Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.mosuljournals.com

Workload Forecasting Methods in Cloud Environments: An Overview

Samah Fakhri Aziz [1] ,*Manar Y. Kashmola[2]

Department of Computer Science, College of computer science and mathematics, University of Mosul/ working in University of AL-Hamdaniya, Mosul, Iraq[1], College of Information Technology, Ninevah University ,Mosul, Iraq[2]
**Corresponding author. Email: samah.fakhri@uohamdaniya.edu.iq[1]*

Article information

Article history:
Received: 26/3/2023
Accepted: 10/9/2023
Available online:

Abstract

Cloud computing is becoming increasingly popular due to its on-demand resource allocation and scalability. It is essential to precisely anticipate workload as applications and users on cloud-based services increase to distribute resources effectively and avoid service interruptions. We present an overview of approaches for workload forecasting in cloud systems in this study. We explore more sophisticated approaches like algorithms for deep learning (DL) and machine learning (ML) in addition to more conventional approaches like analysis of time series and models of regression. We also discuss difficulties and unresolved research questions in the area of workload forecasting for cloud settings. Cloud service providers may allocate resources wisely and guarantee good performance and accessibility for their clients by being aware of these techniques and problems. Cloud computing with virtualization and customized service is crucial to improving the service provided to customers. Accurate forecasting of workload is key to optimizing cloud performance. In this study, we discuss some methods of predicting workload in cloud environments.

This study presents an overview of workload prediction techniques in cloud systems, with a special emphasis on long short-term memory (LSTM) networks. We go through the fundamental ideas behind LSTM networks and how well they can detect long-term relationships in data from time series. We also examine the particular difficulties and factors involved in LSTM-based workload forecasting implementation in cloud systems. We also examine previous research and methods that have employed LSTM networks to forecast workload in cloud systems. We examine the benefits and drawbacks of different methods, focusing on their effectiveness, scalability, and interpretability.

.

Keywords:
long short term memory (LSTM) ; machine learning (ML); deep learning (DL) ; neural networks; review; cloud computing; workload forecasting.

Correspondence:
Author: Samah Fakhri A
Email: samah.fakhri@uohamdaniya.edu.iq

1. INTRODUCTION

Forecasting workloads is crucial in cloud systems for efficient resource management and optimization. Cloud suppliers can decide wisely on the allocation of resources, capacity scheduling, and efficiency optimization by precisely forecasting future workloads. Insights into various workload forecasting techniques utilized in cloud systems are intended to be provided through this overview. Forecasting the quantity of computing power, such as memory and central processing units, that will be needed to handle incoming activities or workloads is known as workload forecasting. The efficiency, cost-effectiveness, and user happiness of the cloud infrastructure are directly impacted by how accurately these forecasts are made.

Statistical time series evaluation techniques, such as automated regression integrated moving averages (ARIMA) and exponential normalization, are frequently used in conventional workload forecasting methodologies. The dynamic and erratic nature of cloud workloads, however, necessitates the employment of more sophisticated and adaptable strategies.

Several models, including regression-based designs, are examples of typical ML-based approaches. These models employ regression algorithms to forecast future workload levels using previous workload data and other pertinent variables. It also makes use of a time series analysis technique that concentrates on examining workload data in the period domains while accounting for elements like lagged data and autocorrelation. Deep learning methods have also shown potential in forecasting workload, particularly recurring neural networks (RNN) and LSTM networks. These models are good in capturing non-linear patterns and long-term relationships in workload data. [1]

It should be mentioned that workload forecasting is a discipline that is always creating new methods in order to increase flexibility and accuracy. The kind of workload, the data at hand, the available computing power, and the needed prediction horizon all play a role in the forecasting technique selection.

Currently, the huge increase in the use of cloud computing in various fields of business is causing cloud environment providers to rise. The main reason is that cloud data centers are treated as a primary source of energy for the power users. Therefore, cloud environments are treated as causing environmental, economic and energy impacts. Workload forecasting model using LSTM networks is discussed. [2]

Forecasting methods usually have to make restrictive assumptions in operating environments or algorithms for future workload, because cloud environments are difficult to configure and complex.

In this study, we also discuss some of the emerging challenges of cloud computing, and the focus is on forecasting workloads, including effective analysis of some algorithms for workloads, and then some prediction models are compared to the current workloads.

The rest of this study is planned as follows: Section II lists previous work on cloud workload prediction methods. Section III presents the workload forecasting activity. Section IV explains the findings and assessments. Finally, we end with emerging challenges and a conclusion in Section V.

2. Related Works

The authors [3, 4] discuss several methods for predicting cloud services in several aspects such as workload and cloud service work. Various resource saving issues and resource-saving models in the cloud environment are demonstrated. Discuss several prediction methods in ML for cloud services and provide a typology of prediction approaches. The authors presented in [5] a method for predicting cloud workload using moving average (MA) and automatic regression (AR) methods. The author [6], using neural network (NN) and linear regression (LR) models, developed a predictive framework to know the expected value of the application workload. The authors in [7] developed a hybrid prediction method to improve the accuracy of the prediction process. This model is a combination of the moving average methods used in the first stage and the rules of fuzzy neural networks in the second stage, in order to predict the fluctuations of the workload and improve the degree of prediction. The author proposed in [8] a model to perform pre-measurements using genetic methods and backpropagation method for network optimization.

Kumar et al. [9], they proposed an algorithm for predicting the workload of a neural network, using an x-y-z neuron structure, where x, y & z are the amount of neurons in the input sequence.

In [10] the authors developed an ARIMA statistical forecasting approach for workload prediction and estimation of allocated cloud storage resources. A recurrent neural network LSTM (RNN-LSTM) prediction approach is discussed in [11] to predict the load of cloud work on data center providers for resource allocation and reallocation. The researchers compared their proposed schemes with the ARIMA statistical prediction model and demonstrated that the LSTM approach has more effective prediction accuracy.

Herbst et al. [12] Developed a business model for dividing and anticipating the workload, which is a technique for managing self-projected resources.

Table 1. Comparison of some previous studies of workload prediction models

Researchers Year	**Model/ Method**	**Workflow /Strategy**	**Datasets**	**Expected parameters**	**Application/ Simulation tool**	**Outputs**
Zhang et al. [10] (2019)	CP auto-encoder	Multi-parallel decomposition compresses features and learns the Patattern stack auto-encoder for prediction	Planet-Lab traces	CPU utilization	MATLAB	Reach higher training efficiency and prediction accuracy for industrial workloads
Gao et al. [14] (2020)	Bi-LSTM	Classification of training data based on the input layer, one output layer, two layers Bi-LSTM and the LR layer during the learning stages.	55,55,55 tasks traces	Task failure rate	Tensor-flow in Python	Predictions were about 95% accurate and the task failed 90% correctly
Kumar et al. [15] (2018)	LSTM-RNN	The network packets in the chain contain the learning data, and perform a specific step to achieve the predicted network results.	Web server HTTP traces	The order number per unit of time	MATLAB	RMSE lowered to 0.00317
Bi et al.[16] (2019)	SG-LSTM	The S-G model provides loop data to LSTM to predict the desired accuracy	Google tracking	Memory, CPU	Python	SG-LSTM and LSTM were performed
Kardani et al. (2021) [17]	ADRL: Deep RL+Q-Learning	RL algorithm that relies on deep learning Q-learning to solve the problem of memory bottleneck and CPU decision-making.	Rice University Web-Based Bidding System (RUBiS)	CPU, Memory, Response-time	Python with Java-based Cloud-Sim	Improved QoS and stability
Shuvo et al . [18] (2020)	LSRU: LSTM + GRU	Some statistical strategies such as ARMA and ARIMA for forecasting and plugging into the combined GRU and LSTM to improve the prediction score.	Bit-brains	CPU, Disk, memory, bandwidth	Kaggle	Reducing prediction errors across LSTM and GRU
Bi et al. [19] (2019)	SGW-S	Integration of random configuration networks with SG filter and wavelet decomposition to predict workload with great accuracy	Google Cluster traces	task arrival rate	Not mentioned	The SG filter supports wavelet analysis to increase the efficiency of prediction accuracy
Kumar et al. [20] (2020)	E-ELM	Local predictors using the ELM approach are trained using the weight update method with the help of a heuristic model.	Data Flow Statistics traces	Network resource traces	Cloud-Sim	Effective in forecasting required resources

The authors in [13] tackled the problems of workload prediction with the help of developing two methods: date-based prediction and homologous prediction.

A multi-layered task failure prediction system built around Bi-directional long-term memory (Bi-LSTM) was reported by Gao et al. in a paper [14]. A single input level, two Bi-

LSTM levels, an output level, and a layer for logistic regression (LR) are all included in this model to determine if tasks succeeded or failed. Bi-LSTM acts using the forward as well as backward stages in contrast to regular LSTM, which only employs the forward state. This enables a more precise calculation of the values of the closest and furthest input characteristics.

A three-layer neural network with a feed-forward algorithm undergoes training utilizing self-adaptive differential evolution (SaDE) as part of an evolutionary neural network (ENN)-based cloud workload prediction technique that was introduced in [15]. The community of networks is updated during the learning process by performing exploration and exploitation procedures utilizing three mutation techniques sparingly followed by routine intersection. In order to increase the neural network's learning effectiveness, Kumar et al. [15] suggested an alternating training-based Biphase Adaptive Differential Evolution (BaDE) neural network model that incorporated dual adaptation at the interface level during the process of utilization and mutation in the exploration phase. This work, therefore, fared better in regards to accuracy in predicting than SaDE.

Bi et al.'s suggested "BG-LSTM" integrated deep learning approach, which combines Grid-LSTM and Bi-LSTM to provide excellent resources and workload prediction, was presented in paper form by Bi et al. [16]. Before smoothing the workload, it employs a Savitzky-Golay (SG) filtering during preprocessing to lower the standard deviation. For a comparatively long time series, it can successfully extract complicated and nonlinear characteristics while achieving high prediction accuracy.

Regarding flexible resource scalability in a cloud setting, Kardani et al. [17] created an anomaly-aware hybrid capacity scale using Anomaly-aware Deep Reinforcement Learning-based Resource Scaling (ADRL). It presents an approach of deep reinforcement Q-learning driven by decision-making that detects abnormalities in the system and initiates appropriate responses. Both global and local decision-makers are involved in this activity to regulate the crucial measuring methods. With very simple adjustments, ADRL boosts system stability and enhances service quality.

To increase forecast accuracy, Shufu et al. [18] suggested a brand-new hybrid technique termed "LSRU." For short-term forecasting as well as long-term pre-prediction with a spike in workload, LSRU integrates LSTM and Gated Recurrent Units (GRU).

A DNN-based workload model for forecasting with a logarithmic operation before the job is smoothed to lower the standard deviation was put out by Bi et al. [19]. The original data sequence is then subjected to a Savitzky-Golay (S-G) filtering to eliminate points of extreme intensity and overlapping noise. Complex characteristics of significant time series are extracted using DNN-based LSTM (SG-LSTM). The gradient burst problem is eliminated when the Back Propagation Through Time (BPTT) technique is applied, and the resulting model can be improved with the algorithm known as Adam.

A training-based pre-imaging workload paradigm known as "E-ELM" was introduced by Kumar et al. [20] that employs extreme learning machines and related predictions that are evaluated by a voting mechanism. In this study, the best weights were chosen using a metaheuristic algorithm that was motivated by the black hole theory. Regarding Google Clustering trace CPU and memory requests, as well as PlanetLab VM trace CPU consumption, model correctness.

The literature study on techniques for workload prediction in cloud systems indicates a range of solutions with various advantages and disadvantages. The continuous endeavor to create precise and effective workload forecasting algorithms that can accommodate the constantly changing characteristics of systems using cloud computing is highlighted by this. To enhance forecasts for workloads in cloud systems, future research areas might concentrate on investigating hybrid approaches, adding more sophisticated ML methods, and utilizing big data analytics.

3. Workload forecasting activity

Cloud computing provides an opportunity for all online transactional systems to automatically scale resources, which is one feature that distinguishes the cloud model from computing frameworks in general. However, designing such a virtual model in the cloud is challenging, as hosting platforms introduce some lag times when using their own resources. For this reason, analytical prediction of proprietary resource usage is the only solution for virtual machine promotion and some model configuration decisions, for example, automatic rule generation in the cloud and workload management. Figure 1 shows the general model for predicting cloud data center workload. Huge groups of users send requests online in cloud data builder. The response process for requests takes place there and most requests that arrive within a documented forecast period are collected as archived data, which is then used to forecast expected workloads. The archived data is processed and aggregated to normalize it. Benchmarked data is passed to the forecasting framework to predict the expected workload. The projected workload will provide preliminary reports on the upcoming workload, and this gives time for energy management to decide and save its resources. [27]

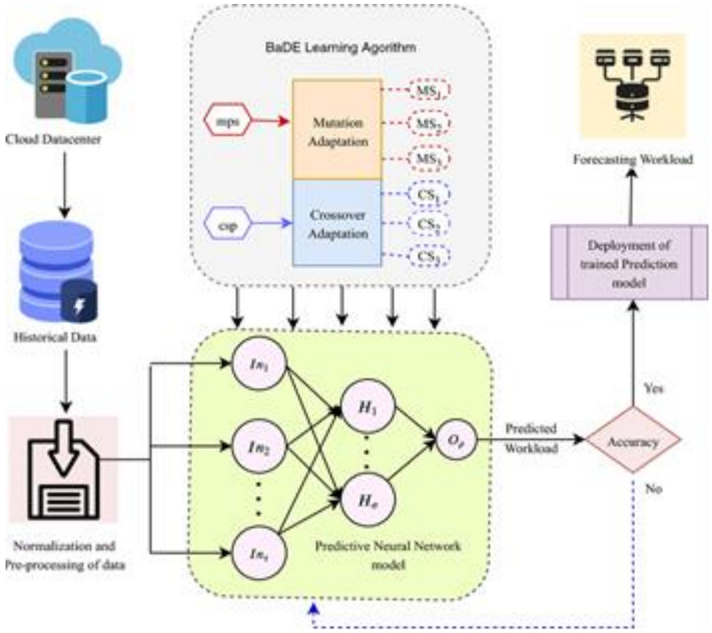

Figure 1: General Workload Forecasting Model [28]

3.1. Deep learning for load prediction

In this section, we present a simplified analysis of a set of workload prediction models based on deep learning (DL) and its advanced systems used in the workload prediction model. Figure 2 shows the distribution of traditional workload forecasting models in the cloud depending on the class of models used in the workload forecasting process.

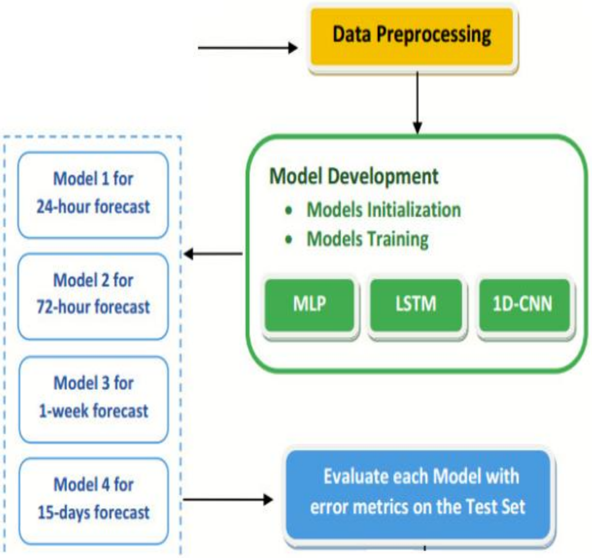
Data Preprocessing
Model Development
Models Initialization
Models Training
MLP
LSTM
1D-CNN
Model 1 for 24-hour forecast
Model 2 for 72-hour forecast
Model 3 for 1-week forecast
Model 4 for 15-days forecast
Evaluate each Model with error metrics on the Test Set

Figure2: Diagram of the workload forecasting system

In this part, we used ML, optimization and some DL models to predict the workload forecasting needs of STLF. For a layered perceptron (MLP), an iterative backpropagation model was used to optimize the feature (tool) by adjusting the amount of neurons within the imaginary layer, this quantity must be large to put the algorithm of question. In addition, to reduce the cost function related to delivery weights, a gradient descent model with backpropagation model is used. [29, 30].

3.1.1. Workload descriptions

In this section each component of the workload is decomposed into other major components as necessary. In web-based environments, users interact through sessions. The core parts of the workload in e-business platforms are characterized by transaction fulfillment criteria and service request rates for each requested resource. Where the navigation algorithm is customized for each client. Examples of e-business functions are: online shopper, product search process based on entered word (key) syllables, product selection to get more information about the product, user registration, purchase process, and account verification.[31]

3.1.2. Forecasting using long term memory (LSTM)

Forecasting information about workload is one of the reasons for sourcing sizing. Optimized resource scaling reduces system cost. A good resource expansion approach helps reduce energy consumption by closing out of useless resources. So the current system is environmentally friendly.

As shown in Figure 3, the predictive module results are fed using a resource management tool that takes into account the specific pattern of the data center before making workload adjustment decisions. If the available resources are insufficient for the expected workload, the resources can swell. While if the amount of resources is more than required, then it can be reduced.

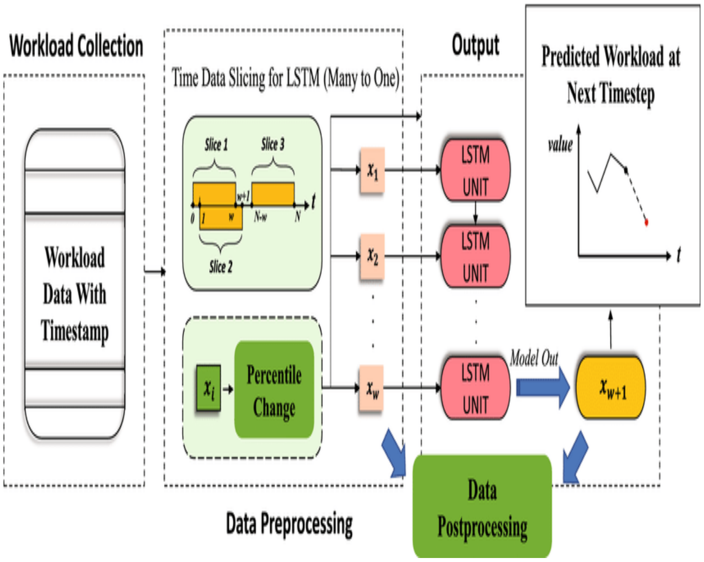

Figure3: A workload time series forecast volume model based on LSTM

3.2. Workload forecasting

A crucial component of successfully managing cloud infrastructures is workload forecasting. In order to appropriately manage resources and guarantee optimal utilization, it entails anticipating the demand for computing resources. This review attempts to give a thorough grasp of workload forecasting methods and the significance of those methods in cloud computing settings.

Recent years have seen a meteoric rise in cloud usage, which has resulted in an exponential rise in the amount of individuals and applications served on platforms in the cloud. To avoid the inadequately or over-provision of capital in response to this rise in demand, precise workload forecasting is necessary. The incorporation of advanced algorithms that modify and automatically adjust depending on actual and historical information is another effort to increase prediction accuracy. Cloud service providers may efficiently distribute resources, guarantee excellent system performance, boost cost-effectiveness, and play a crucial role in workload forecasting for cloud-based systems by properly anticipating workload trends. [32]

The expected workload of web platform services on cloud computing is essential to maintaining service effectiveness. It is of great importance in the operation of web platform services as high pressure on the service can slow down dedicated servers. Speed and support begin to decline as more customers try to access the web platform's service. This method reflects the importance of great forecasting and regulation for cloud web platforms. [33]

4. Discussion

In cloud systems, forecasting workloads is crucial for efficient resource management and allocation. Cloud service providers may assign resources depending on the anticipated demand thanks to accurate workload forecasting, ensuring that customers experience sufficient performance. Cloud providers are better able to scale, provide, and schedule resources by having a thorough understanding of workload patterns and how they change over time. This is necessary to maintain service quality, achieve high resource utilization, and be cost-effective. However, workload planning in cloud settings is fraught with several difficulties. First off, forecasting is challenging since workloads in cloud systems are dynamic and unpredictable. Workloads can differ significantly depending on a number of variables, such as the time of day, the day during the week, and seasonally. Workload planning is made more difficult by the absence of previous information for brand-new and continually changing cloud apps. The variety of cloud workloads presents another problem, making it difficult to construct general-purpose forecasting systems that can precisely anticipate the behavior of various workload types.

Researchers have put forth a number of methods for workload prediction in cloud systems to address these issues. These methods may roughly be divided into three categories: statistical techniques, ML computations, and hybrid strategies. Time-series analysis is one statistical technique that uses previous workload data to simulate and forecast future workload trends. The neural networks along with support vector neural networks are two examples of machine learning techniques that use past workload data to develop predictive algorithms that can predict future workloads. Hybrid methods integrate both machine learning and statistical methods to increase workload forecasting's precision and effectiveness. [34]

Preliminary results from the investigated workload prediction methods extend the exploration of the field further. Based on previous studies, we see that workload prediction algorithms not only help smart resource scaling decisions, but help scale cloud computing by reducing the amount of efficient hardware.

Different architecture packages were also analyzed to find an optimized package, where the quality of service that users hope for is analyzed. The final results of this study motivate the management to make the decision regarding the various work objectives and plan for the future.

Always we need to look for new ways of forecasting the workload in order to reduce the cost and avoid the drawbacks of insufficient basic capacity planning activity.

Cloud computing plays a huge role in enterprise departments in a variety of ways. It affects every model that is developed, such as confidentiality, security, big data, and others. Workload expectation greatly affects user behavior, due to the significant development in cloud computing, as well as the use of IT resources. Workload can be considered as an order, loads and volume of data stored and so on.

Providing workload forecasting has become the norm, as most businesses use the cloud. Inaccurate workload forecasting can have a positive or negative impact on resource savings. If the cloud workload is accurately predicted, the cost and benefits can be reduced for each facility. Forecasting software workload varies by application. For example, the workload needs of an e-business are different from the workload needs of finance and banking. The phases introduced by Almeida [35] can serve as a basis for predicting the cloud capability of any model**.**

4.**Emerging Challenges and Conclusion**

The following challenges can be summarized to develop and increase the effectiveness of more effective management of elastic resources in the cloud environment:

a. Use of a single resource while load balancing: Most of the current elastic resource allocation methods consider a single resource i.e. CPU only, while in the current world cloud resource management all resources matter as a dispute over which one: RAM, used disk space, may be A large load on the processor leads to a deterioration in the overall performance of the system.

b. Improved Workload Forecasting: Accuracy is paramount in forecasting workloads and current models to deliver results remain 100% accurate. An accurate estimate of future resource requirements is essential in achieving highly efficient cloud environments while managing resources efficiently.

c. Sudden changes in workload: When we assess our own workload the categories of factors that affect the workload can be identified.

d. Privacy and Security: Data security is key to moving to the cloud. User data is very sensitive and needs a high degree of security. One of the challenges of computer cloud security is electronic fraud with account theft, infection with malicious applications, and much more, which leads to a loss of trust among application users. [36, 37]

The changing nature of demand patterns makes workload forecasting in cloud settings the most challenging. It is challenging to anticipate future workloads in cloud systems because of user demand variations, seasonal shifts, and the launch of fresh offerings or apps. Furthermore, unforeseen occurrences like abrupt traffic surges or hardware malfunctions can make workload forecasting much more difficult. Traditional workload forecasting techniques could find it difficult to adjust to the cloud environments' quick rate of change. To overcome these difficulties, though, modern technologies like time series analysis and machine learning techniques may be applied. These techniques can assist in identifying patterns, trends, and linkages in historic workload data and may offer forecasts using that data. [38]

The levels of accuracy and reliability of prediction algorithms effectively influence final decisions in cloud data center locations. In the proposed research, a modern technique for predicting future workload in cloud data centers is presented. Cloud data centers offer a range of features including mobility, disaster recovery, and on-demand resources. Flexibility, for example, is one of the modern advantages of the cloud environment that gives the program room to meet resource requirements at any time.

In this study we analyzed several methods for predicting workload in cloud environments. Some of the resulting challenges such as high energy consumption, quality of service violations, waste of resources and other problems were also presented. [39]

This work can be developed in the future to unify the cloud management system and the proposed machine learning model developed using the micro-services architecture model and integrate them into a single system. A model can also be developed based on various features of the workload. [40]

References

[1] Almalaq A.; Zhang J.J. (2020) Deep learning application: Load forecasting in big data of smart grids. In Deep Learning: Algorithms and Applications; Springer: Cham, Switzerland.

[2] D. Saxena, I. Gupta, J. Kumar, A. K. Singh, and X. Wen, (2021) "A secure and multiobjective virtual machine placement framework for cloud data center," IEEE Systems Journal.

[3] Maryam and L. Mohammad (2018) "Survey on prediction models of applications for resources provisioning in cloud", Journal of Network and Computer Applications.

[4] Kumar K.D and E. Umamaheswari (2019) "Prediction methods for effective resource provisioning in cloud computing: A survey", Multi-agent and Grid Systems.

[5] N. Bonvin, T.G. Papaioannou and K. Aberer (2013) "Autonomic SLA drove provisioning for cloud applications", in Proceeding of 11th International Symposium on Cluster, Cloud Grid Computing.

[6] S. Islam, J. Keung, K. Lee and A. Liu (2015) "Empirical prediction models for adaptive resource provisioning in the cloud", Future Generation Computer System.

[7] Z. Chen et al. (2016) "Self-adaptive prediction of cloud resource demands using ensemble model and subtractive-fuzzy clustering based fuzzy neural network", Computational Intelligence and Neuroscience.

[8] T. Dang et al. (2018) "A proactive cloud scaling model based on fuzzy time series and SLA awareness", Procedia Computer Science.

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

[10] Zhang G.; Zhu X.; Yan H.; Bao W.; Tan D. (2019) Local Storage-Based Consolidation with Resource Demand Prediction and Live Migration in Clouds. IEEE Access.

[11] Sudhakar, C.; Kumar, A.R.; Siddhartha, N.; Reddy, S.V. Workload (2019) Prediction using ARIMA Statistical Model and Long Short-Term Memory Recurrent Neural Networks. In Proceedings of the 2019 International Conference on Computing, Power and Communication Technologies.

[12] Nikolas Roman Herbst, Nikolaus Huber, Samuel Kounev, and Erich Amrehn (2015). Self-Adaptive Workload Classification and Forecasting for Proactive Resource Provisioning.

[13] Ardagna D. et al. (2012) Dual time-scale distributed capacity allocation and load redirect algorithms for cloud systems.

[14] J. Gao, H. Wang, and H. Shen, (2020) "Task failure prediction in cloud data centers using deep learning," IEEE trans-actions on services computing.

[15] J. Kumar, R. Goomer, and A. K. Singh, (2019) "Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters," Procedia Computer Science, vol. 125.

[16] J. Bi, S. Li, H. Yuan, Z. Zhao, and H. Liu, (2019) "Deep neural networks for predicting task time series in cloud computing systems," in 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC).

[17] S. Kardani-Moghaddam, R. Buyya, and K. Ramamoha-narao, (2020) "Adrl: A hybrid anomaly-aware deep reinforce-ment learning-based resource scaling in clouds," IEEE Trans. on Parallel and Distributed Systems, vol. 32.

[18] M. N. H. Shuvo, M. M. S. Maswood, and A. G. Alharbi, (2020) "Lsru: A novel deep learning based hybrid method to predict the workload of virtual machines in cloud data center," in 2020 IEEE Region 10 Symposium (TENSYMP).

[19] J. Bi, H. Yuan, and M. Zhou, (2019) "Temporal prediction of multi-application consolidated workloads in distributed clouds," IEEE Trans. on Automation Science and Engineering.

[20] J. Kumar, A. K. Singh, and R. Buyya, (2020) "Ensemble learning based predictive framework for virtual machine resource request prediction," Neuro-computing, vol. 397.

[21] X. Dutreilh, A. Moreau, J. Malenfant, N. Rivierre and I. Truck (2018) "From data center resource allocation to control theory and back", in IEEE 3rd International Conference on Cloud Computing.

[22] M. Dabbagh, B. Hamdaoui, M. Guizani, and A. Rayes, (2017) "An energy efficient vm prediction and migration framework for overcommitted clouds," IEEE Transactions on Cloud Computing, vol. 6.

[23] N. K. Sharma and G. R. M. Reddy, (2017) "Multi-objective energy efficient virtual machines allocation at the cloud data center," IEEE Transactions on Services Computing, vol. 12.

[24] M. Amiri, L. Mohammad-Khanli, and R. Mirandola, (2019) "An online learning model based on episode mining for workload prediction in cloud," Future Generation Computer Systems, vol. 87.

[25] Hochreiter, S., Schmidhuber, J. (2001): Long short-term memory. Neural Comput.

[26] Sutskever I., Vinyals O., Le Q.V. (2015): Sequence to sequence learning with neural networks. Adv. Neural Inf. Proces. Syst.

[27] H. Rong, H. Zhang, S. Xiao, C. Li, and C. Hu, (2017) "Optimizing energy consumption for data centers," Renewable and Sustainable Energy Reviews, vol. 58.

[28] Daradkeh, T.; Agarwal, A.; Zaman, M.; Manzano, R. Analytical Modeling and Prediction of Cloud Workload. In Proceedings of the IEEE ICC 2021 Workshop - I-CPSaaS: Sensing-as-a-Service for Industrial Cyber Physical Systems, Virtual, 14–23 June 2021.

[29] G.-F. Fan, L.-L. Peng, W.-C. Hong, (2019) Short term load forecasting based on phase space reconstruction algorithm and bi-square kernel regression model, Appl. Energy.

[30] A. E. Khantach, M. Hamlich, N. E. Belbounaguia, (2019) Short-term load forecasting using machine learning and periodicity decomposition, AIMS Energy.

[31] Sahi S.K. and Dhaka V.S. (2016) "Study on predicting for workload of cloud services using Artificial Neural Network": Computing for Sustainable Global Development (INDIACom).

[32] Deepika Saxena, Ashutosh Kumar Singh et al. (2021) Workload Forecasting and re-search management models based on machine learning for cloud computing environment.

[33] Maryam and L. Mohammad (2017) "Survey on prediction models of applications for resources provisioning in cloud", Journal of Network and Computer Applications.

[34] Naruei, I.; Keynia, F. (2021) Wild horse optimizer: A new meta-heuristic algorithm for solving engineering optimization problems. Eng. Comput. 38, 3025–3056.

[35] Virgilio A.F. Almeida (2016) "capacity planning for Web Services Techniques and Methodology".

[36] John Grady (2015) "Major Current Trends in Cloud Computing".

[37] M. Dabbagh, B. Hamdaoui, M. Guizani, and A. Rayes, (2016) "Exploiting task elasticity and price heterogeneity for maximizing cloud computing profits," IEEE Transactions on Emerging Topics in Computing, vol. 6.

[38] C. Griner, J. Zerwas, A. Blenk, M. Ghobadi, S. Schmid, and C. Avin, (2022) "Cerberus: The power of choices in datacenter topology design-a throughput perspective," Proceedings of the ACM on Measurement and Analysis of Comp. Systems, vol. 5, no. 3, pp. 1–33.

[39] D. Saxena, A. K. Singh, C.-N. Lee, and R. Buyya, (2023) "A sustainable and secure load management model for green cloud data centers," Scientific Reports.

[40] Z. Ren, J. Wan, and P. Deng, (2023) "Machine-learning-driven digital twin for lifecycle management of complex equipment," IEEE Trans. on Emerg. Topics in Comp.

طرق التنبؤ بحجم العمل في البيئات السحابية: نظرة عامة

ا.د. منار يونس كشمولة
كلية تقنية المعلومات
جامعة نينوى
الموصل، العراق
Manar.kashmola@uomosul.edu.iq

سماح فخري عزيز
قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل،
تعمل في جامعة الحمدانية, الموصل،
العراق
samah.fakhri@uohamdaniya.edu.iq

تاريخ الاستلام:26/3/2023 تاريخ القبول:10/9/2023

الملخص

أصبحت الحوسبة السحابية ذات شعبية متزايدة بسبب تخصيص الموارد حسب الطلب وقابلية التوسع. من الضروري توقع عبء العمل بدقة مع زيادة التطبيقات والمستخدمين على الخدمات المستندة إلى السحابة لتوزيع الموارد بشكل فعال وتجنب انقطاع الخدمة. نقدم لمحة عامة عن أساليب التنبؤ بعبء العمل في الأنظمة السحابية في هذه الدراسة. نحن نستكشف أساليب أكثر تطورًا مثل خوارزميات التعلم العميق (DL) والتعلم الآلي (ML) بالإضافة إلى أساليب أكثر تقليدية مثل تحليل السلاسل الزمنية ونماذج الانحدار. نناقش أيضًا الصعوبات والمسائل البحثية التي لم يتم حلها في مجال التنبؤ بعبء العمل لإعدادات السحابة. يمكن لمقدمي الخدمات السحابية تخصيص الموارد بحكمة وضمان الأداء الجيد وإمكانية الوصول لعملائهم من خلال إدراك هذه التقنيات والمشاكل. تعد الحوسبة السحابية مع المحاكاة الافتراضية والخدمة المخصصة أمرًا بالغ الأهمية لتحسين الخدمة المقدمة للعملاء. يعد التنبؤ الدقيق بعبء العمل أمرًا أساسيًا لتحسين أداء السحابة. سنناقش في هذه الدراسة بعض طرق التنبؤ بعبء العمل في البيئات السحابية.

تقدم هذه الدراسة لمحة عامة عن تقنيات التنبؤ بأعباء العمل في الأنظمة السحابية، مع التركيز بشكل خاص على شبكات الذاكرة طويلة المدى (LSTM). نتناول الأفكار الأساسية وراء شبكات LSTM ومدى قدرتها على اكتشاف العلاقات طويلة الأمد في البيانات من السلاسل الزمنية. نحن ندرس أيضًا الصعوبات والعوامل الخاصة التي ينطوي عليها تنفيذ التنبؤ بأعباء العمل المستندة إلى LSTM في الأنظمة السحابية. نقوم أيضًا بفحص الأبحاث والأساليب السابقة التي استخدمت شبكات LSTM للتنبؤ بعبء العمل في الأنظمة السحابية. نحن ندرس فوائد وعيوب الأساليب المختلفة، مع التركيز على فعاليتها، وقابلية التوسع، وقابلية التفسير.

الكلمات المفتاحية : الذاكرة طويلة المدى (LSTM) ؛ التعلم الآلي(ML) ؛ التعلم العميق (DL) ؛ الشبكات العصبية؛ مراجعة؛ حوسبة سحابية؛ التنبؤ بحجم العمل.