International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 02, No. 01, Dec 2021-Jan 2022
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.21.1.10

Load Prediction Techniques in Cloud Environment

Esraa Mohammad Ahmad Jaradat*

**Research and development manager in AL mi'rad municipality, Jordan*

*Corresponding Email: *jaradat_israa@yahoo.com*

Received: 18 September 2021 **Accepted:** 01 December 2021 **Published:** 02 January 2022

Abstract: ***Businesses and websites have rapidly increased their energy consumption, necessitating the development of data centres tailored to the cloud. Predicting when a system's resources will be needed means you can allocate them more efficiently and save money in the cloud. Predictive accuracy may be increased by classifying loads first. In this research, we offer a new method for predicting future demand for cloud-centric data centres. The Phase Space Reconstruction (PSR) and Extended Approximation-Group Method of Data Handling (EA-GMDH) methods are compared to the Bayesian model for predicting the mean load over a long-term time period. Multi-step ahead CPU load prediction using Support Vector Regression is very stable, i.e., its prediction error increases quite slowly as the predicted steps increase; this is in contrast to a neural network, which predicts the future load based on the past historical data and is distinguished by the presence of hidden layers***

Keyword: load, Forecasting, Techniques and Cloud Environment.

1. INTRODUCTION

With cloud computing, customers only pay for the resources they actually use, making it a cutting-edge innovation in the IT industry. Users are able to get to their data whenever and wherever they choose because of this (**Vasu**, Nehru, & Ramakrishnan, 2016). In contrast to the usual computer configuration, in which the user must be physically present at the same place as the data storage facilities, the cloud separates the two functions. Chandini, Pushpalatha and Ramesh (2016) defines "cloud computing" as "the delivery of computer services and applications via the Internet from remote data centres." So, cloud computing may be seen as a client-server arrangement, with the user controlling the workload in terms of resource demands and the frequency with which tasks are submitted. Because of its versatility, the cloud framework is ideal for running a broad range of workloads and applications. In 2018, cloud computing's 3.6 billion consumer users will rely heavily on cloud-based interactive apps (**Lobert**, 2019). With an ever-increasing user base, cloud service providers need to balance prioritising the satisfaction of a growing customer base with the need to generate revenues while minimising costs. Proactively provisioning resources is seen as a crucial activity to guarantee the on-demand availability of resources and the observance of the SLA. Its goal is to allocate resources to the most productive workloads based on estimates of their future demands on the server. However, predicting server loads in the cloud is a tricky endeavour that calls for a thorough familiarity with cloud workload characteristics. Workloads in the cloud are often more dynamic and changeable than those in traditional computing.

Concept of Cloud Computing

The term "cloud computing" refers to the practice of using remote servers for the purposes of data storage, management, processing, and transmission. These remote databases are kept on remote servers rather than on your local machine. All sorts of things fall under this category, including email servers, software, data storage, and even boosting your computer's processing capacity. Cloud computing, or a cloud-based system, is the provision of hosted services through the internet, as stated by Marielle (2022). The availability of a cloud service might be restricted to a select group of users or made available to the general public. In a public cloud, services are offered to anybody with an internet connection. A private cloud is a secure, permissioned network or data centre that only a select few users have access to. Both public and private cloud computing strive to make IT resources and services readily available via the Internet. The phrase "cloud computing" is used to describe any scenario in which a service is provided via a remote server connected to the internet. Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) are the three primary classifications of cloud services (Chai & Bigelow 2021). Cloud computing refers to the practise of making available, through the Internet, a wide range of computer resources and services, including but not limited to servers, storage, databases, networking, software, analytics, and intelligence.

The use of cloud services continues to rise. It's the latest fad that's inspiring businesses to use the phrase in their advertisements. Cloud computing, on the other hand, is when resources for computing are located elsewhere and accessible remotely over networks. The term "cloud computing" refers to the practise of providing various services through the Internet. Data storage, servers, databases, networking, and software are all examples of such resources. People and organisations alike are increasingly turning to cloud computing due to its many benefits, such as lower overhead and more productivity, as well as improved speed, efficiency, performance, and safety (**Frankenfield**, Mansa, & Schmitt, 2022). The term "cloud computing" is used to describe the practise of providing various types of computing resources (such as servers, storage, databases, networking, software, analytics, and intelligence) to users via the Internet ("the cloud") in order to facilitate increased rates of innovation, greater adaptability of resources, and greater economies of scale. You only pay for the cloud services you actually use, which helps you save money, improves efficiency, and allows you to scale up or down as your company requires.

Factors That Affect Load Prediction Techniques In Cloud Computing

Economic: There are a number of economic elements that may significantly alter the load pattern, including the customer base (residential, agricultural, commercial, industrial), population, per capita income, GDP growth, national economic development, social activities, etc. Most noticeably, these economic considerations have an impact on load forecasting over the longer term.

Weather: The weather (including dry bulb and wet bulb temperatures, humidity, cloud cover, etc.) has a significant impact on load forecasting. The temperature is the most crucial element of the weather. The load required for winter heating and summer cooling has been greatly impacted by the modifications. Load forecasting takes into account things like humidity, especially in places that are hot and humid, rain, thunderstorms, wind speed, and how bright the day is.

Random Disturbances: A significant change in the load pattern might result from a random disturbance in the power supply. Some examples of these random problems are factories that close or open without warning, large-scale strikes, weddings, special events, etc.

Conceptual Explanation of Load Forecasting

Using load forecasting, utilities may reduce their exposure to risk by planning for future demand for the commodities they transport and provide. Methodologies such as price elasticity, weather and demand response/load analysis, and predictive modelling for renewable power are used (Gartner, 2022). Time-series customer load profiles and regional data on customer loads are required for accurate forecasting. Seasonal changes are necessary for precise forecasting. As part of the distribution circuit load measurements, it is necessary to reconcile the distribution load prediction with the distribution network design. Predicting how much electricity will be needed in the near future, the intermediate future, and the far future is the goal of load forecasting. Service providers may better manage operations and supply to consumers thanks to the predictions. According to Techopedia (2022), power and energy providers use load forecasting to estimate how much energy will be required to satisfy consumer demand. The operational and management burden of a utility business depends heavily on the accuracy of its forecasts. Electric utilities use a method called load forecasting to plan for the constant balancing of supply and load demand (Anwar, Sharma, Chakraborty, & Sirohia, 2018). It's essential to the success of the electrical sector. Short-term (lasting just a few hours), intermediate-term (lasting several weeks to an entire year), and long-term (lasting more than a year) are all possible categories.

Both the deregulated and regulated energy markets now heavily rely on accurate load forecasting for operating and planning purposes. Electric load forecasting may be broken up into three distinct time scales: the immediate future, the intermediate future, and the far future. Using the current system condition and historical data, load forecasting attempts to predict the expected electric load over a specified time range. Data collection, input set selection based on patterns, load forecasting, and outcome analysis are the standard phases in load forecasting (Dai & Wang, 2007).

Short Term Load Forecasting (STLF): Short-term load forecasting is used to estimate the need for electricity anywhere from one day to many weeks in the future. By estimating load flows, we can protect the electricity grid from overload and save money.

Medium-Term Load Forecasting (MTLF): Medium-term load forecasting is useful for planning and operating power systems since it estimates load demand over the next month to many years.

Long-Term Load Forecasting (LTLF): Long-term load forecasting is used mostly in power system planning to estimate future load demands anywhere from one year out to twenty years out.

Techniques for Host Load Prediction

1. Neural Network Load Prediction

A neural network is a computer programme that mimics the way the human brain processes information in order to solve a problem. A digital computer is used to either build the network physically or make an electronic copy of it.

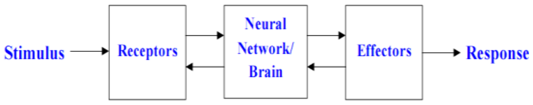

Fig 1. Block Diagram of a Human Nervous System.

Environmental data is gathered by the receptors. The environment is influenced by the effectors, which create interactions. Directional arrows depict the transmission of data and the triggering of events.

Multilayer Feed forward Networks

One of the defining characteristics of a feedforward neural network is the inclusion of one or more hidden layers, the computational nodes of which are known as hidden neurons. The hidden neuron's job is to mediate between the outside input and the network's final result. Neurons in the second layer receive the input signal. Each successive layer in the network receives its input from the previous layer's output. [3] [4]

Back propagation algorithm

Training them supervised by the widely used error back-propagation method has shown effective in applying many layers to handle various challenging and diverse tasks. This algorithm uses the principle of mistake correction to learn. In error back-propagation learning, errors are propagated both forward and backward via the network's layers. The forward pass involves feeding an input vector into the network's nodes and seeing its impact ripple across the system layer by layer. At last, a collection of outputs is generated as the network's real reaction. All network weights are kept constant throughout the forward pass. Error correction weights are applied to each weight during the backward pass. An error signal is generated by subtracting the network's actual response from the anticipated answer. This erroneous signal is subsequently sent backwards through the network, counter to the natural flow of the synapses. The network's actual response is brought closer to the anticipated answer by adjusting the weights. Predicting the demand on each server allows for a more equitable distribution of resources. A model of a neural network with three layers and five input nodes You can see the neural model in Fig. 4. Five pieces of data are sent from the outside world to the neural model's input layer of neurons. At runtime, the network's layers process the input, calculate the result (I), and pass it on to the layer below it.

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

Load forecasting using the input value is accomplished with the help of the aforementioned equation (1). where Yn+1 is the output from the current node, n is the number of nodes in the previous layer, Xi is the input from the previous layer for the current node, b is the bias value, and w is the adjusted weight based on the mean square error and our suggested technique.

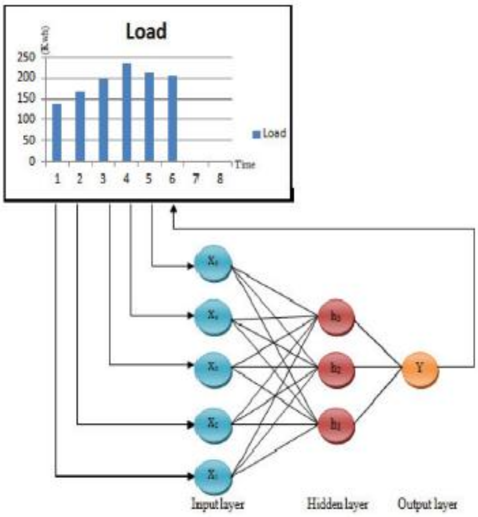

Fig 2. Neural model

Here, we build a neural predictor and show experimentally that it makes very accurate predictions. The neural model is fed data derived from the analysis of a datacenter's load distribution.

Support vector and kalmann smoother

A cloud-friendly CPU load prediction technique that uses support vector regression to look forward several time steps Technology based on Kalman smoothing is included to further lower the prediction error. The reliability and precision of this method's predictions were tested using actual trace data. In this effort, we will concentrate on improving CPU usage via load forecasting. KSSVR combines the SVR algorithm with Kalman smoothing theory. KSSVR is also quite stable, with a very low rate of rise in prediction error as the number of projected steps grows Rongdong Hu, Jingfei Jiang, Guangming Liu, Lixin Wang, (2013); John J. Prevost, KranthiManoj Nagothu, Brian Kelley and Mo Jamshidi, (2011) and and Anwar, T., Sharma, B., Chakraborty, K. & Sirohia, H. (2018).

A. Support Vector Machine

Pattern identification, object categorization, and regression analysis were just some of the numerous machine learning tasks for which SVM was put to use. It is founded on the notion of structural risk reduction, which aims to limit both the extent to which the model complexity may grow and the potential harm that might come from overgeneralization. The concept relies on the fact that the sum of the empirical error plus a confidence interval component that depends on the Vapnik-Chervonenkis (VC) dimension sets an upper constraint on the generalisation error. On the other hand, conventional regression methods, such as conventional artificial neural networks (ANN), are founded on the empirical risk reduction concept, which seeks to reduce the training error alone. It has an overly complicated and inefficient learning process, and it lacks a rigorous theoretical foundation upon which to base decisions on model topologies and parameters. Thus, it might experience over-fitting or under-fitting if the parameters are poorly selected. SVM, on the other hand, has a firm theoretical and mathematical basis that precludes the possibility of local optimization and dimensional catastrophe. For small sample sizes in particular, it may yield improved generalisation performance. There are few options for modelling parameters, and rapid, memory-efficient techniques are available.

B. Kalman Smoother

When it comes to autonomous or aided navigation, the Kalman filter has seen extensive application. Because of its history with estimating the states of dynamic systems over time, the Kalman smoother is well-suited for the assessment of cloud application loads. This method employs a filtering technique to isolate the true primary swings in resource use from the noise introduced by measurement inaccuracy. A greater quality of service (QoS) and more efficient use of cloud resources are essential.

Prediction using Bayesian Model

The long-term mean load, as well as the mean load in the subsequent future time periods, are predicted using a Bayes model-based prediction approach. Create a strategy for predicting cloud loads that can properly forecast host loads over a time span of up to 16 hours. Prediction using a Bayesian model is preferable since it faithfully stores crucial data, such as load fluctuations and noise. In this study, we used a one-month load trace from a Google data centre with hundreds of computers to assess the efficacy of the Bayesian prediction approach Sheng Di, Derrick Kondo1, Walfredo, (2013).

The goal is to forecast long-term changes in host load, and the purpose is dual. The first step is to anticipate the average load during a single period, beginning at the present time. In a second step, estimate the average workload for future time periods. We provide a novel measure, the exponentially segmented pattern (ESP), to quantify the host load variation across time. Prediction intervals may be broken down into a series of sequential segments with exponentially growing lengths. The average workload during each time period is estimated. A visual representation of ESP is shown in Figure 1. The total length of the prediction interval is represented by the symbol s. First, we have the baseline segment (s1), which begins at the present time of t0 and extends to t0 + b. Each subsequent segment (represented by si) has a length of b • 2i2, where I = 2, 3, 4,... If b is 1 hour, then s might be 16 (=1+1+2+4+8) hours over the full prediction period. Estimate the median number of hosts across the segments. Li represents the mean, with I = 1, 2, and 3.

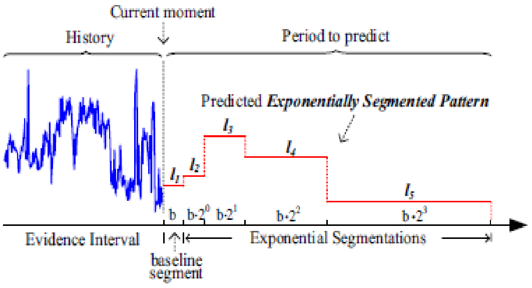

Fig 3. illustration of exponential segmented pattern

Above, we see that our goal is to predict the vector of load values (denoted by l), where each value reflects the average load value along a certain segment. Predictors often make use of recent load samples in order to make accurate predictions. The "evidence interval" or "evidence window" refers to the time frame in which the most recent samples were taken for the purpose of making the prediction. Given the nature of the prediction issue, feedback control is one method for improving the predictive model. Dynamically validating the prediction accuracy during runtime might be done by changing the following interval's predicted values by the inaccuracy in the previous interval. If that happened, the margin of error in our forecasts would shrink to an acceptable level. The feed-back control paradigm, which is often used in the one-step look-ahead prediction setting, is where this idea comes from.

To fine-tune the long-term forecast, it is also possible to employ the short-term forecast's inaccuracy. It is possible to tailor the anticipated values by utilising the prediction error in a shorter time period to foretell the prediction error in a longer time period, say 4 hours. However, this concept does not work for cloud load prediction since short-term forecast error invariably follows long-term error.

Prediction Based on PSR and EA-GMDH

A novel prediction strategy using an evolutionary algorithm-based hybrid of the Phase Space Reconstruction (PSR) and Group Method of Data Handling (GMDH) techniques (EA) In addition to predicting the average load over future time intervals, the suggested technique was able to estimate the loads throughout future time periods. When compared to other approaches, this one has a 60%+ advantage in terms of mean load prediction and works admirably when attempting to predict real loads across varying time periods (0.5h to 3h). The primary concept behind this technique is to forecast host load using an evolutionary algorithm and the PSR method. Since the time series may be reconstructed using the right set of variables, PSR is an essential part of local prediction approaches. It has been shown that the GMDH technique, which is a self-organizing approach, may be used to effectively resolve a wide variety of prediction issues Qiangpeng Yang, Chenglei Peng, Yao Yu, He Zhao, Yu Zhou, Ziqiang Wang, Sidan Du, (2013).

TECHNIQUE	DESCRIPTIONS	MSE	SUMMARY/ FINDINGS	PREDICTION DURATION
Neural Network	based on historical data from the past, predicts the load for the future.	Medium error rate	Suitable for short term period	1 second to 90 second
KSSVR	Support Vector Regression and Kalmann smoother algorithms are the foundation of the multi-step-ahead CPU load prediction approach.	Highest error rate	Suitable for the complicated and dynamic cloud environment features. Given that the prediction errors rise gradually throughout the course of prediction phases, it is stable.	1 to 2 hours
Bayesian	predicts the average load both over a lengthy time and over a series of future time intervals.	Lesser error rate	keeps track of noise and load fluctuations via ESP	Upto 16 hours
GMDH & PSR	predicts both the average load over a lengthy time period and the average load over a series of subsequent future time periods.	Least error rate	PSR: Reconstructing the time series using a collection of suitable variables. Self-organizing GMDH	0.5 to 3 hours

Future of Load Prediction

The task scheduler receives the client's task and sends it to the resource manager in the Future Load Prediction Architecture. Then the virtual machines will receive the tasks from the resource manager, and if any virtual machines become overloaded, we will migrate those virtual machines. The predictor uses a number of components in the scheduler to make predictions about the workload of PMs and the resource demands of VMs based on historical data. a PM's workload by combining the resource usage of its own virtual machines. so the load prediction's specifics The next paragraph will also provide a description of algorithm. The LNM at each node makes an initial effort to meet the new requirements locally by changing the resource distribution of VMs that share the same VMM. By modifying the weights of the VMs in its CPU scheduler, the Xen hypervisor may alter how much CPU is allocated to each one. It is also in charge of changing how much RAM is allocated. Systems use load balancing to manage hierarchical and multidimensional resource constraints, however single-dimensional storage is not feasible because resource constraints are too sluggish. Using the virtual machine monitor in a data centre network, two-dimensional constraints can be discovered. There isn't a distributed computing environment with a server provisioning strategy developed by Chen, Wenbo, Liu, Nath, Rigas, Xiao, and Zhao (2008) that dynamically turns servers on and off in response to changing user loads. In large-scale utility computing, where pay per use is the predominant business model, on-demand processes are not feasible. to decide where a reduction job ought to be executed. Given a reduction job, Padala, Hou, Shin, Zhu, Uysal, Wang, Singhal, and Merchant need to collect map output from a group of nodes by shuffle time and calculation time as map reduce is fetching less in heterogeneous domain network (2009). A framework for distributed computing in a homogenous domain is expected to be viable using map-reduce technology. The client will be put into practise by improving dynamic resource allocation for scheduling applications in public clouds through big data centres mapping is done adaptively so that PMs become minimised as opposed to VMs. When partitioned in a cloud environment, resource management will aggregate all of the demands. Compared to a static consolidation strategy, VM can lower the amount of physical capacity needed to meet a certain rate of SLA (Service Level Agreement) breaches for a particular workload by as much as 50%. Layers between physical machines are simply provided using SLA region. Implementing skewness might be used to load dynamic memory. Vahdat, Thakar, Anderson, Chase, and Doyle (2001).

2. CONCLUSION

This study discusses the classification of load prediction systems and their effect on the cloud. While several of the aforementioned techniques may accurately forecast host load variables like CPU use, others are either too simplistic or too complex. In conclusion, the table below shows that PSR and EA-GMDH are superior to other methods for predicting the dynamic load in clouds. As a long-term load predictor, it performs well, with low error rates and great performance accuracy (MSE).

3. REFERENCES

1. Dai, W. & Wang, P. (2007). Application of Pattern Recognition and Artificial Neural Network to Load Forecasting in Electric Power System. Third International Conference on Natural Computation (ICNC 2007).
2. Anwar, T., Sharma, B., Chakraborty, K. & Sirohia, H. (2018). Introduction to Load Forecasting. International Journal of Pure and Applied Mathematics 119(15) Pp 1527-1538
3. Techopedia (2022). Load Forecasting: What Does Load Forecasting Mean? Available at: https://www.techopedia.com/definition/30629/load-forecasting-electric-power-engineering

4. Gartner (2022). Load Forecasting. Retrieved from: https://www.gartner.com/en/information-technology/glossary/load-forecasting
5. Vasu, R., Nehru, E. & Ramakrishnan, G. (2016). Load Forecasting for Optimal Resource Allocation in Cloud Computing Using Neural Method. Middle-East Journal of Scientific Research 24(6) Pp 1995-2002
6. Marielle, G. (2022). What's a Cloud-Based System and How Does It Work? Retrieved from: https://unity-connect.com/our-resources/tech-insights/what-is-a-cloud-based-system-and-how-does-it-work/
7. Chai, W. & Bigelow, S. (2021). Definition of Cloud Computing. Available at: https://www.techtarget.com/searchcloudcomputing/definition/cloud-computing
8. Fernando, J. (2022). Accounting Explained with Brief History and Modern Job Requirement. Available at: https://www.investopedia.com/terms/a/accounting.asp
9. Iniya, Venkatalakshmi, Ranjithflalakrishnan (2013) "Neural Load Prediction Technique for Power Optimization in Cloud Management System" Proceedings of2013 IEEE Conference on Information and Communication Technologies (IC2013).
10. John J. Prevost, KranthiManoj Nagothu, Brian Kelley and Mo Jamshidi, (2013) Electrical and Computer Engineering "Prediction of Cloud Data Center Networks Loads Using Stochastic and Neural Models" proceedings of 6th international conference, 2011 IEEE.
11. Rongdong Hu, Jingfei Jiang, Guangming Liu, Lixin Wang (2013)"CPU Load Prediction Using Support Vector Regression and Kalman Smoother for Cloud" 2013 IEEE 33rd International Conference.
12. Sheng Di, Derrick Kondo1, Walfredo Cirne (2012)"Host Load Prediction in Google Compute Cloud with a Bayesian Model" France, 2Google Inc., USA, 2012 IEEE.
13. Qiangpeng Yang, Chenglei Peng, Yao Yu, He Zhao, Yu Zhou, Ziqiang Wang, Sidan Du (2013) "Host Load Prediction Based on PSR and EA-GMDH for Cloud Computing System" 2013 IEEE Third International Conference on Cloud and Green Computing.
14. Chandini M.S , Mrs Pushpalatha R, Dr. Ramesh Boraiah, (2016) A Brief study on Prediction of load in Cloud Environment. International Journal of Advanced Research in Computer and Communication Engineering. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 5, May 2016
15. Chase,J. S.; Anderson, D. C.; Thakar, P. N.; Vahdat, A. M.; and Doyle, R. P. (2001)"Managing energy and server resources in hosting centers," in Proc. Of the ACM Symposium on Operating System Principles (SOSP'01).
16. Padala, P. Hou, K.-Y.; Shin, K. G.; Zhu, X.; Uysal, M., Wang, Z., Singhal, S., and Merchant, A., (2009) "Automated control of multiple virtualized resources," in Proc. of the ACM European conference on Computer systems (EuroSys'09), 2009.
17. Chen, G., Wenbo, H., Liu, J., Nath, S., Rigas, L., Xiao, S., and Zhao, F. (2008)"Energy-aware server provisioning and load dispatching for connection-intensive internet services," in Proc. of the USENIX Symposium on Networked Systems Design and Implementation (NSDI'08).