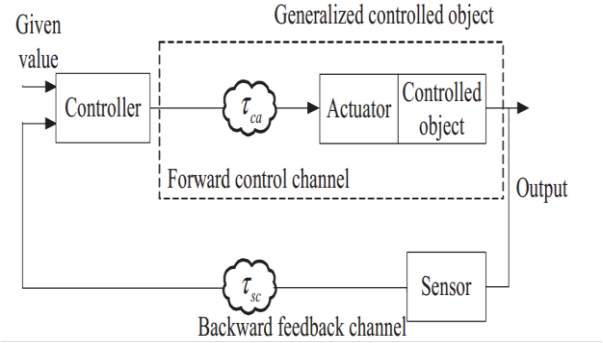

Fig 3 Network Control System [16]

4.3 Support Vector Machines

SVM are supervised learning method used for pattern recognition, classification and regression analysis.SVM can be used even when there are less examples and huge number of attributes.[19] It lowers the problem of over-fitting or local minima. It is based on structural risk minimization but not on empirical risk minimization of neural networks. It finds its use in Regression and is termed as Support Vector Regression (SVR). [17]

4.4 Predictive Bayesian Network

A Bayesian Network is a probabilistic network of history records that shows conditional independence between variables using a Directed Acyclic Graph (DAG). Each node represents a random variable and the edges represent the probabilistic relationship among nodes. BN can be constructed by structural development to identify the structure containing nodes and edges and then parameter learning to know the dependencies between variables. [20]

4.5 Markov Chain Prediction

This is based on the assumption that user demand is ad-hoc. It consists of finite number of states where in future state depends only on present state and is independent of any other previous states. The summary of algorithm is as follows.

- Check the history and divide them into states.
- Create a sequence of states based on the history
- Create probability matrix for each sequence.
- Predict the future state using these matrices.[9]

Given a wide number of models, it is difficult to decide which model best fits the data. Context and objectives are key factors to determine which is the best model . We can use either local models with changing parameters (like state space model) or global models with constant parameters (like ARIMA). There might also be a chance to use more than one model to meet a given objective. Furthermore, different models can be used for different subset of data and different models can be used for different time horizon. [21]

5. COMPARISON

An absolute comparison of the various models is not possible because of the inherent differences in features and requirements of the different systems at different systems. An MA method will have poor results in cloud environment because of the non static nature of the history data . So authors just use this to remove noise or for a basic prediction. Exponential smoothing shows better results comparatively with MA and WMA since it uses current as well as previous data. AR method's performance depends on monitoring-interval length, size of history window and size of adaptation window. History window determines sensitivity of algorithm to local versus global trends and adaptation window shows how far the model applies in future. ARMA predicted values are used to find response time. An optimizer controller takes this value as input and does resource allocation using SLO violations, cost and reconfiguration cost. In neural networks, history window can be used as input. Pattern Matching has two drawbacks: large number of parameters and time required to find past history.

5.1 SVR versus NN versus LR

A research was carried out to evaluate the accuracy of selected machine techniques in forecasting resource usage and to integrate SLA into model. CPU utilization prediction model and SLA model are used to achieve the result [17] . Using the Java implementation of TPC-W online bookshop, and deploying the application on a two-tier architecture, the metrics were employed for response time and throughput, the following responses were obtained as in Fig.4, 5, 6.

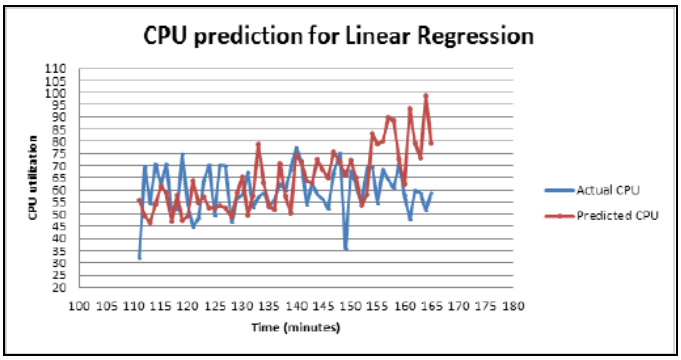

Fig 4 Linear Regression Prediction [17]

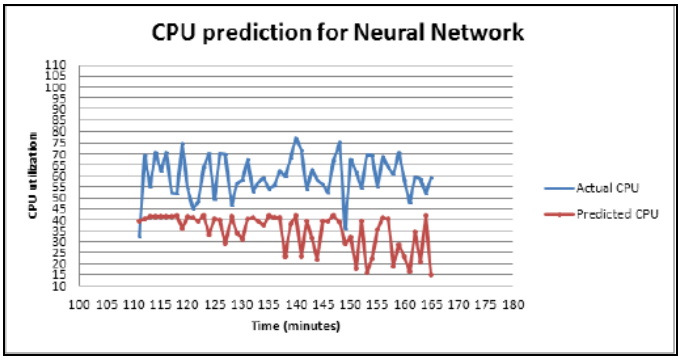

Fig 5 Neural Network Prediction [17]

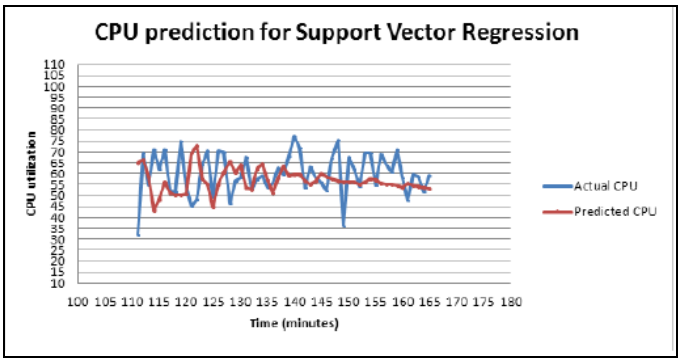

Fig.6 SVM Prediction [17]

From the graphs as shown above the authors have concluded that the SVR approach is most responsive to random or nonlinear user request pattern, thus displaying a strong generalization property and excellent prediction consistency .

5.2 ARIMA versus ARMA

A research was conducted to find a model to efficiently forecast electricity consumption in a household by applying Box and Jenkins. The analysis was conducted as daily, monthly, weekly and quarterly, that is, a short term analysis. The proposed forecasting time series steps are as in the Fig. 4.

6) *Data Processing:* The data set used has a sampling rate in one minute over a long period of time from years 2006 to 2010. In data processing, there are two steps:
- Fill any missing data.
- Make data into suitable format.

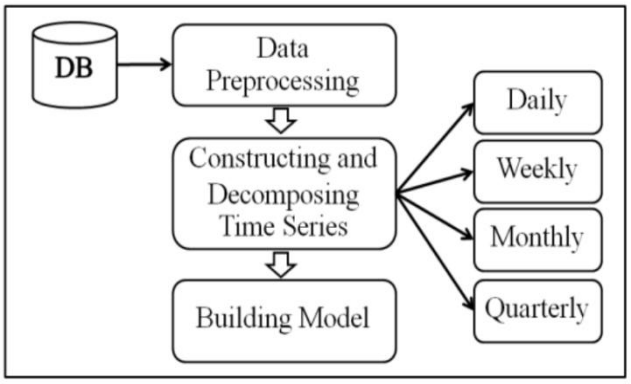

Fig 7 Steps for Time Series Forecast [22]

7) *Constructing and Decomposing Time Series:* Data series which is converted to required format is used to construct time series and decompose it as shown in the Fig. 8- 11.

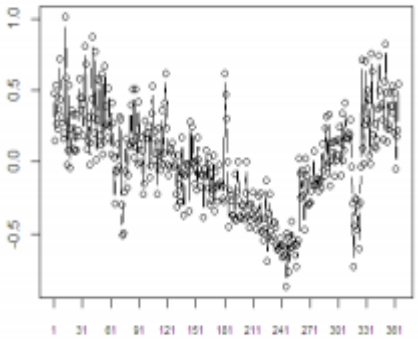

Fig 8 Seasonal Component of day

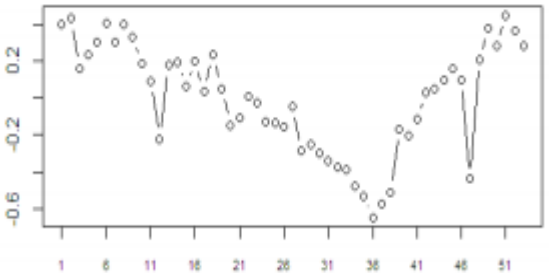

Fig.9 Seasonal Component of week

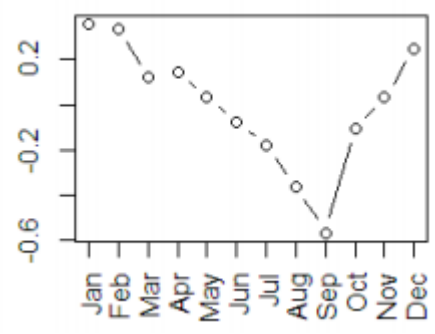

Fig.10 Seasonal Component of month

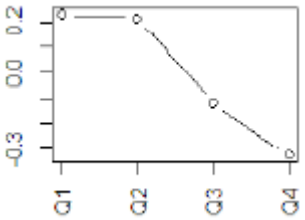

Fig 11 Seasonal Component of Quarter

8) *Building Models:* The models which are constructed are ARIMA and ARMA.

a) ARIMA: The steps for constructing ARIMA are

- Determine order that is p, d and q. Suitable orders which are determined are ARIMA(3,1,3), ARIMA(1,0,1), ARIMA(0,0,0) and ARIMA(0,0,0) respectively for daily, weekly, monthly and yearly data.
- Respective orders were used to construct model as shown in the Fig. 12- 15.

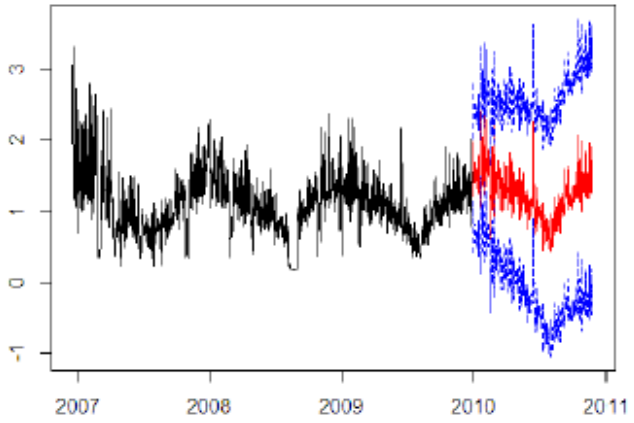

Fig 12 Daily Times series forecasting

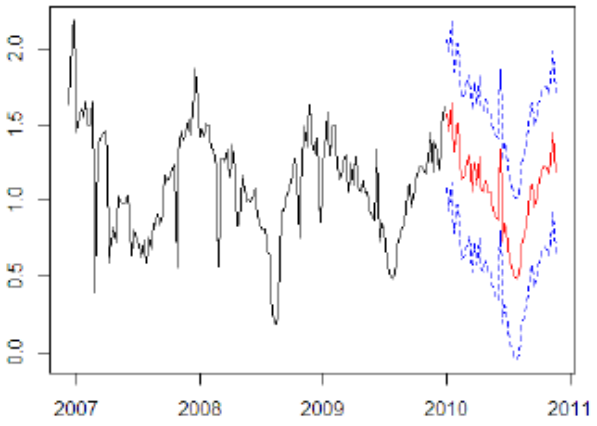

Fig 13 Weekly Times series forecasting

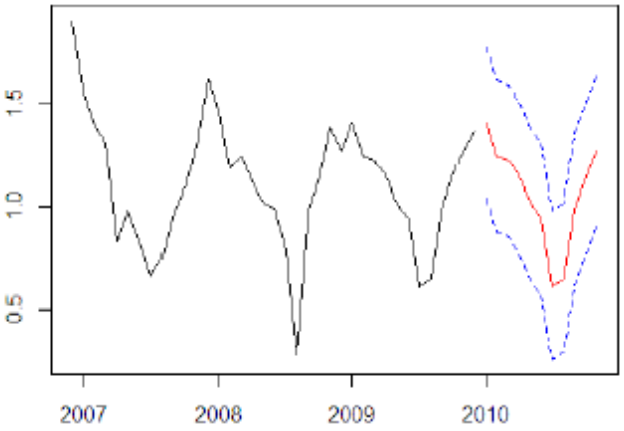

Fig 14 Monthly Times series forecasting

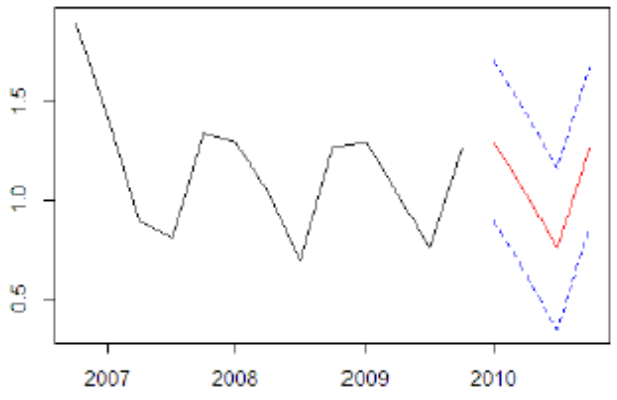

Fig.15 Quarterly Times series forecasting

b) ARMA: Appropriate orders required to construct ARIMA in step 1 above was used to construct ARMA (p,q).[22] as shown in the Fig. 16- 19.

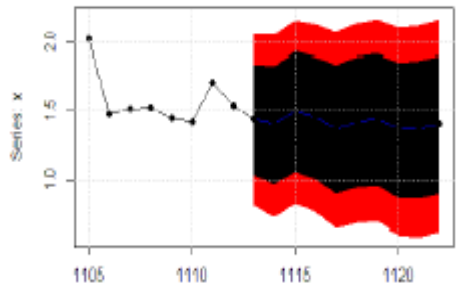

Fig 16 Daily Times series forecasting model

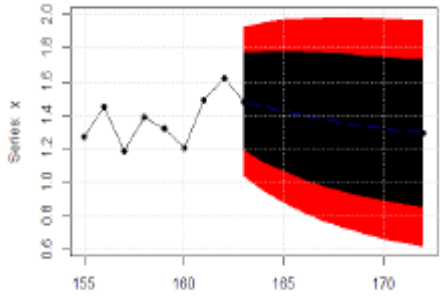

Fig 17 Weekly Times series forecasting model

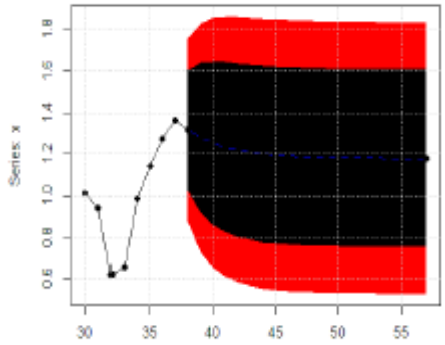

Fig 18 Monthly Times series forecasting model

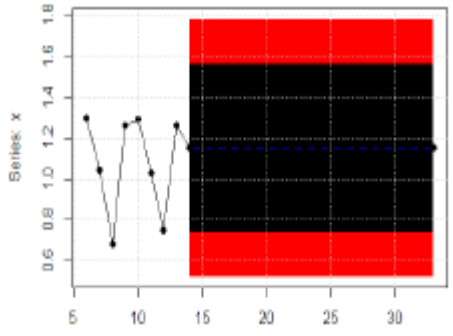

Fig 19 Quarterly Times series forecasting model

The comparison of performances between ARMA and ARIMA models using the AIC measurement.

Table 2 [22]

	ARMA	ARIMA
Daily	615.33	771.54
Weekly	-11.19	6.94
Monthly	1.64	-11.91
Quarterly	11.37	-0.98

6. ERROR MEASURES

There exist some metrics to measure the accuracy of the fitted models, generated from different algorithms. Some of the common metrics are –

1. MAPE (Mean Absolute Percentage Error): It represents the accuracy of values of time series in terms of percentage.

$$MAPE = \frac{\sum_{t=1}^{n} |(Y_t - \widehat{Y}_t / Y_t)|}{n} \cdot 100. \quad (11)$$

2. MAD (Mean Absolute Deviation): It indicates the size of error and is represented as

$$MAD = \frac{\sum_{t=1}^{n} |(Y_t - \widehat{Y}_t)|}{n} \quad (12)$$

3. MSD(Mean Squared Deviation): It is useful in large forecasts as it is more sensitive than MAD.

$$MSD = \frac{\sum_{t=1}^{n} |(Y_t - \widehat{Y}_t)|^2}{n} \quad (13)$$

4. R^2 Prediction Accuracy: It is a measure of the goodness of fit of the prediction model .

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\ddot{y}_i - y_i^2)}{\sum_{i=1}^{n} (y_i - \ddot{y})}$$

Where $\ddot{y} = = \frac{1}{n} \sum_{i=1}^{n} y_i$ and y_i is the actual output ,

$\ddot{y}_i$ is the predicted output and n is the number of observations in the data set. An R^2 prediction accuracy of 1.0 indicates that the forecasting model is a perfect fit.

In the presence of these metrics, it is easy for the users to now employ the different models in forecasting and identify the best model to use. With the best model future requirements can easily be made available for predictive scaling.

7. CONCLUSIONS

With the rapid advances of technologies, there is a demand to analyze the benefits of the different mechanisms for load forecasting in Cloud Computing. A brief description about the models and the approaches of load forecasting essential for dynamic resource allocation has been put forward. Many models with their specific implementation approaches have been discussed. This helps a majority of developers to identify trending techniques for forecasting that will increase the efficiency of the emerging Cloud technology development

REFERENCES

[1] Nilabja Roy, Abhishek Dubey and Aniruddha Gokhale, "Efficient Autoscaling in the Cloud Using Predictive Models for Workload Forecasting", 2013 IEEE Sixth International Conference on Cloud Computing, vol. 0, pp. 500-507, 2011.

[2] Reiss, Charles and Tumanov, Alexey and Ganger, Gregory R and Katz, Randy H and Kozuch, Michael A, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis", Proceedings of the Third ACM Symposium on Cloud Computing, pp. 7, 2012.

[3] Rao, Jia and Bu, Xiangping and Wang, Kun and Xu, Cheng-Zhong, "Self-adaptive provisioning of virtualized resources in cloud computing", Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems, pp.129-130, 2011.

[4] Moreno, Ismael Solis and Xu Jie, "Customer-aware resource overallocation to improve energy efficiency in realtime cloud computing data centers", Service-Oriented Computing and Applications (SOCA), 2011 IEEE International Conference on, pp. 1-8, 2011.

[5] Gong, Zhenhuan, Gu, Xiaohui, Wilkes and John, "PRESS : Predictive elastic resource scaling for cloud systems", Network and Service Management (CNSM), 2010 International Conference on, pp. 9-16, 2010.

[6] Hagan, Martin T and Behr, Suzanne M, "The time series approach to short term load forecasting", Power Systems, IEEE Transactions on, vol. 2, pp. 785-791, 1987.

[7] Hahn, Heiko and Meyer-Nieberg, Silja and Pickl, Stefan, "Electric load forecasting methods: Tools for decision making", European Journal of Operational Research, vol. 3, pp. 902-907, 2009.

[8] Victor, Ogunoh Arinze and Daniel, Ezeliora Chukwuemeka and Chuka, Chinwuko Emmanuel and others, "Regression and Time Series Analysis of Petroleum Product Sales in Masters Energy oil and Gas", International Journal of Science, Engineering and Technology Research, vol. 2, pp. 1264, 2013.

[9] Mark, Ching Chuen Teck and Niyato, Dusit and Chen-Khong, Tham, "Evolutionary optimal virtual machine placement and demand forecaster for cloud computing", Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on, pp. 348-355, 2011.

[10] Huang, Jinhui and Li, Chunlin and Yu, Jie, "Resource prediction based on double exponential smoothing in cloud computing", Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on, pp. 2056-2060, 2012.

[11] Rosadi, D and others, "New procedure for determining order of subset autoregressive integrated moving average (ARIMA) based on over-fitting concept", Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on, pp. 1-5, 2012.

[12] Fang, Wei and Lu, ZhiHui and Wu, Jie and Cao, ZhenYin, "Rpps: A novel resource prediction and provisioning scheme in cloud data center", Services Computing (SCC), 2012 IEEE Ninth International Conference on, pp. 609-616, 2012.

[13] Web site http://www.ltrr.arizona.edu/~dmeko/notes_5.pdf as accessed on 4th March, 2013.

[14] Huang, Chao-Ming and Huang, Chi-Jen and Wang, Ming-Li, "A particle swarm optimization to identifying the ARMAX model for short-term load forecasting", Power Systems, IEEE Transactions on, vol. 20, pp. 1126-1133, 2005.

[15] Yang, Hong-Tzer and Huang, Chao-Ming and Huang, Ching-Lien, "Identification of ARMAX model for short term load forecasting: an evolutionary programming approach", Power Industry Computer Application Conference, 1995. Conference Proceedings, pp. 325-330, 1995.

[16] Liu, Sheng and Jing, Qi, "Research on ARMAX model generalized predictive control of online delay", Electrical and Control Engineering (ICECE), 2011 International Conference on, pp. 4670-4673, 2011.

[17] Bankole, Akindele A and Ajila, Samuel A, "Predicting cloud resource provisioning using machine learning techniques", Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on, pp. 1-4, 2013.

[18] Aranildo Rodrigues, LJ and de Mattos Neto, Paulo SG and Ferreira, T, "A prime step in the time series forecasting with hybrid methods: The fitness function choice", Neural Networks, 2009. IJCNN 2009. International Joint Conference on, pp. 2703-2710, 2009.

[19] Niehorster, Oliver and Krieger, Alexander and Simon, Jens and Brinkmann, Andre, "Autonomic resource management with support vector machines", Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing, pp. 157-164, 2011.

[20] Li, Jian and Shuang, Kai and Su, Sen and Huang, Qingjia and Xu, Peng and Cheng, Xiang and Wang, Jie, "Reducing operational costs through consolidation with resource prediction in the cloud", Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on, pp. 793-798, 2012.

[21] Chatfield, Chris, "Model uncertainty", Encyclopedia of Environmetrics, 2006.

[22] Website::http://www.iaeng.org/publication/IMECS2013/IMECS2013_pp295-300.pdf accessed on 4th March, 2013.

[23] Herbst, Nikolas Roman and Huber, Nikolaus and Kounev, Samuel and Amrehn, Erich, "Self-adaptive workload classification and forecasting for proactive resource provisioning", Proceedings of the ACM/SPEC international conference on International conference on performance engineering, pp. 187-198, 2013.

[24] Araujo, Jean and Matos, Rubens and Maciel, Paulo and Vieira, Francisco and Matias, Rivalino and Trivedi, Kishor S, "Software rejuvenation in eucalyptus cloud computing infrastructure: A method based on time series forecasting and multiple thresholds", Software Aging and Rejuvenation (WoSAR), 2011 IEEE Third International Workshop on, pp. 38-43, 2011.