

Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management

Dr. Ranjan Kumar Mondal

Assistant Professor, Department of CSE, Swami
Vivekananda University, Kolkata, WB

ranjankm@svu.ac.in

Abstract

Cloud computing has become more prevalent in recent years. Cloud computing is the use of centralized computer resources. Users may utilize these resources to finish calculations, and the cloud computing center will provide the user with the program's processing results. Cloud computing is helpful for both consumer and business users. Users may save computing expenditures by avoiding the need to acquire several PCs by investing in a cloud server. China Economic News Network reports that the country's cloud computing market is worth 209.1 billion yuan. China's more established cloud service providers include Huawei Cloud, Baidu Cloud, Ali Cloud, etc. In light of this, this study suggests a novel method for handling challenging issues related to scheduling and managing cloud computing resources using machine learning optimization approaches. This study provides a comprehensive solution, including optimization techniques like deep learning and genetic algorithms, to improve system performance and efficiency and thus bring breakthroughs and progress in cloud computing resource management. Rational allocation of resources plays a crucial role in cloud computing. The study thoroughly examines issues like low resource utilization and unbalanced load in the cloud environment. Users come in order at the cloud computing center, which has restricted cloud resources for resource allocation. Each user requests a specified quantity of cloud resources to be used by the cloud computing center at a given time.

Keywords: artificial intelligence, cloud computing, resource allocation, and machine learning optimization

Introduction

Cloud computing has been employed extensively in recent years. Cloud computing is the use of centralized computer resources. Users may utilize these resources to finish calculations, and the cloud computing center will provide the user with the program's processing results. Cloud computing is helpful for both consumer and business users. Users may save computing expenditures by avoiding the need to acquire several PCs by investing in a cloud server. Cloud service providers may dynamically schedule computing resources based on customers' access needs to optimize computer resource consumption efficiency. China Economic News Network reports that the country's cloud computing market is worth 209.1 billion yuan. China's more established cloud service providers include Huawei Cloud, Baidu Cloud, Ali Cloud, etc.

In cloud computing, rational resource distribution is essential. Users come in order at the cloud computing center, which has restricted cloud resources for resource allocation. Each user requests a specified quantity of cloud resources to be used by the cloud computing center at a given time. The cloud computing service center must decide whether to approve or deny the user's request. Under resource constraints, judicious use of cloud resources may optimize service user quality and boost user happiness. Heuristic algorithms are used by the majority of current resource allocation techniques to allocate resources, but heuristic algorithms struggle to provide accurate outcomes in complex circumstances.

Users' different demands must be considered while allocating cloud computing resources. The resources in the cloud are required differently by various users. Resources must be distributed by the cloud computing center to user demands. Two limitations may be applied to users' demand for cloud resources: hard limits and soft constraints. Problematic limitations are those that need to be adhered to. For instance, a user can ask for many VMS rather than just one VM.

Each virtual machine (VM) must be situated in a distinct fault domain (a server sharing a single point of failure) to guarantee high availability and prevent financial losses from the cloud computing center failing. Soft constraints are either satiable or unsatisfactory, yet satiating a restriction may significantly increase the quality of the service. For instance, placing cloud virtual machines following one another in the network topology helps lower network latency and speed up processing during the neural network training process.

Figure 1. Policy networks using reinforcement learning techniques (Figure: Mao et al. (2016))



Figure 1. Policy networks using reinforcement learning techniques (Figure: Mao et al. (2016))

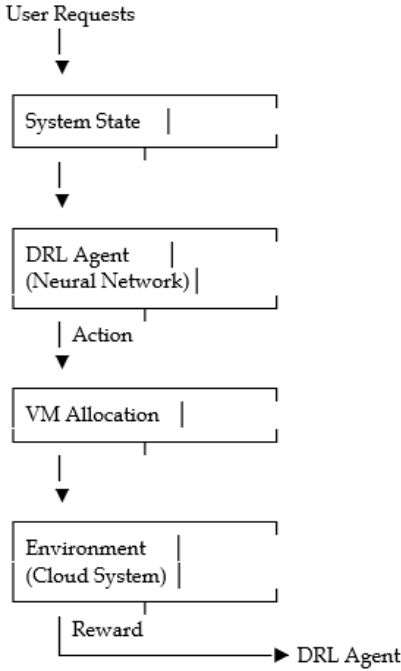
As seen in Figure 1, the action is the neural network's output, while the system's current state serves as its input. The environment will give the appropriate reward once the system accepts the action. System optimization's objective is to.

To increase the user's waiting time, decrease it.

As user requests are processed, VMS that are Being physically near to users may enhance user experience and speed up communication between the cloud computing center and users.

This paper primarily suggests a deep reinforcement learning-based cloud resource allocation strategy that considers user requirements. The primary goal of this research is to enhance customer service quality by allocating users to nearby servers. On the other hand, if users are paired with nearby servers, there's a chance that some of them may be overloaded, leading to lengthy wait times. As a result, this paper examines a deep

reinforcement learning-based approach to cloud resource allocation. With deep reinforcement learning, resources may be dynamically allocated based on the system's present state, optimizing cloud resource consumption and reducing user wait times.



Nearby VM → Lower Latency

Overloaded VM → Higher Waiting Time

Related Works

Deep reinforcement learning for resource scheduling

Mao et al. (2016) investigated resource allocation in cloud computing systems using deep learning techniques. The strategy gradient approach is primarily used in their study to achieve dynamic resource allocation. The strategy gradient method is a reinforcement learning technique that maximizes the intended cumulative reward by optimizing the strategy

for effective resource allocation. This method allows for effective resource use in cloud computing settings that are constantly changing, in addition to adapting to various conditions and needs. Mao et al.'s study offers novel concepts and approaches for resolving resource allocation issues and significant references and inspiration for using deep learning in cloud computing resource management.

Task allocation and unloading under priority restrictions, Optimization Issue To improve computational efficiency, complicated jobs are divided into smaller tasks called task flows, which are then assigned to processors for parallel processing in the cloud computing resource allocation and task scheduling system. Directed acyclic graphs (DAGs) are helpful for modeling and abstracting workflows when jobs have priority restrictions, and some calculations rely on the outcomes of earlier tasks. Nodes in this model represent subtasks, while priority restrictions among subtasks are shown by edges connecting nodes. A tree model of the DAG scheduling issue is given in Literature [4], which considers variables like interconnectivity between processors, processor resource constraints, and communication time between processes. Algebraic Model and Associated Techniques

The restrictions on the order of priority between tasks may be written as follows: for any $i \rightarrow j$, $EST(j) < CT(i)$, where $CT(i)$ is the task's completion time, $i \rightarrow j$ means that task i must be finished before task j , and $EST(j)$ is the earliest time job j can start processing. Regarding the issue of cloud offloading on

CPUs in mobile terminals are capable of handling many jobs at once. A deterministic delay-constrained task segmentation technique based on dynamic programming is introduced in the literature, and its suboptimality is shown. Assuming that processors are limited to one job at a time, the topology of DAG remains unrestricted in light of the deadline limitations. A mixed-integer programming paradigm that considers cooperation between edge computing nodes and distant cloud nodes is presented in the literature [5]. It is a generic allocation issue with NP-hard features for which a polynomial-time optimization technique is unavailable. Books

[6] turns the issue into a convex optimization problem using a relaxed integer programming model with 0-1 variables. A heuristic method is then designed. The literature [7] proposes a mixed-integer programming approach to address the multi-DAG mobile terminal offloading issue. It determines whether to upload work to the cloud and minimize energy usage under deadline limitations.

Heuristic Approaches

Heuristic methods, intelligent algorithms, and hybrid algorithms are the critical optimization techniques for DAG scheduling. List scheduling, clustering, and job replication are the main categories of heuristic methods. List scheduling assigns the highest-priority job from the pending tasks to the relevant processor by arranging the tasks according to priority. Tasks are grouped in clusters until the number of categories and processors is equal. To reduce processing delay, jobs containing important data are replicated across many processors using task replication.

Literature [6] uses forward ranking as the criteria for list scheduling in “cloud-edge-end” systems to minimize computation, communication costs, and latency of cloud and edge computing nodes. In the literature [7], techniques like lower bound estimates are used to calculate nodes for DAG subtasks and assign deadlines while considering cost and deadline. To maximize total time performance, literature [8] assigns virtual machines according to task deadline order, considering the earliest job completion time.

Sensible Algorithms

Intelligent algorithms strive for global optimization performance, as opposed to heuristic guidelines. Genetic algorithms are used in the literature [8] to maximize task-edge node group assignments. Literature [9] uses probability to describe the spatial connection between activities. Following heuristic-based DAG pre-segmentation, the literature

[9] ranks jobs and optimizes edge node energy consumption and total application completion time using bivariate correlation distribution estimation techniques. The Estimation

of Distribution Algorithm (EDA) is used in the literature [10] to minimize overall delay while considering task deadline information. Regarding the assignment issue of task-node, the literature

[10] employs the particle swarm optimization algorithm to maximize the weighted cost and completion time goals. The techniques above are logically linked to the real needs of task scheduling and cloud computing resource allocation.

ACHIEVE DYNAMIC RESOURCE ALLOCATION

One essential component of resource management in cloud computing settings is the dynamic distribution of resources. To do this, several crucial elements and procedures must be taken into account:

Algorithm for Resource Scheduling

Dynamic resource allocation is based on the resource scheduling algorithm. It assesses several variables, including resource availability, kind, and demand, to determine the best scheduling plan. Static allocation, dynamic allocation, and load balancing are standard algorithms. Every algorithm considers specific needs and limitations to guarantee practical resource usage.

Resource Observation

Real-time detection of changes and abnormalities requires resource monitoring. Resources of all kinds need to be regularly observed to track utilization trends and spot problems. By monitoring their resources, organizations may proactively handle variations in demand and adjust their resource allocation methods.

Forecasting Resources

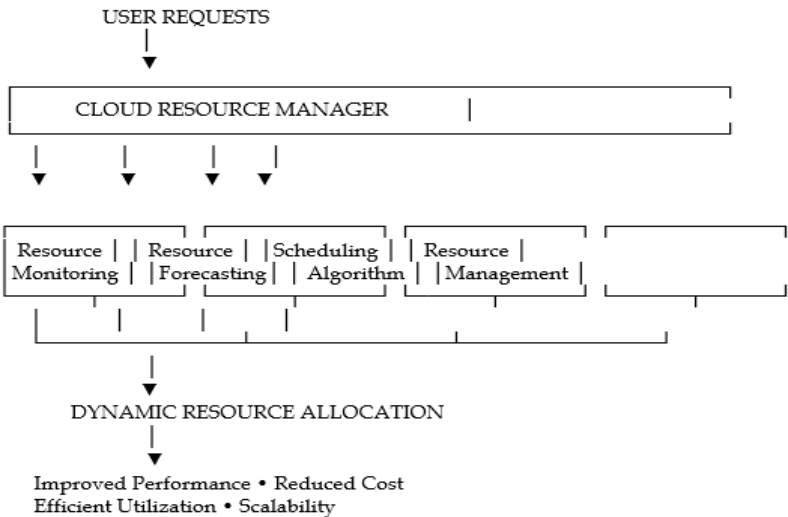
Effective resource allocation depends on predicting future patterns in resource demand and utilization. Future resource needs are expected via prediction tools and historical data analysis. Organizations may avert shortages and maximize efficiency by anticipating resource demands and allocating

resources in advance. This minimizes waste and maximizes usage.

Management of Resources

Enabling dynamic resource allocation requires efficient resource management. Appropriate labeling, classification, and organization of resources are necessary for adequate scheduling and placement. Robust resource management techniques can help organizations maximize performance, optimize resource use, and simplify operations.

Organizations may fully benefit from dynamic resource allocation in cloud computing environments by combining these elements into a unified architecture. This strategy ensures that resources are distributed effectively and flexibly, and by changing needs and circumstances, eventually boosting output and propelling company success.



RESULTS

This research concludes with a thorough strategy for resolving the difficulties associated with resource scheduling and management in cloud computing settings. The suggested approach notably improves system effectiveness and

performance by utilizing machine learning optimization methods, including deep reinforcement learning. It is clear from thorough testing and analysis that the proposed algorithm performs better in terms of time, cost efficiency, service quality, and system resource load than conventional techniques like ant colony optimization and simulated annealing. This highlights its potential to transform resource management in cloud computing and lead to new developments in the industry.

Deep Learning and Cloud Computing Together for Scheduling

Combining deep learning methods with cloud computing scheduling might lead to more intelligent and adaptable resource allocation strategies. This research shows that deep reinforcement learning allows for dynamic resource allocation depending on the system's current state, which maximizes resource usage efficiency and minimizes user waiting times. Deep learning-based methods may provide customized and optimal solutions to complicated scheduling issues in cloud settings by considering system limitations and user demands in an all-encompassing way.

Benefits of Deep Learning for Scheduling in Cloud Computing

Cloud computing scheduling benefits incredibly from deep learning. First, it can adjust to changing cloud settings because of its capacity to extract intricate patterns and correlations from data. Second, deep learning models enable more precise and sophisticated decision-making by handling high-dimensional and non-linear data. Furthermore, deep learning algorithms can learn and grow over time, which may increase scalability and performance for cloud scheduling jobs. Deep learning and cloud computing scheduling integration have much potential to manage contemporary cloud settings' growing complexity and dynamic nature.

Prospects for AI and Cloud Computing Integration in the Future

Significant breakthroughs in several sectors are anticipated due to the confluence of cloud computing with artificial

intelligence (AI). More intelligent and independent features, such as anomaly detection, self-optimizing systems, and predictive resource provisioning, will be available with AI-powered cloud services. Furthermore, companies can make better choices and get a competitive edge thanks to AI-driven insights from massive cloud data volumes. Combining AI technology with cloud computing will bring in a new age of innovation and industry change as these technologies develop.

Clearing the path for digital ecosystems that are more intelligent, effective, and resilient.

References

X. Mo and J. Xu, "Energy-efficient federated edge learning with joint communication and computation design," *Journal of Communications and Information Networks*, vol. 6, no. 2, pp. 110-124, 2021.

Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with cpu-gpu heterogeneous computing," *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 7947-7962, 2021.

Yu, L., Liu, B., Lin, Q., Zhao, X., & Che, C. (2024). Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method. *arXiv preprint arXiv:2401.06782*.

Hussain H, Malik S U R, Hameed A, et al. A Survey on Resource Allocation in High-Performance Distributed Computing Systems[J]. *Parallel Computing* (S0167-8191), 2013, 39(11): 709-736.

[9]. Bellendorf J, Mann Z Á. Classification of Optimization Problems in Fog Computing[J]. *Future Generation Computer Systems* (S0167- 739X), 2020, 107(1): 158-176.

Liu, B., Zhao, X., Hu, H., Lin, Q., & Huang, J. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. *Journal of Theory and Practice of Engineering Science*, 3(12), 36-42. [https://doi.org/10.53469/jtpes.2023.03\(12\).06](https://doi.org/10.53469/jtpes.2023.03(12).06)

Brogi A, Forti S, Guerrero C, et al. How to Place Your Apps in the Fog: State of the Art and Open Challenges[J]. *Software: Practice and Experience* (S0167-739X), 2019, 1(1): 1-8.

Wu C, Li W, Wang L, et al. Hybrid Evolutionary Scheduling for Energy-efficient Fog-enhanced Internet of Things[J]. *IEEE Transactions on Cloud Computing* (S2168-7161), 2018, 1(1): 1-1.

Atzori L, Iera A, Morabito G. The Internet of Things: A Survey[J]. *Computer Networks* (S1389-1286), 2010, 54(15): 2787-2805.

Bonomi F, Milito R, Natarajan P, et al. Fog Computing: A Platform for Internet of Things and Analytics. N. Bessis, C. Dobre. *Big Data and Internet of Things: A roadmap for intelligent environments*[M]. Cham: Springer, 2014, 546: 169-186.

Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." *arXiv preprint arXiv:2312.12872* (2023).

Xiaoxi Zhang, Jianyu Wang, Li-Feng Lee, Tom Yang, Akansha Kalra, Gauri Joshi, Carlee Joe-Wong, "Machine Learning on Volatile Instances: Convergence, Runtime, and Cost Trade-offs," *IEEE/ACM Transactions on Networking*, 30(1):215–228, 2022

Yichen Ruan, Xiaoxi Zhang, Shu-Che Liang, Carlee Joe-Wong, "Towards Flexible Device Participation in Federated Learning for Non-IID Data," *International Conference on Artificial Intelligence and Statistics(AISTATS)*, 2021

Yichen Ruan, Xiaoxi Zhang, Carlee Joe-Wong, "How Valuable Is Your Data? Optimizing Device Recruitment in Federated

Learning", submitted to ToN, preliminary results are published in *WiOpt* 2021

Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." *arXiv preprint arXiv:2312.12872* (2023).