JOURNAL OF SOUTHERN COMMUNICATION ENGINEERING
ISLAMIC AZAD UNIVERSITY BUSHEHR BRANCH
E-ISSN: 2980-9231
https://jce.bushehr.iau.ir
https://doi.org/10.30495/jce.2025.1993480.1327

Vol. 14/ No. 54/Winter 2025

Research Article

Spot Price Prediction of Resources in Cloud Computing by Proposing a New Structure in Deep Learning Method Considering the Level of QOS

Seyed Soroush Nezamdoust, Ph.D. Student[1] | Mohammad Ali Pourmina, Associate Professor[2] | Farbod Razzazi, Associate Professor[3]

[1]Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University (IAU), Tehran, Iran, soroush.nezamdoust@srbiau.ac.ir

[2]Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University (IAU), Tehran, Iran, pourmina@srbiau.ac.ir

[3]Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University (IAU), Tehran, Iran, razzazi@srbiau.ac.ir

Correspondence
Mohammad Ali Pourmina, Associate Professor of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University (IAU), Tehran, Iran,
pourmina@srbiau.ac.ir

Received: 5 November 2023
Revised: 13 January 2024
Accepted: 17 February 2024

Abstract

Cloud computing is a computing model that uses three instance, on-demand, reserved, and spot, to provide resources to users. The price of spot instances is on average lower than other patterns and fluctuates based on supply and demand. When a user requests a spot instance, they must provide an offer. Only if the price offered by the user is higher than the spot price, the user can use this type of resources. Therefore, predicting the price of spot instances is very important and challenging. Forecasting such dynamic time series that follow the nonlinear model requires intelligent tools such as neural networks to be able to predict the future values with the least error by observing the values of a time series. Therefore, the reliability and as a result the quality of the service is improved. For this purpose, we considered Amazon EC2 as an experimental platform and used the spot price history to predict the future price by building a new model based on deep learning. The obtained results showed that the model presented in the article based on the proposed structure of MGRU(modified GRU) can well predict nonlinear values and perform better than other methods used in this field.

Keywords: Spot price prediction, Cloud computing, Deep neural network, Modified GRU(MGRU).

Highlights

- Examining deep learning structures for predicting time series.
- Providing an efficient and powerful algorithm to analyze the historical developments of Amazon EC2 spot prices and predict the future price of resources.
- Presenting a proposed architecture based on modified GRU (MGRU).
- Forecasting price trends in the future with the aim of improving the quality of services.
- Accurate prediction of real-world time series with highly volatile data.

Citation: SS. Nezamdoust, MA. Pourmina, and F. Razzazi, "Spot Price Prediction of Resources in Cloud Computing by Proposing a New Structure in Deep Learning Method Considering the Level of QOS," *Journal of Southern Communication Engineering*, vol. 14, no. 54, pp. 1–16, 2025, doi:10.30495/jce.2025.1993480.1327, [in Persian].

1. Introduction

Cloud Computing is a structure that enables easy access to resources based on user demand through a network infrastructure without the need for initial capital investment. Clouds offer a scalable and flexible environment with diverse and cost-effective payment models. As a result, in recent years, cloud computing has experienced significant growth due to its advantages and has become a suitable model for implementing applications with high reliability and security. Today, cloud computing has become a notable and appropriate option for accessing infrastructure, software platforms, and software as a service for commercial companies and research topics [1].

The Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) models provide computing capabilities and storage space as standard services over the network. In this way, users receive these resources as infrastructure services from the cloud service provider instead of purchasing servers, data center space, and networking equipment. Amazon Web Services (AWS) is a set of cloud services offered by Amazon. In this model, the IaaS service provider is recognized as the owner of the equipment and is responsible for providing the physical location, setup, and maintenance of the equipment [2].

In the Platform as a Service (PaaS) model, the responsibility for executing and maintaining the software system and the underlying computing infrastructure falls on the cloud service provider. This service provides a software layer in a packaged format that can be used to produce higher-level services. In this model, users are given the opportunity to directly design, develop, and test their desired application in the cloud. An example in this area is the Google App Engine, which allows for the implementation and execution of applications through the infrastructure created by Google [2-3].

Software as a Service (SaaS) refers to providing application programs on demand over the internet. These applications do not require installation and setup on customer computers. SaaS delivers software as a service over the internet, allowing users to connect and benefit from it. This way, software service delivery on the cloud is very easy, with updates, maintenance, and support handled centrally in the cloud. However, customization and modification of these applications are not available. Examples of this method in current use include CRM, Salesforce, YouTube, and Gmail [2-3]. Currently, the revenue from cloud services and forecasts related to it indicate that IaaS is the fastest growing segment, a model that provides users with infrastructure resources based on virtual machines and allows the rental of virtual machines with varying capacities in an elastic manner [3]. Computing services are offered to users with diverse pricing models. The challenge of pricing services, with an approach that allows the service provider to achieve maximum benefit while maintaining service quality and user satisfaction, is one of the most significant fundamental challenges for cloud service providers.

In the static pricing model, demanded resources are provided for a specified fee over an agreed period, which can include an on-demand pricing mechanism. In the demand-based pricing model, users pay a fixed rate depending on the type of region and area available, with costs varying based on the number of required processor cores, processor execution speed, memory size, and other influencing parameters. Fixed-price virtual machines have high reliability and availability but are priced higher than other models [4]. Additionally, presenting a computational model based on supply and demand and the conditions prevailing in the environment introduces a new concept called dynamic pricing. Spot pricing is based on the dynamic pricing model and aims to incentivize users to utilize excess resources. The provider (such as Amazon) reprices its unused surplus resources and rents them out at the lowest possible discount (base price) to profit while preventing additional maintenance costs for the resources.

The pricing process for these resources occurs once every hour, and interested users must submit a bid for their desired resource. If the submitted bid exceeds the base price, the resource is allocated to the user for one hour. At the end of the hour, a new pricing evaluation is conducted, and if the user's previous bid is higher than the new price, the resource rental continues for another hour until the user's bid falls below the base price. If the user's price drops below the new base price, the resources are reclaimed. This situation is known as a "bid failure," and Amazon activates a warning two minutes prior to this event. Such interruptions can lead to data loss and undermine the reliability of this method, presenting various challenges [4-5]. Consequently, in various proposals in recent years, more attention has been given to the methods of acquiring these resources and strategies to enhance their performance.

One of the strategies employed in recent years has been the accurate prediction of spot pricing. Generally, accurate forecasting of this pricing model leads to reduced risk, increased reliability, enhanced availability, and consequently improved quality of service for the spot pricing model. Optimizing the timing of purchases with an awareness of future pricing trends offers another advantage of accurate price forecasting. As a result, users' inclination to utilize this model has increased, allowing them to benefit from lower costs while maintaining service quality.

On the other hand, the topic of forecasting has been explored in many studies. Among these, time series forecasting techniques are categorized into various groups. Traditional statistical models, such as the Autoregressive Integrated Moving Average (ARIMA) model, are less commonly used in the real world because they involve linear components. Conversely, for forecasting time series with nonlinear patterns, several nonlinear statistical methods have been proposed, such as Autoregressive Conditional Heteroskedasticity (ARCH) models and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models, which are usually suitable for specific nonlinear models. The process of finding an appropriate model for time series in the real world is complex. In cloud computing systems, there is also significant variance in data centers; therefore, effective forecasting methods must be employed that are compatible with the highly variable data of the cloud computing environment [6-7].

In recent years, computational intelligence techniques such as Artificial Neural Networks (ANNs) have been used for time series forecasting issues. ANNs have enhanced the landscape of information technology, and due to their general approximation capabilities, data-driven nature, and ability to model nonlinear patterns, they have led to significant advancements in time series forecasting [8]. Recurrent Neural Networks (RNNs) are a crucial component of neural networks and are very effective in time series forecasting problems due to their excellent ability to process sequences. However, RNNs struggle to effectively learn long-term memory dependencies due to the vanishing gradient problem, which poses a significant challenge in forecasting spot prices in cloud computing. The introduction of Long Short-Term Memory (LSTM) models, Gated Recurrent Units (GRUs), and Transformers has resolved many of the issues associated with recurrent neural networks because of their extensive capacity for

managing information. Nevertheless, these methods also have limitations in time series forecasting, leading to various studies addressing these important issues [9-12].

For spot price prediction, Singh, and Dutta [13] suggested an autoregressive-based model. In the complicated cloud space, linear models do not function well. Moving average techniques (such as basic, weighted, and exponential) are used by the authors in [14] to predict next-hour spot prices using estimates, and Alkharif et al. use a Seasonal-Arima (SARIMA) model in [15]. Spot instances have been offered to improve the accuracy of price prediction. Due to the extreme volatility of spot prices on the cloud platform, these models are often incompatible. The k-Nearest Neighbors (kNN) regression model was introduced in the article [16] to achieve the best performance in spot price prediction. Despite its benefits, kNN often comes with a significant computational cost.

To overcome the RNN challenge, LSTM networks with the ability to learn long-term dependencies for time series sequences were proposed. In [17] used LSTM deep neural network structure in Amazon EC2 to predict spot prices. Similarly, Kong and Dong in [18] proposed an LSTM-based model for estimating spot prices using the 90-day past price data history provided by Amazon. GRU is another model derived from RNN that, despite its structural advantages and computational efficiency, little research has been conducted on using it to predict cloud computing [19]. Overall, limited research on spot price assessment and prediction has not been able to comprehensively and accurately assess the challenges in this area, which can lead to unnecessarily time-consuming, high computational loads and reduce the accuracy of the expected price. Accordingly, we designed a framework that takes into account a variety of concerns and leads to improved prediction accuracy and computational efficiency in a cloud computing environment. The accuracy of the proposed approach is confirmed by comparison with several other models.

2. Innovation and contributions

In this paper a comprehensive review of the structures of neural networks has been conducted to effectively leverage the capabilities of these networks in data forecasting. In this context, an efficient and powerful algorithm has been proposed in the field of forecasting to effectively analyze historical price fluctuations of Amazon EC2 spot prices and predict future resource prices. Among the innovations of this paper, the following can be stated: The proposed architecture, based on a modified Gated Recurrent Unit (GRU), aims for adaptive and more accurate forecasting of spot prices, allowing users to manage the high volatility of spot instances and consequently provide precise predictions of future price trends. This approach enables optimal timing for purchases, offering the best price while avoiding excessively high bids that lead to increased costs or bids that are too low, which may result in an inability to use the instances.

3. Materials and Methods

Historical Price Information from the cloud data center is used in the proposed prediction model. To make collected data more applicable, the raw data $\vec{X} = (x_1, x_2, x_3, \dots, x_n)$ must first be pre-processed to length n. Data cleaning and normalization are two important data preparation tasks that should be implemented in the data preparation stage of the proposed prediction method.

Before the training phase in the proposed model, a subset of values for each hyper parameter must be defined, and a combination of the best hyper parameters is estimated in each iteration. Adjusting and finding the hyper parameter values is a powerful process for identifying the best possible values to achieve optimal modeling results. Using grid search, we determine six parameters before training the proposed model, including the number of samples in each batch, the number of hidden layers, the number of neurons in each layer, the learning rate, the number of epochs, and the update gate coefficient.

Next, we proceed to train the proposed structure of the paper. Given the features and capabilities of the GRU model discussed earlier, we propose a modified structure known as MGRU (Modified GRU) by implementing structural changes and necessary adjustments. Two solutions are proposed in this paper to address the linear limitations and increase the learning rate of the basic GRU model. Finally, after making the necessary changes, the function of the proposed MGRU model is defined as follows:

$$C_t = \gamma \Gamma_u \odot C_{t-1} + \sqrt{1-\gamma^2 {\Gamma_u}^2} \odot \hat{C}_t \qquad \gamma \epsilon (0,1) \tag{1}$$

4. Results and Discussion

To evaluate the performance of the proposed model, the ARIMA, RNN, Transformer, LSTM, GRU models, and the model presented in this thesis were compared using similar datasets. Initially, we compared the performance of different models in forecasting spot prices for three diverse virtual machines using two metrics: MAE (Mean Absolute Error) and RMSE (Root Mean Square Error). The outputs indicated lower errors and, consequently, better performance for the proposed method. Subsequently, we assessed the accuracy of the examined structures using the R^2 evaluation criterion. Analysis of the results and accuracy levels of the various models indicates that forecasting spot prices using the proposed method achieves higher accuracy compared to the other examined methods. The results showed that computational intelligence techniques have strong capabilities in predicting real-world data, with the model presented in this paper demonstrating the best performance in various aspects. Moreover, as the forecasting horizon increases, the prediction error in the proposed model showed less growth, indicating its high capability in addressing issues with long-term memory dependencies. This feature can help maintain dynamic change conditions throughout price history periods and enable accurate forecasting.

5. Conclusion

Using an efficient method enables users to overcome the high volatility of spot instances and, consequently, accurately predict future price trends. This allows for optimal timing in purchasing, helping to avoid excessively high bids that lead to increased costs, or bids that are too low, which may result in an inability to utilize the instances. Thus, it can be concluded that the higher accuracy and lower error of the proposed model enhance the reliability and availability characteristics of the spot pricing pattern. The improvement of these two features will lead to an overall enhancement in service quality.

6. Acknowledgement

The authors would like to acknowledge the valuable comments and suggestions of the reviewers, which have improved the quality of this paper.

7. References

[1] L. Teylo, L. Arantes, P. Sens, and L. Drummond, "A dynamic task scheduler tolerant to multiple hibernations in cloud environments," Cluster Computing, vol. 24, no. 2, pp. 1051-1073, 2021, doi: 10.1007/s10586-020-03175-2

[2] J.P.A. Neto, D.M. Pianto, C.G. Ralha, "A prediction approach to define checkpoint intervals in spot instances," In: 2018 11th International Conference on Cloud Computing (CLOUD SCF). Springer, vol. 10967, pp 84–93, 2018,doi: 10.1007/978-3-319-94295-7_6.

[3] J. Lancon, J. Kunwar, D. Stroud, M. McGee, R. Slater, "AWS EC2 instance spot price forecasting using LSTM networks," SMU Data Science Review, vol. 2, no. 2, 2019.

[4] V. K. Singh and K. Dutta, "Dynamic Price Prediction for Amazon Spot Instances," 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 2015, pp. 1513-1520, doi: 10.1109/HICSS.2015.184.

[5] P. Varshney and Y. Simmhan, "AutoBoT: Resilient and Cost-Effective Scheduling of a Bag of Tasks on Spot VMs," in IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 7, pp. 1512-1527, July 2019, doi: 10.1109/TPDS.2018.2889851.

[6] M. Khashei, M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," Applied Soft Computing, vol. 11, no. 2, pp. 2664-2675, 2011, doi: 10.1016/j.asoc.2010.10.015.

[7] Y. Liu, Z. Wang and B. Zheng, "Application of Regularized GRU-LSTM Model in Stock Price Prediction," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 2019, pp. 1886-1890, doi: 10.1109/ICCC47050.2019.9064035.

[8] B. Song, Y. Yu, Y. Zhou, Z. Wang and S. Du S, "Host load prediction with long short-term memory in cloud computing," The Journal of Supercomputing, vol. 74, no. 12, pp. 6554–6568, 2018, doi: 10.1007/s11227-017-2044-4.

[9] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural Comput, vol. 9, no. 8, pp.1735–1780, doi: 10.1162/neco.1997.9.8.1735.

[10] H. Abbasimehr, R. Paki, "Improving time series forecasting using LSTM and attention models," J Ambient Intell Human Comput, vol. 13, no. 1, pp. 673-691, 2022, doi: 10.1007/s12652-020-02761-x.

[11] Cho K, van Merrienboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y, "Learning phrase representations using RNN encoder decoder for statistical machine translation," 2014, doi: 10.48550/arXiv.1406.1078.

[12] Z. Chen, J. Hu, G. Min, A. Y. Zomaya and T. El-Ghazawi, "Towards Accurate Prediction for High-Dimensional and Highly-Variable Cloud Workloads with Deep Learning," in IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 4, pp. 923-934, 1 April 2020, doi: 10.1109/TPDS.2019.2953745.

[13] V. K. Singh and K. Dutta, "Dynamic Price Prediction for Amazon Spot Instances," 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 2015, pp. 1513-1520, doi: 10.1109/HICSS.2015.184.

[14] J.L. Lucas-Simarro, R. Moreno-Vozmediano, R.S. Montero, I.M. Llorente, "Cost optimization of virtual infrastructures in dynamic multi-cloud scenarios," Concurr Comput Pract Exp, vol. 27, no. 9. pp. 2260–2277, doi: 10.1002/cpe.2972.

[15] S. Alkharif, K. Lee and H. Kim, "Time-series analysis for price prediction of opportunistic Cloud computing resources," In 2018 7th International Conference on Emerging Databases. Springer, vol. 461, pp. 221–229, 2018, doi: 10.1007/978-981-10-6520-0_23.

[16] W. Liu, P. Wang, Y. Meng, C. Zhao and Z. Zhang, "Cloud spot instance price prediction using kNN regression," Hum Cent Comput Inf Sci, no. 10, no. 1, pp.10–34, 2020, doi: 10.1186/s13673-020-00239-5.

[17] H. Al-Theiabat, M. Al-Ayyoub, M. Alsmirat and M. Aldwair, "A Deep Learning Approach for Amazon EC2 Spot Price Prediction," 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), Aqaba, Jordan, 2018, pp. 1-5, doi: 10.1109/AICCSA.2018.8612783.

[18] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu and Y. Zhang, "Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network," in IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 841-851, Jan. 2019, doi: 10.1109/TSG.2017.2753802.

[19] Y. Guo and W. Yao, "Applying gated recurrent units pproaches for workload prediction," NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, Taipei, Taiwan, 2018, pp. 1-6, doi: 10.1109/NOMS.2018.8406290.

Declaration of Competing Interest: Authors do not have conflict of interest. The content of the paper is approved by the authors.

Publisher's Note: All content expressed in this article is solely that of the authors, and does not necessarily reflect the views of their affiliated organizations or the publisher, editors, and reviewers. Any content or product that may be reviewed and evaluated in this article is not guaranteed or endorsed by the publisher.

Author Contributions: All authors reviewed the manuscript.

Open Access: Journal of Southern Communication Engineering is an open access journal. All papers are immediately available to read and reuse upon publication.