International Journal of Infrastructure Research and Management
Vol. 12 (1) June 2024, pp. 44 - 51

ADAPTABILITY ANALYSIS OF CLOUD ENVIRONMENT AND LOAD PREDICTION ALGORITHM

Jin Yuanrong[1], Li Zhenxiang[1], Wang Haipei[1], Liang Zhantu[1,2] & Tadiwa Elisha Nyamasvisva[1]
[1]*Infrastructure University Kuala Lumpur, MALAYSIA*
[2]*Dongguan University of Technology, CHINA*

ABSTRACT

In the current cloud environment, resource scheduling is an important research field aimed at effectively managing and allocating cloud computing resources to meet user needs and optimize system performance (Yu, 2021). However, resource scheduling and load prediction are two closely related concepts that influence and depend on each other in the cloud environment (Kumar & Sharma, 2020). Load prediction provides an important reference for resource scheduling (Niri et al., 2020; L. Zhang et al., 2021a). By accurately predicting the load situation, resources can be allocated and adjusted in advance before load fluctuations occur, avoiding problems of resource shortage or waste. At the same time, load prediction can also help resource scheduling algorithms better understand load patterns and trends, thereby formulating more reasonable scheduling strategies. It can be said that to a certain extent, load prediction is the basis for resource scheduling. How to carry out precise load prediction has become a typical challenge faced by current research on cloud computing scheduling optimization. This paper first analyses the characteristics of the cloud environment and finds that there are problems such as increasingly obvious dynamic load characteristics, diversified resource requirements, and poor reliability of workflow task execution (Saif et al., 2021; Zhou et al., 2020). Then, starting from the dynamic characteristics of the cloud environment, this paper summarizes and analyzes its impact on cloud resource scheduling (Cao et al., 2022; Peng et al., 2020), and outlines the limitations of traditional load prediction methods (Sideratos et al., 2020; L. Zhang et al., 2021b)in view of the non-stable characteristics of dynamic changes in resource utilization in the cloud environment. The contribution of this paper is to propose a decomposition-prediction algorithm that reduces the impact of the above uncertainties on scheduling by predicting the host load.

Keywords:
Cloud environment, dynamic load characteristics, resource scheduling, load prediction methods, prediction algorithm

INTRODUCTION

Cloud computing, as a new computing model and service category, provides flexible demand allocation, scalable computing services, and elastic resource scheduling for enterprises and users through virtualization, distributed computing, and dynamic scheduling technologies (Bello et al., 2021). It effectively solves problems such as uneven resource sharing (Barrouillet et al., 2007) and low storage efficiency (Abdalla et al., 2022; Nannai John & Mirnalinee, 2020), greatly improving the availability of computing resources. More and more users choose to migrate their applications or data to the cloud to accept its computing or storage services. As the physical carrier of cloud computing, the scale of cloud data centers is expanding (Gao et al., 2022), making the load dynamics of the cloud environment obvious (J. Chen et al., 2023; Rani & Geetha Kumari, 2021).

Load prediction is the process of predicting and estimating the load situation in a future period of time (Fatin et al., 2022; Saripalli et al., 2011; Singh et al., 2021). This makes load prediction that conforms to the characteristics of cloud computing particularly important. Reviewing traditional load prediction algorithms, Moving Average (Schaffer et al., 2021), Exponential Weighted Moving Average (Sukparungsee Id et al., 2020), Autoregressive Moving Average (Prado et al., 2020), and Neural Networks (Chicco, 2021; Gawlikowski Student Member et al., 2021; Li et al., 2022) all have a strong dependence on historical data, which does not fit well with the dynamic load of the cloud environment.

Therefore, this study uses Multiple Prediction Combination Methods to overcome the limitations of traditional methods. It is expected that the method proposed in this study will be better adapted to the characteristics of cloud computing. This paper mainly introduces the research topic, research motivation, problem statement, and conclusion.

RESEARCH MOTIVATION

The motivation for this study lies in the continuous development of cloud computing technology, which has higher requirements for the adaptability of load prediction methods.

Moving Average is a simple and commonly used real-time load prediction algorithm (Prado et al., 2020). It predicts future loads based on the average value of historical load data. The moving average algorithm is simple to use, has low computational complexity and real-time performance, and is suitable for stationary or slowly changing load situations. However, the moving average algorithm has poor adaptability to rapidly changing and nonlinear load patterns.

Exponential Weighted Moving Average (EWMA) is a real-time load prediction algorithm based on exponential weighting (Nyamasvisva et al., 2022; Sukparungsee Id et al., 2020). It performs a weighted average of historical load data, with newer data having higher weights. The EWMA algorithm can adapt to changes in load more quickly and has a certain degree of real-time performance and accuracy. However, the EWMA algorithm is sensitive to sudden changes or abnormal data in the load, which may cause error accumulation.

The Autoregressive Moving Average (ARMA) model combines the characteristics of autoregression (AR) and moving average (MA) for real-time load prediction (Schaffer et al., 2021). The ARMA model considers the historical data and error terms of the load, and predicts future loads through parameter estimation and model fitting. The ARMA model is suitable for load data with certain autocorrelation and trends. However, the parameter estimation and model fitting of the ARMA model are relatively complex and need to be adjusted and optimized according to specific situations.

Neural network models are also widely used in real-time load prediction (Chicco, 2021; Li et al., 2022). Among them, recurrent neural networks (RNN) and long short-term memory networks (LSTM) are common models. These models can capture the temporal characteristics and complex relationships of load data, and have strong nonlinear modeling capabilities. Neural network models can achieve relatively accurate real-time load prediction, but require a large amount of training data and computational resources, and the adjustment of hyperparameters and model optimization are relatively complex.

The Kalman filter algorithm has the advantages of efficiency and accuracy, making it suitable for state estimation and prediction in dynamic systems (Khodarahmi, et al., 2022). It can also be combined with other algorithms for applications in emerging fields like cloud computing. However, the algorithm relies on linear assumptions and noise models, with its performance being significantly affected by initial conditions and parameter settings. Additionally, the computational cost cannot be ignored when dealing with large-scale complex systems. Therefore, it is essential to leverage its strengths and address its weaknesses by optimizing algorithm parameters and models to enhance its effectiveness in specific application scenarios.

From Table 1, it can be observed that the moving average method is simple and easy to use, suitable for stable loads, but weak in adapting to rapid changes. The Exponentially Weighted Moving Average (EWMA) responds quickly to load changes, but is sensitive to abrupt data and tends to accumulate errors. The Autoregressive Moving Average (ARMA) model is suitable for autocorrelated loads, but parameter fitting is complex. Neural network models, such as RNN and LSTM, can precisely capture complex load relationships, but have high training costs and complex optimization. Therefore, the Kalman filter algorithm aligns well with the characteristics of cloud environments.

Table 1: SWOT Analysis of Existing Load Prediction Algorithms

	Algorithm	Strengths	Weaknesses	Opportunities	Threats
1	Moving Average (MA) (Prado et al., 2020)	Based on the average of historical load data to predict future load, easy to use, fast calculation	Poor adaptability to rapid changes and nonlinear load patterns	Suitable for flat or slowly changing loads	Poor adaptability to dynamic data
2	Exponential Weighted Moving Average，（EWMA) (Nyamasvisva et al., 2022; Sukparungsee Id et al., 2020)	The weighted average of historical load data, the more recent data has a higher weight, with a certain real time and accuracy.	More sensitive to load mutations or abnormal data	Able to adapt to changes in load faster	It may result in accumulation of errors.
3	Autoregressive Moving Average，（ARMA）(Schaffer et al., 2021)	Taking into account the historical data and errors of loads and predicting future loads through parameter estimates and models.	Parameter estimates and model adaptation are more complex	Applicable to load data with a certain relevance and trend	Need to be adjusted and optimized according to specific load conditions
4	Neural Networks (NN) (Chicco, 2021; Li et al., 2022)	Able to capture timing characteristics and complex relationships of load data	It requires a lot of training data and computational resources	Strong non-linear modelling capabilities	Adjustment and model optimization for super-parameters are more complex
5	Kalman Filter Algorithm (Khodarahmi, et al., 2022)；	Has the advantages of efficiency and accuracy	The algorithm relies on linear assumptions and noise models	Can be combined with other algorithms for applications in emerging fields like cloud computing	The computational cost cannot be ignored when dealing with large-scale complex systems

STATEMENT OF THE PROBLEM – Load Prediction Algorithm

Many scholars have pointed out that due to the high scalability and flexibility of cloud computing, it has received increasing attention, and cloud services supported by it have become a new IT service model (Javadpour et al., 2022; Mapetu et al., 2021; Zhu et al., 2019). More and more users choose to migrate applications or data to the cloud to accept its computing or storage services. The scale of cloud data centers, as the physical carrier of cloud computing, is expanding, making the load dynamics in the cloud environment obvious.

Other scholars pointed out that workload prediction algorithms based on statistical methods lack adaptability to highly variable workloads (Y. Chen et al., 2020; Gao et al., 2020). In addition, some scholars also pointed out that workload prediction algorithms based on classical machine learning require manual feature extraction and model parameter adjustment, which is both difficult and time-consuming (Gao et al., 2020; Zhu et al., 2019). Additionally, scholars pointed out that workload prediction algorithms based on deep learning do not require manual feature extraction, but their prediction accuracy is limited (Gao et al., 2020; Toumi et al., 2019).

There are also scholars who pointed out that neural network algorithms or linear regression methods cannot predict real loads with large fluctuations well (Toumi et al., 2019; Xu et al., 2022). Although the use of ensemble learning has a more accurate final learning effect, the nonlinear characteristics of the load sequence cannot achieve satisfactory real value prediction, and the prediction time is too long to predict real-time loads.

In summary, with the development of cloud computing and cloud data centers, the cloud environment is becoming more complex. Cloud environment workload prediction faces problems such as obvious dynamic characteristics of the load, low prediction accuracy, and poor real-time performance of prediction algorithms.

PROPOSAL

As described earlier, it is particularly important to propose a load forecasting algorithm that can better adapt to cloud environments. Therefore, this article proposes combining the Kalman filter algorithm with the EMD algorithm, aiming to better adapt to the characteristics of cloud environments.

The Kalman Filter Algorithm has good performance in linear system models and real-time application scenarios (Khodarahmi, et al., 2022). Through optimal estimation and recursive updating, it provides accurate state estimation and prediction results. It also has dynamic model adaptability and low computational complexity, and is suitable for many application fields that require real-time, accurate and efficient filtering.

The EMD algorithm has the advantages of adaptability, being data-driven, flexibility, no prior assumptions, and time locality. These characteristics make the EMD algorithm widely used in signal processing, vibration analysis, modal analysis, and other fields, providing more accurate, comprehensive, and reliable signal decomposition and feature extraction results (Quinn et al., 2021; Y. Zhang et al., 2022).

This study uses the prediction method of multiple prediction combination methods to propose a decomposition-prediction method. The schematic diagram is shown in Figure 1. By processing the original dynamic data through the EMD algorithm and then predicting the load through the Kalman Filter Algorithm, it aims to both adapt to the dynamic load characteristics of the cloud environment and ensure real-time prediction accuracy.

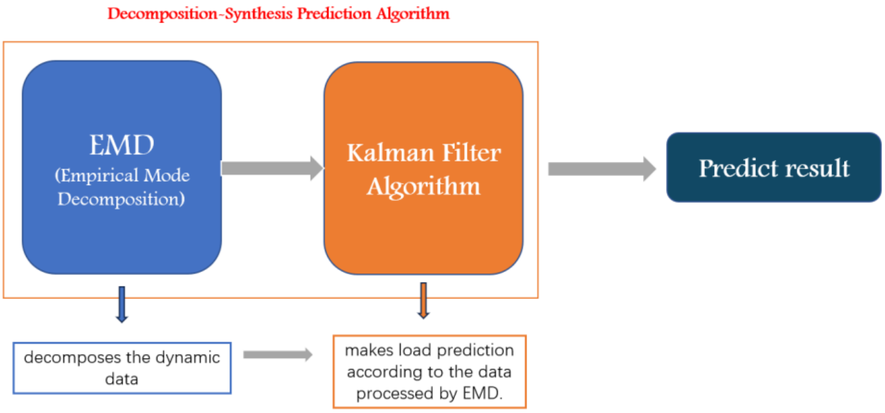

Figure 1: Decomposition-Synthesis Prediction Algorithm diagram.

CONCLUSION

This article analyzes the importance of load forecasting technology and the relationship between resource scheduling and load forecasting. It also identifies existing problems. The SWOT method is used to evaluate different load forecasting methods and analyze the advantages and disadvantages of algorithms. A solution is proposed: the prediction method of multiple prediction combination methods to propose a decomposition-prediction method.

A good load forecasting algorithm should be able to accurately and adaptively predict the future trend and pattern of load changes, while having scalability, robustness, interpretability, and comprehensive performance. Such an algorithm can provide strong support for resource scheduling and load balancing in a cloud environment, improving system performance and efficiency.

AUTHORS BIOGRAPHY

Jin Yuanrong is student of the postgraduate programme PhD (Information Technology) at Infrastructure University Kuala Lumpur (IUKL) Faculty of Engineering, Science and Technology. Her research interests include Cloud Computing and Load Prediction algorithms.
Email: 222923382@s.iukl.edu.my

Li Zhenxiang is student of the postgraduate programme PhD (Information Technology) at Infrastructure University Kuala Lumpur (IUKL) Faculty of Engineering, Science and Technology. His research interests include Blockchain and Cloud Computing. *Email: 222923380@s.iukl.edu.my*

Wang Haipei is student of the postgraduate programme PhD (Information Technology) at Infrastructure University Kuala Lumpur (IUKL) Faculty of Engineering, Science and Technology. Her research interests include Cloud Computing and cyber security. *Email: 223923726@s.iukl.edu.my*

Liang Zhantu is a Ph.D. in IT candidate at IUKL. His research direction is in human action recognition in the field of artificial intelligence. Liang Zhantu is currently working as an Information and Technology teacher at DGUT in China. *Email: 223923795@s.iukl.edu.my*

Tadiwa Elisha Nyamasvisva, PhD is a member at the Faculty of Engineering and Science Technology in IUKL. His research interests are in Computer Algorithm Development, Data Analysis, Networking and Network Security, and IT in Education. *Email: tadiwa.elisha@iukl.edu.my*

REFERENCES

Abdalla, A., Arabi, M., Nyamasvisva, T. E., & Valloo, S. (2022). ZERO TRUST SECURITY IMPLEMENTATION CONSIDERATIONS IN DECENTRALISED NETWORK RESOURCES FOR INSTITUTIONS OF HIGHER LEARNING. International Journal of Infrastructure Research and Management, 10(1), 79–90. https://iukl.edu.my/rmc/publications/ijirm/

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and Cognitive Load in Working Memory. Journal of Experimental Psychology: Learning Memory and Cognition, 33(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570

Bello, S. A., Oyedele, L. O., Akinade, O. O., Bilal, M., Davila Delgado, J. M., Akanbi, L. A., Ajayi, A. O., & Owolabi, H. A. (2021). Cloud computing in construction industry: Use cases, benefits and challenges. Automation in Construction, 122, 103441. https://doi.org/10.1016/J.AUTCON.2020.103441

Cao, B., Zhang, J., Liu, X., Sun, Z., Cao, W., Nowak, R. M., & Lv, Z. (2022). Edge-Cloud Resource Scheduling in Space-Air-Ground-Integrated Networks for Internet of Vehicles. IEEE Internet of Things Journal, 9(8), 5765–5772. https://doi.org/10.1109/JIOT.2021.3065583

Chen, J., Han, P., Liu, Y., & Du, X. (2023). Scheduling independent tasks in cloud environment based on modified differential evolution. Concurrency and Computation: Practice and Experience, 35(13), e6256. https://doi.org/10.1002/CPE.6256

Chen, Y., Wang, L., Chen, X., Ranjan, R., Zomaya, A. Y., Zhou, Y., & Hu, S. (2020). Stochastic Workload Scheduling for Uncoordinated Datacenter Clouds with Multiple QoS Constraints. IEEE Transactions on Cloud Computing, 8(4), 1284–1295. https://doi.org/10.1109/TCC.2016.2586048

Chicco, D. (2021). Siamese Neural Networks: An Overview. Methods in Molecular Biology, 2190, 73–94. https://doi.org/10.1007/978-1-0716-0826-5_3/COVER

Fatin, F., Majid, S., Syafiq, M., & Mohamed, N. (2022). DEVELOPMENT OF SURVEILLANCE SYSTEM WITH AUTOMATED EMAIL AND TELEGRAM NOTIFICATION USING OPEN-SOURCE APPLICATION PROGRAMMING INTERPHASE (API). International Journal of Infrastructure Research and Management, 10(2), 39–49. https://iukl.edu.my/rmc/publications/ijirm/

Gao, J., Wang, H., & Shen, H. (2020). Machine Learning Based Workload Prediction in Cloud Computing. Proceedings - International Conference on Computer Communications and Networks, ICCCN, 2020-August. https://doi.org/10.1109/ICCCN49398.2020.9209730

Gao, J., Wang, H., & Shen, H. (2022). Task Failure Prediction in Cloud Data Centers Using Deep Learning. IEEE Transactions on Services Computing, 15(3), 1411–1422. https://doi.org/10.1109/TSC.2020.2993728

Gawlikowski Student Member, J., Rovile Njieutcheu Tassi, C., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung Member, P., Roscher Member, R., Shahzad, M., Yang Senior Member, W., Bamler Fellow, R., & Xiang Zhu Fellow, X. (2021). A Survey of Uncertainty in Deep Neural Networks. https://arxiv.org/abs/2107.03342v3

Javadpour, A., Abadi, A. M. H., Rezaei, S., Zomorodian, M., & Rostami, A. S. (2022). Improving load balancing for data-duplication in big data cloud computing networks. Cluster Computing, 25(4), 2613–2631. https://doi.org/10.1007/S10586-021-03312-5/METRICS

Khodarahmi, M., & Maihami, V. (2022). A Review on Kalman Filter Models. Archives of Computational Methods in Engineering, 30(1), 727–747. https://doi.org/10.1007/s11831-022-09815-7

Kumar, M., & Sharma, S. C. (2020). PSO-based novel resource scheduling technique to improve QoS parameters in cloud computing. Neural Computing and Applications, 32(16), 12103–12126. https://doi.org/10.1007/S00521-019-04266-X/METRICS

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Mapetu, J. P. B., Kong, L., & Chen, Z. (2021). A dynamic VM consolidation approach based on load balancing using Pearson correlation in cloud computing. Journal of Supercomputing, 77(6), 5840–5881. https://doi.org/10.1007/S11227-020-03494-6/METRICS

Nannai John, S., & Mirnalinee, T. T. (2020). A novel dynamic data replication strategy to improve access efficiency of cloud storage. Information Systems and E-Business Management, 18(3), 405–426. https://doi.org/10.1007/S10257-019-00422-X/METRICS

Niri, M. F., Bui, T. M. N., Dinh, T. Q., Hosseinzadeh, E., Yu, T. F., & Marco, J. (2020). Remaining energy estimation for lithium-ion batteries via Gaussian mixture and Markov models for future load prediction. Journal of Energy Storage, 28, 101271. https://doi.org/10.1016/J.EST.2020.101271

Nyamasvisva, T. E., Abdalla, A., & Arabi, M. (2022). A COMPREHENSIVE SWOT ANALYSIS FOR ZERO TRUST NETWORK SECURITY MODEL. International Journal of Infrastructure Research and Management, 10(1), 44–53. https://iukl.edu.my/rmc/publications/ijirm/

Peng, Z., Lin, J., Cui, D., Li, Q., & He, J. (2020). A multi-objective trade-off framework for cloud resource scheduling based on the Deep Q-network algorithm. Cluster Computing, 23(4), 2753–2767. https://doi.org/10.1007/S10586-019-03042-9/METRICS

Prado, F., Minutolo, M. C., & Kristjanpoller, W. (2020). Forecasting based on an ensemble Autoregressive Moving Average - Adaptive neuro - Fuzzy inference system – Neural network - Genetic Algorithm Framework. Energy, 197, 117159. https://doi.org/10.1016/J.ENERGY.2020.117159

Quinn, A. J., Lopes-dos-Santos, V., Dupret, D., Nobre, A. C., & Woolrich, M. W. (2021). EMD: Empirical Mode Decomposition and Hilbert-Huang Spectral Analyses in Python. Journal of Open Source Software, 6(59), 2977. https://doi.org/10.21105/JOSS.02977

Rani, D. R., & Geethakumari, G. (2021). A framework for the identification of suspicious packets to detect anti-forensic attacks in the cloud environment. Peer-to-Peer Networking and Applications, 14(4), 2385–2398. https://doi.org/10.1007/S12083-020-00975-6/METRICS

Saif, M. A. N., Niranjan, S. K., & Al-ariki, H. D. E. (2021). Efficient autonomic and elastic resource management techniques in cloud environment: taxonomy and analysis. Wireless Networks, 27(4), 2829–2866. https://doi.org/10.1007/S11276-021-02614-1/METRICS

Saripalli, P., Kiran, G. V. R., Shankar, R. R., Narware, H., & Bindal, N. (2011). Load prediction and hot spot detection models for autonomic cloud computing. Proceedings - 2011 4th IEEE International Conference on Utility and Cloud Computing, UCC 2011, 397–402. https://doi.org/10.1109/UCC.2011.66

Schaffer, A. L., Dobbins, T. A., & Pearson, S. A. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. BMC Medical Research Methodology, 21(1), 1–12. https://doi.org/10.1186/S12874-021-01235-8/FIGURES/5

Sideratos, G., Ikonomopoulos, A., & Hatziargyriou, N. D. (2020). A novel fuzzy-based ensemble model for load forecasting using hybrid deep neural networks. Electric Power Systems Research, 178, 106025. https://doi.org/10.1016/J.EPSR.2019.106025

Singh, A. K., Saxena, D., Kumar, J., & Gupta, V. (2021). A Quantum Approach towards the Adaptive Prediction of Cloud Workloads. IEEE Transactions on Parallel and Distributed Systems, 32(12), 2893–2905. https://doi.org/10.1109/TPDS.2021.3079341

Sukparungsee Id, S., Areepong, Y., & Taboran, R. (2020). Exponentially weighted moving average-Moving average charts for monitoring the process mean. https://doi.org/10.1371/journal.pone.0228208

Toumi, H., Brahmi, Z., & Gammoudi, M. M. (2019). RTSLPS: Real time server load prediction system for the ever-changing cloud computing environment. https://doi.org/10.1016/j.jksuci.2019.12.004

Xu, M., Song, C., Wu, H., Gill, S. S., Ye, K., & Xu, C. (2022). esDNN: Deep Neural Network Based Multivariate Workload Prediction in Cloud Computing Environments. ACM Transactions on Internet Technology (TOIT), 22(3). https://doi.org/10.1145/3524114

Yu, H. (2021). Evaluation of cloud computing resource scheduling based on improved optimization algorithm. Complex and Intelligent Systems, 7(4), 1817–1822. https://doi.org/10.1007/S40747-020-00163-2/FIGURES/5

Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., & Livingood, W. (2021a). A review of machine learning in building load prediction. Applied Energy, 285, 116452. https://doi.org/10.1016/J.APENERGY.2021.116452

Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., & Livingood, W. (2021b). A review of machine learning in building load prediction. Applied Energy, 285, 116452. https://doi.org/10.1016/J.APENERGY.2021.116452

Zhang, Y., Li, C., Jiang, Y., Sun, L., Zhao, R., Yan, K., & Wang, W. (2022). Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model. Journal of Cleaner Production, 354, 131724. https://doi.org/10.1016/J.JCLEPRO.2022.131724

Zhou, Z., Li, F., Zhu, H., Xie, H., Abawajy, J. H., & Chowdhury, M. U. (2020). An improved genetic algorithm using greedy strategy toward task scheduling optimization in cloud environments. Neural Computing and Applications, 32(6), 1531–1541. https://doi.org/10.1007/S00521-019-04119-7/METRICS

Zhu, Y., Zhang, W., Chen, Y., & Gao, H. (2019). A novel approach to workload prediction using attention-based LSTM encoder-decoder network in cloud environment. Eurasip Journal on Wireless Communications and Networking, 2019(1), 1–18. https://doi.org/10.1186/S13638-019-1605-Z/FIGURES/12