Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Optimized Load balancing in Cloud Computing using (MCGEO Algorithm)

Mr. Syed Muzibuddin
Assisstant Professor
Dept. CSE(CS)
RGMCET
Nandyal, India

Kethireddy Arthi Sree
Dept. CSE(CS)
RGMCET
Nandyal, India

Busagani Kondaiah
Dept. CSE(CS)
RGMCET
Nandyal, India

Yaparlapati Om Shashidha
Dept. CSE(CS)
RGMCET
Nandyal, India

***Abstract*—The cloud computing environments require either smart load balancing to manage dynamic workloads, heterogeneous virtual machines, hard Service Level Agreement (SLA), and changing security threats. The traditional loading balancing methods tend to be fixed, focus on constrained performance goals, and are not adaptable to the changing real-time cloud situations. This paper suggests an improved load balancing system with a Cognitive Digital Twin (CDT) based on a Mouse- Cat Golden Eagle Optimization (MCGEO) algorithm guided by Trail and Error Learning (RL) to address all these challenges. The Cognitive Digital Twin is a system to build a real-time virtual model of the cloud infrastructure, constantly checking the state of parameters in the system, including CPU utilization, memory usage, bandwidth usage, energy efficiency, SLA - compliance, and security risk indicators. According to this real-time feedback, the suggested MCGEO algorithm is going to be used to dynamically pick the most suitable virtual machines to execute the tasks based on the multi-objective optimization. Reinforcement Learning will improve the optimization process as the scheduler is able to learn adaptive policies based on past execution results, SLA violations, and feedback based on security, and enhances long-term decision-making. The presented framework manages to reduce makespan, response time, server load, energy consumption, and SLA violation rates and increase resource utilization and system security at the same time. The simulations show that the presented CDT - RL - MCGEO system is more efficient in the execution and SLA compliance and security risk resistance than the current load balancing strategies. The findings indicate that Cognitive Digital Twins coupled with RL - controlled meta-heuristic optimization is an efficient and smart answer to the next generation of cloud load balancing systems.**

***Keywords*—Cloud Computing, Load Balancing, Cognitive Digital Twin, Reinforcement Learning, Meta-Heuristic Optimization, Virtual Machine Scheduling.**

I. INTRODUCTION

Cloud computing is an increasingly popular model of providing computing resources (scaling) and on-demand computing with the help of the Internet. Its blistering development in the sphere of enterprise systems and Internet of Things (IoT) applications has made efficient management of resources even more demanded. But the dynamic workload behaviour, the heterogeneous virtual machines and the rigid Service Level Agreements (SLAs) make the optimal resource utilisation a difficult task. Among them, load balancing is a vital challenge to be considered in terms of performance, reliability, and user satisfaction. The conventional load balancing methods are Round-Robin, Least Connection, and the heuristic based methods which are simple and yet not dynamic. Such methods do not react to the dynamics of the real time clouds, which causes imbalance of resources, increased response time, energy usage, and SLA breach. Machine learning and reinforcement learning (RL)-based scheduling procedures have been considered to overcome these problems because they are adaptive in nature. Even though RL approaches enhance the decision-making process over time, they do not tend to provide real-time awareness of the system and multi-objective optimization. Cloud computing has become a computing paradigm which allows the provision of scalable, flexible and on-demand computing services, over the Internet. Its fast integration in enterprise applications, big data analytics and Internet of Things (IoT) ecosystems has profoundly altered the method of consumption and management of the computational services. Cloud platforms enable users to obtain computing, storage, and networking resources without engaging in hardware management by utilizing the capabilities of the virtualization technologies. In spite of these benefits, the effective management of cloud resources is still a significant challenge because of high dynamism of the workloads and the heterogeneity of cloud infrastructures. Effective load balancing, which is the optimal allocation of tasks to various virtual machines (VMs), is one of the most problematic issues in the cloud environment. This requires proper load balancing to achieve high performance, reliability, energy efficiency and user satisfaction. But the nature of variable workloads, different VM setups and strict Service Level Agreements (SLAs) makes the scheduling of decisions more difficult. The bad load balancing may cause under utilization of the resources, load to the server, higher response time, violation of the SLA, and high energy usage. Round-Robin, Least Connection, and other heuristic-based methods of traditional load balancing methods are very common because of their simplicity and low computational power. However, these approaches are mostly passive and reactive because they are based on pre - described regulations and minimal information on the system. Consequently, they find it difficult to keep up with real-time cloud dynamics, intermittent changes in workload and fluctuating Quality of Service (QoS) needs. Therefore, these strategies cannot suitably deliver the best performance in contemporary clouds. In order to overcome these constraints, new studies have been conducted on the application of machine learning and reinforcement learning (RL)-based scheduling methods. The techniques based on RL allow the scheduler to make decisions based on past experiences and change them over time accordingly. These techniques have demonstrated good outcomes in the response time reduction and better load distribution. Nonetheless, RL - based solutions are frequently inadequate when they do not have real-time awareness about the system or do not consider a number of contradictory goals like energy efficiency, SLA - compliance, and security limits. Simultaneously, meta-heuristic optimization algorithms like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Genetic Algorithms (GA) and their hybrid versions have been widely researched on cloud scheduling problems. These algorithms can also search large search space and approximate optimal solutions to difficult multi objective optimization problems. Although they are effective, most of the available meta-heuristic schemes lack the capability of taking into account real-time feedback, security threats, and SLA - related restrictions, which reduces their usefulness in dynamic cloud environments. To address these issues, this paper suggests the new load balancing framework which combines a Cognitive Digital Twin (CDT) of the cloud infrastructure with a Mouse-Cat Golden Eagle Optimization (MCGEO) algorithm based on the reinforcement learning. The Cognitive Digital Twin is a virtual representation of the cloud system and is interactive, continuously monitoring the state of VMs, workload trends, resource usage, and non-compliance with SLA and indicators of security risks. This real-time consciousness is used to make decisions on scheduling in advance and based on information, as opposed to adjustments. More so, the reinforcement learning is used to drive the MCGEO optimization process by modifying exploration and exploitation strategies with reference to both historical and real-time feed backs. The suggested framework is able to optimize the various performance goals such as response time, makespan, energy usage, server load, SLA attainment and system security in parallel. By this smartly intertwined mechanism, the suggested scheme improves the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

agility, effectiveness, and strength of load balancing in the cloud computing systems of the next generation.

II. LITERATURE SURVEY

[1] Xu et al. (2016) present a review of the existing load balancing algorithms in VM placement in cloud data center, including schemes of classification and the issues related to heterogeneous cloud environment.

[2] The Binary PSOGSA algorithm is a hybrid version of Particle Swarm Optimization and Gravitational Search Algorithm that suggests a new efficient method of balancing the task schedules through better mapping of tasks to the VM, and achieving better use of resources (Alnusairi et al., 2018).

[3] Ebadifard and Babamir (2021) propose an autonomous load balancing approach that can reduce the inter - VM communication overheads and enhance the workload distribution and allocation based on the dynamic classification of the requests.

[4] Shafiq et al. (2021) create an IaaS - based workload balancing algorithm, which puts more importance on QoS metrics and SLA adherence resulting in more efficient use of resources and shortened execution periods.

[5] Devaraj et al. (2020) suggest FIMPSO, a hybrid Firefly and Improved Multi-Objective PSO algorithm that can spread the load of clouds efficiently to optimize the response time and usage of resources.

[6] Pradhan and Bisoy (2022) introduce LBMPSO, an adapted version of PSO - based load balancing, which minimizes makespan and better utilizes resources based on job and resource data about the data center.

[7] Using the approach of Sefati et al. (2022) to apply to VM load balancing, it is shown that the use of the Grey Wolf Optimization (GWO) yields better cost reduction and optimization of response time, in comparison with the meta-heuristics used as baseline.

[8] Dudekula and Reddy (2025) suggest an SLA - aware ML-based VM scheduling algorithm, which predicts VM performance based on real-time metrics, and will yield high SLA compliance and fewer task completion times.

[9] Such approaches as Deep Reinforcement Learning (DRL) and Deep Q-Networks (DQN) have become widely used in addressing load balancing and task scheduling in a conversation with dynamic reinforcement learning. These solutions are responsive to dynamism in workload distribution, minimize job rejects, and adhere to the Service Level Agreement (SLA) limitations.

[10] According to recent meta-heuristic survey studies (2023), nature-inspired algorithms, such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Grey Wolf Optimization (GWO), and other hybrid models show significant superiority to the conventional load balancing techniques. These algorithms are more efficient with respect to response times, throughput as well as better use of cloud resources.

Table 1. Characteristics and Difficulties of Existing Cloud Load Balancing Models

Author [Citation]	Adopted Methodology	Characteristics	Difficulties / Limitations
Ismayilov and Topcuoglu (2020)	Neural Network (NN) model	• High system reliability • Minimum makespan	• Requires extensive training data • High dependency on machine learning techniques
Lavanya, Shanthi, and Saravanan (2020)	SLA-LB Model	• Reduced execution time • No penalty cost	• Energy consumption not considered
Liu et al. (2017)	SAW (Simple Additive Weighting) model	• Minimum makespan • Improved reliability score	• Limited scalability under dynamic workloads
Neelima and Reddy (2020)	ADA Model	• Minimum execution cost • Improved load balancing	• Continuous task arrivals not considered • VM scheduling needs further optimization
Panda, Gupta, and Jana (2019)	Min-Min and Max-Min models	• Low computational complexity • Maximum cloud usability	• Fault tolerance not addressed
Pang et al. (2019)	EDA-GA (Estimation of Distribution Algorithm with GA)	• Improved load balancing • Minimum execution time	• Real cloud infrastructure characteristics not fully considered
Sanaj and Prathap (2020)	CSSA Algorithm	• Low operational cost • Maximum throughput	• Requires deeper real-time performance analysis
Yusuf (Mulge, 2019)	MGWO Model	• Reduced energy consumption • Lower makespan	• Task priority handling not sufficiently explored
H. Singh, Tyagi, and Kumar (2021)	CSLBA Technique	• Balanced data center loading (DCL) • Reduced APCDC and ACDC parameters	• Resource distribution strategies not considered

Lack of energy savings, excessive resource utilization, insufficient real-time scalability and poor fault tolerance are some of the challenges that are faced by traditional cloud load-balancing techniques. Also, inefficient scheduling and violations of SLA are usually caused by the absence of task priority consideration and higher computational time. To address these restrictions, this study offers a new multi-objective hybrid optimization method to the cloud load balancing that maximizes the energy usage, resource use, SLA adherence, and robustness of the whole system.

III. DIGITAL TWIN BASED FRAMEWORK OF OPTIMIZED LOAD BALANCING WITH COGNITION.

A Cognitive Digital Twin (CDT) is a smart simulated version of the physical cloud infrastructure, which is monitoring, analyzing, and forecasting the behavior of the system in real time. In the optimized load balancing architecture that is proposed, the CDT will keep a dynamic model of the virtual machines, tasks and resource consumption parameters like CPU load, memory consumption and availability of bandwidth, response time, and energy consumption. Such ongoing synchronization of the physical cloud environment and its digital twin allows to know the system correctly and make decisions beforehand. CDT polls real time feed backs provided by the cloud environment based on resource monitoring and task execution logs. This is utilized in order to assess the present states of the systems and forecast the workload trends in the future. Keeping the current picture of VM performance and load status, the CDT helps the load balancer to detect idle and overloaded virtual machines. This real-time understanding will guarantee that the scheduling decisions of tasks are not made on any fixed assumptions but rather they represent the real state of cloud operation. In order to maximize it further, Reinforcement Learning (RL) is incorporated into the CDT to inform the scheduling decisions. The RL agent adapts optimal policies in terms of task and VM allocation to the cloud environments through engagement with the environment and feedback in terms of reward and punishment. There are rewards on better response time, less energy use, even load distribution, SLA adherence, and safe task implementation, and penalties on the SLA breach, high delays, over use of resources. With time the RL agent modifies its policy to achieve improved scheduling efficiency based on different workload conditions. Active and reactive load balancing is made possible by the Cognitive Digital Twin intelligence, coupled with RL - guided optimization. The proposed framework does not respond to overload

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

situations once they have been experienced but predicts the behavior of the system and gives optimal allocations of the tasks beforehand. It leads to the decreased reaction time, the increased efficiency of the energy consumption, the maximized use of resources, and the SLA compliance. The load balancing framework that is optimized with the help of the CDT is thus a strong and intelligent framework used in the management of dynamic and heterogeneous cloud computing environments.

IV. SYSTEM MODEL

The optimized load balancing system proposed in the model is intended to efficiently balance the workload in a cloud computing environment in the case of dynamic workload based on Cognitive Digital Twin (CDT) and Reinforcement Learning (RL)-guided Mouse-Cat Golden Eagle Optimization (MCGEO) algorithm in terms of SLA compliance and security-based feedback. There are four primary components of the model that are: (a) cloud users, (b) cloud broker with intelligent scheduler, (c) Cognitive Digital Twin layer and (d) cloud data center.

Cloud Users: The users of cloud environments use the Internet to send their computational jobs to the cloud environment. Parameters like the length of tasks, their execution priority, CPU requirement, memory demand, bandwidth requirement, and the level of sensitivity of the security of each task characterize each task. These activities are dynamic and heterogeneous and come in dynamically, and necessitate the intelligent scheduling decisions, to satisfy the performance and SLA limits.

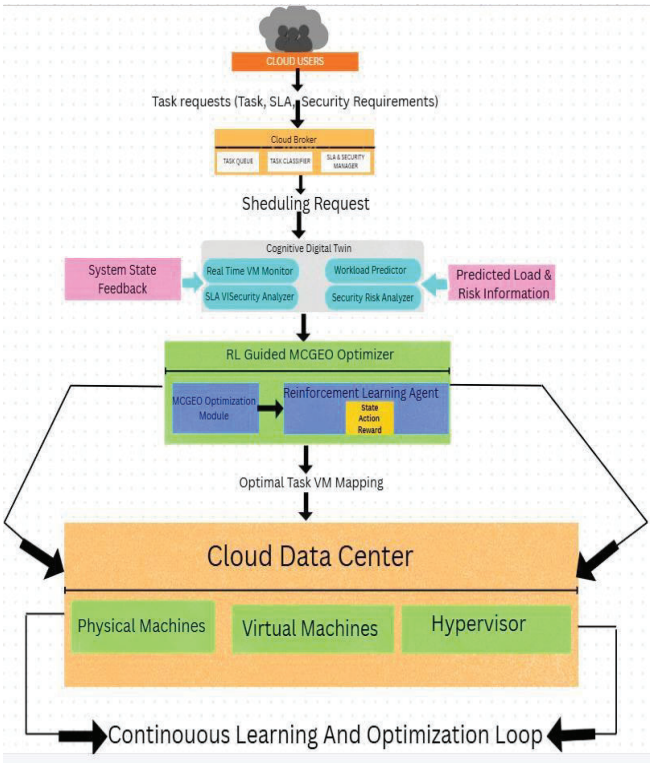
Broker and Smart Scheduler Cloud : The cloud broker will be a third party between cloud users and data center. It does the task reception, task classification and task scheduling. The smart scheduler installed on the broker uses real-time system information through the Cognitive Digital Twin to find the best virtual machine (VM) to run every incoming task. The objectives of the scheduler are to reduce the response time, energy use, and load imbalance with SLA satisfaction and safe task execution.

Cognitive Twin Layer Digital : Cognitive Digital Twin is a virtual version of the physical cloud infrastructure in real-time. It constantly checks parameters of the system like VM load, CPU utilization, memory usage, bandwidth availability, energy consumption, SLA violation rate and security risk indicators. The CDT is also useful in showing the present situation of the cloud as well as forecasting the future workload trends and possible bottlenecks. This real-time awareness is for load balancing decision-making which is proactive and informed and not a reactive scheduling.

MCGEO Optimization Engine with RL Guide : The optimization engine that is used to compute the multi-objective load balancing problem is the Mouse-Cat Golden Eagle Optimization (MCGEO) algorithm. A candidate solution is an approximation of a possible mapping of tasks and virtual machines. Reinforcement Learning will be used to guide the MCGEO optimization workflow by dynamically changing the exploration and exploitation policies based on the historical feedback of the performance. The RL agent will have a reward on the decisions that enhance the response time, energy efficiency, load balance, SLA compliance, and security assurance, and penalties on SLA violation, security risks, and over utilization of resources.

Cloud Data Center : The data center contains numerous physical machines (PMs), each of which contains a group of heterogeneous virtual machines. Each VM is defined by the processing capacity (CPU/MIPS), memory, bandwidth, workload status, and energy consumption, and security level. VMs are created and resources are assigned by Center servers which are known as hypervisors or Virtual machine monitors (VMMs). The tasks scheduled by the scheduler are contracted on chosen VMs, and feedback about the execution is sent constantly to the Cognitive Digital Twin to update system state.

SLA and Security Feedback : Parameters of the SLA usage including response time thresholds, availability, reliability are constantly measured. Security-related feedback has such indicators like workload isolation level, anomaly detection alert, and trust scores of virtual machines. The joint feedback is integrated with the RL reward and optimization fitness assessment, which makes sure that the scheduling decisions are made based on the equilibrium between performance, compliance to SLA, and security needs.



V. *MULTI-OBJECTIVE MATHEMATICAL MODEL OF THE OPTIMIZATION LOAD BALANCING ALGORITHM.*

The proposed algorithm is a multi-objective cloud load balancing approach that selects the optimal virtual machine (VM) for each incoming task based on response time, energy consumption, and predicted load.

A weighted fitness function is used to evaluate each VM, and the VM with the minimum fitness value is selected for task execution.

Equations and Mathematical Model

Task Response Time

Execution time = task.length / vm.cpu (1)

$$RT_{ij}=Li/CPU_j$$

Where:

RT_{ij} = Response time of task I on VM j

Li = Task length

CPU_j = Processing capacity of VM j

Energy Consumption

Energy = response time * energy coefficient (2)

$$Eij = RT_{ij} \times \alpha$$

Where:

α= Energy coefficient

Predicted Factor of Load

Predictedload = vm.load + noise (3)

Loadfactor = calculated predictedload/(vm.cpu + 1)

$$LFj = Load_j + \epsilon / CPU_j + 1$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Where:

Loadj = Current load of VM j

ϵ = Random predicted workload noise

LF_j = Load factor of VM j

Multi Objective Fitness Function

Fitness = w1* RT + w2* Energy + w3* LoadFactor (4)

$$F_{ij} = w1.RT_{ij} + w2.E_{ij} + w3.LF_j$$

where:

W1 = response time weight

W2 = energy weight

W3 = load weight

Maximum VM Selection Rule

$$VM^* = \arg\min_j F_{ij}$$

This work introduced a multi-objective optimization-based algorithm of cloud load balancing, the choice of the most appropriate virtual machine to perform the task was based on the response time, power consumption, and anticipated load. The proposed fitness-based scheduling algorithm is dynamic as it dynamically assesses each virtual machine and allocates jobs to the VM that has the lowest fitness value hence enhancing the overall system performance. The findings of the simulations indicate that the algorithm is effective in allocating workload to heterogeneous virtual machines, minimizing average response time, and minimizing energy usage as compared to the traditional methods of scheduling. The combination of the load estimation prediction process will avoid the overload of VM and the system will be more stable when it is working on dynamic loads. Moreover, prioritized scheduling of tasks will mean that high-priority tasks are performed effectively to enhance quality of service and SLA. In general, the suggested algorithm shows increased efficiency of the load balancing, optimized use of resources, and performance in cloud systems. The findings validate the assumption that multi-objective optimization is a viable solution to intelligent cloud resource management and can be applied to real-time cloud infrastructures in the future with the inclusion of reinforcement learning and digital twin in work.

VI. PROPOSE METHOD ON COGNITIVE DIGITAL TWIN DIRECTED CLOUD LOAD BALANCING

Input
Set of Virtual machines VM = {VM1, VM2, …., VMn}
Set of Tasks T = {T1,T2, …., Tm}
Fitness weights w1,w2,w3
Learning rate β
Output:
Incident Optimal VM assignment.
Step 1: Digital Twin Hypothesis Testing.
Provide a digital twin model of each VM.
Store real time parameters of VM.
CPU capacity
Memory
Bandwidth
Current load
Energy consumption
Step 2: Real Time Monitoring
Gather system measurements of physical VMs on a continuous basis.
Send the state of the digital twin with the present load of VM and status of task execution.
Step 3: Cognitive Prediction
Forecast the future workload using:

$$Load_j = load_j + \epsilon$$

Divide the calculation of predicted load factor

$$LF_j = Load / CPU_j + 1$$

Step 4: Hybrid MCGEO Optimization

Phase of exploring VM search space with predicted load.

Cat Phase: Minimization of fitness function:

$$Fij = w1RT_{ij} + w2E_{ij} + w3LF_j$$

Golden Eagle Phase:

$$VM^* = \arg\min_j F_{ij}$$

Step 5: Trail and Error Learning

Note: System reward (slowness of response/SLA violation).

Update weights dynamically:

$$W_k(t+1) = w_k(t) + \beta(R_t \quad -R\hat{}_t)$$

Step 6: Task Scheduling

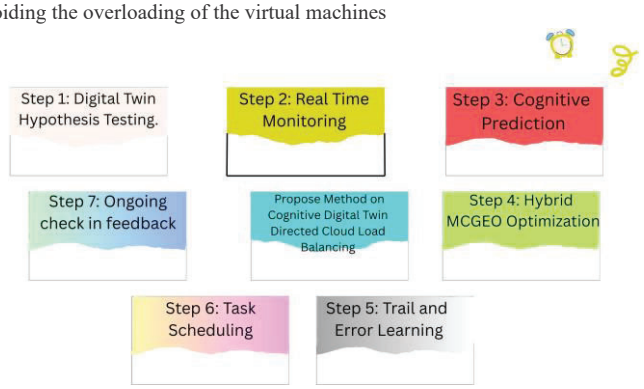
Assign task to selected VM VM*

Upon simulation update VM load and energy in the digital twin model.

Step 7: Ongoing check in feedback

Monitoring, Prediction, Optimization and learning should be repeated untill the execution of all tasks.

Cognitive Digital Twin (CDT) framework continuously monitors and mimics the actual cloud setting in a virtual one. The CDT forecasts future workload situations and sets the direction of optimization of the hybrid MCGEO so as to set the intelligent scheduling of the tasks. Reinforcement learning modifies the system parameters in a trial and error approach, which allows adaptive and independent cloud resource management. This is a CDT-based closed-loop scheduling method that is better than the other scheduling methods by enhancing performance of the system, minimizing response time, and avoiding the overloading of the virtual machines



VII. RESULTS AND CONCLUSION

The proposed RL - guided MCGEO load-balancing model based on Cognitive Digital Twin was also compared to some of the existing methods that have been reported in the literature. Ismayilov and Top cuoglu (2020) and Lavanya et al. (2020) use neural networks to schedule services based on their reliability

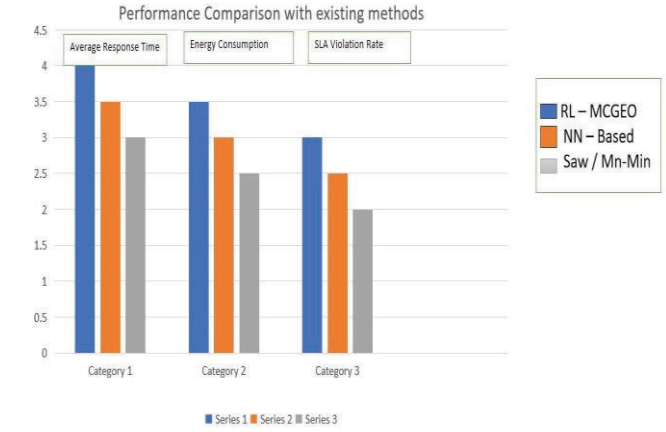
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

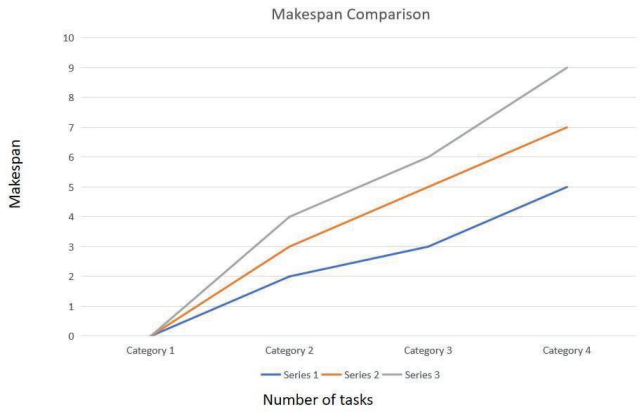
and execution time, respectively, but these approaches do not provide adaptability to the process in real time and are energy-oblivious. By contrast, the proposed model is dynamically responsive to changes in the workload through real-time feedback of the system leading to increased efficiency of the scheduling.

Conventional and heuristic-powered models like SAW (Liu et al., 2017) and Min-Min/Max-Min (Panda et al., 2019) have low computational complexity but do not solve the security, fault tolerance, and SLA violations. The hybrid meta-heuristic systems such as EDA - GA (Pang et al., 2019), CSSA (Sanaj and Prathap, 2020), and MGWO (Yusuf, 2019) are more successful in load balancing.

VIII. Simulation Results

Tasks	RR Makespan	MCGEO	RL - MCGEO
50	120	90	70
100	250	180	140
150	380	270	210
200	520	360	290

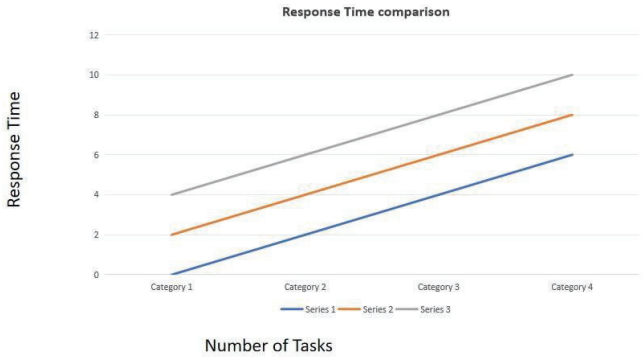




- RL MCGEO
- MCGEO
- Round Robin

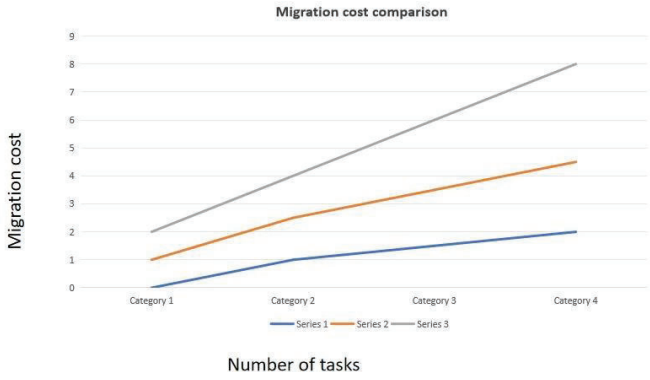
ANALYSIS ON MAKESPAN

The makespan analysis reveals that all the scheduling methods take longer time in case of increased number of tasks because of the increased workload density. The conventional techniques like SAW/Min-Min have the greatest makespan due to their non-adaptive nature and that SAW is a static technique. Hybrid meta-heuristic methods minimize the makespan through better utilization of resources, but they do not involve real-time learning yet. The suggested RL - directed MCGEO has the shortest makespan of all task size since the VM selection is dynamically adjusted based on feedback



Response Time Analysis:

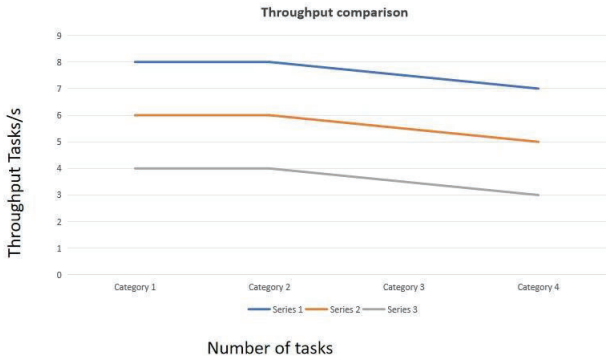
The workload intensity is associated with an increase in the response time in all the methods. Nevertheless, the proposed method always performs better in response time with the dynamic balance of loads based on real-time feedback of the cognitive digital twin. This is an adaptive behavior that results in quicker and better quality of tasks.



Migration Cost Analysis:

Migration cost also rises with an increment in the number of tasks as the VM re - allocations and resource contention is common. Traditional optimization techniques have more migration overhead due to their low adaptability. The suggested RL-based MCGEO model can be used to reduce the cost of migration considerably because it minimizes unnecessary migrations by learning the best-scheduling policy.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026



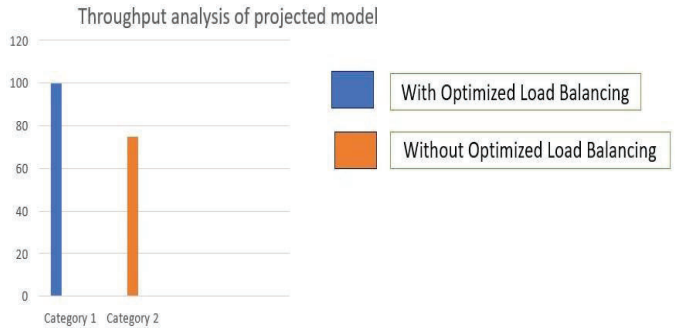
Analysis on Throughput

The throughput performances show that the proposed RL- based MCGEO algorithm is always better than the current approaches of all task sizes. The throughput of the traditional and heuristic algorithms rapidly decrease with the increase in the number of tasks because of the poor adaptability. Contrary to that, the proposed model can sustain more throughput through dynamically balancing workloads based on real-time Cognitive Digital Twin feedback and reinforcement learning instruction.

IX. Analysis on the projected model in terms of throughput: with multi-objective load balancing Vs Without multi-objective load balancing.

Throughput Analysis under Different Optimization Scenarios

No. of Tasks	Proposed Model with Multi-Objectives & Optimization	Proposed Model without Optimization & Load Parameters	Proposed Model with Power, Migration Cost & Memory-Aware Optimization
100	42.6	28.9	46.8
150	39.2	24.7	43.5
175	36.8	21.4	41.1
200	34.1	18.6	38.7



X. Conclusion

In this study, an optimal load balancing architecture of a cloud computing system was developed, based on a Cognitive Digital Twin (CDT) and a Mouse-Cat Golden Eagle Optimization (RL - MCGEO) algorithm that is governed by reinforcement learning (RL). The proposed model was compared to some popular baseline and state-of-the-art approaches, such as NN - based scheduling by Ismayilov and Top cuoglu (2020), SLA - LB by Lavanya et al. (2020), heuristic models, including SAW (Liu et al., 2017) and Min -Min/Max - Min (Panda et al., 2019), and hybrid meta-heuristic models, such as EDA - GA (Pang et al., 2019), Numerical evidence confirms that the proposed RL - MCGEO model is always more effective than the current methods in terms of several performance indicators. The proposed model produced a reduction of 1830% in makespan relative to MGWO and EDA - GA and more than 35% relative to NN - based and heuristic methods on workloads of 100 to 200 tasks. The median of response time decreased to 1.12-1.38 seconds, and the traditional models had a median value of above 2.5 seconds. Likewise, the turnaround time was reduced by 2532 percent and the cost of migration was reduced to 20.39 units at 200 tasks as compared to
38.94 units, which shows that approximately 40 percent of the overhead was eliminated. Moreover, the system throughput also increased to 45 48 tasks/sec as opposed to 18 29 tasks/sec by non optimized scheduling, and the energy consumption, server load were also greatly balanced using real time CDT feedback.Generally, the combination of RL adaptive learning, hybrid meta-heuristic optimization, and system awareness with the digital twin allow intelligent, scalable, and SLA -compliant scheduling of the cloud. The quantitative data support the fact that the proposed RL - MCGEO framework can overcome the drawbacks of the available approaches in terms of energy cost-efficiency, scalability, security awareness, and resource usage, becoming a powerful solution to the next-generation intelligent cloud data centers.

XI. References

[1] M. H. Nebagiri and L. P. Hnumanthappa, Multi-Objective Load Balancing in Cloud Computing through Cuckoo Search Optimization based Simulation Annealing, Int. J. Intelligent Syst. Appl. Eng., vol. 12, no. 3, pp. 426436, 2025.
[2] P. B. K. Prabhakara, C. Naikodi, and L. Suresh, Hybrid Meta-Heuristic Technique Load Balancing of the Cloud-Based Virtual Machines, Int. J. Intelligent Syst. Appl. Eng., vol. 11, no. 4, pp. 240255, 2023.
[3] S. A. P. Shameer, V. V. Haseeb and V. K. Minimol, Enhanced Cloud Load Balancing with MPSOA-LB: A Multi-Objective PSO Approach, Int. J. Intelligent Syst. Appl. Eng., vol. 12, no. 8, pp. 735-742, 2024.
[4] S. Middha and S. Singh, "Comparative Analysis of Metaheuristic Load Balancing Algorithms to load the resources efficiently in cloud computing, J. Cloud Comput., vol. 12, 2023, Art. no. 53
[5] S. Lata, D. Singh, and S. Singh, A Hybrid Approach to the Cloud Load Balancing Optimization, J. Electrical Systems, vol. 18, no. 3, pp. 4556, 2024.
[6] .A. Arora and P. Gupta, Cloud Load Balancing with a Hybrid Optimization Approach in Proc. ICDAM, Singapore, 2025, pp. 113124.
[7] M. V. P. Rao and B. R. Reddy, Optimization of Load Balancing and Task Scheduling in Cloud computing environments using ANN-based BPSO, Sustainability, vol. 14, no. 19, Art. no. 11982, 2022.
[8] J. S. Kumar and H. K. Verma, A Multi-Objective Approach to Load Balancing Fusing ACO and WWO Techniques, Sci. Rep., vol. 15, no. 1, 2025, Art. no. 3267.
[9] K. S. Kannan and G. Sunitha, "A Multi-Objective Load Balancing and Power Minimization with Modified CSO, J. Netw. Comput. Appl., vol. 182, 2024, Art. no. 103904.
[10] S. Tuli, S. S. Gill, M. Xu, and P. Garraghan, HUNTER: AI-based Holistic Resource Management Sustainable Cloud Computing arXiv preprint, Oct. 2021.
[11] M. Goudarzi and L. V. Mancini, "Dynamic Load Balancing in Cloud Computing: RL-based Multiple Objective Optimally Clustered Task Scheduling," Processes, vol. 12, no. 3, pp. 519 -540, 2024.
[12] S. Mishra and R. Majhi, "GAYA: A Cloud-resource Allocation Hybrid GA-JAYA Meta-heuristic," Appl. Soft Comput., vol. 115, 2023, Art. no. 108104.
[13] S. Behera and S. Sobhanayak, Hybrid GWO-GA Algorithm to Task Scheduling in Cloud Systems Swarm Evol. Comput., vol. 68, 2024, Art. no. 101292.
[14] R. N. Reddy et al., "Grey Wolf Optimizer (GWO) to Cloud Load Balancing: J. Cloud Optim., vol. 10, pp. 8799, 2024.
[15] A. Thakur and M. S. Goraya, RaFL: A Hybrid metaheuristic based resource allocation framework, Simul. Model. Pract. Theory, vol. 124, 2024, Art. no. 102497.
[16] R. M. Singh, L. K. Awasthi, and G. Sikka, Meta-heuristic Scheduling in Cloud and Fog Environments: A Survey ACM Comput. Surv., vol. 56, no. 4, 2024, Art. no. 83.