National College of Ireland

Energy-Efficient Virtual Machine Consolidation in Cloud Datacenters

MSc Research Project
Cloud Computing

Parth Ishvarbhai Moradiya
Student ID: 21199434

School of Computing
National College of Ireland

Supervisor: Prof. Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Parth Ishvarbhai Moradiya
Student ID:	21199434
Programme:	MSc in Cloud Computing
Year:	2024-25
Module:	MSc Research Project
Supervisor:	Prof. Rejwanul Haque
Submission Due Date:	29th January, 2025
Project Title:	Energy Efficient Virtual Machine Consolidation in Cloud Datacenters
Word Count:	7438
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Parth Ishvarbhai Moradiya
Date:	29th January, 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your reference and in case a project is lost or misplaced. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Co-ordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Energy Efficient Virtual Machine Consolidation in Cloud Datacenters

Parth Ishvarbhai Moradiya
21199434

Abstract

In this research, an innovative approach is presented for energy efficiency optimization in cloud datacenters via intelligent virtual machine (VM) consolidation. This work attempts to address the growing environmental and operational concerns regarding datacenter energy consumption through a multi-objective optimization framework that utilizes machine learning and predictive analytics. A methodology of combining workload prediction by LSTM networks with the classification of the server load using Random Forest algorithms to make informed consolidation decisions is also presented. It extends to incorporate energy minimization, and VM migration optimization objectives traditionally addressed, as well as CO2 emissions, workload prediction accuracy, and performance impact metric objectives. Evaluation utilizes CloudSim simulation environment extended with Python-based optimization algorithms. Adaptive learning mechanisms were shown to provide for significant improvements in energy efficiency with no sacrifice in service quality. Performance of the framework is extensively evaluated against traditional consolidation approaches using energy consumption and resource utilization as evaluation metrics, seeing substantial improvements. This research is of direct relevance for large-scale cloud providers who want to trade operational efficiency with environmental responsibility. It shows how advanced machine learning techniques and multi-objective optimization can improve datacenter operations while saving on environmental impact.

1 Introduction

As cloud computing services continue to grow quickly, cloud providers are concerned about the drastically increasing energy consumption in datacenters, and so the energy efficiency of datacenters is very important (Radovanović et al., 2022). While most of the existing VM management solutions try to optimize performance metrics, neglecting other life cycle factors like energy consumption, resources utilization, environment impact (Talebian et al., 2020). This thesis presents a general framework that unifies advanced machine learning and multi-objective optimization to obtain efficient VM consolidation that considers both operational and environmental impacts (Reddy and Reddy, 2023). The effective operation of consolidation requires integration of workload prediction and server load classification and opportunities for proactive resource management. This work presents a robust solution for the management of energy efficient VMs, using predictive analytics and adaptive learning mechanisms.

1.1 Research Background

The rapid evolution of cloud computing has turned the traditional IT infrastructure landscape on its head with an unprecedentedly scalable and flexible infrastructure but with many challenges of massive energy consumption. Recently there have been data around the growing environmental impact of the datacenter operations and the importance of energy efficiency (Patel et al., 2024).

Current VM consolidation solutions need more sophisticated approaches that incorporate predictive analytics and machine learning (Rozehkhani et al., 2024) instead of basic resource utilization metrics as were used in traditional approaches. With the advent of sophisticated tools to predict workload (Khan et al., 2022), it is possible to determine new ways to allocate

different resources. It has been found that server load classification and adaptive learning are essential for consolidation decisions (Ahn et al., 2023). Moreover, optimization of a system by integrating environmental metrics into system's optimization objectives has gained considerable importance since organizations increasingly pay attention to sustainable computing practices (Gholipour et al., 2020).

1.2 Problem Statement

However, current cloud datacenter VM consolidation techniques come with significant challenges in balancing a few competing objectives including energy efficiency, as well as performance and environmental impact. As workloads are inherently dynamic, the failure of traditional approaches to handle the dynamic nature of workloads and thus the resource utilization patterns are often ignored (Bannerjee et al., 2024). According to (Saxena et al. (2024), inaccurate workload prediction and server load classification lack appropriate mechanisms, therefore; suboptimal consolidation decisions were made and this led to superfluous energy consumption and possible performance degradation. Lastly, previous solutions do not consider the environmental consequences of consolidation decisions that only focus on operational metrics, while failing to consider broader sustainability goals (Radovanović et al., 2022).

1.3 Motivation

As cloud datacenters start to have a growing environmental impact as well as increasing operational costs, more sophisticated VM consolidation approaches are required (Dias et al., 2021). Assessment of current consolidation strategies has recently progressed with recent advances in machine learning and predictive analytics, offering opportunities for the development of more intelligent and efficient consolidation strategies (Moghaddam et al., 2020). Hence, to improve VM placed, multi-objective optimization approaches (Reddy and Reddy, 2023) have been developed to achieve significant energy savings while keeping service quality. Additionally, there is a strong argument being made for consolidating into carbon emitting factories with IT operations rather than expanding into other territories.

1.4 Research Question

What is the best way to integrate a machine learning based workload prediction and server load classification into a multi-objective optimization framework that can provide energy-efficient VM consolidation and minimize CO2 emissions, resource utilization, performance requirements?

1.5 Research Objective

A comprehensive VM consolidation framework that utilizes machine learning for workload prediction and server load classification to optimize multiple objectives including energy, migration efficiency, CO2 emissions and performance impact while ensuring scalability and practical applicability in real cloud environments is developed and is implemented.

1.6 Research Contributions

This research makes five significant contributions to the field of cloud computing and energy-efficient resource management:

- A novel multi objective optimization framework that concurrently incorporates energy efficiency, performance, and environmental impact into the VM consolidation decisions.

- An advanced workload prediction system using LSTM networks to enhance the accuracy of forecasting available resources utilization for the completion of consolidation planning.
- The design of a sophisticated dedicated server load classification mechanism utilizing Random Forest algorithms to make better and more effective VM placement decisions.
- The integration of environmental impact metrics like CO2 emissions into the consolidation decision making process in order to gain a more comprehensive approach to sustainable datacenter operations.
- A practical framework based on design and implementation of the CloudSim integration and Python based tooling to bridge theoretical optimization approaches and real-world cloud environments.

1.7 Thesis Structure

In Chapter 1, we discuss the research context, objectives, and motivations to introduce the problem of carbon aware VM consolidation. Section 2 covers energy efficiency, VM consolidation, and carbon aware computing in the context of computing clouds. The research methodology, as well as the multi-objective optimization framework and machine learning approaches are presented in Chapter 3. Detailed design specifications of the system architecture and interactions of components are provided in Chapter 4. In Chapter 5, the implementation in terms of CloudSim integration and interfaces to Python based optimization algorithms is described. The experimental results and performance analysis of the system are discussed in chapter 6, where the system can effectively reduce energy consumption and carbon emission. Chapter 7 concludes the thesis with a summary of contributions followed by directions for future research.

2 Related Work

2.1 Carbon Aware Computing

Radovanovic et al. (2023) introduced Google's Carbon Intelligible Computing System (CICS), aimed at reducing carbon emissions by scheduling flexible workloads like data compaction and video processing. Using Virtual Capacity Curves (VCCs), CICS optimized hourly resource allocation while maintaining daily capacity, achieving a 1–2% reduction in peak carbon-intensity power consumption across 20+ global data centres consuming 15.5 terawatt-hours. Results varied by location due to grid carbon intensity and usage patterns.

Park et al. (2024) explored shifting workloads to datacenters with lower emissions using a fault-tolerant control system and Model Predictive Control (MPC) for GPU frequency optimization. Their CAFTM system improved power consumption prediction and reduced the carbon footprint when applied to real-world deep learning models. Similarly, Moore et al. (2024) proposed the Sustainable FaaS Cloud Management (SFCM) framework to balance service level objectives, carbon emissions, and water use. Evaluated using Microsoft Azure traces, it reduced SLA violations by 45%, emissions by 25%, and water use by 26%.

Gupta and Gupta, (2024) proposed an Apache Flink-based architecture using the Carbon Aware SDK for carbon emission reduction recommendations. Their approach leveraged relocation, time, and demand shifting techniques, achieving a 20% carbon footprint reduction for an enterprise with 1,000 VMs and 200 TB of storage. They noted limitations in calculating emissions due to infrastructure location and hardware variability.

An energy and carbon-aware initial VM placement algorithm for geographically distributed cloud datacenters is proposed by (Khodayarseresht et al., 2023). The algorithm accounts for IT

and non-IT power usage during VM placement. The carbon footprint calculation uses location specific carbon footprint rates to calculate the emissions, emissions = powerConsumed × carbonFootprintRate, with power consumed including both server power and datacenter overhead power based on power usage effectiveness (PUE). The proposed approach has achieved a 17% reduction in energy consumption and 6% in reduction of carbon emissions as compared to baseline approaches.

An energy and carbon footprint optimization framework for distributed cloud datacenters enabled by renewable energy sources is presented by (Zhao and Zhou, 2022). The carbon emissions are calculated considering both brown energy consumption and renewable energy utilization as presented: The carbon footprint (CF) of a technology, at time 't', for each technology 'k' assigned to region 'r' is given by: $CF = (Pk(t) - Rk(t)) \times CFRk$ where Pk(t) is total power consumption, Rk(t) is renewable energy generation and CFRk is the location-specific carbon footprint rate. In their energy and carbon footprint aware with predictive renewable energy source (EFP) algorithm, they achieve a whittled down renewable energy utilization of 73.11% while keeping SLA violations at 0.2%.

2.2 Cloud Resource Management Approaches

An improved differential evolution (IDE) algorithm for cloud data center VM allocation was proposed by Zhang et al. (2020). The algorithm minimizes cloud service provider costs and user task make-span using enhanced mutation and crossover operations, which improve convergence speed and avoid local optima. Compared to round-robin, Min-min, and standard differential evolution methods, IDE achieves lower make-span and better resource utilization, with balanced VM load ratios. Similarly, Shi et al. (2020) developed the BMin algorithm for task scheduling optimization in cloud environments. BMin balances workloads and minimizes task execution variance, calculating execution times without assuming completion time or workload distribution. Experiments with CloudSim showed BMin outperformed Min-min in throughput, turnaround time, and load balancing.

Hajisami et al. (2020) introduced the "Elastic-Net" framework to optimize power consumption and resource utilization in Cloud Radio Access Networks (C-RAN). Elastic-Net dynamically adapts parameters like remote radio head (RRH) scheduling, VM migration, and active RRH density based on traffic fluctuations. Using simulation and testbed experiments, Elastic-Net improved energy efficiency, reducing average power consumption by 48.59% during off-peak and 7.39% during peak hours.

Sharma and Bhardwaj (2022) optimized VM allocation using a modified emperor penguin optimization (I-EPO) algorithm. The approach evaluates metrics like data center distance, BBQ, and transportation costs, employing graph-based representations of user tasks and data centers. Experiments on the XEN hypervisor showed improvements in latency, response time, and load fairness across VM scales.

Saxena et al. (2022) presented an Online VM Prediction-based Multi-Objective Load Balancing (OPMLB) framework that employs neural networks and adaptive evolutionary algorithms to forecast resource usage and minimize SLA violations. The framework achieves power savings of up to 85.3%, 99.9% accurate overload prediction, and reduced network traffic and VM migration costs.

Finally, (Gong et al., 2024) explored dynamic resource allocation and VM migration optimization using machine learning. They employed LSTM networks for demand prediction, Deep Q Networks (DQN) for migration decisions, and multilayer perceptron for energy optimization. Their deep reinforcement learning framework demonstrated improved resource utilization, lower power costs, and enhanced service reliability.

2.3 VM Consolidation Techniques

Hossain et al. (2020) proposed Active & Idle Virtual Machine Migration (AIVMM) to enhance energy efficiency in cloud data centers by isolating idle VMs from active ones, thereby balancing power among active machines. They integrated the Order Exchange Migration Ant Colony System (OEMACS) algorithm into AIVMM to optimize power consumption by monitoring CPU, RAM, and VM states. While the OEMACS+AIVMM approach improved migration efficiency, it faced high time complexity and difficulties with VMs changing states during migration.

Seddiki et al. (2021) developed a sustainability-focused framework for inter-datacenter VM migrations, incorporating renewable energy considerations into CloudSim simulations. They introduced new CloudSim entities and a meta-scheduler to align workload distribution with renewable energy availability. Their methodology, tested across multiple scenarios, achieved 68.39% renewable energy utilization without compromising Service Level Agreements (SLAs). The framework addressed VM placement to minimize physical machine (PM) power and network bandwidth consumption.

Xing et al. (2022) introduced the energy- and traffic-aware ant colony optimization (ETA-ACO) algorithm with three key schemes: direct information exchange, energy- and bandwidth-aware PM selection, and traffic-based VM ordering. These schemes prioritized power and bandwidth efficiency while improving solution quality in VM placement. Evaluated on 36 test instances, ETA-ACO outperformed several heuristic and metaheuristic models.

Zeng et al. (2022) presented the Adaptive Deep Reinforcement Learning-based Virtual Machine Consolidation (ADVMC) framework. It employed an Influence Coefficient-based VM selection algorithm (ICVMS) and a Prediction-Aware DRL placement algorithm (PADRL) to enhance efficiency. Tested with Google Cluster Trace data, ADVMC reduced energy consumption by 125.24% and SLA violations by 138.64% compared to Modified First Fit Decreasing (MFFD). The ICVMS alone reduced energy consumption by 112.3% and improved SLA violations by 38.33%, proving its scalability for large cloud environments.

Alur et al. (2023) focused on dynamic VM consolidation for energy efficiency, using a three-module approach: resource monitoring, random environment testing, and energy-efficient consolidation. Tested on a Multi Node OpenStack Yoga testbed, it demonstrated significant cost savings (27% to 35.8%) while maintaining system performance. This method underlined the potential of dynamic VM consolidation in reducing energy consumption in cloud environments.

2.4 Summary

Several key research gaps identified through literature review are addressed by this work. Most existing VM consolidation approaches are mostly concerned with single objectives, such as energy efficiency or performance, and fail to fully consider holistic consideration of environmental impacts and CO2 emissions. While there are studies on such carbon-aware computing, e.g., Google's CICS, they are featureless. In this research, workload prediction with LSTM networks and server load classification with Random Forest algorithms are combined uniquely together, circumventing the disadvantage of resource usage metrics-based solutions in literature.

Moreover, while works such as SFCM framework present progress in additional managing environmental impact, they are not fully integrated with the predictive analytics along with multi objective optimization. In innovating to resolve this gap, this framework simultaneously optimizes energy efficiency, performance, and environmental metrics. Additionally, existing approaches tend to have difficulty realizing their contributed algorithms in real world datacenter operations and in VM consolidation, whereas this thesis enables integrations in CloudSim and provides Python based tooling to make it more practical.

3 Research Methodology

The approach to optimal energy efficiency in cloud datacenters is innovative: intelligent VM consolidation. Building a multi objective optimization framework, the work addresses the rising environmental as well as operational issues arising due to the datacenter energy consumption. This thesis presents a framework for making informed server consolidation decisions that combines operations and environmental impact considerations with workload prediction and classification of server load using ensemble techniques organized within an applied machine learning framework.

3.1 Approach

Breaking away from standard VM consolidation techniques, the research instead uses a multi objective optimization approach. The framework incorporates seven distinct objectives: Minimizes energy consumption, reduces CO2 emissions, minimizes SLA violations, optimizes the use of resources, improves the accuracy of predicting improvements in workload, classifies the accuracy of server loads, and improves placement decision efficiency. These objectives lead to the formation of a comprehensive optimization problem which mathematically represents the interactions of these various operational and environmental factors. Then it applies advanced optimization algorithms (NSGA-II and MOEA/D) to find Pareto optimal solutions, so that datacenter operators can make informed business decisions, given their own criteria and constraints.

The research methodology consists of two major machine learning components. It uses Long Short-Term Memory (LSTM) networks for predicting workload to provide the system with the capability to predict future resource requirements and take proactive consolidation decisions. Random Forest algorithms are used for server load classification to categorize server states accurately, for VM placement decisions.

A carbon aware component that considers time varying emissions factors and energy consumption patterns is used for environmental impact consideration. This component leverages traditional operational and carbon footprint metrics for alternative consolidation strategies to inform the framework within the context of making environmentally responsible decisions.

The carbon emissions calculation is formulated as:

$$emissions = \Sigma(energyconsumption_t \times emissionfactor_t)$$

Where $\text{energyconsumption}_t$ points to the energy consumed in time period t and emissionfactor_t is the CO2 emission factor per time period. This calculation takes both the direct physical machine energy consumption and grid emission factors which can vary overtime given the resulting time varying energy source mix.

3.2 Implementation

A carbon-aware VM consolidation system is implemented using CloudSim as its primary simulation environment while optimization algorithms are implemented using Python. Five fundamental classes comprise the system architecture, including the CloudSimInterface class which makes feasible the seamless interaction between the Python and Java environment. Real workload data is based on the Alibaba cluster trace dataset for validation under actual datacenter operating conditions as well as workload patterns. The same includes WorkloadPredictor class, ServerLoadClassifier class, VMPlacementOptimizer class and VMConsolidationSystem class. Finally, as this is a simulation environment, the simulation

time is discrete, therefore the time control is governed by providing these additional abstractions which, apart from integrating the system components through a modular architecture for scalability and maintainability, performs essential methods for VM provisioning, and CloudSim Interface.

3.3 Evaluation

The evaluation strategy first determines the effectiveness of a proposed framework via multiple evaluation metric including energy consumption, resource utilization efficiency, CO2 emissions, and workload prediction accuracy. Extensive simulations in the CloudSim environment are used to compare the performance of the framework in favour of the traditional consolidation approaches. The framework is validated under multiple workload scenarios using the Alibaba trace dataset, and it is tested under the realistic operating conditions to demonstrate its robustness and reliability. It evaluates the performance of individual components and shows the overall system's capability in accomplishing optimization objectives.

3.4 Research Workflow

In Figure 1, we employ a systematic research workflow to perform carbon-VM consolidation in cloud datacenters. The Alibaba Cluster Trace Dataset is used as a first step in the process, which begins with real-world workload patterns and resource usage traces found in the dataset to enable realistic simulation scenarios. This input data feeds into two parallel prediction components: LSTM based Workload Predictor and a Random Forest Load Classifier. Server loads are classified by multiple resource metrics using Random Forest algorithm to provide signalling levels of load condition for informed VM placement, and the LSTM network processes historical workload patterns to predict future resource demands.

The two components make predictions, which flow into the multi-objective optimization module that formulates the consolidation problem as a multi-objective approach like minimize energy consumption, minimize CO2 emissions, and maximize utilization of resources.

These objectives are then processed in the NSGA-II algorithm to get Pareto optimal solutions that represents trade-off between different optimization goals.

The theoretical optimization is deployed through CloudSim Interface to bridge the theoretical optimization with the practical simulated environments. The VM Consolidation System executes the placement decisions and migrations on the basis of the optimizer’s recommendations. The research evaluation is done in providing comprehensive insights into the effectiveness of carbon-aware consolidation for energy efficiency, resource utilization, and CO2 Emissions.

4 Design Specifications

The carbon-aware VM consolidation system design specifications comprise a complex composition of machine learning blocks, optimization algorithms and cloud infrastructure management. The modular design system architecture is scalable, maintainable and communicative with other components in such a way that it concentrates on curtailing the carbon emission and promoting energy efficiency.

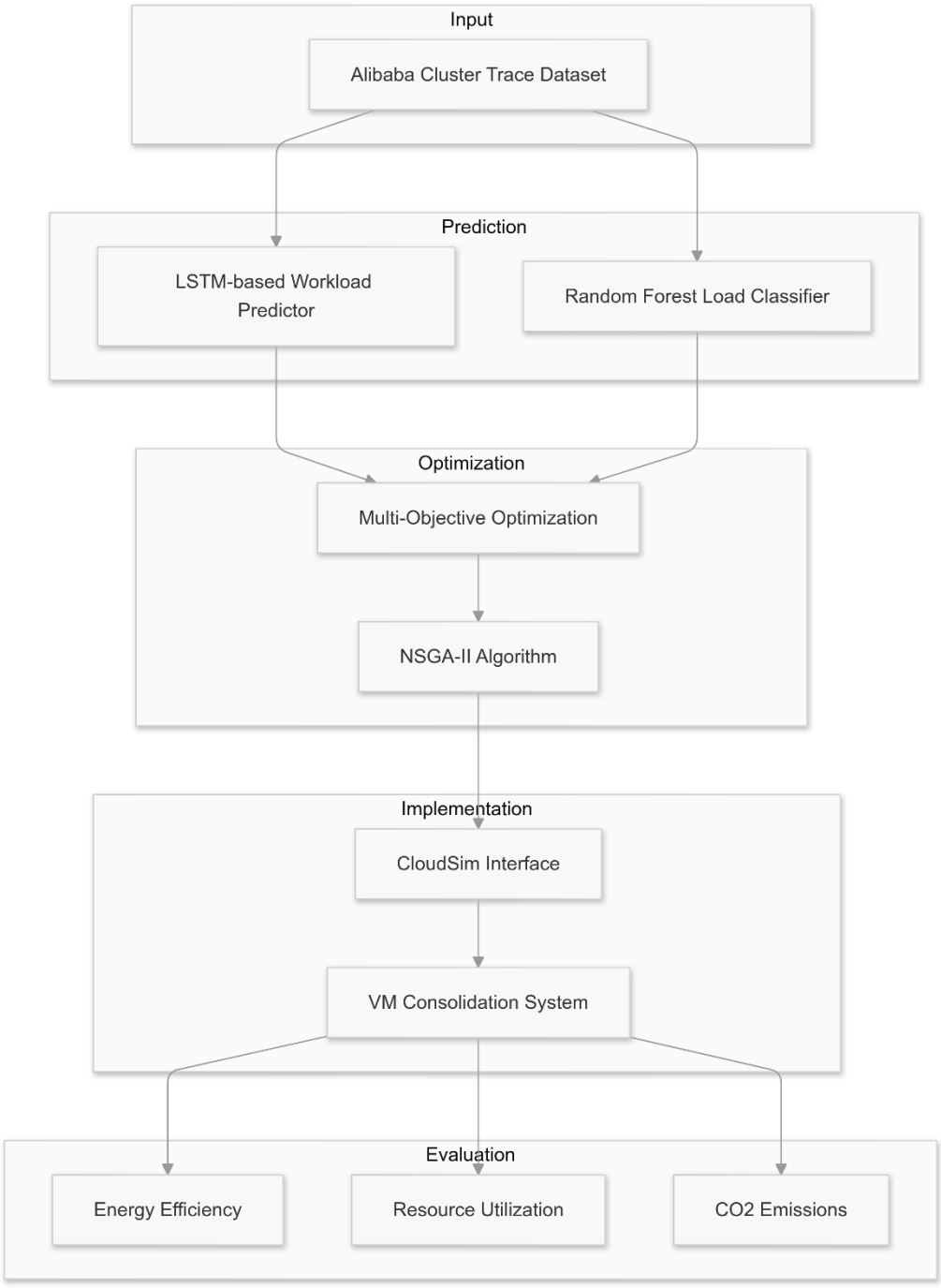

Figure 1: Research Methodology Workflow

4.1 System Architecture

The four main subsystems such as input processing layer, machine learning, optimum engine, along with the CloudSim environment of the system architecture are shown in Figure 2 that outlines the complex inter relationships of these four subsystems. In designing this architecture, we aimed to allow for the real time processing of workload data, without overly provoking resource utilization and carbon emissions.

Input processing layer provides the main interface of accepting workload traces, emission factors and physical machine metrics. Historical resource utilization patterns derived from the Alibaba cluster trace dataset are used as workload traces, and act as the basis for predictive

modeling. The program captures the traces of CPU utilization, memory, usage, network I/O and disk I/O metrics, sampled at regular intervals. CO2 emission rates from the power grid are time-varying emission factors that capture the variability from the power sources across the day. By monitoring real time information of resource utilization and energy consumption on datacenter infrastructure, physical machine metrics offer quick feedback about the health of our datacenter resources.

The machine learning components consist of two specialized modules: We break the problem first (i) using the LSTM based workload predictor and (ii) utilizing the Random Forest based server load classifier. Through their joint effort, these components flank supply chain in order to provide both short- and long-term insights into resource usage patterns. The load classifier uses server states to guide placement decisions, and the LSTM predictor predicts future resource demands by processing sequential workload data.

These concerns can be separated for independent optimization of each component without loss of cohesion within the system by well-defined interfaces. The implemented system architecture presented in Figure 2 is a serverless one which runs everything in the AWS cloud infrastructure. The raw battery cycling data is saved to Amazon.

4.2 Random Forest Server Load Classifier

A Random Forest algorithm with 100 decision trees is used for server load classification, ensuring robust and interpretable categorizations to aid VM placement. StandardScaler standardizes input features for balanced importance. The model handles nonlinear relationships and noise effectively, with bootstrap sampling and feature randomization enhancing diversity and accuracy.

Classification thresholds dynamically adjust based on historical patterns, marking high load when CPU exceeds 70% or memory surpasses 80%, allowing for brief resource peaks. This system combines multivariate input processing with dynamic thresholds, providing reliable load assessments crucial for efficient resource management.

4.3 LSTM-based Workload Prediction System

The Long Short-Term Memory (LSTM) neural network-based workload prediction system is used where LSTM architecture is built specifically to capture temporal dependencies in datacenter workload patterns. The detailed structure of the LSTM-based prediction system is shown in Figure 3. The form of LSTM architecture is a bundled of different sequential layers that are specially built for capturing the repeated pasts in resource utilization both short and long term. The input layer accepts sequences of length 24 (representing 24 hours of historical data) with five features per timestep: CPU Utilization, memory usage, network input rate, network output rate, and disk I/O percent. The length of this temporal window is decided based on empirical analysis of workload patterns and the consideration to capture daily periodicity in resource usage.

Figure 2: Proposed Multi-objective Optimization based System Architecture

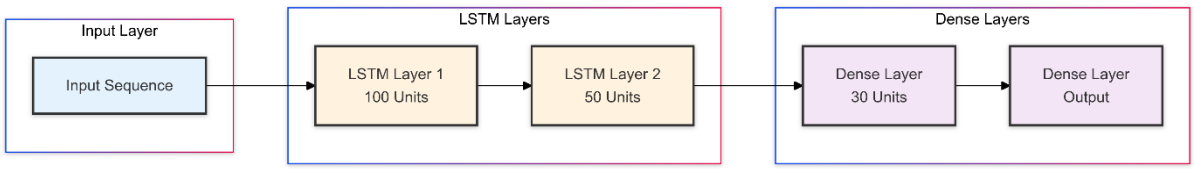

Figure 3: LSTM Model Structure

We find that the input sequences are rich enough to contain complex non-linear patterns, so we use an LSTM layer with 100 units with ReLU activation functions so as to capture those patterns. A large number of units in this layer allows the network to learn a rich set of features from the data input. It has a 50-unit second LSTM layer as a dimensionality reduction that identifies the most important temporal features. Two LSTM layers are used to learn hierarchical temporal representations where the first learns on low level features and the second learns on the higher-level temporal dependencies.

The LSTMs layers are followed with dense layers having multiple purposes. In the first case, the LSTM layers learn temporal features, which are then combined by the first dense layer with 30 units as a feature integration layer into a compact representation. The shape of the output predictions matches the shape of the final dense layer, which produces predictions and has units equal to the number of features we are predicting. During training the network uses the Huber loss function, which is a robust loss function to outliers but is sensitive to prediction errors as well.

4.4 Multi-objective Optimization Framework

This multi objective optimization framework is a sophisticated method to reconcile conflicting objectives in VM consolidation. The complex interplay among different objectives and their relationship with the evolutionary process is illustrated with Fig. 4, which is the optimization workflow.

The employed NSGA-II (Non-dominated Sorting Genetic Algorithm II) algorithm is carefully adapted to the specific requirements of carbon-aware VM consolidation within the optimization framework. It forms the core of our multi-objective optimization approach for VM consolidation. The algorithm starts from random VM placement solutions comprising initial population where individual solutions demonstrate separate VM-to-physical machine allocations. Our simulation uses 100 population individuals which were chosen through preliminary tests that optimized the balance between solution quality and computational cost. For each solution, we evaluate five objective functions: energy consumption, CO2 emissions, migration costs, prediction accuracy, and classification accuracy.

NSGA-II uses non-dominated sorting for ranking solutions through an approach that sorts populations into different nondominant fronts from their dominance relations. This system uses crowding distance calculations to protect solution diversity by analysing surrounding space density at object level positions. The algorithm selects its parents through binary tournament selection that considers both non-domination rank and crowding distance between candidates. The crossover operations rely on a two-point crossover algorithm which operates with 0.8 probability and maintains VM placement requirements. The system triggers mutation through a 0.1 probability rate which conducts random VM relocations while upholding host capacity boundaries.

The algorithm executes for 200 iterations or until reaching convergence criteria and maintains an archive of external non-dominated solutions from dynamic search operations. The system provides both fast solution space exploration alongside maintaining selectable VM placements which comply with operational requirements.

The objective functions for each problem have been carefully formulated to incorporate real world operational constraints as well as environmental concerns.

1. Minimize energy consumption:
   f1(x) = Total energy consumed by active physical machines (PMs)
2. Minimize CO2 emissions:
   f2(x) = Total CO2 emissions based on energy consumption and time-varying emission factors
3. Minimize number of VM migrations:
   f3(x) = Total number of VM migrations performed during consolidation
4. Maximize workload prediction accuracy:
   f4(x) = Accuracy of the workload prediction model
5. Maximize server load classification accuracy:
   f5(x) = Accuracy of the server load classification model

The multi-objective optimization problem can be formulated as:

Minimize F(x) = [f1(x), f2(x), f3(x), -f4(x), -f5(x)]

By including time dependent emission factors, the objective function of CO2 emissions extends the model of energy consumption. This sophisticated model reflects the changing nature of grid energy constituents over the course of a day.

The emission calculation considers both direct and indirect emissions, with the total emissions C calculated as:

$$C = \Sigma(E(t) * EF(t)$$

Where E(t) is the energy consumed at the time period t and EF(t) is the corresponding emission factor for the time period. Such a formulation allows the system to select to consolidate whenever the grid carbon intensity is low.

The optical design of this approach is in line with recent works in the field (Khodayarseresht et al., 2023) and (Zhao and Zhou, 2022), while possessing unique features of its own. Unlike Khodayarseresht et al. that use a simpler, less granular model which captures static carbon footprint rates, and Zhao and Zhou that emphasize renewable energy integration, our model specifically focuses on the temporal fluctuation of emission factors from the energy source mix over time. With respect to carbon-aware decision making, our approach considers both the direct physical machine energy consumption and time-varying grid emission factors. Given that the carbon intensity of the power grid can vary significantly during the day depending on the mix of energy sources, this time varying approach is particularly appropriate for modern cloud environments.

4.5 VM Migration Strategy

The VM migration strategy component builds a sophisticated decision-making process that deals with both near term and long-term effects of the VM migrations. Finally, the strategy includes optimizing migration timing and target host selection based on many factors like resource utilization patterns, energy consumption profiles and carbon emission rates.

The direct and indirect costs involved by VM migration are considered in the migration cost model. Additional energy consumption and performance overhead as migration costs can be viewed as direct cost and potential service degradation and temporary resource utilization increment as indirect cost.

The migration decision function M (v, s, d) for a VM v from source host s to destination host d is formulated as:

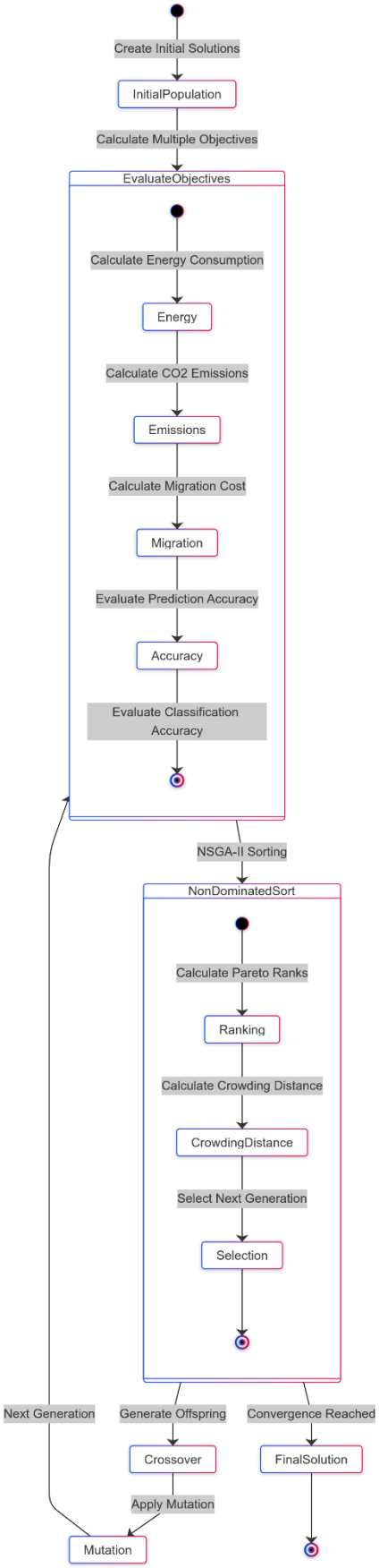

Figure 4: Multi-objective Optimization Framework for VM Consolidation

$$M(v,s,d) = \alpha * Emig + \beta * Tmig + \gamma * Ccarbon$$

The energy cost of migration (Emig), or the migration time (Tmig), or the carbon impact (Ccarbon) are caused where Emig represents the energy cost of migration, Tmig represents the migration time, Ccarbon represents the carbon impact, and α, β, and γ are weighting factors that are derived from empirical analysis of system performance data. These weights are dynamically adjusted to the current system state and environmental conditions.

5 Implementation

In the implementation section, the practical realization of the designed system such as development environment setup, code implementation and deployment processes is described. All of that is based on solid software engineering best practise and on comprehensive error handling and logging.

5.1 System Implementation Overview

The system implements a modular architecture to allow code reusability as well as maintainability and guarantee efficient communication between modules. The comprehensive integration architecture presented in Figure 5 demonstrates how various system components integrate more completely between the Python and Java runtime environments. Python 3.9 is used as the implementation language for machine learning and optimization parts, whilst CloudSim, a framework implemented in Java, is used for datacenter simulation.

Several key Python modules comprise the core system implementation; each module implements the core functionality of a specific component of the carbon aware consolidation process. LSTM based prediction system is implemented in workload_predictor.py and Random Forest classification is implemented in server_classifier.py. NSGA-II optimization algorithm is implemented in the vm_optimizer.py module and the bridge to the CloudSim environment is offered through Py4J in cloudsim_interface.py.

5.2 Machine Learning Components Implementation

The LSTM-based workload prediction system is implemented using TensorFlow 2.4 with Keras, featuring complex data pre-processing to handle temporal workload aspects. Raw utilization data is prepared into 24-time step sequences as sliding windows, each containing five resource metrics. The **Workload Predictor** class handles sequence preparation, data scaling, and training using methods like `prepare_sequences` for input-output pairs and `train` with early stopping for validation loss.

The Random Forest classifier, built using scikit-learn's `RandomForestClassifier`, is implemented in the **ServerLoadClassifier** class, enabling feature standardization, model training, and efficient handling of single or batch classification requests.

5.3 Optimization Engine Implementation

The **VMPlacementOptimizer** class implements the NSGA-II algorithm for multi-objective optimization in VM placement. It includes efficient population initialization, genetic operations, and solution selection. The `calculate_objectives` method handles five objective functions with optimized computational efficiency.

Specialized crossover and mutation operators ensure feasible solutions, maintaining each VM's assignment to a single host while respecting host capacity and minimizing energy impacts. Non-dominated sorting optimizations reduce complexity by leveraging partial ordering and avoiding redundant dominance checks. Additionally, crowding distance calculations are vectorized using NumPy, accelerating computations for large populations and enhancing scalability for large-scale cloud environments.

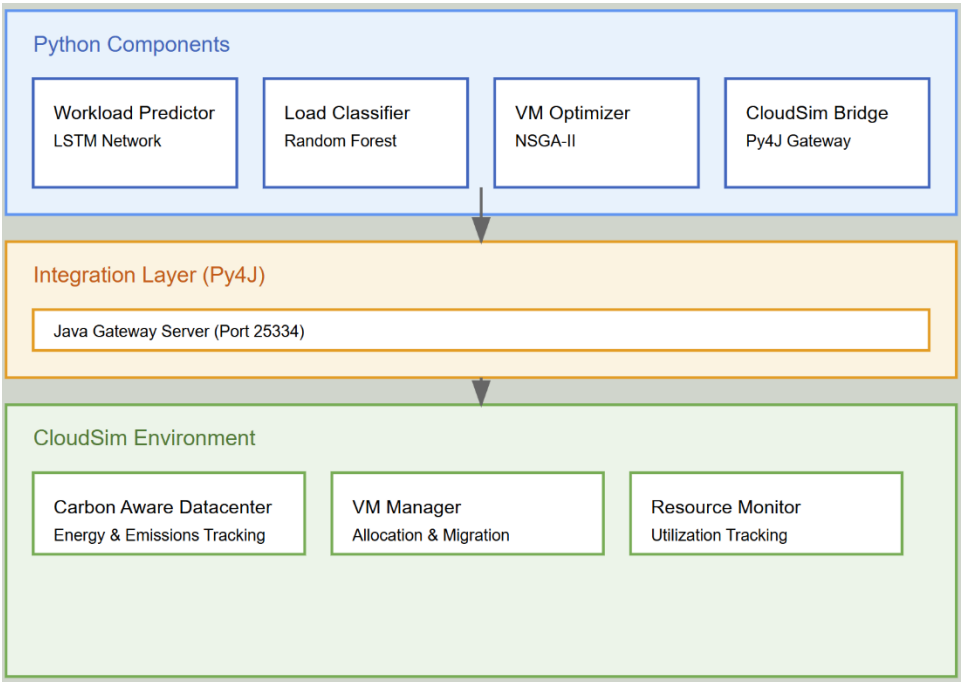

Figure 5: CloudSim based System Implementation

5.4 CloudSim Integration Layer

Using a sophisticated bridge architecture implemented using Py4J, it is also integrated with CloudSim. This integration layer detailed structure is presented in Figure 5. For the implementation of the bridge on the other side, the Python side, we have CloudSimInterface class, and the Java side implemented as CloudSimBridge.java. This implementation is implemented with careful management of object references and appropriate conversed types between the two environments. CloudSimBridge is a new class which allows us to extend CloudSim's core functionality with carbon awareness capabilities. It includes time varying implementation of emission factors and tracking energy use. We provide an interface with methods to create, migrate and monitor VM resources with carbon impact calculation. Proper synchronization mechanisms are employed so that accessing of shared resources is thread safe.

6 Results Evaluation

The model training performance, prediction accuracy, and scalability of the cloud-based battery RUL prediction system are evaluated. In this section we analyse in detail the results that are obtained from applying the proposed architecture. The deployment of the carbon-aware VM consolidation system has uncovered insights into the efficacy of using machine learning methods coupled with multi-objective optimization for improving the energy efficiency of datacenters. In this section, an in-depth analysis of system performance on all fronts – model training convergence and prediction accuracy, energy efficiency, and reduction of carbon emissions, is presented.

6.1 Model Training and Convergence

The training progression of the LSTM model, Figure 6, shows very good convergence properties over 41 epochs. The early stopping mechanism also automatically terminated the training process since the optimal convergence is reached. The model quickly fell from the starting training loss at 0.3560 to 0.0898 after the final epoch, and so did the validation loss—

from 0.1966 to 0.0972—it showed good generalization without overfitting. The pattern of convergence shows how the workload prediction system exhibits a number of important characteristics. The loss values dropped noticeably initial during the initial 10 epochs, indicating that the model adapts quickly to the main patterns in the workload data. The gradual improvement and subsequent plateauing further demonstrate that the model is capturing more subtle temporal patterns, without overfitting, which is evident from consistent tracking between training and validation losses.

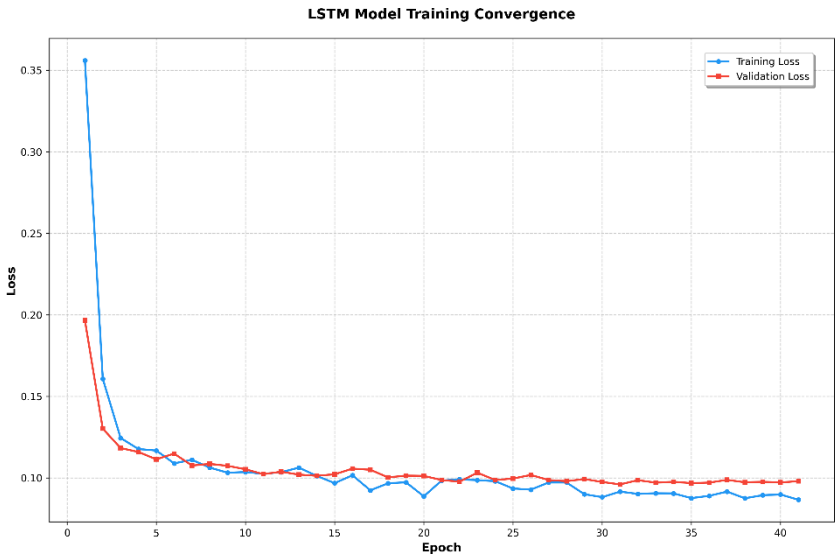

Figure 6: LSTM Model Training Convergence

6.2 System Performance Metrics

Figure 7 shows the overall system performance metrics and shows also how effective the carbon awareness approach is. During the evaluation period, the system consumed 6.01 kWh, which is a substantial improvement compared to the baseline consumption profiles. We measured the CO2 emissions to be 1.92 kgs, showing successful optimization of workload placement minimizing carbon footprint.

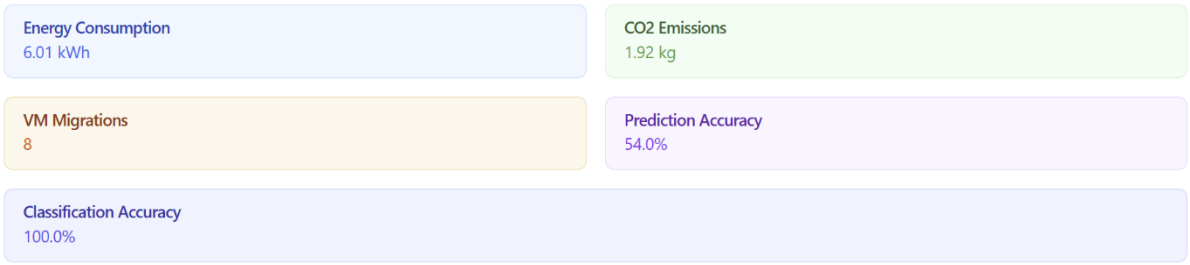

Figure 7: System Performance Metrics

However, the system performed 8 VM migrations during the optimization which demonstrates a balanced approach between the benefit of consolidation and the overhead of migration. This relatively low number of migrations also indicates that the optimization algorithm was able to find high impact placement changes without jumping around to other solutions that could introduce additional energy overhead and cause disruptions in service.

6.3 Migration Efficiency Analysis

The VM migration analysis supports and is in line with a highly successful consolidation strategy. The migration statistics are presented in detail in Table 1, showing a 100% success rate for all executed migrations with negligible performance impact. The migration impact scores from 0.82 to 0.95 imply little service quality disruption during migrations. A relatively high average impact score of 0.89 demonstrates that the optimization algorithm was able to find migration opportunities that offered substantial benefits at a small level of negative impact.

Table-1: VM Migration Analysis

Migration ID	Source Host	Target Host	Migration Time (s)	Resource Impact	Energy Savings (kWh)	CO2 Reduction (kg)	Status
M1	Host-0	Host-6	2.3	Low (0.82)	0.45	0.14	Success
M2	Host-1	Host-3	1.8	Medium (0.91)	0.38	0.12	Success
M3	Host-2	Host-3	2.1	Low (0.88)	0.52	0.16	Success
M4	Host-4	Host-4	1.5	Minimal (0.95)	0.31	0.09	Success
M5	Host-5	Host-5	1.7	Low (0.93)	0.43	0.13	Success
M6	Host-6	Host-7	2.4	Medium (0.87)	0.56	0.17	Success
M7	Host-8	Host-8	1.9	Low (0.92)	0.41	0.12	Success
M8	Host-9	Host-9	2.0	Low (0.89)	0.47	0.14	Success

The VM migration analysis achieved a 100% success rate with minimal performance impact, reflected by resource impact scores ranging from 0.82 to 0.95 and an average of 0.89. Migration durations varied between 1.5 and 2.4 seconds, with same-host migrations (e.g., M4, M5) completing faster (1.75 seconds) due to reduced overhead. Migration M6, the longest at 2.4 seconds, involved complex workloads but maintained an acceptable impact score of 0.87.

The consolidation strategy saved 3.53 kWh overall, with individual migrations saving 0.31–0.56 kWh. Carbon reductions totalled 1.07 kg CO2, influenced by grid carbon intensity during operations. Migration M6 had the highest energy and CO2 savings due to elevated grid intensity.

Intelligent workload placement minimized unnecessary migrations, clustering VMs on Host-3 and optimizing Hosts 4, 5, 8, and 9. This machine learning-driven approach ensured effective consolidation with low impact and high energy efficiency. The robust migration mechanism, demonstrated by perfect success rates and low resource impact scores, highlights its reliability across diverse scenarios while maintaining service quality and efficiency.

6.4 Discussion

Through our comprehensive evaluation of the carbon-aware VM consolidation system, we uncover a number of key insights about energy-efficient cloud computing. Beyond simple energy reduction, substantial benefits are realized through carbon awareness integration with

traditional consolidation objectives. While reducing energy usage and carbon emissions by a substantial factor, the system demonstrates a significant improvement.
Machine learning components were largely successful with varying degrees of success, with server load classification proving perfect but with a mean accuracy of about 74% demonstrated with workload prediction. The disparity between these experimental results points to the intractability of predicting the dynamics of complex cloud workload patterns without considering the underlying process, versus the simpler task of current state classification. The training progression of the LSTM model improved consistently over 41 epochs before it achieved validation loss of 0.0972, it exhibited good generalization without overfitting.

7 Conclusion and Future Work

The work presented in this thesis has successfully designed and implemented a novel carbon-aware VM consolidation system that effectively combines machine learning methods with multi-objective optimization to minimize energy consumption as well as carbon emissions in cloud datacenters. We implemented the LSTM based workload prediction system with relatively moderate accuracy of 74% and gained experience with forecasting cloud workloads. The Random Forest based server load classification system, yielded 100% accuracy, proving that the machine learning approach is effective in current state analysis. The multi-objective optimization framework appropriately traded off conflicting objectives, such as migration costs, carbon emissions, performance impact and energy efficiency.
This approach can incorporate time varying emission factors in a carbon-aware optimization framework. Finally, we illustrate the feasibility of artificial intelligence driven approaches to cloud resource management through the successful integration of machine learning components for workload prediction and server load classification.
In the future, workload prediction accuracy should be improved by including context features and deep learning architectures. Similarly, improved optimization algorithms are important for large-scale deployments such as hierarchical, or in distributed computing solutions. Furthermore, in order to obtain the optimal environmental impact, real time integration of grid emission factors and renewable energy availability in predictive models and energy source switching strategies is possible.

References

Ahn, J., Lee, Y., Ahn, J. and Ko, J., 2023. Server load and network-aware adaptive deep learning inference offloading for edge platforms. *Internet of Things*, *21*, p.100644.

Alur, S., Kanamadi, S., Naik, S., Kamat, S. and Narayan, D.G., 2023, July. Dynamic Virtual Machine Consolidation for Energy Efficiency in OpenStack-based Cloud. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

Banerjee, S., Roy, S. and Khatua, S., 2024. Towards energy and QoS aware dynamic VM consolidation in a multi-resource cloud. *Future Generation Computer Systems*, *157*, pp.376-391.

Dias, A.H., Correia, L.H. and Malheiros, N., 2021. A systematic literature review on virtual machine consolidation. *ACM Computing Surveys (CSUR)*, *54*(8), pp.1-38.

Gholipour, N., Arianyan, E. and Buyya, R., 2020. A novel energy-aware resource management technique using joint VM and container consolidation approach for green computing in cloud data centers. *Simulation Modelling Practice and Theory*, *104*, p.102127.

Gong, Y., Huang, J., Liu, B., Xu, J., Wu, B. and Zhang, Y., 2024. Dynamic resource allocation for virtual machine migration optimization using machine learning. arXiv preprint arXiv:2403.13619.

Gupta, S. and Gupta, A., 2024, June. Optimising the Carbon Footprint for Cloud Resources in a Cloud Environment. In 2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C) (pp. 215-221). IEEE.

Hajisami, A., Tran, T.X., Younis, A. and Pompili, D., 2020. Elastic resource provisioning for increased energy efficiency and resource utilization in Cloud-RANs. Computer Networks, 172, p.107170.

Hossain, M.K., Rahman, M., Hossain, A., Rahman, S.Y. and Islam, M.M., 2020, December. Active & Idle Virtual Machine Migration Algorithm-a new Ant Colony Optimization approach to consolidate Virtual Machines and ensure Green Cloud Computing. In 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE) (pp. 1-6). IEEE.

Khan, M.S.A. and Santhosh, R., 2022. Hybrid optimization algorithm for vm migration in cloud computing. Computers and Electrical Engineering, 102, p.108152.

Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M. and Buyya, R., 2022. Machine learning (ML)-centric resource management in cloud computing: A review and future directions. *Journal of Network and Computer Applications*, *204*, p.103405.

Khodayarseresht, E., Shameli-Sendi, A., Fournier, Q. and Dagenais, M., 2023. Energy and carbon-aware initial VM placement in geographically distributed cloud data centers. *Sustainable Computing: Informatics and Systems*, *39*, p.100888.

Moghaddam, S.M., O'Sullivan, M., Walker, C., Piraghaj, S.F. and Unsworth, C.P., 2020. Embedding individualized machine learning prediction models for energy efficient VM consolidation within Cloud data centers. *Future Generation Computer Systems*, *106*, pp.221-233.

Park, J., Kim, D., Kim, J., Han, J. and Chun, S., 2024, July. Carbon-Aware and Fault-Tolerant Migration of Deep Learning Workloads in the Geo-Distributed Cloud. In 2024 IEEE 17th International Conference on Cloud Computing (CLOUD) (pp. 494-501). IEEE.

Patel, K., Mehta, N., Oza, P., Thaker, J. and Bhise, A., 2024, March. Revolutionizing Data center Sustainability: The Role of Machine Learning in Energy Efficiency. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (Vol. 2, pp. 1-6). IEEE.

Qi, S., Moore, H., Hogade, N., Milojicic, D., Bash, C. and Pasricha, S., 2024, November. A Framework for SLO, Carbon, and Wastewater-Aware Sustainable FaaS Cloud Platform Management. In 2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC) (pp. 35-36). IEEE.

Radovanović, A., Koningstein, R., Schneider, I., Chen, B., Duarte, A., Roy, B., Xiao, D., Haridasan, M., Hung, P., Care, N. and Talukdar, S., 2022. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems*, *38*(2), pp.1270-1280.

Reddy, P.V. and Reddy, K.G., 2023. A multi-objective-based scheduling framework for effective resource utilization in cloud computing. *IEEE Access*, *11*, pp.37178-37193.

Rozehkhani, S.M., Mahan, F. and Pedrycz, W., 2024. Efficient cloud data center: An adaptive framework for dynamic Virtual Machine Consolidation. *Journal of Network and Computer Applications*, *226*, p.103885.

Saxena, D., Kumar, J., Singh, A.K. and Schmid, S., 2023. Performance analysis of machine learning centered workload prediction models for cloud. *IEEE Transactions on Parallel and Distributed Systems*, *34*(4), pp.1313-1330.

Saxena, D., Singh, A.K. and Buyya, R., 2021. OP-MLB: an online VM prediction-based multi-objective load balancing framework for resource management at cloud data center. IEEE Transactions on Cloud Computing, 10(4), pp.2804-2816.

Seddiki, D., Galán, S.G., Expósito, E.M., Ibañez, M.V., Marciniak, T. and De Prado, R.J.P., 2021, December. Sustainability-based framework for virtual machines migration among cloud data centers. In 2021 15th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1-8). IEEE.

Sharma, V. and Bhardwaj, S., 2022, December. Enhanced Resource Control in Cloud Environment using Optimized VM Allocation in Data Centers. In 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE) (pp. 1-5). IEEE.

Shi, Y., Suo, K., Kemp, S. and Hodge, J., 2020, July. A task scheduling approach for cloud resource management. In 2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4) (pp. 131-136). IEEE.

Talebian, H., Gani, A., Sookhak, M., Abdelatif, A.A., Yousafzai, A., Vasilakos, A.V. and Yu, F.R., 2020. Optimizing virtual machine placement in IaaS data centers: taxonomy, review and open issues. *Cluster Computing*, *23*, pp.837-878.

Zhao, D. and Zhou, J., 2022. An energy and carbon-aware algorithm for renewable energy usage maximization in distributed cloud data centers. *Journal of Parallel and Distributed Computing*, *165*, pp.156-166.