A survey of the workload forecasting methods in cloud computing

Archana Yadav[1], Shivam Kushwaha[2], Jyoti Gupta[3], Deepika Saxena[4],

Ashutosh Kumar Singh[5]

National Institute of Technology, Kurukshetra, Haryana, India

Abstract

A few years ago, cloud computing had widely altered the way of computation and storage. It is quite challenging for cloud service providers to maintain the required Quality of Service (QoS) standard without violating a service level (SLA) agreement. To improve the performance of the cloud, accurate workload prediction plays a vital role. Cloud computing promises scalability, on-demand service, and virtualization to improve the Quality of Service (QoS) offered to end-users. In this paper, we present a comprehensive survey of workload prediction approaches in cloud environments. It also highlights the emerging challenges like resource wastage, excess power consumption, and quality of service violations, etc.

Keywords: Cloud Computing · Workload Prediction · Scalability · Classification

1. Introduction

Cloud computing continues as a model for IT service delivery at an astonishing speed driven by a wide range of interactive and interactive features. It is widely used for computation and storage over the internet for various domains like Marketing, Business, Education, Banking, Entertainment, etc Instead of storing files on a portable hard drive or local storage device, cloud-based storage makes it easier to store them in a remote database. The popularity and usage of cloud computing are increasing day by day, cloud-based applications make people highly dependent on it for their day-to-day activities. It offers several advantages compared to traditional computer models along the way including reduced costs and increased flexibility. Cloud computing is a much-needed delivery of IT resources to the Internet at pay-as-you-go rates. Cloud is present at remote locations rather than local servers. It serves several benefits like Broad Network, Elasticity, Data Security, Scalable, On-Demand Service, Virtualization, etc. Fig 1 illustrates the various types of characteristics of Cloud Computing. We can process and access data and applications from servers via the Internet. It can also help save our planet by providing a complete computing environment. Cloud computing can be described as a new computer-style where powerful and easy-to-use resources are offered as an Internet service. Cloud computing will redesign information technology (IT) processes and the IT market and is also a powerful computer infrastructure for many applications. The purpose of using the cloud is to allow users to benefit from all of these technologies, without the need for in-depth information about each of them. The cloud aims to reduce costs and help users focus on their core business instead of being blocked by IT barriers. Cloud computing creates privacy concerns because a service provider can access data in the cloud at any time. It may alter the information by mistake or intentionally or delete the information. The main motivation is to reduce the error in workload prediction by extracting knowledge from previous existing models so that we could identify current as well as predict future workloads in the cloud environment.

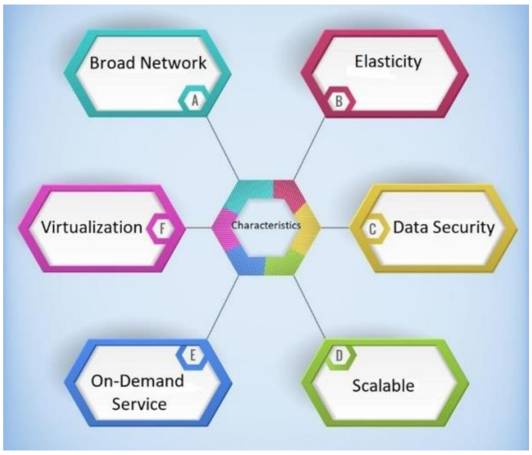

Figure 1. Characteristics of Cloud Computing

This document presents an in-depth investigation of the technical predictions of the project, their strategies used, and motivations to conduct them. The main contributions of this paper are as follows:

- We are describing what is cloud, what is cloud computing, and how it works.
- The benefits and highlighting emerging challenges of cloud computing.
- What is workload prediction including a survey of various existing workload prediction models.
- Comparison among various existing workload prediction models in a tabular form.

The rest of the paper is organized as follows: Section 2 discusses related activities. Comparison of different approaches is presented in Section 3 then we have provided emerging challenges in Section 4 and finally, we end up with a conclusion in section 5.

Workload Prediction

Forecasting takes information available in the present and uses it to predict the future. Workload prediction is a technique that is used to improve efficiency and reduce the operational cost of the cloud. Application management and resource management are the two very basic requirements that motivate prediction. Fig. 2 illustrates the general scenario of workload forecasting at the cloud data center. Servers receive millions of requests sent by users. Applications are processed there as well all requests that come at a certain time of speculation are compiled as historical data, which is later used to predict future work. Historical data is collected and processed in advance for standardization. Then normal data transferred to a forecasting system to predict future workload.

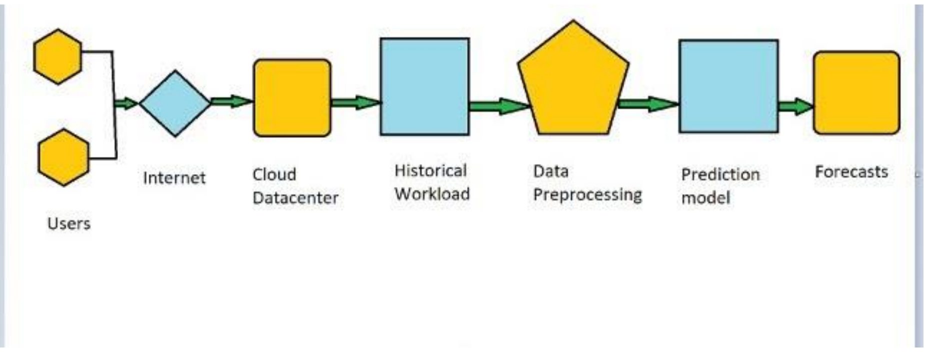

Figure 2: General Workload Forecasting Model

2. Related Work

2.1 Neural Network-based approaches:

Saxena et al. [1], have proposed a model of network load variability to predict the average load between consecutive forecast periods. The variable load prediction model shows the predictable demand for demand in cloud data centers. This data is used to train and evaluate data in the neural network predictive system. The AADE learning algorithm then trains the neural network. If the predicted workload is almost equal to the actual workload then these data are sent to trained workload prediction models otherwise it is again sent to the AADE learning algorithm.

Lu et al. [2], have proposed a model of a predictive load forecast to predict the future value of work for cloud conservationists. The projection model for the proposed novel, called RVL BPNN is based on the BP neural network algorithm and predicts a much larger future load in terms of how to use the internal relationships between the incoming workloads. The test results show that the proposed RVL BPNN model is gaining clarity and efficiency.

Saxena et al. [3], have proposed the provision of energy-efficient resources and a distribution framework to predict the changing needs of future applications. Predicting the use of multiple resources simultaneously, A predictive model for the Online Multi-Resource feed-forward Neural Network (OM-FNN) is proposed. The Tri-adaptive Differential Evolution (TaDE) learning algorithm is designed for optimal use of OM-FNN predictions. simultaneously with future applications, automated VM deployment based on the integration of specific resource requirements and reduced VM allocation enables PMs.

Kumar et al. [4], have developed a workload prediction model using a neural network and a different evolution algorithm. .In the author's way, using a structure containing n-p-q neurons.where n, p & q are the number of neurons in the sequence of input, hidden and output.

A workload forecasting framework based on a neural network model with supervised learning technique is proposed in [5]. To upgrade the network learning process a Bi-Phase adaptive differential evolution (BaDE) learning algorithm is introduced to train the model. The BaDE learning algorithm is used to train the proposed framework for predicting the network load (WFNN) function. To train the neural predictive network, historical data is actively provided by the cloud data keeper. The proposed activity is measured by two real-world stocks, Saskatchewan HTTP and NASA.

Bi et al. [6], have presented an integrated forecasting system, provided by sound filtering and data representation, known as Savitzky-Golay and Wavelet supported by Stochastic Configuration Networks (SGW-SCN), to indicate the amount of workload in the future.

2.2 Heuristic approaches:

Kumar et al. [7], have proposed a performance management system (SDWF) that identifies the practice of predicting errors by incorporating the latest forecasts and using it to increase the accuracy of other predictions. The model uses a healing method based on dark conditions in training neurons. This paper offers to create a way to predict the burden of learning from past mistakes and deal with them appropriately. It also develops one of the meta-heuristic optimization algorithm proposed by black-hole scenarios to achieve the most accurate network tools.

Nguyen et al. [8], have established a predictive model for a series of novel timelines, based on an advanced reading machine. Due to Uncertainty in the weights of installing over-the-counter learning equipment requires a large no. Hidden neurons to achieve better results for this reason the model weight is increased. Winning the issue is a new war of attrition based on opposition to extreme learning machines, choosing the right input tools, and hidden bias. So in this model two algorithms are used for extreme learning machines (ELM) and opposition warfare (OTWO). ELM was raised to overcome the regression of media-based approaches in a single hidden layer of supplying neural networks through two phases.

Zhu et al. [9], have raised the cache memory network (LSTM) encoder network. The model has two components: the LSTM-based network encoder-decoder network.

First, the data is encrypted In context, the vector then selects the code to produce the middle prediction result of the output layer. Finally, the output layer displays the predictive values of the workload. The result was satisfactory in predicting mixed loading in the computer environment.

2.3 Recurrent neural networks based approaches:

Gao et al. [10], have compared the making of predictable methods of independent artwork. They suggest how to make predictions for a while before a predetermined time point to allow for sufficient time for work planning based on the predicted workload. Continuing to improve the accuracy of the forecast, they introduced a combined method of predicting the load, which begins to combine all the tasks into several stages and trains the forecast model for each phase in sequence.

Kumar et al. [11], have analyzed and compared the accuracy of the assumptions for different machine learning algorithms aimed at predicting the load of server logs. The proposed speculation model for the comparison study was used using Linear Regression (LR), K- Nearest Neighbors (KNN), Support Vector Machine (SVM), ARMA, ARIMA, and Support Vector Regression (SVR) for web applications to select the appropriate algorithm depending on the characteristics of each load.

Singh et al. [12], have developed a dynamic predictive model using linear regulation, ARIMA and supported vector regression for web applications. In addition to the cloud-based information discussed in web applications with a new weather module called technocrat load forecaster. In the cloud environment, the start of the VM takes 5-10 min, so the task of predicting work requires a solution in the first step to help prepare the required VM for the incoming workload.

2.4 Hybrid Approaches :

Chandy et al. [13], have proposed machine learning methods to manage resource allocation to cloud computing for large data processing systems, the simulation of the proposed model using network simulator 2 enables better performance and resource utilization at a higher cost, time, power, and memory usage. A random forest algorithm used uninstall features that use the bootstrap reset process and upgrade the resolution tree for all visual effects.

Saxena et al. [14], have proposed predictive resource management models in cloud environments. In the conceptual framework for cloud resource management, m
users request different applications to be executed at a data center. Each application had a specific resource requirement that is to be fulfilled by the data center. The workload forecasting accuracy and virtual machine resource prediction should be increased to ignore SLA violations.

Kumar et al. [15], have proposed a test of six different ways to predict the real-world tracking of web servers and cloud created. All analyzes were performed three times as three different tasks were used to measure the deviation of the predictions. A three-step guessing error, means a complete error, means a complete error and a root means a double error.

Kumar et al. [16], compared the predictive accuracy of the various machine learning algorithms proposed to predict server load uploads. Line reversal is often the basic strategy used in statistical analysis where all the attributes included in the expectations are numeric.

3. Comparison table

Author, Year	**Algorithm**	**Dataset**	**Programming language/Tool**	**Advantage**	**Disadvantage**
Saxena et al., 2020	Auto-adaptive learning based	NASA & Saskatchewan HTTP traces	Python 3	Optimal prediction accuracy	Hardware dependence
Jing Bi et al., 2018	Integrated machine learning (SGW-SCN)	Google workload datasets	Savitzky-Golay filter	Better forecasting performance	It's never completely accurate
Kumar et al., 2020	Self-directed learning based	HTTP weblogs from three different World Wide Web ser	MATLAB R2017a	Provides more accurate weights of the network	It's never completely accurate
Nguyen et al., 2020	Extreme learning machine and enhanced tug of war optimization	The internet traffic (in Megabyte (from EU) & in Bytes (from the UK))	Opposition-based tug of war optimization (OTWO)	Good accuracy	Over-fitting problem
Zhu et al., 2019	Attention-based LSTM encoder-decoder network	Alibaba and dinda workload traces dataset	ARIMA, PSR+ EA-GMDH	Controls prediction accuracy	Require large amounts of memory bandwidth
Jayakumar et al.,	Self-optimized generic workload prediction framework (LOADdynamics)	Azure and LCG workloads	The inference & training were executed on a 16-core Intel Xeon Platinum 8153 CPU.	The prediction error is very low	Dependency on long short term memory (LSTM)
Hu et al., 2016	Elastic mechanism	CPU workload time series	Time series approach, Kalman filter model	Easy scalability	Data loss
Singh et al., 2018	Technocrat ARIMA and SVR model	ClarkNet and NASA	MAE, MSE, RMSE, MAPE	Improves resource provisioning oscillation	Workload is non-linear
Qiu et al., 2016	Deep learning	Workload datasets	EWMA model	Good performance	Require a large amount of data
Kumar et al., 2018	Artificial neural network and adaptive differential	NASA and Saskatchewan servers' HTTP	MATLAB	Fast evolution	Hardware dependence

	evolution	traces			
Kumar et al., 2017	Long short term memory recurrent neural network (LSTM-RNN)	HTTP traces of, Calgary server, & Saskatchewan server	Python along with Keras library	Remembers information for a long time	Prone to overfitting
Lu et al., 2016	Random variable learning rate back-propagation neural network (RVLBPNN)	Publicly available google workload traces	MATLAB 7.14	High accuracy	Less secure
Linma et al.,	Query-based model	Three real-world database traces	MySQL and PostgreSQL	Improves system performance	Require more powerful hardware

4. Emerging Challenges

Cloud computing has placed many challenges in different aspects of data handling. Some of these challenges are :

1) Decreased performance and overuse - Performance is an important factor when considering cloud-based solutions. If cloud performance is unsatisfactory, it can drive users away and reduce profits. Even a slight delay in loading an application or web page can lead to a significant decrease in the percentage of users.

2) Adaptability- The word "adaptability" means the quality of being able to adjust to new conditions. The prediction model should be adaptable to behavior changes.

3) Security and privacy – Data security is a major problem when switching to cloud computing. The information held by the user or organization is sensitive and confidential. Cloud security issues include identity theft, data breach, malware infection, and many more that ultimately reduce trust among users of your apps

4) Workload fluctuation – Evaluate your workload and determine a pattern of factors that influence workload.

5)Accurate workload prediction – Accuracy is a key factor in the forecasting of workloads and existing methods remain to produce 100% accurate results.

5. Conclusion:

Providing a variety of remote services and services to its customers is a key objective of the cloud computing paradigm. Problems such as under-provision or over-provision could be caused by the error of predicting that the past reduces the lead to SLA violations and cloud performance, and this ultimately leads to the problem of resource wastage. Various load forecasting schemes are provided in the literature relating to the lack of accurate workload forecasting based on job history data and handling issues such as Slashdot results, and workload fluctuations. This paper begins by introducing the basic concepts and challenges to the burden forecasting systems. After that, they passed the research on the proposed methods of forecasting the load and described their main contribution, and used an algorithm to make predictions. A comparison table is included in this paper to compare algorithms, data sets, etc.

References:

[1] Deepika Saxena, Ashutosh Kumar Singh, et al. Auto-adaptive learning-based workload forecasting in dynamic cloud environment.2020;

[2] Yau Lu, John panneerselvam, LucLiu, and Yan Wu, et al. RVLBPNN: A workload forecasting model for smart cloud computing.2016;

[3] Deepika Saxena, Ashutosh Kumar Singh, et al.A proactive auto-scaling and energy-efficient VM allocation framework using online multi-resource neural network for cloud data center.2020;

[4] Jitendra Kumar, Ashutosh Kumar Singh, et al. Workload Prediction In Cloud Using Artificial Neural Network And Adaptive Differential Evolution.2018;

[5] Jitendra Kumar, Deepika Saxena, Ashutosh Kumar Singh, et al.BiPhase adaptive learning-based neural network model for cloud datacenter workload forecasting.2020;

[6] Jing Bi, Haitao Yuan, LiBo Zhang, Jia Zhang, et al. SGW-SCN: An integrated machine learning approach for workload forecasting in geo-distributed cloud data centers.2018;

[7] Jitendra Kumar, Ashutosh Kumar Singh, Rajkumar Buyya, et al. Self-directed learning-based workload forecasting model for cloud resource management.2020;

[8] Thieu Nguyen, Bao Hoang, Giang Nyuyen, Binh Minh Nguyen, et al.A new workload prediction model using extreme learning machine and enhanced tug of war optimization. 2020;

[9] Yonghua zhu, Weilin Zhang, Yihai Chen, and hanghao Gao, et al. A novel approach to workload prediction using attention-based LSTM encoder-decoder network in cloud environment.2019;

[10] Jiechao Gao, Haoyu Wang, and Haiying Shen, et al. Machine Learning-Based Workload Prediction in Cloud Computing.2020;

[11] Krishan Kumar, K. Gangadhara Rao, Suneetha Bulla, D Venkateswarulu, et al. Forecasting of Cloud Computing Services Workload using Machine Learn ing.2021;

[12] Parminder Singh, Pooja Gupta, Kiran Jyoti, et al.TASM: technocrat ARIMA and SVR model for workload prediction of web applications in the cloud.2018;

[13] Dr. Abraham Chandy, et al. Smart Resource usage prediction using cloud computing for massive data processing system.2019;

[14] Deepika Saxena, Ashutosh Kumar Singh et al. Workload Forecasting and research management models based on machine learning for cloud computing environment.2021;

[15] Jitendra Kumar, Ashutosh Kumar Singh, et al.Performance Assessment Of Time Series Forecasting Models For Cloud Datacenter Network Workload Prediction.2020;

[16] Krishan Kumar, K. Gangadhara Rao, Suneetha Bulla, D Venkateswarulu, et al. Forecasting of Cloud Computing Services Workload Using Machine Learing.2021;