A Review of Machine Learning-driven Approaches for Workload Allocation in Cloud Environments

Nisha Devi
Dept. of Computer Science and Applications
Maharshi Dayanand University
Rohtak, India
nisha1074.rs.dcsa@mdurohtak.ac.in

Sandeep Dalal
Dept. of Computer Science and Applications
Maharshi Dayanand University
Rohtak, India
Sandeepdalal.80@gmail.com

Abstract **-** Cloud computing, an infrastructure that offers adaptable and extensible computing resources, encounters a significant obstacle in task scheduling and virtual machine (VM) migrations, directly impacting system reliability and overall client satisfaction. Due to NP-completeness, solving the task scheduling dilemma is complicated. Resource distribution refers to the action of assigning resources to various user tasks. This occurs at two levels: the virtual machine (VM) level and the host level. Scheduling at the virtual machine level entails allocating VMs to physical computers based on resource requirements and availability. At the host level, the scheduling process involves effectively overseeing resources, such as CPU, memory, and storage, across multiple physical machines to optimize performance and guarantee high availability. Virtual machine migration is essential for maximizing resource utilization and maintaining effective load balancing in cloud systems. Virtual machine management focuses on the strategic allocation of VMs to physical host machines and the seamless transfer of VMs to enhance energy efficiency and effectively resolve resource conflicts. Performing numerous live migrations in any sequence can lead to a decrease in service quality. This review paper examines the various issues related to task scheduling and load balancing. In addition, this study provides a comprehensive examination of various strategies to identify patterns, challenges, tools, benchmark datasets, and potential future scopes related to load balancing.

Keywords - Virtual Machine Migration, Machine learning, Load Balancing, Cloud computing, Resource allocation.

INTRODUCTION

Cloud computing has transformed the way software and IT facilities are used, making computing the fifth utility. Cloud computing has facilitated business growth, scientific progress, and diverse computation models. Three primary service models facilitate its widespread use: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) (Buyya et al., 2018). Cloud computing provides rapid access to a flexible pool of shared resources over the Internet. Workload distribution enables the dynamic allocation of tasks on virtual machines (VMs), hence improving user accessibility. An important challenge in this domain is attaining optimal load balancing, which is crucial for evenly distributing jobs across multiple processors. Load balancing is a technique that evenly distributes tasks among virtual machines (VMs) by utilizing a virtual machine manager (VMM). It aids in managing multiple kinds of workloads, including CPU, network, bandwidth, and memory requirements (Mishra & Majhi, 2020). Figure 1 illustrates the scheduling process at the VM level and host level in cloud data centres. We use various techniques such as First Come First Serve (FCFS), Shortest Job First (SJF), Min-min, Max-min, Round Robin (RR), and Particle Swarm Optimization (PSO) (Vergara et al., 2023) to effectively distribute workloads across cloud centres and avoid situations of overloaded, underloaded, or idle nodes. Energy optimization and workload mapping are crucial for ensuring a high level of service quality. Load balancers, which use virtualization via hypervisors on VMMs, play a critical role in spreading client requests among multiple servers while ensuring an effective and superior cloud computing experience. Load balancers, which use virtualization via hypervisors on

VMMs, play a critical role in spreading client requests among multiple servers while ensuring an effective and superior cloud computing experience.

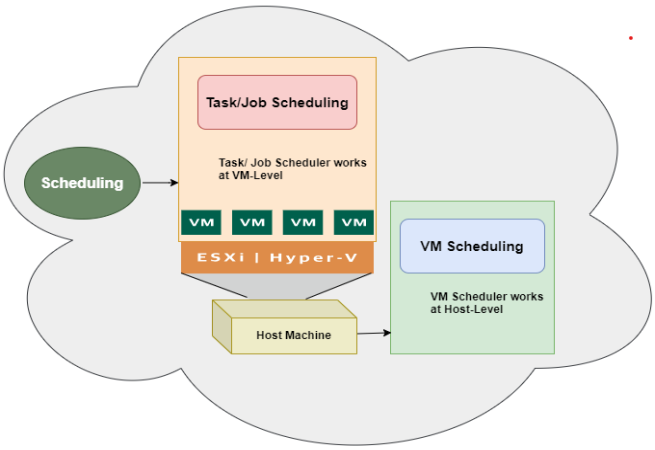

Figure 1. Scheduling at VM-level and Host-level in Cloud Datacenter.

The other load-balancing approaches are shown in Figure 2. Load balancing enables fast data transmission and reception, while also ensuring an equitable distribution of workload among all resources (Upreti, 2024). Efficient task scheduling assigns requests to appropriate virtual machines (VMs) based on the availability of resources, threshold value, processing capacity, and distance. Load balancing, on the other hand, redistributes workloads over the entire cloud architecture.

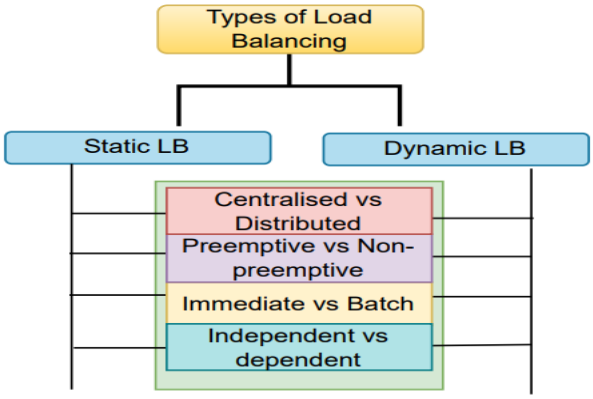

Figure 2 Types of Load balancing approaches.

Prevalent research issues of cloud platform

The cloud platform offers a ubiquitous computing solution that is easily accessible and flexible to all customers over the web. It helps in reducing startup and ongoing expenses by adopting a subscription-based service model. Cloud data centre's scalability feature optimizes IT resource utilization and integrates remote computing across the internet, enabling applications to be scaled. This method reduces the risk of resource allocation errors due to inaccurate workload estimation or power outages and addresses other management-related risks. Dynamic or Adaptive provisioning and virtualization technologies have propelled cloud services to the forefront of developing computing paradigms. Besides the above benefits, some challenges need to be discussed:

- Traffic management
- Reduction in power consumption
- Data Leakage detection

- Task scheduling
- Elasticity management
- Workload prediction
- Load distribution
- VM consolidation and migration management
- Fault tolerance predictions
- Automatic resource provisioning

MOTIVATION AND CONTRIBUTION

The attractive aspects of cloud computing, such as its low start-up costs and high scalability, attract enterprises, academics, research institutions, and all types of public or private organizations. Examples of daily life activities that utilize technology include social media, e-governance, online purchasing, and others. The resource requirements fluctuate across various time intervals, such as an hour, day, week, month, and year, depending on the nature of the workload and execution deadline specified by the user. Hence, it becomes challenging to accurately predict future resource requirements and determine the allocation of resources. Research has shown that a common issue in resource management is the inefficient use of resources, often caused by an imbalanced distribution of virtual machines. This can result in both wasted resources and a decrease in performance. Even during periods of inactivity, servers still consume a substantial amount of electricity. With this goal in mind, we began our review of the relevant literature and presented this review to identify the current research gaps, shortcomings, and strengths of previous studies undertaken in this area. We have selected a period spanning from 2018 to 2024 to choose the study material. In our review article, we have provided the following contributions:

- We have attempted to discuss challenges, comparative analysis, research gaps and contributions of ML-based techniques in load balancing and VM scheduling tabulated in Table 2.
- We have also highlighted the details of datasets and simulation tools used in the selected articles.
- A thorough analysis is included in Table 2, which includes potential areas for future research under examination.

Objectives of the review article

The objectives of this study are as follows:

1. To examine the existing ML techniques utilized for workload distribution.
2. To identify tools and benchmarked datasets used in the reviewed articles.
3. To investigate research gaps and future research prospects.

The following structure governs the subsequent sections of the paper: Section 1 contains the introductory information about the title. Section 2 presents the study methodology. Section 3 provides a technological survey. Section 4 evaluates the findings and concludes the paper.

RESEARCH METHODOLOGY

This survey encompasses both quantitative and qualitative research articles published between 2018 and 2024. The selection criteria for the reviewed paper include keywords "virtual AND machine AND migration AND using AND machine AND learning AND for AND load AND balancing" with citations of more than 8 from 2018 to 2024.

A. Inclusion exclusion criteria for selection of papers

We have employed an inclusion-exclusion criteria to include a research article in our literature review. The criteria are explained below in Table 1:

Table 1 Inclusion-exclusion criteria.

Inclusion criteria	Exclusion criteria
Those articles that cover the period from 2018 to 2024.	Not covering a period from 2018 to 2024.
A study is included if written in the English language.	Excluded if not written in the English language.
All papers having no. of citations more than zero are included.	All papers having zero citations are excluded.
All papers that contain the selected keywords and are relevant to the title are included.	A study is excluded if not contain selected keywords or is not relevant to the topic.
All the papers that are not commercial or available free of cost without subscription.	A study is excluded if a subscription is needed to access it.

LITERATURE REVIEW

The arena of cloud computing and data centre management has seen a growth in interest regarding the role of virtual machine (VM) migration in reducing maintenance-related downtime and disruptions, particularly concerning load balancing. (Buyya et al., 2018) conducted research that underscores the significance of virtual machine migration in reducing downtime. We achieve this by smoothly transferring virtual instances from servers undergoing maintenance to healthier nodes. This process of reallocating resources ensures that there are no service interruptions and allows for effective resource utilization throughout maintenance cycles. Furthermore, (Imran et al., 2022) conducted a study that demonstrated the importance of implementing live migration strategies to mitigate the adverse impact of downtime during critical updates or system maintenance. This, in turn, improves the availability and dependability of the system.

In their study, Zhang et al. (2018) explore the importance of VM migration in automating and optimizing maintenance operations using load-balancing strategies. Redistributing workloads among servers in an efficient manner before maintenance procedures helps minimize downtime and maintain system efficiency. Proactive migration options help reduce service disruptions while improving maintenance processes and resource management. Virtual machine migration can lead to resource optimization within a cloud data centre. Load balancing is a critical issue. (Li, 2019) emphasizes how VM migration enables resource consolidation by dynamically reallocating VMs to underutilized physical servers, thus promoting optimal utilization of computing resources. This consolidation reduces energy consumption and enhances the operational efficiency of the data centre infrastructure. Similarly, the study conducted by (Mishra & Majhi, 2020) underlines the significance of resource consolidation in the context of load balancing, demonstrating how VM migration aids in balancing the system load by efficiently redistributing workloads across servers. The resource consolidation approach contributes to cost savings and ensures a more sustainable and resilient cloud environment. The implementation of cloud computing in industrial IoT has made single-objective optimization insufficient for meeting industrial requirements. As a result, there has been a growing focus on multi-objective optimization. (Ni et al., 2021)presented a multi-objective task scheduling algorithm based on the Gaussian cloud Whale optimization strategy. The suggested model for whale-Gauss cloud scheduling consists of three layers: the user task layer, the task allocation layer, and the data centre layer. (Khan, 2024) investigated various machine learning algorithms to anticipate workloads in nonlinear settings. The features used in their study include both provisioned and utilized resources from a remote data centre in a cloud hosting environment. The features encompass performance metrics such as allocated CPU, allotted memory, CPU utilization, memory utilization, disk throughput, and network throughput. The authors also present an ensemble learning approach to cluster similar groups of VMs that can minimize energy consumption. For task scheduling, (Behera & Sobhanayak, 2024)suggested a hybrid technique. The authors propose a combined algorithm that integrates Genetic Algorithm (GA) and Grey Wolf optimization (GWO) for task scheduling in cloud computing systems. This method addresses early convergence issues, increases solution

exploitation, and speeds up the search procedure, resulting in faster optimal solutions and faster exploration within a reasonable timeframe. Table 2 depicts a comparative analysis of existing surveys presented by various authors.

TABLE 2 ANALYSIS OF TECHNOLOGY, CONTRIBUTION, RESEARCH GAP, AND DATASET USED.

Author name	Metrics	Technology used	Compared to	Research gap	Dataset & tool used	Future scope
(Paulraj et al., 2018) (Supervised Learning)	EC, NoM	Combined Forecast Load-Aware technique	ALM technique	Comparative analysis is done with only one technique.	Google Workload and Random Workload CloudSim tool	To reduce network bandwidth utilization and traffic during VM migration
(Li, et al., 2019) (Re-enforcement Learning)	SLAV, VM migrations	Markov decision processes-based adaptive overload threshold selection algorithm	MAD, IQR, Logistics Regression, ST	Due to limited overload thresholds, the algorithm seems insufficient and lacks reliability.	PlanetLab trace workload dataset CloudSim tool	To resolve the practical application issues for large overload threshold values.
(Sui et al., 2019) (Unsupervised)	EC, NoM, Cost, RU, Performance interference	LB based on GA, K-Means, Optimized Minmax and adaptive DE	Bayesian ridge, decision tree and SV Regression	Lack of reliability, the computational overhead is more, longer training time.	Eclipse 4.5.1, JDK1.8.0_111 and CloudSim 4.0	-
(Mashhadi Moghaddam et al., 2020) (Supervised)	EC, SLAV	Energy-aware VM consolidation algorithm	RF, AdaBoost, Linear regression	Suitable for small data centres. Lack of reliability due to simulated work.	The CoMon project Dataset, workload traces from PlanetLab	To use RNNs to increase the precision of VM CPU prediction for mapping
(Kaur et al., 2020) (DL)	Makespan and cost	Deep Learning	HDD-PLB (CWS)	Ignorance task priority comparative analysis.	The Genome workflow CloudSim tool	To work on Machine learning As a Service architecture.
(Ni et al., 2021) (Hybrid heuristic)	Cost, Load ET	GCWOAS2	ACO, PSO, WOA	Poor achievement of Cost efficiency. Lack of fault tolerance aspect.	MatlabR2016b	To decrease operating costs and examine heuristic algorithms.
(Motaki et al., 2021) (Supervised)	MAE, RMSE, ET	Ensemble-learning Strategy	KNNR, SVR, Ridge regression, etc.	Lack of scalability, heterogeneity and limited dataset application.	OLTPBench, Mplayer, SPECWeb workload	To employ a deep autoencoder for better feature extraction
(Dey et al., 2023) (Metaheuristic)	EC, NoM, ET, SLAV	Improved ACO Method	LRR_MU, IQR_MMT etc.	Increased Complexity	10 PlanetLab datasets CloudSim tool	Fewer SLA violations and VM migration
(Dubey et al., 2023) (Hybrid)	Cost, load, makespan	GA-WPC algorithm	GA, SA, PSOCognet, WPCO	Poor relevance to machine failure and real-time job allocation aspect.	Amazon EC2 model VM workload CloudSim tool	To explore work on real-time task allocation challenges.
(Behera & Sobhanayak, 2024) (Hybrid)	Makespan, cost, EC	GWO-GA	GWO, GA, and PSO	Poor consideration of fault tolerance, task precedence and load balance parameters.	NASA Ames iPSC/860 real workload, HPC2N (JDK 8.0, CloudSim)	Resolving issues of load balancing, task precedence and fault tolerance.
(Wang et al., 2024) (Deep learning)	Load balancing, cost and response time	DRLIS	NSGA2, NSGA3, DQN, Q-learning	Limited coverage on energy efficiency and fault tolerance aspect.	Nectar Cloud infrastructure	To employ DRL for improving edge performance.

Abbreviations for Metrics: EC- Energy consumption, NoM- No. of Migrations, SLAV-service level agreement violation, RU- Resource utilization, ET- execution time, RMSE- root mean square error, MAE- mean absolute error.

RESULTS & ANALYSIS

This literature review examined a diverse range of cutting-edge technologies and methodologies employed in data centre optimization and resource management. The studies showcased a variety of innovative approaches, including Deep reinforcement learning (DRL), Supervised techniques, and Metaheuristic algorithms. We have analyzed the selected articles to study research gaps and future scopes. The following observation is inferred from the above technological survey:

Analysis of benchmarked datasets

During the literature review, we examined several benchmarked datasets. Some of the most used datasets available online for academics include PlanetLab's workload, Google workload traces, and Genome traces. Table 2 presents comprehensive statistics on the use of datasets, technology utilizing the observed datasets and research gaps in these studies. This material will assist novice researchers in identifying workload datasets and comprehending the parameters they can achieve with their use.

Analysis of Simulation Tools and metrics.

We conducted an analysis of simulation tools to gain insight into how the cloud environment can be simulated and which tools have been utilized by researchers. During the period from 2018 to 2020, the preferred choice for cloud computing technology was the CloudSim tool. However, as cloud computing has progressed, the choice of tool has likewise changed. With the growing popularity of machine learning (ML) and deep learning (DL) technologies, Python has emerged as the top choice for simulation purposes. It has been observed that metrics like delay or latency, fault tolerance, execution time SLA violation and energy consumption need to be focused more in future research.

Utilization Analysis of ML-based techniques

We used the frequency analysis method to identify the strategy that was most used by authors from 2018 to 2024. We classified the methods into six categories: supervised, unsupervised, reinforcement learning, deep learning, heuristic, metaheuristics, and hybrid. Table 2 presents a summary of the technological review of different methodologies. The authors have applied supervised machine learning more followed by deep learning and hybrid techniques, as seen in Figure 3.

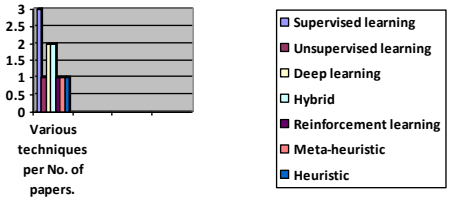

Figure 3Utilization of various algorithms per no. of papers.

FUTURE SCOPE OF THE STUDY

After conducting an extensive analysis of the chosen research publications, we have identified many persistent issues encountered by our researchers. Below, we explain some of the problems that can serve as potential areas for future research (Paulraj et al., 2018), (Dey et al., 2023), (Wang et al., 2024):

- Energy consumption or efficiency
- QoS performance metrics such as fault tolerance and applicability.
- Big Data Integration
- Dynamic task scheduling and VM Migrations
- SLA Breaches
- Workload predictions and decision making

REFERENCES

Behera, I., & Sobhanayak, S. (2024). Task scheduling optimization in heterogeneous cloud computing environments: A hybrid GA-GWO approach. Journal of Parallel and Distributed Computing, 183. https://doi.org/10.1016/j.jpdc.2023.104766

Buyya, R., Srirama, S., Casale, G., Calheiros, R., Simmhan, Y., Varghese, B., Gelenbe, E., Javadi, B., Vaquero, L., Netto, M., Toosi, A., Rodríguez, M., Llorente, I., Vimercati, S., Samarati, P., Milojicic, D., Varela, C., Bahsoon, R., Assuncao, M., & Shen, H. (2018). A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade. ACM Computing Surveys, 51. https://doi.org/10.1145/3241737

Dey, N., Gunasekhar, T., & Purnachand, K. (2023). ACO-Inspired Load Balancing Strategy for Cloud-Based Data Centre with Predictive Machine Learning Approach. Computers, Materials and Continua, 75(1), 513–529. https://doi.org/10.32604/cmc.2023.035139

Dubey, K., Sharma, S. C., Kumar, M., Kumar, P., & Nasr, A. A. (2023). A secured GA-WPC framework for scheduling the independent tasks in cloud environment. Journal of Ambient Intelligence and Humanized Computing, 14(9), 13003–13015. https://doi.org/10.1007/s12652-022-04207-y

Imran, M., Ibrahim, M., Salah Ud Din, M., Atif Ur Rehman, M., & Kim, B. (2022). Live virtual machine migration: A survey, research challenges, and future directions. Computers and Electrical Engineering, 103, 108297. https://doi.org/10.1016/j.compeleceng.2022.108297

Kaur, A., Kaur, B., Singh, P., Devgan, M., & Toor, H. (2020). Load Balancing Optimization Based on Deep Learning Approach in Cloud Environment. International Journal of Information Technology and Computer Science, 12, 8–18. https://doi.org/10.5815/ijitcs.2020.03.02

Khan, A. R. (2024). Dynamic Load Balancing in Cloud Computing: Optimized RL-Based Clustering with Multi-Objective Optimized Task Scheduling. Processes, 12(3). https://doi.org/10.3390/pr12030519

Li, Z. (2019). An adaptive overload threshold selection process using Markov decision processes of virtual machine in cloud data center. Cluster Computing, 22, 3821–3833. https://doi.org/10.1007/s10586-018-2408-4

Mashhadi Moghaddam, S., O'Sullivan, M., Walker, C., Fotuhi Piraghaj, S., & Unsworth, C. P. (2020). Embedding individualized machine learning prediction models for energy efficient VM consolidation within Cloud data centers. Future Generation Computer Systems, 106, 221–233. https://doi.org/10.1016/j.future.2020.01.008

Mishra, K., & Majhi, S. K. (2020). A state-of-art on cloud load balancing algorithms. International Journal of Computing and Digital Systems, 9(2), 201–220. https://doi.org/10.12785/IJCDS/090206

Motaki, S. E., Yahyaouy, A., & Gualous, H. (2021). A prediction-based model for virtual machine live migration monitoring in a cloud datacenter. Computing, 103(11), 2711–2735. https://doi.org/10.1007/s00607-021-00981-3

Ni, L., Sun, X., Li, X., & Zhang, J. (2021). GCWOAS2: Multiobjective Task Scheduling Strategy Based on Gaussian Cloud-Whale Optimization in Cloud Computing. Computational Intelligence and Neuroscience, 2021. https://doi.org/10.1155/2021/5546758

Paulraj, G. J. L., Francis, S. A. J., Peter, J. D., & Jebadurai, I. J. (2018). A combined forecast-based virtual machine migration in cloud data centers. Computers and Electrical Engineering, 69, 287–300. https://doi.org/10.1016/j.compeleceng.2018.01.012

Sui, X., Liu, D., Li, L., Wang, H., & Yang, H. (2019). Virtual machine scheduling strategy based on machine learning algorithms for load balancing. Eurasip Journal on Wireless Communications and Networking, 2019(1). https://doi.org/10.1186/s13638-019-1454-9

Upreti, K. (2024). A Performance Comparison of Load Balancing in Cloud Computing Techniques. https://doi.org/10.1007/978-981-99-9179-2_24

Vergara, J., Botero, J., & Fletscher, L. (2023). A Comprehensive Survey on Resource Allocation Strategies in Fog/Cloud Environments. Sensors, 23(9). https://doi.org/10.3390/s23094413

Wang, Z., Goudarzi, M., Gong, M., & Buyya, R. (2024). Deep Reinforcement Learning-based scheduling for optimizing system load and response time in edge and fog computing environments. Future Generation Computer Systems, 152, 55–69. https://doi.org/10.1016/j.future.2023.10.012