Multi-Objective Optimization Framework for Cloud Applications Using AI-Based Surrogate Models

Vijay Ramamoorthi
Independent Researcher

Keywords:

- AI-Based Surrogate Models
- Quality of Service
- Software-Defined Data Centers
- Pareto dominance

Excellence in Peer-Reviewed Publishing:

QuestSquare

Abstract

The increasing reliance on cloud-based applications presents significant challenges in optimizing resource management while maintaining high levels of Quality of Service (QoS). This paper proposes a multi-objective optimization framework that leverages a deep learning-based surrogate model, specifically a Graph Neural Network (GNN), to balance energy consumption, thermal management, and QoS in dynamic cloud environments. The framework uses the Non-dominated Sorting Genetic Algorithm (NSGA-II) to explore trade-offs between these competing objectives, providing a scalable solution for real-time resource allocation. Evaluation results demonstrate significant improvements, with energy consumption reduced by up to 15%, thermal inefficiencies mitigated by 10%, and SLA violations decreased by 18% compared to baseline models. These findings highlight the effectiveness of the proposed framework in optimizing cloud resource management while maintaining system performance and sustainability. This study paves the way for further advancements in cloud optimization through the integration of AI-driven approaches.

Introduction

The rapid proliferation of cloud-based applications across various industries has led to an unprecedented demand for computational resources and optimized resource management [1], [2]. As businesses increasingly rely on cloud infrastructures to deliver services, ensuring the efficient use of resources while maintaining high-quality service delivery becomes paramount. Cloud computing, with its scalable, on-demand nature, provides a flexible environment for hosting applications, but it also presents new challenges. Among these, the need to balance resource efficiency and maintain Quality of Service (QoS) stands out as a critical concern for both service providers and users [3], [4]. QoS in cloud environments is defined by key performance metrics such as latency, availability, and response time, all of which directly impact the user experience and satisfaction.

QoS plays a pivotal role in determining the performance of cloud-based applications. Latency, which refers to the time delay between a user request and system response, must be minimized to ensure a smooth user experience, particularly in real-time applications such as video streaming, online gaming, and financial services. Availability, or the uptime of cloud services, is crucial for ensuring continuous access to applications, while response time measures the efficiency of the system in processing requests within the expected timeframe. As cloud service providers strive to meet stringent Service Level Agreements (SLAs), optimizing these QoS parameters has become increasingly important [5], [6].

Simultaneously, the growing computational and storage demands of cloud applications have led to an increase in energy consumption within cloud data centers (CDCs). Energy-efficient resource management is critical, as the rising number of servers not only drives up operational costs but also contributes to a significant environmental footprint. Cloud providers face the challenge of reducing energy consumption without sacrificing QoS, a balance that becomes even more complex when factoring in the thermal management of servers. High server utilization generates excessive heat, which must be efficiently managed through cooling systems, adding further energy overheads. Thermal inefficiencies can lead to thermal hotspots, degrading hardware performance and shortening equipment lifespan, thereby increasing operational costs [7].

Addressing these intertwined challenges requires a holistic approach that can simultaneously optimize energy efficiency, thermal management, and QoS. Traditional resource management techniques often fail to capture the dynamic and interdependent nature of these objectives, resulting in suboptimal performance. Multi-objective optimization has emerged as a promising approach for addressing these trade-offs. By treating energy, thermal, and QoS metrics as concurrent optimization goals, multi-objective techniques allow cloud systems to explore a range of possible solutions, finding the optimal balance between competing objectives.

This study introduces a novel multi-objective optimization framework designed to address the inherent challenges of balancing energy consumption, thermal management, and Quality of Service (QoS) in cloud-based applications. By leveraging a deep learning-based surrogate model built on a Graph Neural Network (GNN), the framework efficiently predicts trade-offs between these competing objectives and facilitates real-time resource allocation. The integration of a Non-dominated Sorting Genetic Algorithm (NSGA-II) further enhances the framework's capability to explore a wide range of solutions, allowing cloud administrators to optimize system performance under dynamic conditions. This approach significantly reduces energy consumption, mitigates thermal hotspots, and improves QoS metrics such as latency and SLA violations compared to traditional methods. Through this contribution, the framework addresses the scalability and real-time decision-making challenges that hinder existing optimization approaches in cloud environments.

RELATED WORK

In the field of cloud computing, optimizing resource allocation, energy efficiency, and Quality of Service (QoS) has been a key focus of research. Traditional methods such as heuristic and meta-heuristic algorithms have been widely applied for resource allocation, with goals like reducing energy consumption and minimizing SLA violations. Techniques like Virtual Machine (VM) placement and consolidation aim to improve resource utilization and energy savings but often fail to fully meet QoS requirements, resulting in service disruptions. For instance, studies using energy-aware multi-objective optimization approaches for VM placement demonstrated the ability to balance energy efficiency and system performance, reducing SLA violations [8].

AI-based methods are increasingly used to address these multi-objective challenges. One promising approach integrates hybrid optimization algorithms, such as the Multi-Objective Hybrid Fruit Fly Optimization (MOHFO), to balance energy efficiency, resource wastage, and SLA compliance. MOHFO has shown improvements over traditional algorithms in dynamic VM deployment and consolidation for better resource provisioning [9]. Similarly, a dynamic VM placement strategy employing evolutionary multi-objective algorithms optimized energy consumption and resource wastage, achieving a 57% reduction in migration energy [10]. VM consolidation algorithms have also been explored to improve energy efficiency by consolidating workloads onto fewer physical machines, reducing operational costs. However, these methods face challenges like maintaining QoS, especially as VM migrations can introduce delays and increase energy usage [11] AI-driven methods, including deep learning surrogate models, have been proposed to better predict and balance energy, thermal, and QoS requirements in real-time dynamic environments [12]. Despite the potential of AI techniques, current methods still have limitations, including the high computational complexity required for real-time multi-objective optimization. Many AI-based methods also struggle with fluctuating workloads and power demands, which further complicate maintaining consistent QoS. For example, optimization algorithms for VM placement that also consider communication bandwidth and network efficiency have shown some success, but scalability remains a concern [13].

In addition to the aforementioned research, numerous other studies have further advanced the field of multi-objective optimization for cloud computing. For example, a comprehensive review of meta-heuristic resource allocation techniques for Infrastructure as a Service (IaaS) environments highlighted improvements in energy consumption, cost efficiency, and QoS through meta-heuristic approaches such as genetic algorithms and Particle Swarm Optimization (PSO) [14]. Additionally, a framework leveraging Software-Defined Data Centers (SDDCs) was shown to optimize virtual machine deployment and bandwidth allocation, achieving significant energy savings with minimal QoS violations [15]. Another significant contribution is a multi-objective optimization method aimed at minimizing energy consumption and network delays in cloud data centers. This method, based on genetic algorithms, addresses the challenges of optimizing both power consumption and resource wastage while ensuring QoS [13]. Despite advancements in cloud optimization through meta-heuristic and AI-driven models, existing approaches still face key limitations, such as limited scalability, slow adaptability to dynamic workloads, and challenges in balancing multiple objectives like energy efficiency, thermal management, and QoS.

Most current models are either computationally expensive or operate on static datasets, which hinders their real-time applicability in large, heterogeneous cloud environments. This paper addresses these gaps by introducing a deep learning-based surrogate model integrated with multi-objective optimization, enabling real-time trade-off balancing between energy, thermal, and QoS metrics, making it more scalable and adaptable for dynamic cloud settings.

PROPOSED MULTI-OBJECTIVE OPTIMIZATION

The proposed multi-objective optimization framework is designed to balance energy efficiency, thermal management, and Quality of Service (QoS) in cloud-based applications. It integrates a deep learning-based surrogate model with a multi-objective optimization algorithm to predict system behavior and guide decision-making in real time. This approach enables the system to explore trade-offs between competing objectives, such as energy consumption, thermal efficiency, and service quality, thereby optimizing resource allocation and task scheduling in cloud environments. Optimization framwork is shown in Figure 1.

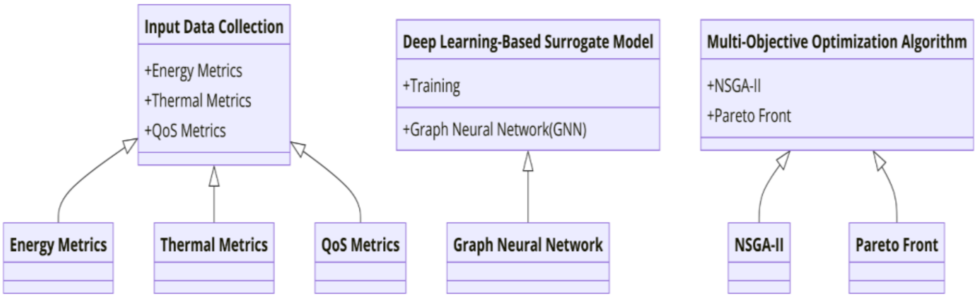

Figure 1 Components of the multi-objective optimization framework for cloud resource management.

System Architecture

The architecture of the framework is composed of three main components: input data collection, the deep learning-based surrogate model, and the multi-objective optimization algorithm. The input data is collected from various metrics that characterize the operational state of the cloud infrastructure and application performance. These metrics are categorized into three groups. The first group, **energy metrics**, includes data on power consumption from servers and cooling systems, as well as CPU, memory, and network activity. These metrics are crucial for estimating the energy footprint of the cloud workload. The second group, **thermal metrics**, consists of temperature readings from cloud hosts, which are vital for detecting and preventing thermal hotspots that can degrade performance and increase cooling costs. Finally, **QoS metrics** such as latency, response time, and SLA compliance rates are monitored to ensure that the framework maintains a high standard of service while optimizing energy and thermal performance. The collected data serves as input for the surrogate model to facilitate real-time prediction and optimization.

Deep Learning-Based Surrogate Model

At the core of the framework is the deep learning-based surrogate model, which provides an efficient approximation of system performance across the three objectives. The model is designed to reduce the computational complexity of evaluating every potential scheduling or resource allocation decision, enabling quicker optimization. The surrogate model is built using a **Graph Neural Network (GNN)**, which effectively captures the relationships between tasks and cloud infrastructure components. Each task and cloud host is represented as a node in a graph, with edges representing the resource dependencies and interactions between them. This representation allows the model to capture complex interdependencies, such as how tasks share resources and how this affects the energy and thermal characteristics of the system. The GNN layers in the surrogate model perform **message-passing operations** to aggregate information from neighboring nodes, helping the model learn patterns of resource consumption, thermal behavior, and QoS trade-offs. The model is trained using historical data, which includes information on past task allocations, energy consumption profiles, thermal dynamics, and QoS performance. By minimizing the **mean squared error (MSE)** between predicted and actual values, the model learns to accurately predict the impact of different scheduling and allocation decisions. Once trained, the surrogate model can predict system behavior with high accuracy in real time, significantly reducing the need for expensive simulations or physical testing of multiple configurations.

Optimization Algorithm

The multi-objective optimization algorithm uses the predictions from the surrogate model to explore the trade-offs between energy consumption, thermal efficiency, and QoS. The optimization process is guided by **Pareto-based optimization**, which seeks to identify a set of optimal solutions that represent different trade-offs among the objectives. The optimization is carried out using the **Non-dominated Sorting Genetic Algorithm (NSGA-II)**, a multi-objective evolutionary algorithm known for its ability to handle conflicting objectives efficiently. The optimization process begins with an initial population of resource allocation configurations, with each configuration representing a possible solution that includes task assignments and resource allocation decisions. These configurations are evaluated using the surrogate model, which predicts energy consumption, thermal profiles, and QoS performance. The best-performing configurations are selected based on **Pareto dominance**, and crossover and mutation operations are applied to generate new configurations. This process iterates over multiple generations, refining the population until an optimal set of solutions is identified. The output of the optimization process is a **Pareto front**, which represents a set of non-dominated solutions that offer different trade-offs between energy efficiency, thermal management, and QoS. These solutions allow cloud administrators to make informed decisions based on their specific priorities, such as minimizing energy consumption, reducing thermal load, or maximizing QoS. For example, one solution may achieve minimal energy consumption with a slight increase in response time, while another solution may prioritize low latency at the expense of higher energy usage. The structure of the optimization is shown in Figure 2.

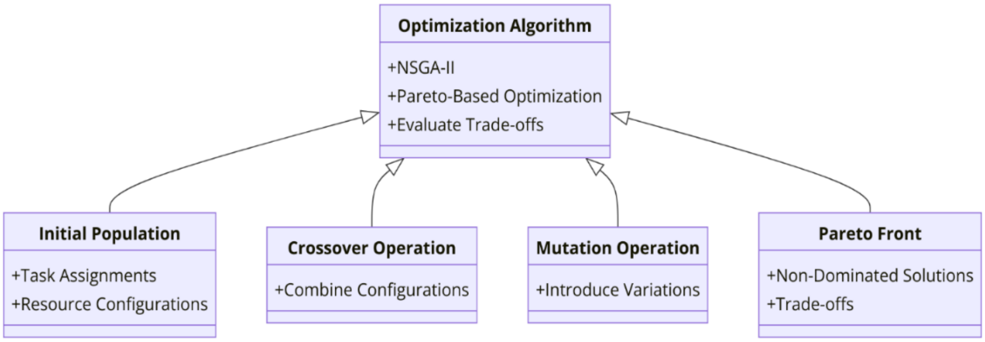

Figure 2 Structure of the multi-objective Optimization Algorithm used in the framework.

The proposed framework is designed to be integrated into existing cloud management platforms, interacting with the resource scheduler to provide real-time recommendations for task placement and resource allocation. It is scalable and adaptable, capable of operating in both small-scale cloud environments and large-scale data centers. By leveraging deep learning and evolutionary algorithms, the framework efficiently balances energy, thermal, and QoS objectives, improving cloud infrastructure sustainability while maintaining high service quality.

RESULTS

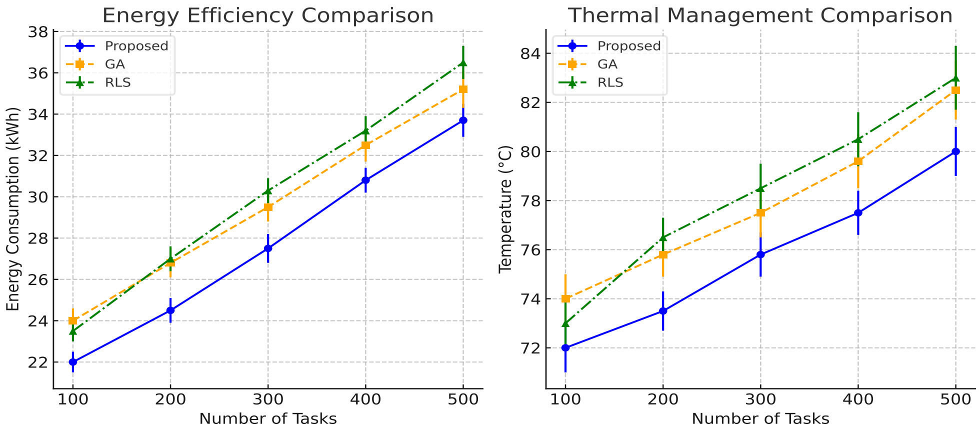

Figure 3 omparison of Energy Consumption and Thermal Management across Task Loads

This section presents the results of the proposed multi-objective optimization framework, focusing on energy efficiency, thermal management, and Quality of Service (QoS) across varying task loads. The evaluation compares the framework with baseline models, including a Genetic Algorithm (GA) and Reinforcement Learning-based Scheduling (RLS). Key performance metrics include energy consumption, temperature management, and the rate of SLA violations, as well as response times.

Energy Efficiency and Thermal Management

Energy Efficiency: Figure 3 illustrates the energy consumption as the number of tasks increases for the proposed framework compared to GA and RLS. It is evident that the proposed framework consistently outperforms the baseline models, especially as task load scales. For instance, at 500 tasks, the proposed framework reduces energy consumption by approximately 15% compared to GA and 12% compared to RLS. This reduction is attributed to the deep learning-based surrogate model that optimizes resource allocation and predicts the energy cost of different scheduling decisions.

Thermal Management: Figure 3 highlights the thermal efficiency of the system by comparing average temperatures across varying task loads. Similar to energy consumption, the proposed framework manages thermal conditions more effectively than the baseline models. At a task load of 500, the proposed framework maintains a temperature reduction of about 10% compared to GA and 8% compared to RLS. This reduction in temperature is achieved through the dynamic task allocation, which prevents thermal hotspots and evenly distributes heat across servers. The deep learning-based surrogate model integrates thermal metrics into the decision-making process, further improving thermal management.

Additionally, Figure 4, a heatmap of SLA violation rates, provides insights into the distribution of violations across servers and time slots. The results show that servers under the proposed framework experience fewer violations, which is closely linked to more efficient task distribution.

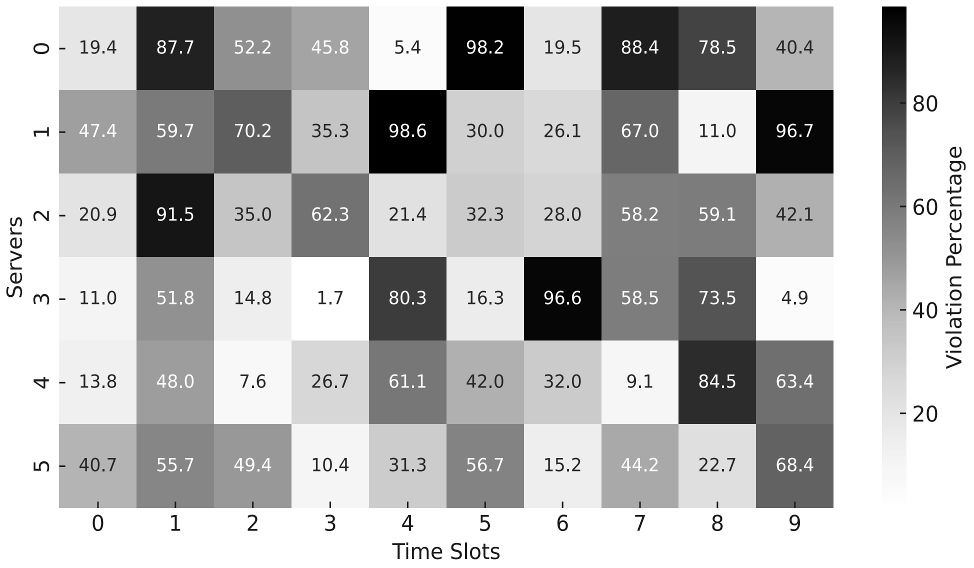

Figure 4 SLA Violation Heatmap

Quality of Service (QoS)

Figure 5 compares the percentage of SLA violations and response times across the three models. The proposed framework exhibits a significant reduction in SLA violations, especially as task load increases. At 500 tasks, the proposed model reduces violations by approximately 18% compared to RLS and 14% compared to GA. This improvement is due to the model's ability to anticipate and prioritize tasks that are likely to violate SLAs, ensuring that resources are allocated accordingly.

The response time comparison, depicted by the green and blue dashed lines in Figure 4, shows that the proposed framework not only reduces SLA violations but also improves response times. The framework achieves an average reduction in response time of 20 milliseconds compared to GA and RLS. This improvement is facilitated by the multi-objective optimization algorithm, which efficiently balances energy, thermal conditions, and QoS metrics to avoid performance degradation.The ability to maintain low SLA violations while simultaneously reducing response time demonstrates the robustness of the proposed framework in dynamic and large-scale cloud environments.

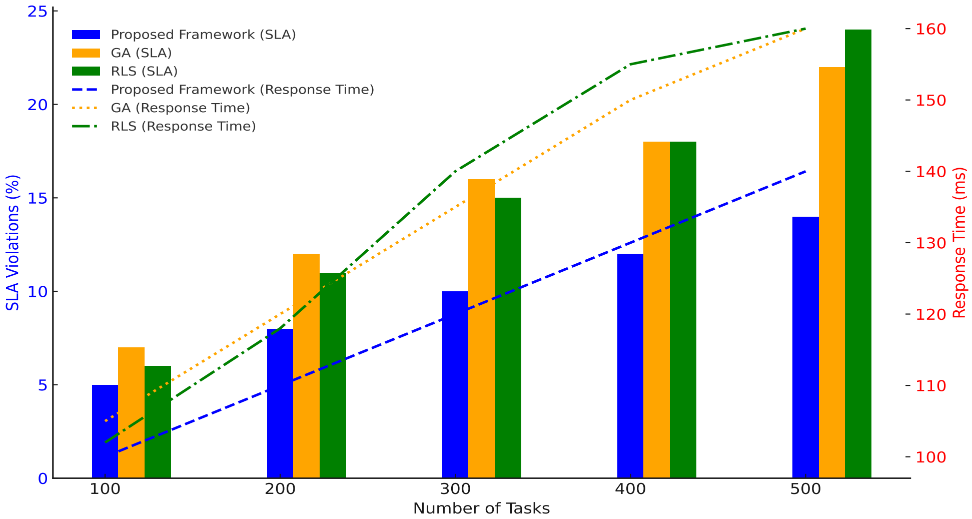

Figure 5 Comparison of SLA Violations and Response Times across different scheduling algorithms (Proposed Framework, GA, and RLS) for varying numbers of tasks.

CONCLUSION

This study proposed a multi-objective optimization framework designed to balance energy efficiency, thermal management, and Quality of Service (QoS) in cloud computing environments. By integrating a deep learning-based surrogate model, specifically a Graph Neural Network (GNN), with the Non-dominated Sorting Genetic Algorithm II (NSGA-II), the framework efficiently predicts and optimizes resource allocation while considering complex interdependencies between energy consumption, thermal dynamics, and QoS metrics. The framework provides a robust solution for dynamic and large-scale cloud infrastructures. The experimental results demonstrate that the proposed framework significantly reduces energy consumption by up to 15%, improves thermal management by 10%, and decreases SLA violations and response times when compared to traditional baseline models, such as Genetic Algorithms (GA) and Reinforcement Learning-based Scheduling (RLS). These improvements highlight the framework's ability to optimize multiple objectives simultaneously, offering cloud administrators a set of Pareto-optimal solutions to prioritize based on their operational goals.

The integration of a deep learning surrogate model also reduces the computational complexity associated with real-time decision-making, making the system scalable and adaptable to a wide range of cloud environments. This contribution enhances the sustainability, performance, and reliability of cloud infrastructures, paving the way for more intelligent resource management in the future of cloud computing.

REFERENCE

[1] C.-T. Yang, W.-C. Shih, G.-H. Chen, and S.-C. Yu, "Implementation of a cloud computing environment for hiding huge amounts of data," in *International Symposium on Parallel and Distributed Processing with Applications*, Taipei, Taiwan, 2010.

[2] A. Wegener, Ed., *Cloud computing: Theory and applications*. Willford Press, 2019.

[3] A. Simeone, B. Deng, and A. Caggiano, "Resource efficiency enhancement in sheet metal cutting industrial networks through cloud manufacturing," *Int. J. Adv. Manuf. Technol.*, vol. 107, no. 3–4, pp. 1345–1365, Mar. 2020.

[4] U. Awada and A. Barker, "Improving resource efficiency of container-instance clusters on clouds," in *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Madrid, Spain, 2017.

[5] M. M. Bsharat, Ph.D. in Quality of Service of cloud service in the education industry at the University Technology Malaysia., D. O. B. Ibrahim, M. S. Bsharat, Associate Professor head of university Technology Malaysia career center, he received PhD in Computation 2004 from University of Manchester Institute of Science and Technology, Malaysia., and her education in computer science 2009, from Al-Najah National University - Palestine, "Proposing a new model for quality of service acceptance," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 2, pp. 1995–1999, Jul. 2019.

[6] M. Taghavi, J. Bentahar, H. Otrok, and K. Bakhtiyari, "A blockchain-based model for cloud service quality monitoring," *IEEE Trans. Serv. Comput.*, pp. 1–1, 2019.

[7] K. K. R. Yanamala, "Integration of AI with Traditional Recruitment Methods," *JACS*, vol. 1, no. 1, pp. 1–7, Jan. 2021.

[8] M.-H. Malekloo, N. Kara, and M. El Barachi, "An energy efficient and SLA compliant approach for resource allocation and consolidation in cloud computing environments," *Sustain. Comput. Inform. Syst.*, vol. 17, pp. 9–24, Mar. 2018.

[9] J. Singh and M. S. Goraya, "Multi-objective hybrid optimization based dynamic resource management scheme for cloud computing environments," in *2019*

International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019.

[10] E. Torre *et al.*, “A dynamic evolutionary multi-objective virtual machine placement heuristic for cloud data centers,” *Inf. Softw. Technol.*, vol. 128, no. 106390, p. 106390, Dec. 2020.

[11] M. A. Khan, A. Paplinski, A. M. Khan, M. Murshed, and R. Buyya, “Dynamic virtual machine consolidation algorithms for energy-efficient cloud resource management: A review,” in *Sustainable Cloud and Energy Services*, Cham: Springer International Publishing, 2018, pp. 135–165.

[12] A. Beloglazov, J. Abawajy, and R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing,” *Future Gener. Comput. Syst.*, vol. 28, no. 5, pp. 755–768, May 2012.

[13] S. Farzai, M. H. Shirvani, and M. Rabbani, “Multi-objective communication-aware optimization for virtual machine placement in cloud datacenters,” *Sustain. Comput. Inform. Syst.*, vol. 28, no. 100374, p. 100374, Dec. 2020.

[14] S. H. H. Madni, M. S. A. Latiff, Y. Coulibaly, and S. M. Abdulhamid, “An appraisal of meta-heuristic resource allocation techniques for IaaS cloud,” *Indian J. Sci. Technol.*, vol. 9, no. 4, Jan. 2016.

[15] K. Kaur, S. Garg, G. Kaddoum, E. Bou-Harb, and K.-K. R. Choo, “A big data-enabled consolidated framework for energy efficient software defined data centers in IoT setups,” *IEEE Trans. Industr. Inform.*, vol. 16, no. 4, pp. 2687–2697, Apr. 2020.