Optimizing Resource Allocation and Load Balancing in Heterogeneous Cloud Computing Environments: A Machine Learning Approach

- Jian Chen, Department of Computer Science and Engineering, Nanyang Technological University, Singapore

The rapid growth of cloud computing has led to the deployment of diverse and heterogeneous computing resources, posing significant challenges in terms of resource allocation and load balancing. Inefficient resource management can result in underutilized infrastructure, degraded application performance, and increased operational costs. This research paper proposes a novel machine learning-based approach to optimize resource allocation and load balancing in heterogeneous cloud computing environments. By leveraging advanced machine learning techniques, such as deep reinforcement learning and graph neural networks, the proposed framework learns to make intelligent decisions based on real-time system state and historical data. The research methodology involves the development of a scalable and adaptive resource allocation algorithm that considers multiple objectives, including maximizing resource utilization, minimizing response time, and ensuring fair distribution of workload across heterogeneous computing nodes. The proposed approach is evaluated through extensive simulations and real-world case studies, demonstrating its effectiveness in improving system performance, reducing resource wastage, and enhancing the overall efficiency of cloud computing environments. The study also presents a comprehensive analysis of the trade-offs between different optimization objectives and provides insights into the scalability and robustness of the proposed framework under dynamic workload conditions. The findings of this research have significant implications for cloud service providers and system administrators, enabling them to make informed decisions regarding resource provisioning, scheduling, and load balancing strategies. By leveraging machine learning techniques, the proposed approach offers a flexible and adaptable solution to the complex challenges of resource management in heterogeneous cloud computing environments. This research contributes to the advancement of intelligent and autonomous cloud computing systems, paving the way for more efficient, cost-effective, and user-centric cloud services.

References

[1] K. Alwasel, Y. Li, P. P. Jayaraman, S. Garg, R. N. Calheiros, and R. Ranjan, “Programming SDN-native big data applications: Research gap analysis,” *IEEE Cloud Comput.*, vol. 4, no. 5, pp. 62–71, Sep. 2017.

[2] M. Abouelyazid, “Forecasting Resource Usage in Cloud Environments Using Temporal Convolutional Networks,” *Applied Research in Artificial Intelligence and Cloud Computing*, vol. 5, no. 1, pp. 179–194, Nov. 2022.

[3] M. Yousif, “Cloud-native applications—the journey continues,” *IEEE Cloud Comput.*, vol. 4, no. 5, pp. 4–5, Sep. 2017.

[4] M. Abouelyazid and C. Xiang, “Architectures for AI Integration in Next-Generation Cloud Infrastructure, Development, Security, and Management,” *International Journal of Information and Cybersecurity*, vol. 3, no. 1, pp. 1–19, Jan. 2019.