A Survey on Load Prediction Techniques in Cloud Environment

Manjunath C R[1], Manaswini C[2], Nilasini Bangar[3]
School of Engineering and Technology
Jain University, Kanakapura, India

Abstract **--- The rapid growth of power demand from business, and Web applications has led to the emergence of cloud-oriented data centres. Load prediction is a significant cost-optimal resource allocation and energy saving approach for a cloud computing environment. Load classification before prediction is necessary to improve prediction accuracy. In this paper, a novel approach is proposed to forecast the future load for cloud-oriented data centres. First, Bayesian model is used to predict the mean load over a long-term time interval which is compared with PSR and EA-GMDH method which combines the Phase Space Reconstruction (PSR) method and the Group Method of Data Handling (GMDH) for effective prediction then Neural Network predicts the future load based on the past historical data which distinguishes itself with the presence of hidden layers followed by ssupport vector and kalmann smoother which is a multi-step-ahead CPU load prediction method based on Support Vector Regression which is very stable, i.e. its prediction error increases quite slowly as the predicted steps increase.**

Index Terms **- Cloud computing, Load prediction, Prediction accuracy.**

I. INTRODUCTION

Cloud computing is a term used to refer to a model of network computing where a program or application runs on a connected server or servers rather than on a local computing device such as a PC, tablet or smartphone. Service delivery in Cloud Computing comprises three different service models, namely Infrastructure-as -a-Service(IaaS),Platform-as-a-Service (PaaS) and Software-as-Service (SaaS). Software-as-a-Service provides complete applications to a cloud's end user. It is mainly accessed through a web portal and service oriented architectures based on web service technologies. Platform-as-a-service comprises the environment for developing and provisioning cloud applications. The principal users of this layer are developers seeking to develop and run a cloud application for a particular platform. The services on the infrastructure layer are used to access essential IT resources that are combined under the heading Infrastructure-as-a-Service (IaaS). These essential IT resources include services linked to computing resources, data storage resources, and the communications channel. Cloud computing platforms are being increasingly utilized by industry, government and academia due to their ability to deliver robust, resilient and scalable computational power. In cloud computing data centres, Giga-bit speed, or faster, networks interconnect both physical and virtual computers. These systems are dynamically provisioned based on a determination of the required computing resources requested by the end user of the cloud application. [1][3][6]

Predicting the processor availability for a new process or task in computer network systems is a basic problem arising in many important contexts. Making such predictions is not easy because of the dynamic nature of current computer systems and their workload. To ensure high scalability, flexibility, and cost effectiveness, cloud platforms need to be able to quickly plan and provide resources, which will ensure that supporting infrastructures can closely match the needs of various applications. Cloud platforms require mechanisms to continuously characterize and predict their loads.

Load prediction is a crucial issue for efficient resource utilization in a dynamic cloud computing environment based on future load prediction and an estimate of the future performance of cloud system. Effective load prediction will help administrators take appropriate actions in preventing the system suffering from traffic surge which is caused by high load. The key to accurate load prediction in cloud computing is proper modelling of the relationship between historic data and future values, and a proper understanding of cloud computing backend workloads.

II. SIGNIFICANCE OF LOAD PREDICTION

Load prediction is an estimation of demand at some future period. This architectural framework is presented in Fig. 1.[3]

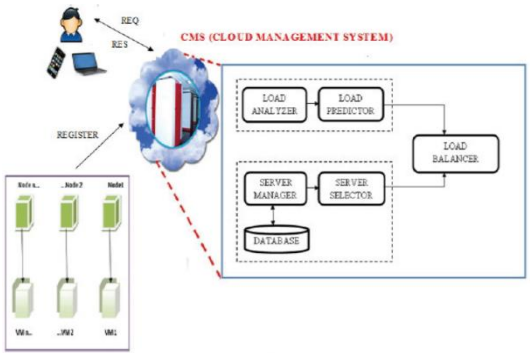

Fig. 1 Cloud architecture framework

The architecture describes how the cloud controller acts as an interface between the cloud service provider and external users which involves Load Analysis & Load Prediction, Management & Selection, Load Balancing.

In the past few years, some studies have been devoted to load prediction in cloud computing environments. This paper gives comparative study of different techniques used for load prediction. Firstly Bayesian method is discussed, which is an effective Cloud load prediction method that can accurately predict host load over a long-term period up to 16 hours in length. Prediction method based on Bayesian model is used to predict the mean load over a long-term time interval, as well as the mean load in consecutive future time intervals. It focuses on CPU. Using a Bayesian model for prediction effectively retains the important information about load fluctuation and noise. [1]

PSR and EA-GMDH is new prediction method which combines the Phase Space Reconstruction (PSR) method and the Group Method of Data Handling (GMDH) based on Evolutionary Algorithm (EA). It predicts not only the mean load in consecutive future time intervals, but also the actual load in each consecutive future time interval. PSR is an important step in local prediction methods because with a set of appropriate variables, we can reconstruct the time series. The GMDH method is a self organizing method and it has been applied to solve many prediction problems with success. [2]

Neural Network is used for load prediction, which predicts the future load based on the past historical data. It is a machine that is designed to model the way in which the brain performs a particular task. A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experimental knowledge and making it available for use. Neural network distinguishes itself by the presence of one or more hidden layers. The input signal is applied to the neurons in the second layer. The output signal of second layer is used as inputs to the third layer, and so on for the rest of the network. [3][4]

Support vector and kalmann smoother is multi-step-ahead CPU load prediction method based on Support Vector Regression which is suitable for the dynamic characteristics of applications and the complex Cloud computing environment. Kalman smoothing technology is integrated to further reduce the prediction error. It is suitable for the complex and dynamic characteristics of the Cloud computing environment. KSSVR is very stable, i.e. its prediction error increases quite slowly as the predicted steps increase. SVM has strict theory and mathematical foundation which could not lead to local optimization and dimensional disaster. [5]

III. TECHNIQUES FOR HOST LOAD PREDICTION

A. *Prediction using Bayesian Model*

Prediction method based on Bayes model is used to predict the mean load over a long-term time interval, as well as the mean load in consecutive future time intervals. Design an effective Cloud load prediction method that can accurately predict host load over a long term period up to 16 hours in length. This approach is to use a Bayesian model for prediction as it effectively retains the important information about load fluctuation and noise. Here Bayesian prediction method is evaluated using a detailed 1-month load trace of a Google data centre with thousands of machines [1].

Objective is to predict the fluctuation of host load over a long-term period, and aim is two-fold. First, at a current time point t0, predict the mean load over a single interval, starting from t0. Second, predict the mean load over consecutive time intervals. A new metric, namely exponentially segmented pattern (ESP), to characterize the host load fluctuation over some time period. For any specified prediction interval, it is into a set of consecutive segments, whose lengths increase exponentially. It predicts the mean load over each time segment. Shown in Figure 1 is an example of ESP. It denotes the total prediction interval length as s. The first segment (denoted by s1) is called baseline segment with length b, starts from the current time point t0 and ends at t0 +b. The length of each following segment (denoted by si) is $b \cdot 2^{i-2}$, where i = 2, 3, 4, · · · . For example, if b is set to 1 hour, the entire prediction interval length s could be equal to 16 (=1+1+2+4+8) hours. For each segment, predict the mean host load. The mean values are denoted by li, where i = 1, 2, 3, · · ·

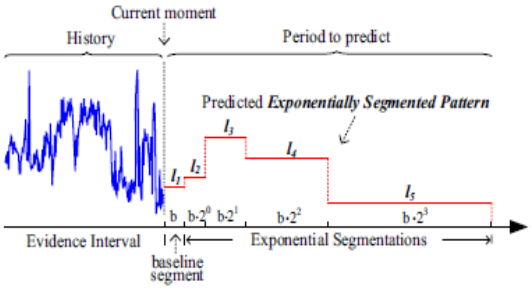

Fig. 2 Illustration of exponential segmented pattern

As illustrated above, aim is to predict the vector of load values (denoted by l), where each value represents the mean load value over a particular segment. To predict load, a predictor often uses recent load samples. The interval that encloses the recent samples used in the prediction is called evidence interval or evidence window. Given the prediction problem, one approach for prediction is to use feedback control. One could dynamically validate the prediction accuracy at runtime, adjusting the predicted values in the next interval by the error in the previous one. Then, prediction error could converge to a low level. This idea is based on the feed-back control model, which is often used in the one-step look-ahead prediction scenario.

Another approach is to use error of short-interval prediction to tune the long-term prediction. For instance, using the prediction error in a 4-hour interval may forecast the prediction error in the 8-hour interval, such that the predicted values could be tuned accordingly. However, this idea is also

inapplicable to Cloud load prediction in that short term prediction error always lags behind long-term error .

B. Neural Network Load Prediction

A neural network is a machine that is designed to model the way in which the brain performs a particular task. The network is implemented by using electronic components or is simulated in software on a digital computer.

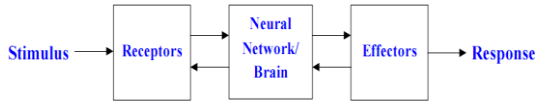

Fig. 3 Block Diagram of a Human Nervous System.

The receptors collect information from the environment. The effectors generate interactions with the environment. The flow of information/activation is represented by arrows.

Multilayer Feedforward Networks: The Feedforward neural network distinguishes itself by the presence of one or more hidden layers, whose computational nodes are correspondingly called hidden neurons. The function of hidden neuron is to intervene between the external input and the network output in some useful manner. The input signal is applied to the neurons in the second layer. The output signal of second layer is used as inputs to the third layer, and so on for the rest of the network. [3][4]

Back propagation algorithm: Multiple layers have been applied successfully to solve some difficult diverse problems by training them in a supervised manner with a highly popular algorithm known as the error back-propagation algorithm. This algorithm is based on the error-correction learning rule. Error back-propagation learning consists of two passes through the different layers of the network: a forward pass and a backward pass. In the forward pass, an input vector is applied to the nodes of the network, and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the forward pass the weights of the networks are all fixed. During the backward pass, the weights are all adjusted in accordance with an error correction rule. The actual response of the network is subtracted from a desired response to produce an error signal. This error signal is then propagated backward through the network, against the direction of synaptic connections. The weights are adjusted to make the actual response of the network move closer to the desired response.

The load on each server is predicted for optimal load balancing. A neural network model consists of three layers with five input nodes. Fig 4 depicts the neural model. The input layer of neurons in the neural model receives five inputs from the external information source. When the network is run, each layer performs the calculation on the input and transfers the result Y_{n+1} I to the next layer.

$$Y_{n+1} = h\left[\left(\sum_{i=1}^{n} X_i w_i + b \right) / n \right] \quad \text{------ (1)}$$

The above equation (1) is used for prediction of future load based on the input value. Where Yn+1 provides the output of the current node and n is the number of nodes in the previous layer, Xi is the input of the current node from the previous layer b is the bias value and w, is the modified weight based on the mean square error and our proposed algorithm.

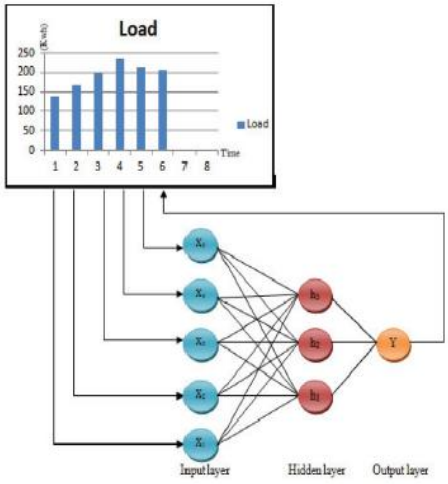

Fig. 4 Neural model

Here the neural predictor is developed and the experiment is performed to prove its highly accurate prediction. A sample load of a datacenter is analysed and given as input for the neural model.

C. Support vector and kalmann smoother

A multi-step-ahead CPU load prediction method based on Support Vector Regression which is suitable for Cloud computing environment. Kalman smoothing technology is integrated to further reduce the prediction error. Real trace data were used to verify the prediction accuracy and stability of this method. The focus of this work is on improving the CPU utilization by load prediction. KSSVR integrates SVR algorithm and Kalman smoothing technology. Furthermore, KSSVR is very stable, i.e. its prediction error increases quite slowly as the predicted steps increase. [5]

- ***Support Vector Machine:*** SVM was used for many machine learning tasks such as pattern recognition, object classification and regression analysis. It is based on the structural risk minimization principle which tries to control the model complexity as well as the upper bound of generalization risk. The principle is based on the fact that the generalization e+-rror is bounded by the sum of the empirical error and a confidence interval term that depends on the Vapnik – Chervonenkis (VC) dimension. On the contrary, traditional regression

techniques, including traditional Artificial Neural Networks (ANN) , are based on empirical risk minimization principle, which tries to minimize the training error only. Its learning process is quite complex and inefficient for modeling, and the choices of model structures and parameters are lack of strict theory. So, it may suffer from over-fitting or under-fitting with ill chosen parameters. In contrast, SVM has strict theory and mathematical foundation which could not lead to local optimization and dimensional disaster. It can achieve higher generalization performance especially for small samples set. It has a limited number of parameters to choose for modeling, and there exist fast and memory-efficient algorithms.

- ***Kalman Smoother:*** The Kalman filter has been widely used in the area of autonomous or assisted navigation. It is Kalman smoother is suitable for the Cloud application's load estimation because it was originally developed to estimate time-varying states in dynamic systems. This approach essentially uses a filtering technique to eliminate the noise of resources usage signal coming from error of measurement technique while still discovering its real main fluctuations in order to achieve a better QoS and higher resource utilization in Cloud.

D. Prediction Based on PSR and EA-GMDH

A new prediction method which combines the Phase Space Reconstruction (PSR) method and the Group Method of Data Handling (GMDH) based on Evolutionary Algorithm (EA). The proposed method could predict not only the mean load in consecutive future time intervals, but also the actual load in each consecutive future time interval. This method outperforms the other methods by more than 60% in mean load prediction, and performs well on actual load prediction over different time intervals, i.e. 0.5h to 3h. The main idea of this approach is to use PSR method and GMDH method based on evolutionary algorithm for host load prediction. PSR is an important step in local prediction methods because with a set of appropriate variables, we can reconstruct the time series. The GMDH method is a self organizing method and it has been applied to solve many prediction problems with success. [2]

The Representation of the EA-GMDH Network : To combine the EA and the GMDH network, we should first consider the representation of the EA-GMDH network. The representation of the EA-GMDH network should contain the number of input variables for each neuron, what is the best type of the polynomials for each neuron, and which input variables should be chosen for each neuron. Therefore, the chromosome for each individual should contain tree sub-chromosomes. Each sub-chromosome is represented as a string of integer

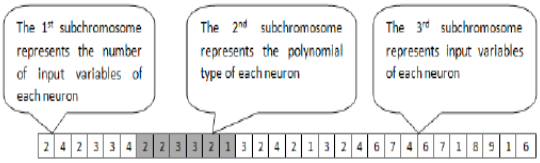

Fig. 5 The chromosome represents the EA-GMDH network

This EA-GMDH network consists of three layers, the number of neurons of each layer are 3, 2 and 1. The number of input variables of each neuron ranges from 2 to 4, and the type of polynomials ranges from 1 to 3.

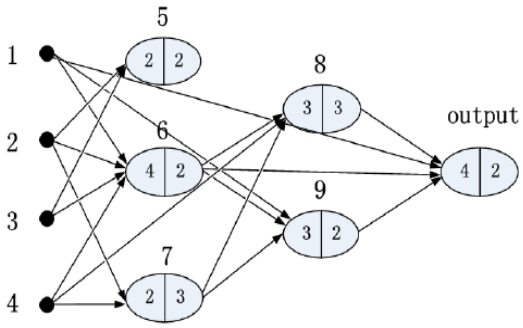

Fig. 6 The structure of the EA-GMDH network

The training set is used to calculate the coefficients of each neuron of the model. The validation set is used to evaluate each individual in each generation according to the fitness function. And the prediction set is used to estimate the performance of the model. The output of this proposed is a vector of the host load, which will not generate cumulative errors regardless of the step length, as the current predict value has nothing to do with the last predict value. We quantified the performance of actual load prediction with mean squared error (MSE).

IV. CONCLUSION

This paper summarizes the classification of load prediction methods and its impact on cloud environment. Some of the methods as discussed in the below table mainly focuses on host load prediction. We conclude that PSR & EA-GMDH outperforms other algorithms for dynamic cloud load prediction.[6] [2] It shows very good performance in long term load prediction with high performance accuracy and least error rate (MSE).

TECHNIQUE	DESCRIPTION	PREDICTION DURATION	MSE	SUMMARY/FINDINGS
Bayesian	Predicts the mean load over a longtime period as well as consecutive future time intervals.	Upto 16 hours	Lesser error rate	Retains information about load fluctuation and noise using ESP
GMDH & PSR	Predicts the mean load over a long time period as well as the mean load in the consecutive future time intervals.	0.5 to 3 hours	Least error rate	PSR: a set of appropriate variables, can reconstruct the time series. GMDH: self organizing.
Neural Network	Predicts the future load based on the past historical data.	1 second to 90 second	Medium error rate	Suitable for short term period
KSSVR	multi-step-ahead CPU load prediction method based on Support Vector Regression and Kalmann smoother methods.	1 to 2 hours	Highest error rate	Suitable for complex and dynamic characteristics of cloud environment. It is stable as the prediction errors increases quite slowly with prediction stages.

Table. 1 Comparision of Different Load Prediction Technique

REFERNCES :

[1] Sheng Di, Derrick Kondo1, Walfredo Cirne "Host Load Prediction in Google Compute Cloud with a Bayesian Model" France, 2Google Inc., USA, 2012 IEEE.

[2] Qiangpeng Yang, Chenglei Peng, Yao Yu, He Zhao, Yu Zhou, Ziqiang Wang, Sidan Du "Host Load Prediction Based on PSR and EA-GMDH for Cloud Computing System" 2013 IEEE Third International Conference on Cloud and Green Computing.

[3] Iniya, Venkatalakshmi, Ranjithflalakrishnan "Neural Load Prediction Technique for Power Optimization in Cloud Management System" Proceedings of2013 IEEE Conference on Information and Communication Technologies (IC2013).

[4] John J. Prevost, KranthiManoj Nagothu, Brian Kelley and Mo Jamshidi, Electrical and Computer Engineering "Prediction of Cloud Data Center Networks Loads Using Stochastic and Neural Models" proceedings of 6th international conference, 2011 IEEE.

[5] Rongdong Hu, Jingfei Jiang, Guangming Liu, Lixin Wang "CPU Load Prediction Using Support Vector Regression and Kalman Smoother for Cloud" 2013 IEEE 33rd International Conference.

[6] Da-yu XU, Shan-lin YANG, Ren-ping LIU "load prediction in cloud-oriented data centers" on Distributed Computing Systems Workshops, Journal of Zhejiang University 2013-14.