IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

# Optimizing Cloud Computing Performance with an Enhanced Machine Learning Algorithm for Load Balancing

**Mr. Thangadurai K[1], Ruchitha S[2], Swetha S[3], Preethi M[4], Sasika E[5]**
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2-5]
Mahendra Institute of Engineering and Technology, Namakkal, India

**Abstract**: *Load balancing (LB) is the process of distributing the workload fairly across the servers within the cloud environment or within distributed computing resources. Workload includes processor load, network traffic and storage burden. LB's main goal is to spread the computational burden across the cloud servers to ensure optimal utilization of the server resources. Cloud computing (CC) is a rapidly growing field of computing that provides computing resources as a product over the internet. This paper focuses on the issues within Cloud Load Balancing (LB) that have attracted research interest. The paper also mainly focused on uncovering machine learning models used in LB techniques. The most common algorithm in the reviewed project included SVM support vector machine. The criteria for LB technique was identified through performance metrics like throughput, response time, migration time, fault tolerance and power saving.*

**Keywords**: Deep deterministic policy gradient-PG, SVM, LSTM, Machine Learning, LASSO

## I. INTRODUCTION

Load balancing refers to the distribution of the computing workload to a group of servers. Load implies CPU load, network traffic burden and server storage capacity. The workload originates from the client requests and is sent to the servers . The concept of load balancing is applied in distributed system administrators to sub-divide, allocate and issue resources between different servers, networks and computers . Load balancing provides the ability to cope with growing hardware architecture and computing needs. The system performance ideally remains relatively independent of the increasing input variables . The importance of load balancing includes equitable distribution of computing resources by ensuring nodes are not overloaded or underloaded . In return these improves the speed and overall throughput of the whole distributed system. Other critical contributions of load balancing include cloud scalability by distributing the new additional workload effectively to the new instances of virtual servers and starts different services to address the growing requests size of information made and controlled will Give at 175 zettabytes by 2025 . This requires more offices and administrations to be set up by cloud sellers. The kinds of these offices and administrations cause more server farms and assets to be provisioned in the loud bringing about more measures of electrical ability to be burned-through. Assets of distributed computing frameworks are accessible for clients' administrations as virtual machines (VMs) that are conveyed and run in server farms. The server farms include numerous actual workers and every worker has a bunch of assets. In this manner, each cloud has countless assets that devour extensive measures of electrical force bringing about undeniable degrees of CO2 emanations. In N. Jones expected that data and coPGrespondence innovation exercises will utilize 20.9% of the worldwide interest for power by 2030Also, she expressed that every year server farms exhaust electrical force of 200 terawatt-hours and they add to the general CO2 emanations by around 0.3%. Likewise, it is normal by 2020 that the business of data and coPGrespondence will create about 12% of the complete carbon dioxide discharges . In regard of the above perceptions, how to acknowledge wanted green processing is as yet an incredible test and an essential woPGy in distributed computing conditions. It addresses a fundamental pattern for suppliers, clients and the climate with the targets of lessening operational expenses and outflow levels of CO2 .The

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

essential objective of green registering is to guarantee better degrees of burning-through electrical energy in processing frameworks like cloud and matrix figuring frameworks. In this vision, the fundamental commitment of this work is to give an energy-productive crossover (EEH) structure for improving the proficiency of devouring electrical energy in server farms. The proposed system relies upon both the booking and solidification approaches and it thinks about that the measure of force devoured by the server farm parts shifts with time. The structure has the accompanying.The rapid growth and adoption of cloud computing have brought about significant advancements in IT infrastructure and services. However, with the ever-increasing complexity and diversity of workloads, the challenge of effectively allocating resources within cloud environments has intensified. This study is driven by several compelling needs:Resource Efficiency: Cloud service providers manage vast aPGays of resources, including virtual machines, storage,and networking components. Efficiently allocating these resources is crucial to avoid wastage due to underutilization or overutilization. Advanced load balancing algorithms can intelligently distribute workloads across resources, ensuring optimal utilization and reducing operational costs.[3]

Performance Optimization: Inadequate resource allocation can lead to uneven distribution of workloads, resulting in longer response times and degraded system performance. Advanced load balancing algorithms can dynamically allocate resources based on real-time monitoring and predictive analytics, thus enhancing the overall responsiveness and throughput of cloud services.

Scalability and Flexibility: Cloud environments are designed to scale resources on-demand. Traditional load balancing techniques struggle to adapt to the dynamic nature of these environments, which often require rapid provisioning and deprovisioning of resources. Advanced algorithms can provide the needed flexibility to seamlessly allocate and deallocate resources, ensuring efficient scaling.

Heterogeneous Workloads: Cloud platforms host a diverse range of applications with varying resource requirements. Customizing resource allocation strategies for different types of workloads is challenging. Advanced load balancing algorithms can account for these differences, ensuring that each workload receives the necessary resources to perform optimally.

Energy Efficiency: Minimizing energy consumption is a critical concern for both economic and environmental reasons. By intelligently distributing workloads and consolidating tasks on fewer resources, advanced load balancing algorithms can contribute to reduced energy consumption, aligning with the principles of green computing.

User Satisfaction: Cloud services cater to a wide range of users with diverse needs. An effective load balancing strategy can contribute to improved user experiences by ensuring prompt response times and efficient service delivery.

## II. RELATED WORKS

Haitao Yuan et al., has proposed in this paper Infrastructure assets in dispersed cloud server farms (CDCs) are shared by heterogeneous applications in an elite and practical manner. Edge registering has arisen as another worldview to give admittance to figuring limits in end gadgets. However, it experiences such issues as burden unevenness, long planning time, and restricted force of its edge hubs. Consequently, smart undertaking planning for CDCs and edge hubs is basically critical to develop energy-productive cloud and edge processing frameworks. Current methodologies can't sagaciously limit the complete expense of CDCs, expand their benefit and improve nature of administration (QoS) of assignments on account of aperiodic appearance and heterogeneity of undertakings. This exposition proposes a class of energy and execution advanced planning calculations based on top of a few insightful enhancement calculations.

Rahul Yadav et al., has proposed in this paper we address the problems of massive amount of energy consumption and service level agreements (SLAs) violation in cloud environment. Although most of the existing work proposed solutions regarding energy consumption and SLA violation for cloud data centers (CDCs), while ignoring some important factor: (1) analysing the robustness of upper CPU utilization threshold which maximize utilization of resources; (2) CPU utilization prediction based VM selection from overloaded host which reduce performance degradation time and SLA violation. In this context, we proposed adaptive heuristic algorithms, namely least medial square regression for overloaded host detection and minimum utilization prediction for VM selection from overloaded hosts. These heuristic algorithms reducing CDC energy consumption with minimal SLA. Unlike the existing algorithms, the proposed VM selection algorithm consider the types of application running and it CPU utilization at

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

different time periods over the VMs. The proposed approaches are validated using the CloudSim simulator and through simulations for different days of a real workload trace of PlanetLab.

Mohammed jodausman et al., has proposed in this paper Cloud computing is a systematic delivery of computing resources as services to the consumers via the Internet. Infrastructure as a Service (IaaS) is the capability provided to the consumer by enabling smarter access to the processing, storage, networks, and other fundamental computing resources, where the consumer can deploy and run arbitrary software including operating systems and applications. The resources are sometimes available in the form of Virtual Machines (VMs). Cloud services are provided to the consumers based on the demand, and are billed accordingly. Usually, the VMs run on various datac enters, which comprise of several computing resources consuming lots of energy resulting in hazardous level of carbon emissions into the atmosphere. Several researchers have proposed various energy-efficient methods for reducing the energy consumption in data centres. One such solutions are the Nature-Inspired algorithms. Towards this end, this paper presents a comprehensive review of the state-of- the-art Nature-Inspired algorithms suggested for solving the energy issues in the Cloud data centers.

A taxonomy is followed focusing on three key dimension in the literature including virtualization, consolidation, and energy-awareness. A qualitative review of each techniques is carried out considering key goal, method, advantages, and limitations. The Nature-Inspired algorithms are compared based on their features to indicate their utilization of resources and their level of energy-efficiency. Finally, potential research directions are identifed in energy optimization in data centers. This review enable the researchers and professionals in Cloud computing data centers in understanding literature evolution towards to exploring better energy-efficient methods for Cloud computing data centers.

Rahul yadav et al., has proposed in this paper In distributed computing, high energy utilization and Service Level Agreements (SLAs) infringement are testing issues considering the interest of computational force is developing quickly, accordingly requiring enormous scope cloud server farms. Despite the fact that, there are many existing energy-mindful methodologies center around limiting energy utilization while overlooking the SLA infringement at the hour of a virtual machine (VM) choice from over-burden has. Additionally, they don't consider the current organization traffic cause execution debasement in this manner may not actually decrease SLA infringement under an assortment of responsibilities.

In this specific circumstance, this paper proposes three versatile models, in particular, inclination plunge based relapse (Gdr), expand relationship rate (MCP), and data transfer capacity mindful determination strategy (Bw), that can essentially limit energy utilization and SLA infringement. Energy-mindful strategies for over-burden have identification and VM choice from an over-burden have are important to improve the energy productivity and SLA infringement of a cloud server farm. Subsequent to moving all VM from underloaded have go to sit have, which change to energy saving mode is likewise advantageous. Gdr and MCP are versatile energy-mindful calculations dependent on the powerful relapse model, for over- burden have identification. A Bw dynamic VM choice approach select VM as per network traffic from the over-burden have under SLAs. Trial results on genuine responsibility follows show that proposed calculations lessen energy utilization while keeping up the necessary execution levels in a cloud server farm. Utilizing a CloudSim test system to approves proposed calculations. energy-mindful calculations proposed to improve the energy proficiency and limit the SLA infringement in cloud climate. The recreation results show that: (1) in regards to the energy productivity, the Gdr have over- burden discovery calculation improving energy utilization better than the MCP calculation; (2) during the VM determination from over-burden have thinking about CPU, memory, and organization traffic factor is more successful than a solitary factor like CPU. Besides, the calculations proposed in this paper are more compelling than the other energy- mindful calculations in any case the responsibility types. In future, We are intending to propose Thermal-mindful calculation for VM position.

## III. METHODOLOGY

Calculations can likewise be utilized to course information to server farms where power is more affordable.

Specialists from MIT, Carnegie Mellon University, and Akamai have tried an energy distribution calculation that effectively courses traffic to the area with the least expensive energy costs. The scientists project up to a 40 percent reserve funds on energy costs if their proposed calculation were to be conveyed. Nonetheless, this methodology doesn't

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

really lessen the measure of energy being utilized; it decreases just the expense to the organization utilizing it. In any case, a comparable methodology could be utilized to guide traffic to depend on energy that is created in an all the more harmless to the ecosystem or effective way. A comparative methodology has likewise been utilized to cut energy utilization by steering traffic away from server farms encountering warm climate; this permits PCs to be closed down to abstain from utilizing cooling.

Bigger worker habitats are some of the time found where energy and land are reasonable and promptly accessible. Nearby accessibility of environmentally friendly power, environment that permits outside air to be utilized for cooling, or finding them where the warmth they produce might be utilized for different purposes could be factors in green siting choices.Ways to deal with really decrease the energy utilization of organization gadgets by appropriate organization/gadget the board methods are reviewed in. The creators gathered the methodologies into 4 primary systems, in particular

• Adaptive Link Rate (ALR),
• IntePGace Proxying,
• Energy Aware Infrastructure, and
• Max Energy Aware Applications

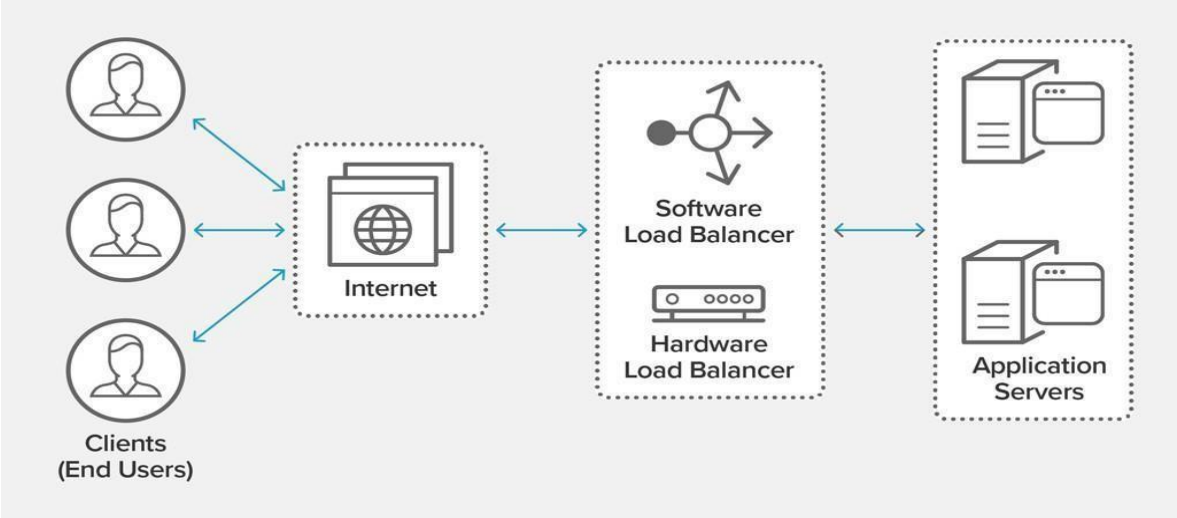


Fig: 1 Load Balancing Architecture

The set of rules enhances the VM choice section primarily based totally on actual time tracking facts collections and evaluation of bodily and digital resources. Our goal is to bolster VM scheduling .In order to include standards associated with the real VM usage stages, so VMs may belocated with the aid of using minimizing the penalization of universal overall performance stages.

The optimization schemes contain analytics to the already deployed VMs to include (a) maximization of usage stages and (b) minimization of the overall performance drops. A tracking engine that permits online aid utilization tracking facts series from VMs. The engine is able to accumulating gadge facts primarily based totally on c programming language and shops it to a web cloud carrier that makes it to be had for facts processing. Data is accrued each a tiny time c programming language (e.g. 1 second) and is saved in a transient neighborhood file.

The goal of this optimization schemes is to outline the burden of the PM in keeping with the useful resource utilization of the VMs. This will display facts approximately the already deployed VMs fame, like indicators that a workload is walking or not. To obtain this we offer optimization schemes. Here category of the VM fame approximately its contemporary useful resource utilization is assessed the usage of the KNN and NB proven in fig 4.1.Initially the digital gadget useful resource utilization dataset is accumulated and monitored after which the accumulated information is assessed the usage of the gadget gaining knowledge of strategies like K-NN and NB. model endeavors to make some

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

inference from noticed qualities. Given at least one sources of info an order model will attempt to anticipate the worth of at least one results. Results are marks that can be applied to a dataset.

There are two ways to deal with AI: managed and solo. In a regulated model, a preparation dataset is taken care of into the arrangement calculation. The k-closest neighbor's calculation (k-NN) is a non-parametric strategy utilized for characterization and relapse. In the two cases, the info comprises of the k nearest preparing models in the element space. The yield relies upon whether k-NN is utilized for arrangement or relapse:

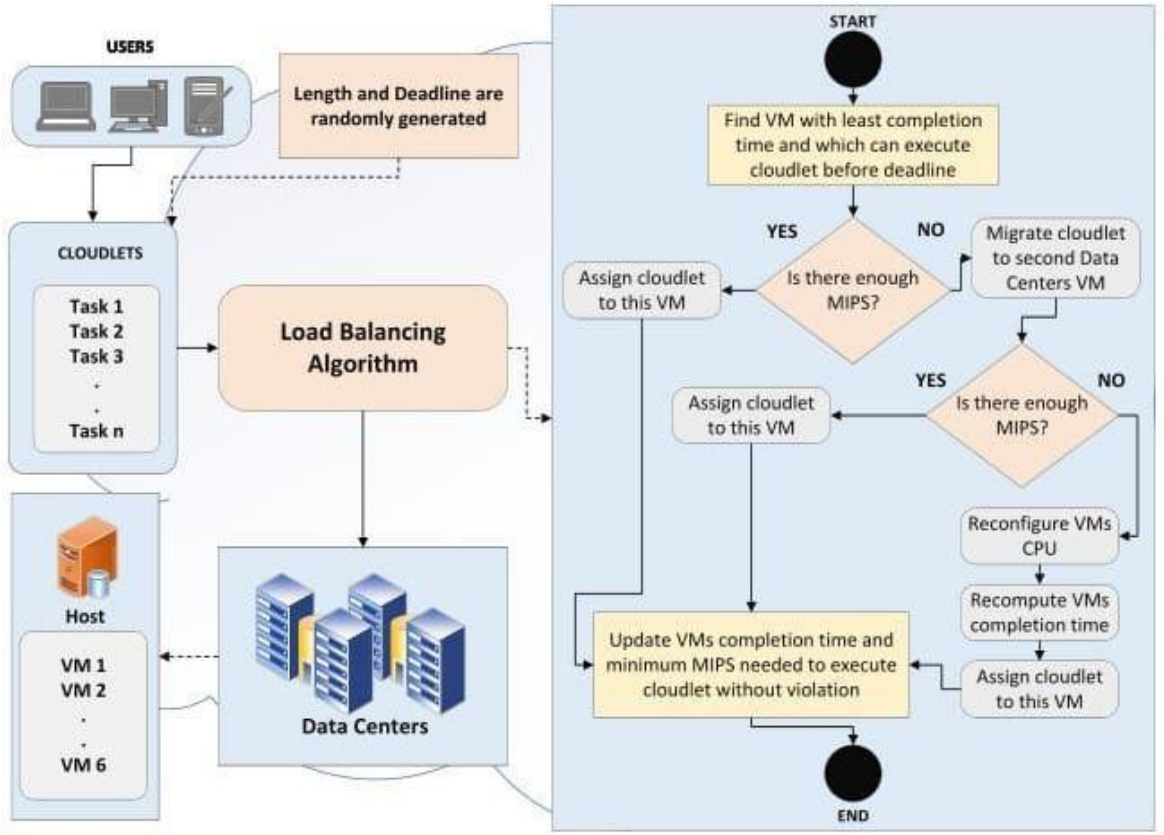


Fig 2: Block Diagram

**A Scheduling and Load Balancing design**

The various modules of the Scheduling and Load Balancing design are discussed below:

• Job Queue: All the client's Job requests are reaching the Request Queue in the order of its arrival. The priority or the length of the jobs has not been considered in the Queue. When a job has been taken-out for VM assignment by the Scheduling Controller using the algorithm of First-In-First-Out (FIFO), it will be moved out of the Request Queue.

• Dependency Task Queue: This queue will contain the tasks, which depends on the other tasks present in the VMs. Once all the child tasks of the tasks present in this queue got completed its execution the this parent task will be taken for the execution by assigning it to the VM.

• Task Manager: This module receives the Job and verifies the job whether it is a complete independent task or it contains multiple tasks. In case, if it contains multiple tasks, then it verifies the inter-dependency between the multiple tasks. Now, all the independent tasks will be directly assigned to the VMs. The dependent tasks will be notified to the scheduler so that parent tasks are scheduled after child tasks are executed.

• Scheduler: The scheduler selects the appropriate VMs based on the configured algorithms. This Scheduler collects the resources information through the Load Balancer from the Resource Monitor. It calculates the processing capacity of each of the VMs and then it applies the configured algorithm to find the appropriate VM for the given job.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, May 2025 Impact Factor: 7.67

• Load Balancer: Load Balancer (LB) calculates the ratio between the number of jobs running and the number of VMs. If the ratio is less than 1, then it communicates the scheduler to identify a VM for the job else it will calculate the load on each of the VM using the Job Execution List of the VMs. If the utilization is less than the 20% then the least utilized VM will be allotted else the scheduler will be communicated to identify the most suitable VM for the job. Once the appropriate VM has been identified the Job will be assigned to that VM.
• Resources: The configured datacenters, hosts and their VM and their Processing Elements form the set of resources available for computing. The resources are probed for idleness and for heavy load so that the job requests are effectively allocated to an appropriate resource

**B Algorithms**

The following scheduling and load balancing algorithms are implemented for functionalities such as scheduling and load balancing. WPG++ is the proposed algorithm. The PG and WPG are existing algorithms which are implemented for carrying out a comparative study.

1) PG: The Round Robin algorithm allocates task to the next VM in the queue iPGespective of the load on that VM. PG works well in most configurations, but could be more effective if the VMs are of roughly equal processing capacity, speed and memory.

2) WPG: Here the VMs are ordered in a circular queue based on weightage assigned to them. The incoming requests are then allocated to the ordered VMs in a circular fashion. WPG does not take into account of the load on the VMs or the length of the tasks allocated to the VMs. Hence, small tasks may be assigned to a VM with high processing capability and vice versa. WPG becomes equivalent to PG when the VM configuration, VMs processing capacity and speed are similar.

3) WPG++: WPG++ is an algorithm which is proposed as an improvement over the existing WPG algorithm. Here the WPG++ considers processing capabilities of the VMs, current load on the VMs and estimated job execution time. The algorithm works with coordination among three major modules of the system. The modules are,

a) Static Scheduler: This module does the functionality of initial job placements by considering the total number of VMs provisioned and the number of job requests.

b) Dynamic Scheduler: The dynamic scheduler takes care of run time job placements by taking into account of current load on the VMs, the nature of the task arrival and the instance at which the task requests are submitted.

c) Load Balancer: The load balancer checks for the current load and remaining time estimated for task completion and balances the load on the VMs by migrating a task from heavily loaded VM to a lightly loaded VM

**C. Mathematical Model**

The problem is to assign dynamically aPGiving dependent /
independent tasks to VMs and balance the load on the VMs to achieve reduced response latency and maximize resource utilization. Let us consider there are n number of VMs and m number of tasks. The set of all VMs are represented as VMjwhere j varies from 1 to n and the set of all task requests are represented as Ti where i varies from 1 to m.

Processing time of all tasks in a VMj can be defined as,

$$\varphi_j = \sum_{i=1}^{m} PT_{ij} \qquad (1)$$

where PTij is the processing time of i th task Ti on j th virtual machine VMj . The factor gives the minimum time the VM is required to be provisioned and is running to execute all the tasks assigned to it.

Processing capacity of a VM can be denoted as follows,

$$\mu_j = n_{(pe)} * mips_{(pe)} \qquad (2)$$

where μj is the processing capacity of the VMj , n(pe) is the number of processing elements in the VM and mips(pe) is the Million Instructions per second of a PE. The earliest start time and latest finish time of a task Ti are represented as tes(i) and tlf(i) respectively. tes(i) is the earliest time a task is able to start, which happens when all its child tasks complete execution as early as possible. It is represented as follows:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

$$t_{lf(i)} = \begin{cases} max\ (t_{lf(children(i))}),\ if\ dependent \\ t_{ls(a)} + t_{burst(a)},\ if\ independent \end{cases} \quad (4)$$

where tls(a) is the latest time a task can be scheduled to a VM without leading to starvation. The Schedule time of task Ti is tsch(i). It is the time at which the task has been scheduled for execution. This parameter can assume any value between tes(i) and tlf(i). Our problem is to identify the right VM and right time to schedule the task such that the VM utilization is high and load on the VMs are balanced well. Formally,

$$util_j = \frac{VM\ CPU\ usage\ (in\ MHz)}{n\ *\ core\ frequency\ (in\ MHz)}$$

## IV. PROPOSED SYSTEM

The goal is to suggest the idea of VM scheduling in step with aid tracking information extracted from beyond aid utilizations and examine the beyond VM usage tiers through the use of type approach along with K-NN and NB as a way to agenda VMs through optimizing overall performance. The proposed VM scheduling algorithm complements the VM choice segment primarily based totally on actual time tracking information collections and evaluation of bodily and digital resources. Our intention is to bolster VM scheduling as a way to include standards associated with the real VM usage tiers, so VMs may be positioned through minimizing the penalization of ordinary overall performance tiers.

The optimization schemes contain analytics on the already deployed VMs to include (a) maximization of usage tiers and (b) minimization of the overall performance drops. The truth that users, have underutilized VMs and do now no longer have the identical aid utilization pattern over the day. Finally, Cloud control processes, along with VM placement, affect already deployed systems (as an example this can contain throughput drop in a database cluster) as properly loaded VMs have a tendency to thieve CPU instances from neighbouring VMs. These constitute easy instances that show the want for a greater refined VM scheduling that would enhance overall performance.

### A. ADVANTAGES:

- Simple to implement
- Flexible to feature / distance choices
- Naturally handles multi-class cases
- Can make probabilistic predictions.
- Handles continuous and discrete data.
- Not sensitive to iPGelevant features

## V. RESULTS AND DISCUSSION

The attention is at the Cloud Sim this is an open supply software program to construct personal and public clouds. Cloud sim default configuration includes putting VMs via way of means of choosing the host with the maximum to be had reminiscence till the VMs range exceeds the limit. Such behaviour overloads effective PMs with inside the stack and leaves low RAM PMs under-utilized. Also the aid analytics primarily based totally on beyond aid utilization via way of means ofgrowing a machine studying version that analyzes PMs and VMs aid utilization on-the-fly. Virtual Machines (VMs) are scheduled to hosts consistent with their immediate aid utilization (e.g. to hosts with maximum to be had RAM) with outthinking about their standard and long-time period usage.

Also, in lots of cases, the scheduling and placement tactics are computational high-priced and have an effect on overall performance of deployed VMs. Thus the conventional VM placement set of rules does now no longer don't forget beyond VM aid usage tiers. To conquer this VM scheduling set of rules is implemented. The idea of VM scheduling consistent with aid tracking facts extracted from beyond aid utilizations (consisting of PMs and VMs) and the aid facts are categorized the usage of the optimization strategies K-NN and NB, therefore acting the scheduling. The set of rules evaluates beyond aid usage tiers and classifies consistent with the general aid utilization.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

| Sl No | Entity | Quantity |
|---|---|---|
| 1 | Data Centre | 24 |
| 2 | Hosts in DC | 100 |
| 3 | Processing Elements (PE) | 24/32 |
| 4 | PE Processing Capacity | 125/355/455 MIPS |
| 5 | Host RAM Capacity | 2/8 GB RAM |
| 6 | VM | 10 to 100 incremented by 10 |
| 7 | No of PE to VM | 1 |
| 8 | VMs PE Processing Capacity | 150/300/90/120/93/112/105/225 TB |
| 9 | VM RAM capacity | 1000 GB |
| 10 | VM Manager | Xen |

Table I. Entities and their Configurations

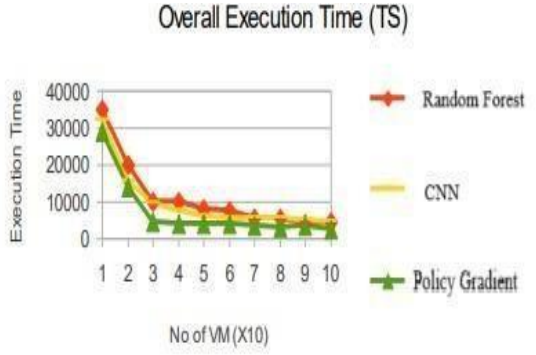


Fig 2. Overall Execution Time (TS)


Fig 3: Overall Execution Time (SS)


DOI: 10.48175/IJARSCT-26212

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

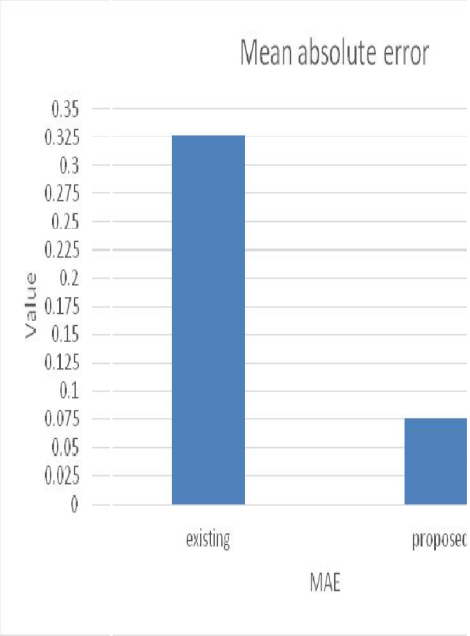


Fig 4: Mean Absolute Error

At the give up the listing of candidate hosts is populated and the assets are ranked accordingly. In detail, via way of means of the usage of this set of rules PMs are re-ranked consistent with the chosen optimization scheme and primarily based totally on their VM utilization. For instance, we use as facts set aid facts from 24 hours tracking and as training set a seven-day aid utilization tracking. The analytics are (a) consistent with usage tiers over the years via way of means of characterizing it as low, medium and heavy and (b) consistent with maintains facts (e.g. reminiscence percentage that will increase over the years). The set of rules plays a weighting technique for the chosen PMs consistent with distinctive features (e.g. CPU, RAM percentage).

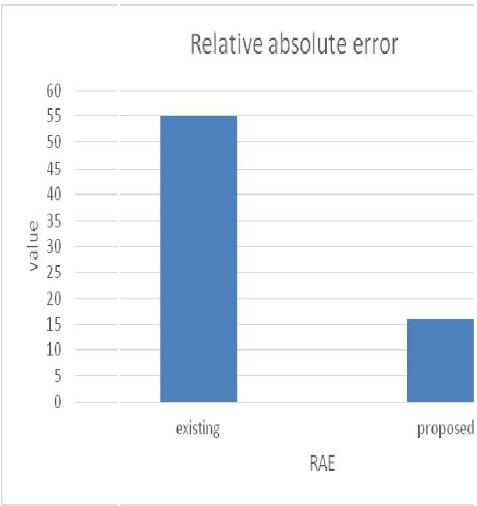


Fig 5: Relative Absolute Error

## IV. CONCLUSION

The focus of this project is cloud load balancing. Load balancing is one of the largest problems in a cloud environment. LB can adversely affect the quality of service and the SLAs hence making the cloud host lose clients. The work of the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

LB component is to share the work burden across the cloud resources to ensure maximum utilization of the resources and efficiency of the growing devices. This proposed machine learning models have finally harnessing of quantum computing power for cloud management An set of rules that permits VM placement in keeping with PM and VM utilization degrees and computational gaining knowledge of method primarily based totally at the idea of studying beyond VM aid utilization in keeping with ancient facts to optimize the PM choice segment changed into delivered. Also, a VM placement set of rules primarily based totally on actual time digital aid tracking changed into delivered where in device gaining knowledge of fashions is used to educate and analyze from preceding digital device sources utilization. Thus, a tracking engine is assumed with aid utilization facts. The depend of the bodily device receives decreased via way of means of four via way of means of the use of knn& PG classifier than Support Vector Machine (SVM) classifier. The undertaking completed via way of means of 28 bodily device while the use of SVM is decreased via way of means of 24 bodily device via way of means of the use of knn& PG classifier set of rules additionally the mistake quotes receives decreased via way of means of 0.025%.

## V. FUTURE WORK

The proposed model allows information processing primarily based totally on a time-frame window to outline the PMs or VMs actual behaviour. In case of VM placement method, end result highlights the major improvements. The destiny studies paintings can be executed with in addition experimentation applicable to numerous systems gaining knowledge of fashions like Deep deterministic policy gradient, selection trees to enhance the performance.

## REFERENCES

**[1].** S. Singh, E. E. Sham, and D. P. Vidyarthi, "Optimizing workload distribution in fog-cloud ecosystem: A Jaya based meta-heuristic for energyefficient applications," Applied Soft Computing, vol. 154, p. 111391, Mar. 2024. doi:10.1016/j.asoc.2024.111391

**[2].** A. Hazra, M. Adhikari, T. Amgoth, and S. N. Srirama, "Fog computing for energy-efficient data offloading of IOT applications in Industrial Sensor Networks," IEEE Sensors Journal, vol. 22, no. 9, pp. 8663–8671, May 2024. doi:10.1109/jsen.2022.3157863

**[3].** I. Z. Yakubu and M. Murali, "An efficient meta-heuristic resource allocation with load balancing in IOT-fog-cloud computing environment," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 3, pp. 2981–2992, Feb. 2023. doi:10.1007/s12652-023-04544-6.

**[4].** A. Najafizadeh, A. Salajegheh, A. M. Rahmani, and A. Sahafi, "Multiobjective task scheduling in cloud-fog computing using goal programming approach," Cluster Computing, vol. 25, no. 1, pp. 141–165, Aug. 2021. doi:10.1007/s10586-021-03371-8.

**[5].** S. Sefati, M. Mousavinasab, and R. Zareh Farkhady, "Load balancing in cloud computing environment using the grey wolf optimization algorithm based on the reliability: Performance evaluation," The Journal of Supercomputing, vol. 78, no. 1, pp. 18–42, May 2021. doi:10.1007/s11227-021- 03810-8.

**[6].** I. Zahraddeen Yakubu and M. Murali, "An efficient IOT-fog-cloud resource allocation framework based on two-stage approach," IEEE Access, vol. 12, pp. 75384–75395, 2024. doi:10.1109/access.2024.3405581.

**[7].** H. Jin, S. Lv, Z. Yang, and Y. Liu, "Eagle strategy using uniform mutation and modified whale optimization algorithm for QoS-aware cloud service composition," Applied Soft Computing, vol. 114, p. 108053, Jan. 2022. doi:10.1016/j.asoc.2021.108053.

**[8].** Z. Tong, X. Deng, H. Chen, and J. Mei, "DDMTS: A novel dynamic load balancing scheduling scheme under SLA constraints in cloud computing," Journal of Parallel and Distributed Computing, vol. 149, pp. 138–148, Mar. 2021. doi:10.1016/j.jpdc.2020.11.007.

**[9].** P. Yadav and D. P. Vidyarthi, "An efficient fuzzy-based task offloading in edge-fog-cloud architecture," Concurrency and Computation: Practice and Experience, vol. 35, no. 26, Jun. 2023. doi:10.1002/cpe.7843.

**[10].** Z. Ning, J. Huang, and X. Wang, "Vehicular fog computing: Enabling real-time Traffic Management for Smart Cities," IEEE Wireless Communications, vol. 26, no. 1, pp. 87–93, Feb. 2019. doi:10.1109/mwc.2019.1700441.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

**[11].** S. B. Akintoye and A. Bagula, "Improving quality-of-service in cloud/fog computing through Efficient Resource Allocation," Sensors, vol. 19, no. 6, p. 1267, Mar. 2019. doi:10.3390/s19061267.

**[12].** B. Negash, A. M. Rahmani, P. Liljeberg, and A. Jantsch, "Fog computing fundamentals in the internet-of-things," Fog Computing in the Internet of Things, pp. 3–13, May 2017. doi:10.1007/978-3-319-57639-8-1.

**[13].** M. Saad, "Fog computing and its role in the internet of things: Concept, security and privacy issues," International Journal of Computer Applications, vol. 180, no. 32, pp. 7–9, Apr. 2018. doi:10.5120/ijca2018916829.

**[14].** J. Li, J. Jin, D. Yuan, and H. Zhang, "Virtual fog: A virtualization enabled fog computing framework for internet of things," IEEE Internet of Things Journal, vol. 5, no. 1, pp. 121–131, Feb. 2018.doi:10.1109/jiot.2017.2774286.

**[15].** H. Atlam, R. Walters, and G. Wills, "Fog computing and the internet of things: A Review," Big Data and Cognitive Computing, vol. 2, no. 2, p. 10, Apr. 2018. doi:10.3390/bdcc2020010.

**[16].** P. Singh and R. Singh, "Energy-efficient delay-aware task offloading in fog-cloud computing system for IOT sensor applications," Journal of Network and Systems Management, vol. 30, no. 1, Oct. 2021. doi:10.1007/s10922-021-09622-8.

**[17].** R. Mahmud, S. N. Srirama, K. Ramamohanarao, and R. Buyya, "Profitaware application placement for integrated fog–cloud computing environments," Journal of Parallel and Distributed Computing, vol. 135, pp. 177–190, Jan. 2020. doi:10.1016/j.jpdc.2019.10.001.

**[18].** P. Rajesh, F. H. Shajin, and G. Kannayeram, "A novel intelligent technique for energy management in smart home using internet of things," Applied Soft Computing, vol. 128, p. 109442, Oct. 2022. doi:10.1016/j.asoc.2022.109442.

**[19].** E. E. Sham and D. P. Vidyarthi, "Cofa for QoS based secure communication using adaptive chaos dynamical system in FOG- integrated cloud," Digital Signal Processing, vol. 126, p. 103523, Jun. 2022. doi:10.1016/j.dsp.2022.103523.

**[20].** M. Taneja and A. Davy, "Resource Aware Placement of IOT application modules in fog-cloud computing paradigm," 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), May 2017. doi:10.23919/inm.2017.7987464.

**[21].** R. Mahmud, K. Ramamohanarao, and R. Buyya, "Latency-aware application module management for Fog Computing Environments," ACM Transactions on Internet Technology, vol. 19, no. 1, pp. 1–21, Nov. 2018. doi:10.1145/3186592.

**[22].** R. O. Aburukba, M. AliKarrar, T. Landolsi, and K. El-Fakih, "Scheduling internet of things requests to minimize latency in hybrid fog– cloud computing," Future Generation Computer Systems, vol. 111, pp. 539–551, Oct. 2020. doi:10.1016/j.future.2019.09.039.

**[23].** P. Hosseinioun, M. Kheirabadi, S. R. Kamel Tabbakh, and R. Ghaemi, "A new energy-aware tasks scheduling approach in fog computing using hybrid meta-heuristic algorithm," Journal of Parallel and Distributed Computing, vol. 143, pp. 88–96, Sep. 2020. doi:10.1016/j.jpdc.2020.04.008.

**[24].** M. Safari and R. Khorsand, "Energy-aware scheduling algorithm for timeconstrained workflow tasks in DVFS-enabled cloud environment," Simulation Modelling Practice and Theory, vol. 87, pp. 311–326, Sep. 2018. doi:10.1016/j.simpat.2018.07.006.