Carbon-Aware Smart Workload Forecasting and Workflow Scheduling for Renewable Energy-based Geo-Distributed Data Centers

MSc Research Project
Programme Name

Forename Surname
Student ID: XXX

School of Computing
National College of Ireland

Supervisor: XXX

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …….………………………………………………………………………………………………………

Student ID: ……………………………………………………………………………………………………..……

Programme: …………………………………………………………… **Year:** …………………………….

Module: …….……….

Supervisor: ……………………………………………………………………………………………………….……….

Submission Due Date: ……………………………………………………………………………………………………….………

Project Title: ……………………………………………………………………………………………………….………

Word Count: …………………………………… **Page Count**……………………………………………….……..

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: ………

Date: ....
....

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 own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

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

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

Carbon-Aware Smart Workload Forecasting and Workflow Scheduling for Renewable Energy-based Geo-Distributed Data Centers

Forename Surname
Student ID

Abstract

1 Introduction

1.1 Background and Motivation

1.2 Problem Statement

1.3 Research Question

1.4 Problem Solution

1.5 Research Objectives

1.6 Challenges and Limitations

2 Related Work

2.1 Carbon-Aware Workload Management

Radovanović et al. (2023) talked about the revolutionary project of Google's Carbon-Intelligent Compute Management system, the first-ever carbon-aware algorithms in full scale and implemented in the cloud data center fleet of Google. This paper explains how reducing electric carbon footprints and operational costs can be achieved by the temporal flexibility in batch workloads through intelligent load shaping. The system's core innovation is introducing Virtual Capacity Curves (VCCs) as the main mechanism, which actively restricts resource use by imposing hourly limits that are computed through a risk-aware day-ahead optimization process that considers flexible and inflexible demand predictions, hourly carbon intensity forecasts from electricityMap.org, and explicit characterization of business and environmental targets. The methodology involves advanced forecasting pipelines that use Exponential Weighted Moving Average (EWMA) for weekly mean predictions in combination with linear regression models to adapt to day-to-day changes, which leads to a high degree of prediction accuracy. The VCC mechanism is used to delay the execution of flexible computing tasks until "greener" times when local grid carbon intensity is lower, and in this way, Service Level Objective (SLO) compliance is ensured with continuous monitoring and feedback mechanisms. The real-world case study shows that during the carbon-intensive periods, VCCs successfully limit the hourly capacity, move the flexible usage from midday to evenings and early mornings when the carbon intensity is the least, and reduce the peak CPU and power consumption on a daily basis. This work gives an example of how datacenter load management can make a contribution to the realization of energy systems that are not just more robust but also more resilient and less costly to operate.

Bahreini et al. (2023) have supplemented theoretical backing with carbon-aware workload scheduling that is developed using canonical approximation algorithms entailing optimal performance guarantees as well as provision for large-scale cloud environments, averting the classic computational intractability problem. This paper accounts for a key optimization problem as that of carbon emissions. The solution for this is based upon the appropriate allocation of diverse and flexible tasks among cloud workloads through the appropriate time and space of the tasks. The authors visualize this complex puzzle as a costly combinatorial optimization problem requiring cloud providers to pick the time and location of task execution while factoring carbon intensity, electric prices, renewables, and quality of service, besides various other factors. In doing so, the authors prove the validity of mixed binary integer linear program formulations with the programming goal of capturing task scheduling, resource assignment, and the datacenter's capacity void and deadline. This research considers realistic carbon intensity with high resolution through hourly data from grid regions, reflects real-time electricity pricing models, and captures workload models that mix temporal flexibility with stringent requirements in latency. Benchmark experiments conducted on state-of-the-art cloud workloads with the help of recent algorithms revealed the efficiency of the approximation algorithms in achieving near-optimal solutions while considerably shortening the time of computation as opposed to exact optimization methods. One thing that is positive is the adapting algorithm framework, even with the implementation challenges such as cloud management system integration, handling uncertainty of carbon intensity forecasts, and tradeoffs among multiple competing objectives.

Khodayarseresht and Shameli-Sendi (2023) consider ECAIVMP (Energy and Carbon-Aware Initial Virtual Machine Placement) as a heuristic algorithm specially oriented to the minimization of both power consumption increases and carbon emission impacts during VM initial placement in geo-distributed cloud datacenters in a two-fold manner. The research considers the common knowledge that typical VM placement algorithms, on the one hand, work on single goals such as energy efficiency and resource utilization, while the inadequate addressal of the environmental issue is on the other hand. In fact, this is due to the geographical areas that are different in terms of electricity grid compositions and of carbon intensity rates. The suggested architecture pertains to serious cloud providers that run numerous geographically scattered data centers, among which each server is outfitted with asymmetric equipment, and each piece of hardware has its own unique power consumption. The methodology is to peruse potentially suitable VM placements through all rather connected datacenters, considering factors such as full load servers, weighing against the costs for marginal loading and initial power yields. Experimental tests with the CloudSim Plus simulator have been completed with real workload traces from MetaCentrum 2 and CIEMAT Euler datasets that demonstrate performance upgrades against rival algorithms gained through those tests. The ECAIVMP practically records total power consumption and carbon emissions by decreasing them massively at the same time, so it increases the quality of service of the VMs in a very clever way, achieved by placing VMs onto the datacenters to exploit the regions with a lower carbon intensity together with a more favourable electricity price.

Li et al. (2022) confront the multi-objective optimization problem that stands in between the execution of scientific workflows in the cloud and the reduction of operational costs and energy consumption while still meeting strict deadline requirements. The authors create an extensive framework for energy cost optimization. Through one of their methods, VMs are rented, which includes a leasing model for the execution time based on the environmental perspective and dynamic power load, while the static power is the relation of idle resources to the total power use of the workflow that satisfies the deadline. A key innovation is the infusion of Dynamic Voltage and Frequency Scaling (DVFS) abilities into the scheduling framework. DVFS allows processors to work for different voltage and frequency levels. The unique optimization technique they use is a two-pronged approach that features an early-stage development based on list scheduling, with the evolution of multi-objective optimization steering the final stage of refinement. A heuristic approach that gives solution vectors with a non-domination rank of 1 identified with a novel fitness function, which on one hand denies the VM rental cost, and on the other hand, effectively uses the VM instances through avoiding over-provisioning, minimizes energy through intelligent DVFS application and VM consolidation, and ensures deadlines through total distribution and critical path. The practical valuation of this approach, made with widely recognized scientific workflow benchmarks (Montage, Epigenomics, CyberShake, Inspiral) discloses that the procedure surpasses previous algorithms in a very significant manner and increases the amount of cost reduction and energy saving compared to the single-objective methods, while maintaining success rates of meeting the deadline.

Lin and Chien (2023) prove a new kind of sustainability model in data centers through the research of data centers on how, by adaptable capacity, they would coordinate with power grids and maximize renewable energy use. The researchers analyze the coordination of resources in different timeframes: from an online to a day-ahead standpoint, from a single datacenter to a global datacenter rate of return, and from local renewable sources to grid-wide energy systems. Using day-ahead optimization, this research looks at optimized energy sharing via forecasted carbon intensity, power prices, and compute backlog on data centers, and shares these daily coordinative plans. The PlanShare scheme is concentrated via joint data centers that capitalize on external information and make plans on capacity, which are distributed in the network, opting for global load optimization and able to follow local constraints and renewable energy availability. The research is driven into several coordination scopes, such as no grid, regional coordination among datacenters within the same grid, and full grid-wide coordination employing renewable energy and carbon intensity across all interconnected grids. The experimental results in this study show that even a local coordination with no grid cuts modest carbon emissions. While the local coordination can be quite effective, geo-distributed coordination (HybridGrid) offers significant performance benefits, and in the end, comprehensive grid-wide coordination with full day-ahead capacity planning achieves the best result with emission reduction and better cost performance.

Sharma et al. (2024) provide a fully integrated cloud computing platform that is sustainable by using workload classification combined with energy-efficient resource allocation in geo-distributed datacenters. The research introduces a novel application of unsupervised machine learning techniques implemented for the task of clustering heterogeneous cloud workloads from the Google Cluster Dataset so as to obtain workload characteristics and to create the right clusters for resource scheduling. The procedure entails data normalization followed by random sample selection for the clustering process, with machine learning algorithms successfully determining appropriate clusters that enable further energy-efficient resource scheduling decisions. A key innovation is the demonstration of workload heterogeneity, which enables the change of the CPU, memory, and storage requirements, necessitating intelligent preprocessing through clustering before scheduling algorithms can be effectively applied. This paper shows workload segregation based on resource requirements, which the scheduler could in turn use to make more informed decisions on datacenter placement and resource allocation, particularly based on the availability of renewable energy across different geographical locations. The authors manage clustering techniques, standard planning for workloads, and machine learning algorithms such as data normalization to cleave the workloads along lines of energy-efficient data centers. The authors' recent work has a strong focus on workload classification rather than dealing with scheduling.

Song et al. (2024) showcase CloudSimPer, which has been introduced as a robust simulation framework that is specifically aimed at filling the gap that hitherto existed in the experimental platforms for testing energy-efficient algorithms and solutions in data centers that are powered by renewable energy. As an evolution of CloudSim, CloudSimPer adds new capabilities such as renewable energy modelling of both wind and solar generation. The methodology consists of creating a new layer of software architecture upon CloudSim that is responsible for managing the complexity of interconnections among the renewable energy generators, datacenters, workloads, geographical regions, and the optimization algorithms (broadly called "schedulers"). A strongly integrated simulation case modelling is the new feature that allows the use of standardized performance metrics and a common scenario based on publicly available traces and functions for fair and reproducible comparisons between the different scheduling approaches. CloudSimPer's results correlate strongly with those metrics of interest, including renewable energy utilization, brown energy consumption, and optimization algorithm effectiveness. By presenting an open simulation platform, CloudSimPer, researchers, in turn, have a chance to try out different approaches such as virtual machine (VM) scheduling algorithms, load balancing strategies, renewable energy trading mechanisms, and hybrid energy management policies through reproducible evaluations, which in turn helps draw significant strides in the research of sustainable data centers.

An innovative task scheduling framework that utilizes real-time carbon intensity data has been introduced by (Beena et al., 2025) for improved performance across cloud infrastructures. This represents a significant advancement towards production-ready carbon-aware systems. The research seeks to solve the problem faced by cloud service providers due to the pressure to reduce the carbon footprint without compromising on performance; consequently, they designed a solution that is seamlessly integrated with the existing infrastructures like AWS, Google Cloud, and Azure. The overall framework architecture consists of several interconnected components that include automatic data collection using Kubernetes CronJobs, which is used to fetch carbon intensity data from the Electricity Maps API; implementation of a sophisticated scheduling algorithm via AWS Lambda, which helps find the optimum times and locations among the carbon intensity patterns, user-defined job durations, regional energy availability, and projected trends; and the AWS DynamoDB-based result management, which is aimed at efficient metadata storage and the AWS SNS service for user notifications. The framework's modular architecture enables additional functionalities like ML-based carbon intensity forecasting and reinforcement learning for advanced scheduling optimization without affecting the core components heavily. The experiments carried out show the enhancements as the result of the tasks being allocated strategically to the regions with the lowest carbon footprint, better carbon savings from the use of dynamic scheduling, and the big benefit of scalability to easily handle growth when data streams increase with scheduling requests.

The new paradigm introduced by (Miao et al., 2024) regarding the distributed ML task scheduling across the edge cloud has utilized multi-renewable energy sources. They address the main question of getting the most from renewable sources while still offering the customer the same service quality across the geo-distributed data centers without issues. This research proposes the Energy and Carbon-aware scheduling with Multi-RES (ECMR) algorithm with Istanbul, Skopje, and Vienna as the data centers. The use of the MILP framework allows the formulation of the optimization problem in a way that the objective functions are not only distinct but also overlapping. The framework describes the layout of renewable energy sites along the timeline and across the geographic space, determining the wind generation based on simulated wind speed and the solar generation based on irradiance and temperature. The algorithm for scheduling the data centers involves giving priority to dispatching the virtual machine tasks to those data centers that have access to more renewable energy sources at low carbon intensity while taking into account the locational marginal pricing and carbon intensity discrepancies between the different regions. The experimental environment is set directly from the datasets gathered from Google trace data and different meteorological services, thus showing performance achievements on the highest possible levels. ECMR has better energy utilization with respect to the renewable sources consumed, which is noticeable progress in the direction of the sustainable datacenters' design.

2.2 ML-Based Renewable Energy Forecasting

In the review paper, (Benti et al., 2023) provide an illustration of ML and deep learning (DL) applications in pursuit of renewable energy generation forecasting, a challenging task due to the variability and uncertainty of renewable energy sources like wind and solar. An evolutionary path from traditional statistical forecasting models was presented by covering diverse models from Autoregressive Integrated Moving Average (ARIMA) and physics-based Numerical Weather Prediction (NWP) models, all the way to intricate ML methods that showcase more potential for dealing with complex nonlinear relationships and high-dimensional renewable energy data. The authors have systematically divided ML into supervised learning algorithms (linear regression, support vector machines (SVM), random forests (RF), artificial neural networks (ANNs)), unsupervised learning techniques (clustering, dimensionality reduction), and advanced deep learning architectures (Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and hybrid ensemble methods). Generally, traditional ANNs come out at the top with high accuracy in solar irradiance prediction because the model captures most of the nonlinear relationships in input parameters like temperature, humidity, and cloud cover. As far as the SVMs are concerned, they exhibited their strength in wind speed forecasting by controlling complex nonlinear relationships without explicit feature engineering. The RF algorithms provide ensemble-based predictions with feature importance rankings for reducing overfitting risks. The DL models, especially LSTM networks, are beneficial since they preserve memory of previous observations in the time series of renewable energy generation, resulting in accurate predictions in multi-step forecasts. In this review, the authors propose that the ensemble methods, where the predictions of heterogeneous models are combined via methods like weighted averaging or stacking, are much better than single models by virtue of these complementary strengths.

A comparative study on the solar power generation forecasting within smart city microgrid applications focusing on LGBM and KNN models was delivered by (Suanpang and Jamjuntr, 2024). The need for high-precision solar energy forecasting has been addressed in this study. It is one of the preconditions for the incorporation and management of the renewable energy systems effectively in the cities. This study focuses on Rayong smart city in Thailand with a number of initiatives like transportation, energy management, and environmental sustainability to enhance urban centers. A systematic study was conducted involving the performance evaluation of models based on several criteria, including forecast accuracy, computational efficiency, training time requirements, and patterns of memory consumption. The empirical results achieved highlight the significant differences observed in the model performance. The LGBM outperformed their predictive accuracy with significantly lower RMSE and MAE as well. The accuracy gained directly corresponds to the dependability increase in solar power forecasts, a prerequisite for optimal energy dispatch decisions and grid stability maintenance. LGBM, on the other hand, can show its strength in a variety of temporal conditions by being more stable through the elimination of different periods, the same as in the seasonal variations where it effectively tackles outliers and missing data, which are common in real microgrid operations.

In the area of limited transparency and inherent unpredictability of renewable energy sources, which remain reliant on weather constraints, (Talwariya et al., 2023) present their study through ML-based forecasting methodologies. They propose a CNN architecture for predicting renewable energy generation and LSTM networks for foreseeing energy consumption at the consumer level. The forecasting technology, with an hourly operating cycle, allows for better management of energy in microgrid sustainability. The consumer loads were classified into three classes, namely, fixed loads (baseline usage), non-shiftable loads (time-critical applications), and shiftable loads (flexible workloads that might be subjected to temporal scheduling). The model performance is verified through experiments applying solar photovoltaic and wind generation data. The forecasting is done for generation first, which achieves RMSE values of 2.472 and 3.034 for solar and wind power, respectively. For the load types, the forecasting model performs differently, namely fixed loads with an RMSE value of

0.027, non-shiftable loads with an RMSE value of 0.067, and shiftable loads with an RMSE value of 0.21. The results authenticate that the proposed methodology is fruitful in the area of renewable energy generation forecasting and demand-side management and opens the opportunity for the coordinated energy balancing between generation and consumption.

Symeonides et al. (2024) introduce the CarbonOracle, an automated ML service for renewable energy source and electricity grid carbon emission forecasting, thus enabling carbon-aware datacenter management. The service offsets the flaws in the current forecasting models, which are mostly inaccurate and have short forecasting horizons, limiting the effectiveness of carbon-aware service orchestration. The architecture of CarbonOracle is made around the automation of data extraction from self-hosted renewable energy sources, national energy grid APIs, and meteorological services while providing seamless ML training and forecasting predictions. The system is built on two asynchronous devices: Forecasting Training, which connects to web services to create and update prediction models, and Forecasting Inference, which is fed from the RESTful API. The overall success story is told with the high accuracy rates in the tests carried out for the data center over a year in Cyprus. For self-hosted photovoltaic panel predictions, CarbonOracle achieves an error rate of approximately 9%, while electricity grid carbon emission forecasts demonstrate even greater performance with error rates below 4%. The service is mainly distinguished through its extensibility and configurability, enabling the users to specify the geographic locations, training schedules, and forecasting parameters according to their operational needs.

Oladapo et al. (2024) presented an extensive study on such advanced techniques as LSTM networks, random forests, and SVM, which are employed for the purpose of optimizing energy generation forecasting, grid management, and energy storage. The authors have carried out the necessary data collection and treatment works, including historical and real-time data on solar irradiance, wind patterns, temperature variations, and other weather factors affecting the production efficiency of the renewables. The research particularly focuses on the gradual addition of wind, solar, and hydropower energy to the existing grid. The methodology framework is developed with the recognition of the need for multi-objective optimization, which helps to achieve more accurate energy generation predictions, fewer operational inefficiencies in grid management, and better coordination of intermittent renewable sources with demand profiles. The research views technical optimization as a means of quantifying the environmental impact through metrics of CO_2 emissions reductions, thus becoming a useful source of information for both policymakers and energy stakeholders. The research assumed the role of a connector in the leap from the theoretical world of ML to the real-world application of it, in this case in renewable energy optimization, and pointed out the significance of predictive modelling in demand forecasting and optimization at the system level for a sustainable energy future.

Zhao and Zhou (2022) unveil an innovative energy- and carbon-aware algorithm with a long-term aim of maximizing usage for distributed cloud data centers using renewable energy sources. Their work is recognized not only for its contributions to the development of the respective joint optimization methodology (which is referenced substantially by future works, including the baseline Miao et al. (2024) article) but also for the approaches they have presented. The algorithm seeks to take care of the issue related to the intermittent supply of renewable energy using a smart workload scheduling algorithm that pairs high computational power with the periods of the highest renewable power contribution. The research accounts for diurnal and spatial availability of renewable energy alongside service quality commitments, by means of which economy of green energy and good service quality are achieved in distributed data centers. The work presents a set of critical rules for the carbon-aware scheduling and serves as a genesis of the redefined research topics, which mainly involve understanding the trade-offs between carbon emissions reduction and operating performance requirements.

2.3 DRL-based Resource Management

Jayanetti et al. (2024) make a noteworthy contribution to the scheduling of renewable energy with the introduction of the hierarchical multi-agent deep reinforcement learning (MADRL) paradigm, which is purposely built for geo-distributed cloud data centers running hybrid brown and green energy sources. The research tackles the NP-hard challenge of workflow scheduling that is combined with the variable characteristics of renewable energies and the need for decentralization in multi-cloud environments. Their framework is a multi-agent one that includes a global scheduler and several local schedulers working under the Partially Observable Markov Decision Process paradigm that adds flexibility that makes decisions only based on local observations and shares information during centralized training. The technique takes advantage of the actor-critic approach that has a neighbourhood-based coordination mechanism to punch through the dimensionality curse of multi-agent systems. Through the use of CloudSim, the experiment was conducted on a network of ten data centers with 1000 workflows received from the Pegasus framework, and it was found that the introduction of the proposed algorithm overshadowed the others by cutting energy consumption significantly while still complying with acceptable makespan. The research project was validated through actual solar energy data from the University of Queensland photovoltaic sites and heterogeneous server specifications from SPECpower benchmarks. Hence, the practical demonstration of the applicability of the green power maximization framework while executing workflows within deadlines was made.

Choppara and Mangalampalli (2024) are engaged in providing a solution for the efficient task scheduling problem in heterogeneous cloud-fog computing environments where tasks from IoT devices show heterogeneous computation needs and latency sensitivities. The DRLMOTS scheduler is the first innovation in managing the dynamic nature of fog computing by resolving task characteristics such as length and processing capacity through the intelligent selection of computation allocation between fog nodes and cloud resources. The methodology adopts a Deep Q-Learning Network (DQN) model, which learns the optimal scheduling policy through the continuous interaction with its surroundings and minimizes makespan, energy use, and fault tolerance by a specially designed reward function. The process consists of priority-based task management where the incoming tasks are put into segments and sent to the DQN-based scheduler, which takes the task allocation decisions in real-time, balancing the resource surplus against the task needs. The performance was evaluated on actual Google Jobs workloads, as well as randomized synthetic workloads using the SimPy discrete event simulation framework that was used to compare with baseline algorithms like CNN, LSTM, and GGCN. The experimental outcomes confirmed a substantial decrease in the better constituted makespan, a drop in energy-use efficiency, and an enhancement in fault tolerance along with the basic findings. The research has shown that the utilization of DRL for optimizing diverse components of fog-cloud structures is indeed effective and meaningful.

Mangalampalli et al. (2024) widen DRL application areas in the multi-cloud ecosystems through their Adaptive Task Scheduler with Improved Asynchronous Advantage Actor Critic Algorithm (ATSIA3C). The research is aimed at solving the NP-hard problem of task scheduling in cloud computing, where tasks are distributed from heterogeneous and variable-length sources, which will directly impact the cloud by higher makespan, energy consumption, and resource costs. The ATSIA3C path is introduced by separating incoming tasks into sub-tasks based on the parameters of execution and utilizing the improved A3C algorithm to arrange them on suitable machines across multiple data centers. The method enhances the former A3C algorithm with the addition of adaptive learning rates and better policy gradient estimation, making the process of scheduling action space more efficient. The simulations conducted on CloudSim used both simulated workloads and real-time supercomputing traces to compare against the baseline algorithms. The results showed considerable gains in makespan and resource cost efficiency and a decrease in energy consumption compared to the previous solutions. The research asserts that the employment of A3C methods to cope with the dynamic and unpredictable aspects of multi-cloud task scheduling is spot on, particularly in the context of varying task arrival rates and resource capabilities in different geographical locations.

A special application of MADRL in task scheduling for cloud-based digital twin systems, which are meant for power grid management, was put forth by (Pei et al., 2024). The GD-MA framework applies Multi-Agent Deep Q-Network principles to both response times and operational cost optimization by playing a role in the intelligent allocation of different tasks such as image processing, log analysis, and numerical data computation across the distributed cloud computing nodes. The method is designed with exclusive task characteristics for one agent, who is responsible for scheduling but needs shared reward structures with other agents. In the parallel design of multi-agent systems, where the kind of approaches applied in single-agent models is not proving scalable in distributed systems, parallel decision-making and efficient resource utilization are achieved. For experimental validation, GD-MA was compared against traditional methods of scheduling like Random, First-Come-First-Served, and single-agent Deep Q-Network and Proximal Policy Optimization algorithms, with task arrival rates, node configurations, and workload distributions being the variables in testing. The results showed the GD-MA kept the average response time and operational costs lower through all the scenarios, especially standing out at larger node counts, where the decentralization of decision-making shows the most advantage.

Nieto et al. (2024), using task offloading schemes in hybrid cloud-edge-MEC architectures, explored finding optimal task placements for user equipment, mobile edge computing servers, and cloud resources. The authors introduce the problem of executing heterogeneous and latency-sensitive techniques that require rapid and energy-efficient task handling, mainly in edge nodes, and cloud processing for insensitive tasks. Their distributed decision-making algorithm offered autonomy to each user's equipment to make offloading decisions depending on several elements like the existing battery, base station topology, MEC server resource availability, and wireless communication link quality. The performance evaluation was conducted on various parameters, including stable MEC availability and communication failures, while testing the robustness of offloading decisions as network connectivity degrades. The simulation tests have been done using four different task class distributions comprising varying proportions of delay-sensitive, energy-sensitive, and insensitive tasks, covering the algorithm's robustness to altering characteristics. The results demonstrated the system framework's proper function even under the referred situation of communication disruptions, such as strategically redirecting tasks using the real-time system state. This effort provides the necessary in-depth study so as to offload and safeguard tasks in emerging edge computing.

2.4 Research Gap Analysis

The presented literature review opens up several critical limitations in the current carbon-aware scheduling approaches for cloud data centers, particularly evident in the baseline work by (Miao et al., 2024). The ECMR algorithm presented by (Miao et al., 2024), even though effective in reducing carbon emissions by means of energy-aware task migration, has some fundamental gaps that could constrain its practical applicability. The deterministic MILP formulation presented employs weighted sum optimization, which can lead to risks of overlooking Pareto-optimal solutions when considering multi-objective scenarios involving energy efficiency, carbon reduction, and QoS trade-offs. Also, the fixed three-datacenter architecture limits scalability, leaving no scope for dynamic resource provisioning capabilities, which is indispensable for handling variable workload demands in production environments. Finally, the absence of integrated renewable energy forecasting may not offer proactive scheduling decisions that are aligned with predicted green energy availability. This proposed research intends to address these gaps by proposing a multi-objective reinforcement learning framework with explicit Pareto front exploration, deployed across five geo-distributed European datacenters equipped with auto-scaling capabilities. An adaptive coupling mechanism that incorporates confidence-based switching between renewable energy predictions and resource provisioning decisions is also presented. This study will employ identical performance metrics as the baseline, wherever relevant and feasible, including resource utilization efficiency, carbon intensity reduction, response time, throughput, cost per workload, and auto-scaling efficiency. This enables direct comparative analysis and conclusive evaluation of the proposed enhancements.

3 Research Methodology

3.1 Research Framework Overview

Figure 1: Research Framework Methodology

3.2 Research Process

3.2.1 Data Acquisition and Processing

3.2.2 Model Development

3.2.3 Model Deployment

3.2.4 Evaluation

3.3 Tools and Technology Resources

3.4 Evaluation Metrics

4 Design Specification

The techniques and/or architecture and/or framework that underlie the implementation and the associated requirements are identified and presented in this section. If a new algorithm or model is proposed, a word based description of the algorithm/model functionality should be included.

5 Implementation

You will of course want to discuss the implementation of the proposed solution. Only the final stage of the implementation should be described.

It should describe the outputs produced, e.g. transformed data, code written, models developed, questionnaires administered. The description should also include what tools and languages you used to produce the outputs. This section must not contain code listing or user manual description.

6 Evaluation

The purpose of this section is to provide a comprehensive analysis of the results and main findings of the study as well as the implications of these finding both from academic and practitioner perspective are presented. Only the most relevant results that support your research question and objectives shall be presented. Provide an in-depth and rigorous analysis of the results. Statistical tools should be used to critically evaluate and assess the experimental research outputs and levels of significance.

Use visual aids such as graphs, charts, plots and so on to show the results.

6.1 Experiment / Case Study 1

. . .

6.2 Experiment / Case Study 2

. . .

6.3 Experiment / Case Study 3

. . .

6.4 Experiment / Case Study N

. . .

6.5 Discussion

A detailed discussion of the findings from the N experiments / case studies. Note that this discussion will have a lot more detail than the discussion in the following section (Conclusion). You should criticize the experiment(s), and be honest about whether your design was good enough. Suggest any modifications and improvements that could be made to the design to improve the results. You should always put your findings into the context of the previous research that you found during your literature review

7 Conclusion and Future Work

Restate your research question, your objectives and the work done. State how successful you have been in answering the research question and achieving the objectives. Restate the key findings. Discuss the implications of your research, talk about the efficacy of your research, and discuss its limitations.

Describe any proposals for future work or potential for commercialisation. Present MEANINGFUL future work. Sweeping more parameters in your simulation / model / platform is probably not meaningful. More discuss what could a follow up research project do, to better / differently approach / extend etc. your work.

References

References should be formatted using APA or Harvard style as detailed in NCI Library Referencing Guide available at https://libguides.ncirl.ie/referencing
You can use a reference management system such as Zotero or Mendeley to cite in MS Word.

Beloglazov, A. and Buyya, R. (2015). Openstack neat: a framework for dynamic and energy-efficient consolidation of virtual machines in openstack clouds, *Concurrency and Computation: Practice and Experience* 27(5): 1310–1333.

Feng, G. and Buyya, R. (2016). Maximum revenue-oriented resource allocation in cloud, *IJGUC* 7(1): 12–21.

Gomes, D. G., Calheiros, R. N. and Tolosana-Calasanz, R. (2015). Introduction to the special issue on cloud computing: Recent developments and challenging issues, *Computers & Electrical Engineering* 42: 31–32.

Kune, R., Konugurthi, P., Agarwal, A., Rao, C. R. and Buyya, R. (2016). The anatomy of big data computing, *Softw., Pract. Exper.* 46(1): 79–105.