International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Task Scheduling Optimization in Cloud Environments

Chandragiri Sai Saketh[1], Daddala Pavan[2], Chukka Bhuvana Bharath[3], Seelam Bhanu Kiran[4], I. Shalini[5]

[1,2,3,4,5]Department of CSE (Cyber Security), Geethanjali Institute of Science and Technology, Nellore.

Abstract:
Cloud computing systems depend on effective task scheduling to achieve better performance and efficient use of resources. This study presents a Generalized Priority (GP) scheduling approach designed to improve task execution in cloud environments. The proposed method is evaluated against conventional scheduling techniques such as First Come First Serve (FCFS) and Round Robin (RR). By using queue-based workload analysis, the GP algorithm directs incoming tasks to the least loaded queue, which helps reduce waiting time, response time, and average queue length. In addition, a Skewness-Avoidance Multi-Resource Allocation (SAMR) mechanism is introduced to distribute diverse resources more effectively and maintain balance across physical machines. Simulation and experimental findings indicate that the proposed framework enhances scheduler efficiency and significantly lowers waiting time compared to existing methods. Future improvements may include support for larger workloads and the integration of parallel cloud algorithms to further optimize execution speed.

Keywords: Cloud computing, Skewness-Avoidance Multi-Resource Allocation (SAMR), First Come First Serve (FCFS), Round Robin (RR).

I.INTRODUCTION

Because system performance depends on how well activities are allocated to available resources, task scheduling is a major difficulty in cloud and edge computing. Traditional scheduling techniques frequently fail to maintain appropriate load balance, minimize delays, or make effective use of resources as computing systems become more dynamic and heterogeneous. As a result, recent research has concentrated on intelligent scheduling models that take into account a variety of variables, including task prioritization, workflow dependencies, real-time system circumstances, and energy consumption.

According to current research, scheduling efficiency can be greatly increased by using optimization techniques, machine learning methodologies, and energy-aware resource management tactics. Research on virtual machine management, workflow-based scheduling, multidimensional resource allocation, and reinforcement learning has shown improved reaction time, makespan, scalability, and power consumption. In order to promote the design of effective, scalable, and dependable cloud computing systems, this study focuses on comprehending recent improvements in task scheduling and resource optimization.

Effective task scheduling is necessary in cloud and edge computing systems to enhance performance, minimize delays, and maintain resource balance. In order to improve multidimensional resource balance in edge clusters, Zou et al. [1] introduced an enhanced particle swarm optimization-based scheduling model that takes into account CPU, memory, disk I/O, and network bandwidth. A thorough assessment of multi-objective task scheduling techniques was published by Abraham et al. [2], with a focus on goals like makespan, energy efficiency, and cost optimization. A Soft Actor-Critic-based real-time scheduling method that enhanced reaction time and load balancing under dynamic workloads was presented by Wang

et al. [5]. According to these findings, adaptive scheduling strategies can greatly raise the caliber of cloud services.

For managing complicated applications in distributed systems, workflow scheduling and dependency-aware execution have become crucial. The ADWEH system, created by Krishna and Vali [3], prioritizes workflow activities according to dependencies and runtime conditions by combining deep reinforcement learning with enhanced Harris Hawk Optimization. Their approach enhanced scalability, energy efficiency, and makespan. For heterogeneous datacenters, Stan and Pop [4] presented the 2HD and QL-2HD algorithms, which prioritize early output availability over total execution time minimization. For big task graphs, their work demonstrated improved utility and adaptability. These methods emphasize how crucial intelligent dependency-aware scheduling is.

Sustainable cloud computing requires virtual machine management and energy-efficient resource allocation. Energy-aware resource allocation policies were established by Beloglazov et al. [6] to lower power usage without sacrificing service quality. Additionally, Beloglazov [7] suggested dynamic virtual machine consolidation methods to enhance resource usage via live migration. The advantages of workload-based consolidation were further illustrated by other energy-conscious virtual machine management techniques [8]. While updated VM migration techniques [10] decreased migration time and network overhead, Wang et al. [9] suggested VMPlanner to optimize VM placement and traffic routing for lowering network power costs. When taken as a whole, these studies demonstrate how important effective resource management is to scalable and dependable cloud infrastructure.

II.LITERATURE SURVEY

Effective task scheduling is necessary in cloud and edge computing systems to enhance performance, minimize delays, and maintain resource balance. In order to improve multidimensional resource balance in edge clusters, Zou et al. [1] introduced an enhanced particle swarm optimization-based scheduling model that takes into account CPU, memory, disk I/O, and network bandwidth. A thorough assessment of multi-objective task scheduling techniques was published by Abraham et al. [2], with a focus on goals like makespan, energy efficiency, and cost optimization. A Soft Actor-Critic-based real-time scheduling method that enhanced reaction time and load balancing under dynamic workloads was presented by Wang et al. [5]. According to these findings, adaptive scheduling strategies can greatly raise the caliber of cloud services.

For managing complicated applications in distributed systems, workflow scheduling and dependency-aware execution have become crucial. The ADWEH system, created by Krishna and Vali [3], prioritizes workflow activities according to dependencies and runtime conditions by combining deep reinforcement learning with enhanced Harris Hawk Optimization. Their approach enhanced scalability, energy efficiency, and makespan. For heterogeneous datacenters, Stan and Pop [4] presented the 2HD and QL-2HD algorithms, which prioritize early output availability over total execution time minimization. For big task graphs, their work demonstrated improved utility and adaptability. These methods emphasize how crucial intelligent dependency-aware scheduling is.

Sustainable cloud computing requires virtual machine management and energy-efficient resource allocation. Energy-aware resource allocation policies were established by Beloglazov et al. [6] to lower power usage without sacrificing service quality. Additionally, Beloglazov [7] suggested dynamic virtual machine consolidation methods to enhance resource usage via live migration. The advantages of workload-based consolidation were further illustrated by other energy-conscious virtual machine management techniques [8]. While updated VM migration techniques [10] decreased migration time and network overhead, Wang et al. [9] suggested VMPlanner to optimize VM placement and traffic routing

for lowering network power costs. When taken as a whole, these studies demonstrate how important effective resource management is to scalable and dependable cloud infrastructure.

III.EXSISTING SYSTEM

One of the most crucial methods for enhancing system performance and guaranteeing effective resource use in cloud computing is load balancing. Several servers or virtual machines collaborate to handle a lot of user requests and application duties in a cloud environment. Inadequate workload distribution can cause delays, slow response times, and lower service quality because some nodes may become overworked while others are idle or underutilized. By spreading work equally among all available resources, load balancing helps to avoid bottlenecks and preserve efficient system operation. Effective load balancing is crucial for preserving dependability and consistency under fluctuating workload conditions, but current cloud systems have mostly concentrated on optimizing resource consumption and cutting execution time. A well-thought-out load balancing method boosts flexibility by adjusting to changing resource needs, increases fault tolerance by providing backup and recovery during failures, and improves overall system stability. Additionally, it guarantees that jobs are performed equitably regardless of the source of the requests, lowers response times, and increases throughput. As a result, load balancing is essential to creating a cloud computing environment that is dependable, scalable, effective, and able to satisfy user demands.

IV.PROPOSED SYSTEM

The Improved Weighted Round Robin (IWRR) algorithm enhances traditional load balancing by considering not only the static configuration of servers but also the execution time of incoming tasks. In this approach, each server is assigned a weight based on its processing capability, memory capacity, and overall performance. Unlike the basic Round Robin method, which distributes tasks equally without considering system conditions, the improved algorithm makes more informed scheduling decisions. It analyzes the expected execution time of tasks and matches high-execution-time tasks with servers that have higher weights and greater processing power. This allows resource-intensive tasks to be handled by stronger servers, while lighter tasks are assigned to less loaded machines. As a result, the workload is distributed more evenly across all available servers, reducing the chances of overload on specific nodes. This balanced allocation helps minimize response time, improve resource utilization, increase throughput, and ensure better overall performance of the cloud environment.

SYSTEM ARCHITECTURE

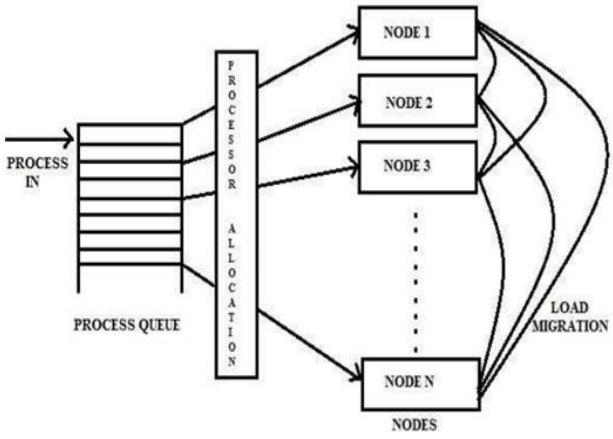

FIG 1. SYSTEM ARCHITECTURE

V.RESULTS & DISCUSSION

When sophisticated scheduling and load balancing strategies were used, experimental study revealed considerable gains in cloud system performance. By allocating jobs according to queue workload, congestion was lessened and more orderly task processing was guaranteed. Compared to conventional techniques like FCFS and ordinary Round Robin, this method decreased average waiting time, shortened response delays, and shortened queue length. Improved task distribution also made better use of the computational resources at hand, enabling the scheduler to handle workloads more skillfully. Workload-aware scheduling can enhance efficiency and service quality in cloud environments, as evidenced by the system's stability despite an increase in incoming workloads.while server capacity and task execution time were taken into account while allocating tasks, further improvements were seen. The burden was more equally spread throughout the system and overload on some machines was avoided by matching lengthier tasks with more competent servers. Higher throughput, better server utilization, and fewer performance bottlenecks were the outcomes of this. Reliability and scalability were further enhanced by the effective distribution of heterogeneous resources, which significantly decreased imbalance among physical machines. Overall, the findings demonstrate that balanced resource allocation and intelligent scheduling can greatly improve cloud performance by lowering latency, boosting responsiveness, and guaranteeing efficient use of infrastructure.

VI.CONCLUSION & FUTURE WORK

Enhancing the performance, dependability, and scalability of cloud computing systems requires effective job scheduling and load balancing. In addition to increasing throughput and overall resource utilization, proper job distribution among available resources helps minimize waiting times, response delays, and workload imbalance. Cloud environments can achieve improved service quality and more efficient task execution by taking into account variables including queue workload, task execution time, server capabilities, and resource availability. These enhancements guarantee that cloud services can efficiently manage changing workloads and user demands while also promoting system stability.\Subsequent research endeavors may concentrate on expanding scheduling methods to accommodate more extensive and intricate workloads in real-time cloud settings. To further increase productivity and shorten execution times, sophisticated techniques including adaptive resource allocation, machine learning-based workload prediction, and parallel task scheduling might be investigated. Furthermore, integrating energy-conscious scheduling, fault tolerance, and security measures can contribute to the development of cloud systems that are more intelligent, dependable, and sustainable for use in the future.

REFERENCES:

1. Z. Zou, Z. Zhai, X. Yan, Z. You and L. Chen, "Multidimensional Resource Task Scheduling Based on Particle Swarm Optimization in Edge Computing," in *IEEE Access*, vol. 13, pp. 116701-116712, 2025, doi: 10.1109/ACCESS.2025.3585628.
2. O. L. Abraham, M. A. B. Ngadi, J. B. M. Sharif and M. K. M. Sidik, "Multi-Objective Optimization Techniques in Cloud Task Scheduling: A Systematic Literature Review," in IEEE Access, vol. 13, pp. 12255-12291, 2025, doi: 10.1109/ACCESS.2025.3529839.
3. M. Shiva Rama Krishna and D. Khasim Vali, "ADWEH: A Dynamic Prioritized Workflow Task Scheduling Approach Based on the Enhanced Harris Hawk Optimization Algorithm," in IEEE Access, vol. 13, pp. 35490-35515, 2025, doi: 10.1109/ACCESS.2025.3543880.
4. R. -G. Stan and F. Pop, "2HD: Hybrid Dynamic Utility-Driven Dependency-Aware Task Scheduling in Heterogeneous Datacenters," in IEEE Access, vol. 13, pp. 212083-212103, 2025, doi: 10.1109/ACCESS.2025.3643945.
5. J. Wang, S. Li, X. Zhang, K. Zhu, C. Xie and F. Wu, "Deep Reinforcement Learning Task Scheduling Method for Real-Time Performance Awareness," in IEEE Access, vol. 13, pp. 31385-31400, 2025, doi: 10.1109/ACCESS.2025.3534980. k

6. W. Lyu et al., "HeroCS: Cooperative Courier Scheduling for Heterogeneous Tasks in Last-Mile Delivery," in IEEE Transactions on Mobile Computing, doi: 10.1109/TMC.2025.3644653.
7. M. Seyfipoor, S. Muhammad Jaffry and S. Mohammadi, "A Hybrid Priority-Laxity-Based Scheduling Algorithm for Real-Time Aperiodic Tasks Under Varying Environmental Conditions," in IEEE Access, vol. 13, pp. 173035-173051, 2025, doi: 10.1109/ACCESS.2025.3612340.
8. T. Li, S. Lin and Y. Sun, "Linear Reciprocating RGV Inbound and Outbound Scheduling With Movable Buffer Zone," in IEEE Access, vol. 14, pp. 2815-2833, 2026, doi: 10.1109/ACCESS.2025.3649645.
9. S. Heng, T. Cheng, J. Song, Z. He, L. Liu and Y. Wang, "Adaptive Dwell Scheduling Based on Dual-Side Time Pointers for Simultaneous Multi-Beam Radar," in Tsinghua Science and Technology, vol. 30, no. 3, pp. 1190-1200, June 2025, doi: 10.26599/TST.2023.9010161.
10. S. Nambi and P. Thanapal, "EMO-TS: An Enhanced Multi-Objective Optimization Algorithm for Energy-Efficient Task Scheduling in Cloud Data Centers," in IEEE Access, vol. 13, pp. 8187-8200, 2025, doi: 10.1109/ACCESS.2025.3527031.