International Journal of Infrastructure Research and Management
Vol. 12 (2), December 2024, pp. 107 - 118

OVERVIEW OF LOAD FORECASTS IN THE CLOUD ENVIRONMENT

Jin Yuanrong, Li Zhenxiang, Wang Haipei, Sun Xuekai & Tadiwa Elisha Nyamasvisva
Infrastructure University Kuala Lumpur, MALAYSIA

ABSTRACT

In a cloud computing environment, resource scheduling and load prediction are closely related concepts that are closely related to each other. The cloud environment is a virtualized, elastic computing platform that provides various computing resources to users through the network. The cloud environment can provide high scalability and flexibility to meet user demand for computing resources. Load prediction is based on historical data and models to predict and estimate the load in the cloud environment. Load usually refers to the user's demand for computing resources, such as CPU utilization, memory usage, network traffic, etc. The purpose of load prediction is to understand the future load situation in advance in order to plan and schedule resources in the cloud environment reasonably. Resource scheduling is the process of reasonably allocating and managing computing resources in a cloud environment. It is based on the results of load prediction, according to user needs and performance goals, dynamically allocating resources such as computing instances and storage to different tasks and users. The goal of resource scheduling is to optimize resource utilization, improve system performance, ensure service quality and meet user needs. Therefore, the cloud environment, resource scheduling and load prediction are interdependent. Load prediction provides information about future load conditions and provides a basis for resource scheduling decisions. Resource scheduling dynamically adjusts resource allocation based on load prediction results to meet the needs of different tasks and users. Through the coordinated work of load prediction and resource scheduling, resource utilization, performance and user experience in the cloud environment can be optimized. It should be noted that load prediction and resource scheduling are dynamic processes. Due to the uncertainty and variability of the load, prediction and scheduling need to be continuously monitored and adjusted to adapt to real-time demand and environmental changes. Therefore, load prediction and resource scheduling are important research and technical fields in the cloud computing environment, which are essential for improving the efficiency and performance of cloud services.

Keywords:
Cloud environment, decision basis, resource scheduling, resource allocation, load prediction

INTRODUCTION

Resource scheduling refers to the reasonable allocation and management of resources in the cloud environment based on the current load situation and load forecasting results, to meet the needs of applications and optimize system performance (Nayak & Shetty, 2023). The goal of resource scheduling is to achieve efficient use of resources under limited resources, ensuring system availability, performance, and user experience.

Load forecasting refers to the process of predicting and estimating the load situation for a future period of time in a computing system (Mamun et al., 2020). Load refers to the workload or request volume borne by the system, which can be computing tasks, network traffic, user requests, etc. The purpose of load forecasting is to effectively plan resources, optimize system performance, improve user experience, and meet the reliability and availability requirements of the system (Xu et al., 2022).

Load forecasting provides an important reference for resource scheduling (Gao et al., 2020; Mapetu et al., 2021). 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 forecasting can also help resource scheduling algorithms better understand load patterns and trends, thereby formulating more reasonable scheduling strategies (J. Kumar & Singh, 2020). It can be said that to a certain extent, load forecasting is the basis for resource scheduling.

PROBLEM STATEMENT

In the cloud environment, resource scheduling and load prediction still face many challenges. The cloud environment usually has the characteristics of large scale, heterogeneity and dynamic changes, so resource scheduling and load prediction face complex environmental and system requirements (Y. Chen et al., 2020; Fu & Zhou, 2020). It is necessary to consider multiple resource types, multiple task requirements, different performance indicators, etc., which increases the complexity of scheduling and prediction (Vashistha & Verma, 2020; Yadav et al., 2021). Load prediction involves estimating future loads, but due to the uncertainty and volatility of the load, accurately predicting future loads is still challenging (Hung et al., 2021; J. Kumar & Singh, 2021). The suddenness, instability and unpredictability of the load will affect the accuracy of load prediction. Load prediction and resource scheduling rely on historical data and real-time monitoring data. However, data quality issues such as incomplete data, data noise, data delay, etc., may affect the accuracy of prediction and the effectiveness of scheduling decisions. Load prediction and resource scheduling involve selecting appropriate algorithms and models and adjusting their parameters to adapt to specific environments and needs(Hou et al., 2021; Shen & Hong, 2020). However, algorithm selection and parameter adjustment may face difficulties. It is necessary to comprehensively consider the performance and adaptability of different algorithms and adjust them according to actual conditions (Karmakar et al., 2020; Lin et al., 2020). The cloud environment usually requires real-time response and adjustment of resource allocation, so load prediction and resource scheduling need to be real-time. However, real-time requirements may increase the complexity of prediction and scheduling, while requiring higher computing and processing capabilities (Nanjappan & Albert, 2022). Resource scheduling needs to balance system performance and cost-effectiveness. For cloud service providers, it is necessary to maximize resource utilization and user experience while controlling costs and the efficiency of resource allocation.

Therefore, this article reviews resource scheduling and load prediction in the cloud environment, finds research gaps and existing problems, and provides a theoretical basis for the feasibility of subsequent research.

LITERATURE REVIEW – Resource Scheduling

Resource scheduling is one of the research hotspots in the field of computer science and distributed systems. With the rapid development of technologies such as cloud computing, big data, and artificial intelligence, resource scheduling is facing increasingly complex and challenging problems (Wadhwa & Aron, 2022). These include the following aspects:

Load imbalance: The goal of resource scheduling is to achieve load balancing, that is, to evenly distribute the load to various resources (Subhash & Udayakumar, 2021; Wadhwa & Aron, 2022). However, due to the dynamic and uncertain nature of the load, resource scheduling may lead to load imbalance problems, where some resources are overloaded while others are underloaded. This can lead to performance degradation, resource waste, and system instability.

Inaccurate prediction: Resource scheduling usually relies on load prediction to predict future load trends and changes (Potu et al., 2021; Yuan et al., 2021). However, load prediction is a complex problem affected by multiple factors such as load fluctuations, seasonal changes, and abnormal events. Inaccurate predictions can lead to unreasonable resource allocation, inability to meet load demand or resource waste.

Data center scale: As the scale of data centers continues to expand, resource scheduling faces greater challenges. Large-scale data centers involve a large number of resources and tasks. Scheduling algorithms and mechanisms need to be able to handle large-scale data and complex scheduling decisions (Geetha & Robin, 2021; Wu et al., 2021). In addition, cross-data center resource scheduling and collaboration are also important issues.

Dynamism and real-time: Resource scheduling needs to monitor the status of loads and resources in real-time and make corresponding scheduling decisions (Geetha & Robin, 2021; Tianqing et al., 2022). However, the dynamic nature of loads and resources makes the scheduling process complex and challenging. Scheduling algorithms need to have real-time response capabilities to adapt to load changes and quickly adjust resource allocation.

Multi-dimensional constraints: Resource scheduling usually involves multiple constraints such as resource capacity, network bandwidth, service quality requirements, etc. Considering multiple constraints at the same time may make the scheduling problem complex and NP-hard (Huang et al., 2022; Singhal & Singhal, 2021). Therefore, designing efficient multi-dimensional scheduling algorithms and strategies is a challenging problem.

High energy consumption and green computing: With the increasing prominence of energy issues, resource scheduling needs to consider energy efficiency and green computing. Unreasonable allocation of resources by resource scheduling may lead to energy waste and high energy consumption (Fan et al., 2021; Strumberger et al., 2019). Therefore, it is necessary to design energy-saving scheduling algorithms and strategies to reduce system energy consumption. At present, many resource scheduling algorithms have been proposed and studied to meet the resource scheduling needs in different scenarios.

Load-balancing-based scheduling algorithms: Load balancing is one of the core goals of resource scheduling (Y. L. Chen et al., 2021; Zheng et al., 2011). Researchers have proposed various load balancing algorithms such as shortest job execution time-based, least connections-based, weighted round-robin-based etc. These algorithms achieve load balancing by reasonably allocating tasks or requests to different resources.

Prediction-based scheduling algorithms: Prediction-based scheduling algorithms use load prediction technology to predict future load conditions and make resource scheduling decisions based on prediction results (Jain et al., 2022). Common prediction algorithms include time series analysis (such as ARIMA models), machine learning (such as regression models, neural networks), and deep learning (such as recurrent neural networks). These algorithms achieve reasonable resource allocation and load balancing by predicting future loads.

Priority-based scheduling algorithms: Priority-based scheduling algorithms prioritize tasks based on their priority or importance and schedule them accordingly (Khodar et al., 2019). These algorithms prioritize high-priority tasks based on factors such as urgency, importance, deadline etc., allocating resources first in order to meet task requirements and system performance requirements.

Heuristic-based scheduling algorithms: Heuristic-based scheduling algorithms use heuristic strategies based on experience and rules for resource scheduling (Javadpour et al., 2022). These algorithms are usually based on some heuristic criteria such as Minimum Slack First (MSF) algorithm or Maximum Slack First (MaxSF) algorithm etc. Heuristic-based scheduling algorithms usually have low computational complexity and fast scheduling speed.

Genetic algorithm-based scheduling algorithms: Genetic algorithm is an optimization algorithm based on evolutionary ideas applied to resource scheduling problems (J. Kumar & Singh, 2020). These algorithms optimize resource scheduling solutions by simulating biological evolution processes. Genetic algorithm has global search capability which can find better solutions for schedule optimization.

Machine learning-based scheduling algorithms: In recent years machine learning technology has been widely used in resource scheduling (Fu & Zhou, 2020). Researchers use machine learning algorithms such as decision trees, support vector machines or random forests etc., learning from historical data about how best to schedule resources.

Table 1: SWOT Analysis of existing Resource Scheduling Algorithm

	Algorithm	**Strengths**	**Weaknesses**	**Opportunities**	**Threats**
1	Scheduling Algorithm Based on Load Balancing (Lavanya et al., 2020)	To achieve a balanced distribution of tasks or requests to various resources, improve resource utilization, reduce the load pressure on individual resources, improve system performance, and have a certain degree of elastic adaptability.	This method may introduce certain scheduling overhead, adaptive limitations, and may lack optimization for other constraints, such as energy efficiency and network bandwidth.	With the development of technologies such as machine learning and deep learning, there is an opportunity to apply these technologies to load balancing-based scheduling algorithms to improve load prediction accuracy and scheduling effectiveness. Optimization is performed on the basis of considering multiple constraints.	Some load balancing algorithms may face challenges in complexity and scalability when dealing with large-scale systems or complex scenarios.
2	Scheduling Algorithm Based on Prediction (Li et al., 2020)	It is possible to accurately predict future load trends based on historical data and predictive models, improving the effectiveness of resource scheduling and planning in advance.	Due to the uncertainty and complexity of load changes, there may be prediction errors, resulting in inaccurate resource scheduling. It is not possible to respond to sudden load changes in real time.	The prediction model can be improved to increase the accuracy and timeliness of predictions. Combined with other constraints, multi-dimensional resource scheduling optimization can be performed.	The prediction model needs to be updated in a timely manner to adapt to changing load patterns and environments, but the cost and frequency of model updates may have a negative impact on system performance.
3	Priority-Based Scheduling Algorithm (Wang et al., 2020)	Resources can be allocated reasonably according to the priority or importance of tasks to ensure that high-priority tasks are met in a timely manner, providing good service quality and meeting the deadline and performance requirements of tasks.	The priorities of different tasks may conflict, requiring a balance of priorities between different tasks, which may result in a decrease in the performance of some tasks, may not be able to flexibly adapt to dynamic loads and changing environments, and limit the	By combining machine learning and adaptive algorithms, the priority of tasks can be intelligently adjusted to adapt to dynamic loads and environmental changes.	This may lead to the problem of unfair resource allocation, where low-priority tasks may be ignored or not met for a long time.

			applicability of the algorithm.		
4	Heuristic Scheduling Algorithm (Zhang et al., 2021)	With lower computational complexity and faster scheduling speed, it can quickly respond to load changes. Based on experience and rules, it usually has good interpretability, easy to understand and adjust.	Usually based on fixed rules and strategies, lacking self-learning and adaptive capabilities, it is difficult to adapt to complex load patterns and environmental changes. It may fall into a local optimum and cannot globally optimize resource scheduling problems.	Combining with reinforcement learning algorithms, the heuristic algorithm has the ability to learn and optimize, improving the quality and adaptability of scheduling decisions. Combining heuristic algorithms with other scheduling algorithms, fully utilizing their respective advantages, and improving scheduling effects and performance.	Performance and effectiveness highly depend on the selection of appropriate heuristic rules and parameters, and improper selection may lead to a decline in performance.
5	Scheduling Algorithm Based on Genetic Algorithm (Saif et al., 2021)	It has global search capabilities and can find better scheduling solutions. Through crossover and mutation operations, it can maintain the diversity of the population and avoid falling into local optima.	It usually takes a long time to evolve and optimize, and is not suitable for real-time scheduling scenarios. The performance of the genetic algorithm highly depends on the selection of appropriate parameter settings, and improper parameter selection may lead to a decline in performance.	By improving the crossover, mutation, and selection operators of the genetic algorithm, as well as optimizing the scheduling strategy, the performance of the genetic algorithm in resource scheduling can be improved. Combining the genetic algorithm with parallel computing and distributed architecture can improve the efficiency and scalability of the algorithm.	The performance and results of the genetic algorithm largely depend on the selection of appropriate genetic algorithm parameters, and improper selection may lead to a decline in performance.
6	Scheduling Algorithm Based on Machine Learning	Able to learn from historical data, extract patterns and rules, and optimize scheduling decisions. Machine	The demand for a large amount of high-quality training data is high. If the data quality is poor or	By using transfer learning techniques, existing models and knowledge can be utilized in	Machine learning algorithms have a high dependence on high-quality training data. If the data quality is

	(Rani & Geethakumari , 2021)	learning algorithms have a certain adaptability and can update and adjust models based on new data and environmental changes.	insufficient, it may affect the performance of the algorithm. Some machine learning algorithms may lack interpretability, making it difficult to explain and understand the decision-making process and results of the algorithm.	different environments and domains to accelerate the training and optimization process of scheduling algorithms.	poor or insufficient, it may affect the performance and prediction accuracy of the algorithm.

The review presented in Table 1 has identified a number of weaknesses and issues with current and existing algorithms. Major issues are listed as:

i. Lack of optimization constraints in energy efficiency and network bandwidth.
ii. Uncertainty and complexity of load changes,
iii. Prediction errors,
iv. Inaccurate resource scheduling.
v. Inability to respond to sudden load changes in real time.
vi. Conflict of priorities and balance between different tasks may conflict,
vii. Decreased adaption flexibility to dynamic loads and changing environments,
viii. Based on fixed rules and strategies,
ix. Lack of self-learning and adaptive capabilities,
x. Risk of falling into a local optimum and might not globally optimize resource scheduling problems.
xi. Long time to evolve and optimize,
xii. Not suitable for real-time scheduling scenarios.
xiii. High dependence on the selection of appropriate parameter settings,
xiv. Lack of interpretability, making it difficult to explain and understand the decision-making process and results.

LITERATURE REVIEW – Load Prediction

Load forecasting refers to the process of predicting the future load situation of a system based on historical data and models. In cloud computing environments, load forecasting is an important component of resource scheduling. The accuracy of load forecasting directly affects the effectiveness of resource scheduling and system performance (Y. Chen et al., 2020) . The research on load forecasting mainly focuses on the following aspects:

i. Prediction methods: There are various methods for load prediction, including time series analysis, machine learning methods, statistical methods, etc. (Yadav et al., 2021; Zhang et al., 2021). Time series analysis methods are commonly used for load forecasting, which can capture the trend and seasonality of load changes over time. Machine learning methods such as neural networks, support vector machines, etc. are also widely used in load prediction. These methods can handle complex load patterns and have high prediction accuracy.

ii. Data source: Load forecasting relies on high-quality data. The data sources usually include historical load data, real-time monitoring data, etc. (M. Kumar & Sharma, 2020; Niri et al., 2020). However, data quality issues such as incomplete data, noise, and latency may affect the accuracy of load forecasting. Therefore, when conducting load prediction, data pre-processing is necessary, such as data cleaning, feature selection, etc.
iii. Prediction accuracy: The accuracy of load prediction is an important indicator to measure its effectiveness. High precision load forecasting can provide accurate information for resource scheduling, thereby optimizing resource allocation and improving system performance (Rani & Geethakumari, 2021). However, due to the complexity and unpredictability of the load, the accuracy of load prediction still faces challenges. In order to improve prediction accuracy, researchers have proposed various improvement methods, such as combination prediction, ensemble learning, etc.
iv. Time range of prediction: Load prediction can be short-term prediction, medium-term prediction, or long-term prediction. The application of predictions in resource scheduling varies across different time ranges (Cao et al., 2022; Zhang et al., 2021). Short term forecasting is typically used for real-time resource scheduling, while long-term forecasting is used for resource planning and management. Choosing an appropriate prediction time range can improve the efficiency and accuracy of resource scheduling.
v. Load pattern recognition: An important task in load prediction is to identify load patterns. Load mode refers to the pattern and characteristics of load changes over time. By identifying load patterns, the dynamic characteristics of loads can be better understood, thereby improving prediction accuracy (Chicco, 2021; Gawlikowski Student Member et al., 2021) . The methods of load pattern recognition include clustering analysis, pattern matching, etc.

METHODOLOGY

This article adopts the literature analysis method to summarize the research status, existing problems, and future research directions of resource scheduling and load forecasting in cloud environments through the study and analysis of relevant literature. Firstly, by searching relevant academic papers, conference papers, technical reports, and other literature, a large amount of information related to resource scheduling and load forecasting was obtained. Then, the content analysis method is used to classify and organize this literature, identifying key issues and challenges in resource scheduling and load forecasting. Finally, based on the characteristics of cloud computing environments, future research directions and possible solutions are proposed.

RESULTS AND DISCUSSION

Through the analysis of literature, this article summarizes several key issues and challenges in resource scheduling and load forecasting in cloud environments. Firstly, the accuracy of load forecasting directly affects the effectiveness of resource scheduling. In order to improve the accuracy of load forecasting, it is necessary to adopt more advanced forecasting models and combine multiple forecasting methods. Secondly, resource scheduling needs to consider multidimensional constraints such as resource capacity, network bandwidth, and service quality requirements. The complexity of these constraints increases the difficulty of resource scheduling. In addition, with the continuous expansion of data center scale, resource scheduling is facing greater challenges. Therefore, future research should focus on developing resource scheduling algorithms and mechanisms that can handle large-scale data and complex constraints.

CONCLUSION

This article provides an overview of resource scheduling and load forecasting in cloud environments, identifies existing research problems and challenges, and proposes future research directions. Firstly, the accuracy of load forecasting is a key factor affecting the effectiveness of resource scheduling. Therefore, future research should focus on developing more accurate predictive models and methods. Secondly, resource scheduling needs to consider multidimensional constraints such as resource capacity, network bandwidth, and service quality requirements. Therefore, researchers need to develop scheduling algorithms that can handle these complex constraints. Finally, with the continuous expansion of data center scale, the challenges faced by resource scheduling are also increasing. Future research should focus on developing resource scheduling mechanisms that can handle large-scale data and complex environments. Overall, by discussing resource scheduling and load forecasting and their related algorithms, the advantages and limitations of different algorithms in practical applications are summarized. The research and development of resource scheduling and load forecasting is crucial for improving the efficiency and reliability of cloud computing systems and is a field that needs to be continuously developed.

AUTHOR 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*

Sun Xuekai 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 the application of blockchain technology in smart education. *Email: 223923752@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/

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. L., Huang, S. Y., Chang, Y. C., & Chao, H. C. (2021). Resource Allocation based on Genetic Algorithm for Cloud Computing. *2021 30th Wireless and Optical Communications Conference, WOCC 2021*, 211–212. https://doi.org/10.1109/WOCC53213.2021.9603125

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

Fan, Y., Wang, L., Wu, W., & Du, D. (2021). Cloud/Edge Computing Resource Allocation and Pricing for Mobile Blockchain: An Iterative Greedy and Search Approach. *IEEE Transactions on Computational Social Systems*, *8*(2), 451–463. https://doi.org/10.1109/TCSS.2021.3049152

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/

Fu, X., & Zhou, C. (2020). Predicted Affinity Based Virtual Machine Placement in Cloud Computing Environments. *IEEE Transactions on Cloud Computing*, *8*(1), 246–255. https://doi.org/10.1109/TCC.2017.2737624

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

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

Geetha, P., & Robin, C. R. R. (2021). Power conserving resource allocation scheme with improved QoS to promote green cloud computing. *Journal of Ambient Intelligence and Humanized Computing*, *12*(7), 7153–7164. https://doi.org/10.1007/S12652-020-02384-2/METRICS

Hou, K., Guo, M., Li, X., & Zhang, H. (2021). Research on Optimization of GWO-BP Model for Cloud Server Load Prediction. *IEEE Access*, *9*, 162581–162589. https://doi.org/10.1109/ACCESS.2021.3132052

Huang, J., Lv, B., Wu, Y., Chen, Y., & Shen, X. (2022). Dynamic Admission Control and Resource Allocation for Mobile Edge Computing Enabled Small Cell Network. *IEEE Transactions on Vehicular Technology*, *71*(2), 1964–1973. https://doi.org/10.1109/TVT.2021.3133696

Hung, L. H., Wu, C. H., Tsai, C. H., & Huang, H. C. (2021). Migration-based load balance of virtual machine servers in cloud computing by load prediction using genetic-based methods. *IEEE Access*, *9*, 49760–49773. https://doi.org/10.1109/ACCESS.2021.3065170

Jain, D. K., Tyagi, S. K. S., Neelakandan, S., Prakash, M., & Natrayan, L. (2022). Metaheuristic Optimization-Based Resource Allocation Technique for Cybertwin-Driven 6G on IoE Environment. *IEEE Transactions on Industrial Informatics*, *18*(7), 4884–4892. https://doi.org/10.1109/TII.2021.3138915

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

Karmakar, K., Das, R. K., & Khatua, S. (2020). Resource Scheduling for Tasks of a Workflow in Cloud Environment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11969 LNCS*, 214–226. https://doi.org/10.1007/978-3-030-36987-3_13

Khodar, A., Al-Afare, H. A. F., & Alkhayat, I. (2019). New Scheduling Approach for Virtual Machine Resources in Cloud Computing based on Genetic Algorithm. *Proceedings - 2019 International Russian Automation Conference, RusAutoCon 2019*. https://doi.org/10.1109/RUSAUTOCON.2019.8867638

Kumar, J., & Singh, A. K. (2020). Cloud datacenter workload estimation using error preventive time series forecasting models. *Cluster Computing*, *23*(2), 1363–1379. https://doi.org/10.1007/S10586-019-03003-2/METRICS

Kumar, J., & Singh, A. K. (2021). Performance evaluation of metaheuristics algorithms for workload prediction in cloud environment. *Applied Soft Computing*, *113*, 107895. https://doi.org/10.1016/J.ASOC.2021.107895

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

Lavanya, M., Shanthi, B., & Saravanan, S. (2020). Multi objective task scheduling algorithm based on SLA and processing time suitable for cloud environment. *Computer Communications*, *151*, 183–195. https://doi.org/10.1016/J.COMCOM.2019.12.050

Li, C., Tang, J., Ma, T., Yang, X., & Luo, Y. (2020). Load balance-based workflow job scheduling algorithm in distributed cloud. *Journal of Network and Computer Applications*, *152*, 102518. https://doi.org/10.1016/J.JNCA.2019.102518

Lin, Y., Wang, X., & Xu, R. (2020). Semi-supervised human resource scheduling based on deep presentation in the cloud. *Eurasip Journal on Wireless Communications and Networking*, *2020*(1). https://doi.org/10.1186/S13638-020-01677-6

Mamun, A. Al, Sohel, M., Mohammad, N., Haque Sunny, M. S., Dipta, D. R., & Hossain, E. (2020). A Comprehensive Review of the Load Forecasting Techniques Using Single and Hybrid Predictive Models. *IEEE Access*, *8*, 134911–134939. https://doi.org/10.1109/ACCESS.2020.3010702

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

Nanjappan, M., & Albert, P. (2022). Hybrid-based novel approach for resource scheduling using MCFCM and PSO in cloud computing environment. *Concurrency and Computation: Practice and Experience*, *34*(7). https://doi.org/10.1002/CPE.5517

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

Nayak, A. A., & Shetty, S. (2023). A Systematic Analysis on Task Scheduling Algorithms for Resource Allocation of Virtual Machines on Cloud Computing Environments. *ICRTEC 2023* -

Proceedings: IEEE International Conference on Recent Trends in Electronics and Communication: Upcoming Technologies for Smart Systems. https://doi.org/10.1109/ICRTEC56977.2023.10111894

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

Potu, N., Jatoth, C., & Parvataneni, P. (2021). Optimizing resource scheduling based on extended particle swarm optimization in fog computing environments. *Concurrency and Computation: Practice and Experience*, *33*(23). https://doi.org/10.1002/cpe.6163

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

Shen, H., & Hong, X. (2020). *Host Load Prediction with Bi-directional Long Short-Term Memory in Cloud Computing*. https://arxiv.org/abs/2007.15582v1

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

Singhal, R., & Singhal, A. (2021). A feedback-based combinatorial fair economical double auction resource allocation model for cloud computing. *Future Generation Computer Systems*, *115*, 780–797. https://doi.org/10.1016/J.FUTURE.2020.09.022

Strumberger, I., Bacanin, N., Tuba, M., & Tuba, E. (2019). Resource Scheduling in Cloud Computing Based on a Hybridized Whale Optimization Algorithm. *Applied Sciences 2019, Vol. 9, Page 4893*, *9*(22), 4893. https://doi.org/10.3390/APP9224893

Subhash, L. S., & Udayakumar, R. (2021). Sunflower Whale Optimization Algorithm for Resource Allocation Strategy in Cloud Computing Platform. *Wireless Personal Communications*, *116*(4), 3061–3080. https://doi.org/10.1007/S11277-020-07835-9

Tianqing, Z., Zhou, W., Ye, D., Cheng, Z., & Li, J. (2022). Resource Allocation in IoT Edge Computing via Concurrent Federated Reinforcement Learning. *IEEE Internet of Things Journal*, *9*(2), 1414–1426. https://doi.org/10.1109/JIOT.2021.3086910

Vashistha, A., & Verma, P. (2020). A literature review and taxonomy on workload prediction in cloud data center. *Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering*, 415–420. https://doi.org/10.1109/CONFLUENCE47617.2020.9057938

Wadhwa, H., & Aron, R. (2022). TRAM: Technique for resource allocation and management in fog computing environment. *Journal of Supercomputing*, *78*(1), 667–690. https://doi.org/10.1007/S11227-021-03885-3

Wang, S., Zhao, T., & Pang, S. (2020). Task Scheduling Algorithm Based on Improved Firework Algorithm in Fog Computing. *IEEE Access*, *8*, 32385–32394. https://doi.org/10.1109/ACCESS.2020.2973758

Wu, C. ge, Li, W., Wang, L., & Zomaya, A. Y. (2021). An evolutionary fuzzy scheduler for multi-objective resource allocation in fog computing. *Future Generation Computer Systems*, *117*, 498–509. https://doi.org/10.1016/J.FUTURE.2020.12.019

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

Yadav, M. P., Pal, N., & Yadav, D. K. (2021). Workload prediction over cloud server using time series data. *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, 267–272. https://doi.org/10.1109/CONFLUENCE51648.2021.9377032

Yuan, M., Cai, X., Zhou, Z., Sun, C., Gu, W., & Huang, J. (2021). Dynamic service resources scheduling method in cloud manufacturing environment. *International Journal of Production Research*, *59*(2), 542–559. https://doi.org/10.1080/00207543.2019.1697000

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

Zheng, Z., Wang, R., Zhong, H., & Zhang, X. (2011). An approach for cloud resource scheduling based on parallel genetic algorithm. *ICCRD2011 - 2011 3rd International Conference on Computer Research and Development*, *2*, 444–447. https://doi.org/10.1109/ICCRD.2011.5764170