DESIGNING ALGORITHM FOR SUSTAINABLE RESOURCE MANAGEMENT IN CLOUD COMPUTING

by

GURPREET SINGH

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA
2025

DESIGNING ALGORITHM FOR SUSTAINABLE RESOURCE MANAGEMENT IN CLOUD COMPUTING

by

GURPREET SINGH

Supervised by

Dr. Kamal Malik

APPROVED BY

Prof.dr.sc. Saša Petar, Ph.D., Chair

RECEIVED/APPROVED BY:

Renee Goldstein Osmic

SSBM Representative

Dedication

To my wife, Harpreet Kaur.

Thank you for the incredible gift of time and peace of mind. While I was lost in research, you masterfully managed our world, shouldering burdens that were mine to carry so that I could see this journey through.

Your sacrifice and love are on every page.

Acknowledgements

I wish to express my heartfelt appreciation to everyone who helped me complete this research.

My deepest gratitude goes to my supervisor, Dr. Kamal Malik. She not only guided me through the various steps of the research process but, most importantly, encouraged me to see this work through to its completion. Her support was essential.

I am also very thankful for my wonderful colleagues and friends at work. Thank you for all the help, the motivating conversations, and for simply being there for me.

This thesis also stands on the shoulders of the many academics and researchers whose papers I have cited. I want to acknowledge their vital contribution to the field and to my work.

ABSTRACT

DESIGNING ALGORITHM FOR SUSTAINABLE RESOURCE MANAGEMENT IN CLOUD COMPUTING

GURPREET SINGH
2025

Dissertation Chair: <Chair's Name>

Co-Chair: <If applicable. Co-Chair's Name>

Sustainability is a term that is widely used in the IT sector to refer to the minimize of the ecological effect caused by anything. From this standpoint, the sustainability is actually representing the present and is promising a better worldwide output-input relationship. IT, at the minimum, deals with two different aspects of sustainability. "Green IT" is a one where green technology is considered the main culprit of excessive power consumption from the process of execution along from the cooling and, therefore, computers are addressed as the major part of the problem. This also goes for cloud clusters, which are invisible but still a significant ecological issue for people. On the other hand, "Green IS" looks technology as a result of its replacing expertise of the modern "develop and destroy" fabrication framework including nature-preserving or regenerative one. The argument for "Green IS" is that cloud platform provide also reach the scaling advantages that conventional IT, either inside a company or by a technical utility, cannot get. The issue of power usage and carbon dioxide emissions comes in as the key concern in the area of SCC (sustainable cloud computing). We are coming up with a hybrid optimization algorithm basically for the objective of reducing energy usage and carbon footprints in SCC environment. The combined optimization approach will be a merging of ant colony and fake normalizing approaches.

TABLE OF CONTENTS

ABSTRACT V

LIST OF TABLES VIII

LIST OF FIGURES IX

CHAPTER 1 1

INTRODUCTION 1

1.1 Background 1
1.2 A Conceptual Model 3
1.3 Issues in Using Cloud as Green Technology 6
1.4 Innovations Related to Sustainable Cloud Computing 8
1.5 Sustainable Development of Cloud Computing 10
1.6 Significance of Sustainable Development in Cloud Computing 13
1.7 Methods for Implementing Sustainability in Cloud 14
1.8 Key Principles of Sustainable Cloud Computing 16
1.9 Implication of Reliability on Sustainability 17
1.10 Ecological Resource Administration in Cloud-Based Settings 17
1.11 Motivation 21
1.12 Challenges 22
1.13 Goal 23
1.14 Visions of different international organizations on sustainable cloud computing 24
1.15. Areas to Investigate 25
1.16 Research Methodology and Taxonomy Design 31
1.17 Research Themes in Sustainable Cloud Commuting 33
1.18 Research Problem 34
1.19 Research Objectives 35
1.20 Significance of the Study 36
1.21 Research Questions 37

CHAPTER 2 38

RELATED WORK 38

2.1 Preliminary Literature Review 38
2.2 Energy Efficiency in Sustainable Cloud Computing 50
2.3 Green Cloud Architectures 60
2.4 AI/ML for Sustainable Cloud Computing 69
2.5 Renewable Energy for cloud computing 76

2.6 Answer to Research Questions 81

CHAPTER 3 83

RESEARCH METHODOLOGY 83

3.1 Methodological Framework for Sustainable Cloud Load Balancing 83
3.2 ACO (Ant Colony Optimization) and its Drawbacks 86
3.3 Hybridization of ACO and Simulated Annealing Algorithm 89
3.4 Proposed Research Architecture 92
3.5 Evaluation Parameters 102
3.6 Design Specification 104
3.7 Implementation 107
3.8 Formulating the Ant Colony Optimization Algorithms 108
3.9 Checking the Initial Simulation Parameters 110
3.10 Hybrid ACO-SA Optimization Framework for Cloud Load Balancing and Scheduling 112

CHAPTER 4 114

RESULT & DISCUSSION 114

4.1 Results 114
4.2 Discussion 161
4.3 Practical Implications for Cloud Service Providers 163

CHAPTER 5 165

CONCLUSION AND FUTURE WORK 165

5.1 Conclusion 165
5.2 Future Work 166

REFERENCES 168

LIST OF TABLES

Table 1 Overview and Comparison of State-of-the-art survey on Energy Efficiency and Resource Optimization in cloud computing 56
Table 2 Overview and Comparison of State-of-the-art survey on Green Cloud Architectures........ 66
Table 3 Summary and Comparison of State-of-the-art survey on AI/ML for Sustainable Cloud........ 73
Table 4 Summary and Comparison of State-of-the-art survey on Renewable Energy for cloud........ 79
Table 5 Response Time Analysis 116
Table 6 Resource Utilization........ 119
Table 7 Makespan Analysis........ 121
Table 8 Energy Consumption Analysis 123
Table 9 Carbon Footprint Reduction Analysis 125
Table 10 SLA Violation Rate Analysis........ 127
Table 11 Virtual Machine Migration Analysis 129
Table 12 Throughput Analysis 131
Table 13 Load Balancing Efficiency and Imbalance Degree Analysis 133
Table 14 Response Time (seconds) Across VM Configurations........ 135
Table 15 Resource Utilization (CPU Units) Across Virtual Machine Configurations 137
Table 16 Energy Consumption Across Datacenter Sizes 139
Table 17 Makespan (seconds) Across Datacenter Size Configurations 141
Table 18 Homogeneous Environment Performance 143
Table 19 Heterogeneous Environment Performance........ 145
Table 20 Steady Workload Pattern Evaluation........ 147
Table 21 Bursty/Peak Workload Pattern (Concentrated Task Arrivals) 149
Table 22 Random Arrival Workload Pattern (Poisson Distribution)........ 150
Table 23 Response Time Scalability (100 to 5000 Cloudlets, 50 VMs) 153
Table 24 Energy Consumption Scalability (100 to 5000 Cloudlets, 50 VMs)........ 155
Table 25 Resource Utilization Scalability (100 to 5000 Cloudlets, 50 VMs)........ 157
Table 26 Overall Performance Analysis........ 159

LIST OF FIGURES

Figure 1 A Theoretical Notion of cloud sustainability 4
Figure 2 Sustainable resource management viewpoint 19
Figure 3 Motivation, challenges, objectives, and approaches 21
Figure 4 Research Areas in Sustainable Cloud Computing 26
Figure 5 Publication of Papers over years 32
Figure 6 Research Themes in sustainable cloud computing 33
Figure 7 Proposed Architecture 93
Figure 8 Flowchart of the proposed ACO and SA algorithm 105
Figure 9 Response Time Analysis 118
Figure 10 Resource Analysis 120
Figure 11 Makespan Analysis 122
Figure 12 Energy Consumption Analysis 124
Figure 13 Carbon Footprint Reduction Analysis 126
Figure 14 SLA Violation Rate Analysis 128
Figure 15 Virtual Machine Migration Analysis 130
Figure 16 Throughput Analysis 132
Figure 17 Load Balancing Efficiency and Imbalance Degree Analysis 134
Figure 18 Response Time (seconds) Across VM Configurations 136
Figure 19 Resource Utilization Across VM Configurations 138
Figure 20 Energy Consumption Across Datacenter Sizes 140
Figure 21 Makespan Across Datacenter Sizes 142
Figure 22 Performance Analysis in Homogeneous VM Environment 144
Figure 23 Performance Analysis in Heterogeneous VM Environment 146
Figure 24 Performance Analysis Under Steady Workload Pattern 148
Figure 25 Performance Under Bursty/Peak Workload Pattern 150
Figure 26 Performance Under Random Arrival Workload Pattern (Poisson Distribution) 152
Figure 27 Response Time Scalability Analysis 154
Figure 28 Energy Consumption Scalability Analysis 156
Figure 29 Resource Utilization Scalability Analys 158
Figure 30 Overall Performance Analysis 160

CHAPTER 1

INTRODUCTION

1.1 Background

Eco-innovation or sustainable technology is a concept that changes the way companies and industries apply technology, focusing on the lowest possible implications on the environment. It stresses the development and application of technology in ways that do not affect the environment (Mentsiev and A., 2022). One of the easiest ways to prove environmentally friendly technologies is the CC. Cloud computing reduces the need for on-premise infrastructure and its associated expenses by allowing organizations to store their data on remote, third-party servers (Junaid et al., 2020a). This facilitates the fight against global warming since it lowers the emission of greenhouse gases that are linked with the use of physical hardware. All data which users access online is stored, processed, or transmitted through hardware. This hardware, like computers, cell phones, servers, antennas, cables, batteries and routers—consumes energy daily, depletes resources during its production and is transported across the globe (Singha et al., 2022). Subscription-based, service-on-demand solutions are the types of cloud computing models which utilize the network to carry out uses and manage clients' workloads (Sudhakar et al., 2018). To ensure that the service is both available and reliable, the core units of the CDCs that consist of CPUs, storage, and networking devices are, in most cases, operating all the time. Large volumes of data are produced by digital activities including data streaming, data transfer, internet browsing, social platforms, electronic commerce and sensor networks, which cloud data centers effectively store and handle (Singh et al., 2023a). Nevertheless, the increase in need for cloud amenities in manifold CDCs drives the need for very large amounts of power that cause a significant amount of carbon output and ecological footprint. In the case of sustainable cloud computing, data centers run on renewable energy rather than conventional grid electricity (i.e., based on fossil fuels), which helps to decrease the carbon emissions (Khosravi and B., 2018). In addition to that, the impetus on instituting energy efficiency measures also helps to ensure that carbon footprints are reduced substantially. Servers' heat sources also contributed to a more sustainable CDC, and another way of achieving that is through the use of free cooling systems. In short, sustainable cloud computing has four key aspects, which involve: i) geothermal power generation as opposed to a fossil-based power grid, ii) the heat reuse potential of servers, iii) cold supply systems (Son et al., 2018), and iv) the installation of energy-saving volumes. All these approaches minimize the carbon footprints, minimize functioning cost, and reduce the general use of energy.

The application model is important for the power saving of CDCs (Patel and Bhalodia, 2019) and power saving of CDCs if well designed. Application models are classified as: data parallel, function parallel and message passing (Bajdor, 2023a). Data parallel means parallelizing which partitions the data among the processors in a computing system in such a way that the processors can work on comparable data. The most typical similar framework examples are the blueprint-Reduce model and the pouch of tasks, or the Limit sweep model (Shao and Chen, 2016). The parallel model deals with the parallelizing of computer code over several processors, with an emphasis on one task which is distributed by the processors or threads running concurrently in various processors (Buyya and S., 2018). Threads and the task models are the only types of function parallel models. MPI supports process communication regardless of whether these processes are distributed or mapped to nodes or servers; thus, it is language-independent. PI's goal is to create a large-scale parallel programming that is both scalable and portable.

A major transformation of sustainability in cloud is the factor of AI. To minimise waste, the latest data centres integrate such technologies with smart power, temperature, lighting, and cooling controls. The rotator cuff complex assessment dorsiflexion comprehensive test is associated with the most in-depth, detailed testing. The highest value selected indicates the most in-depth, detailed testing (Dastagiraiah and K., 2018). AI-driven predictive models, by studying workload requirements and leveraging historical data, enable the data centre to make better energy decisions, leading to more effective and greener data centres. Also, AI algorithms improve data center resource/task assignment, which is productive and consumes less power. AI-enhanced predictive maintenance also amounts to less waste on both parts and maintenance products, which also serves the cause of environmental sustainability. These AI and machine learning applications show how the technologies enhance the sustainability of cloud computing (Stergiou and E., 2018).

1.2 A Conceptual Model

Fig.1 depicts a sustainable cloud architecture, dealing in complete resource management for energy savings and service renewability. The primary features of the recommended architecture have been broadly mentioned below: -

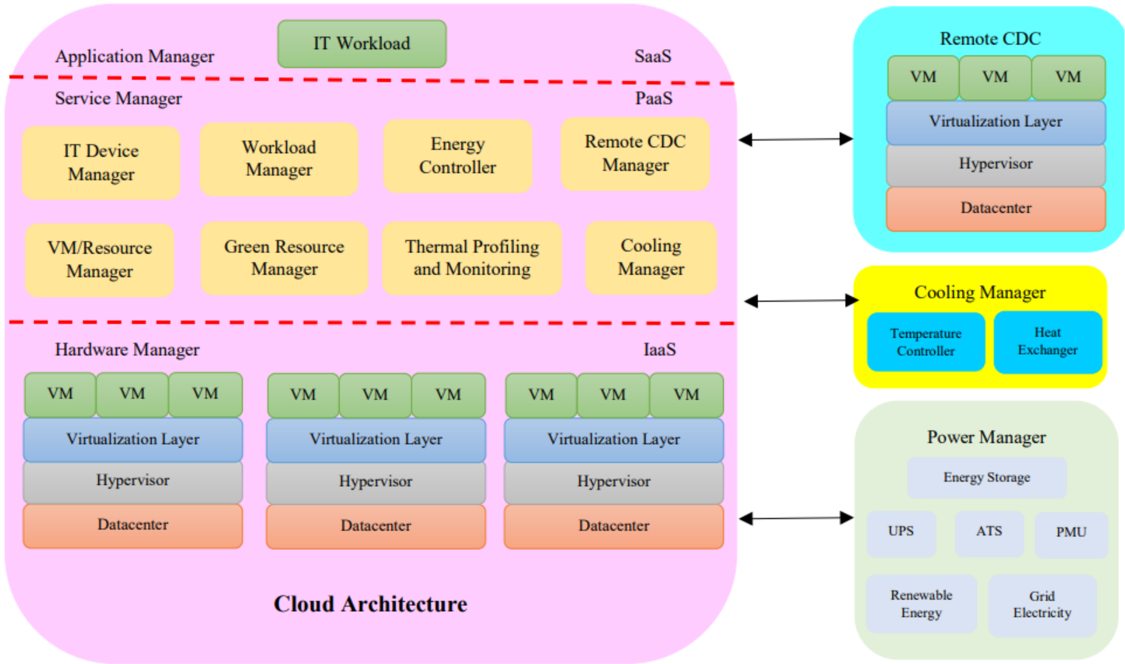

Figure 1
A Theoretical Notion of Cloud Sustainability

I. Cloud Structure

The cloud architecture element has three main sub-components as discussed below:

- *Software as a Service (SaaS)*: In this phase, employment (Tiwari and Joshi, 2016) Administrators generally supervise user activities with the help of application administrators. These activities can be user-based or batch-based. The resource managers are then responsible for dividing the resources among different units.
- *Platform as a Service (PaaS)*: The supervisor or adapter (Joshi et al., 2018) is used at this tier to manage the system's key components. All of the devices connected to the CDCs are managed by the IT device management. Incoming workloads from the application manager are managed by the workload manager, which also determines each workload's QoS needs for proper execution and sends the workload's QoS data to

the virtual machine (VM) or resource manager. To guarantee the sustainability of cloud services, energy controllers regulate the energy usage of cloud datacenters (Makwe et al., 2024). For efficient energy use, the remote CDC manager oversees the workload and virtual machine migrations between local and distant cloud data centers. The VM/resource manager, relying on the QoS needs of the workload, plans and assigns the cloud resources for the execution of the workload by using either virtual or real machines. To achieve an environmentally friendly cloud system, Sustainable Resource Manager manages the power that comes from the energy Manager and chooses clean energy rather than the grid. The changes in the cloud data center are analyzed by the temperature profiling and monitoring method, where thermal sensors record the macroclimate for the CDC (Volkova et al., 2018). At the infrastructure level, the cooling manager regulates the cloud data center's temperature (Navarro Rau et al., 2025).

- *Infrastructure as a Service (IaaS)*: Cloud data centre and VM information are contained at this layer. To ensure that workloads are executed efficiently, virtual machine migrations are carried out for balancing load at the virtualization layer. The temperature variance of various virtual machines (VMs) operating on various cores is tracked by the proactive temperature-aware scheduler. All of the hardware running the virtual machines is powered by an integrated Power Management Unit (PMU). The present states of virtual machines are stored in dynamic random-access memory. Temperature is monitored by a thermal sensor, which sends a message to the heat controller for performing further action if the temperature rises above its threshold.

II. Cooling Manager

In the event that the temperature exceeds the threshold, thermal alarms will be generated, and the heat controller is definitely in charge of adjusting the temperature, but with the smallest possible change in the functioning of the CDC. To regulate the temperature, cooling units are powered by electricity from an Uninterruptible Power Supply (UPS). Through the use of a chiller plant, outside air economizers, and water economizers, district heating management is integrated.

III. Power Manager:

It regulates the electricity produced by fossil fuels and sustainable forms of energy (grid electricity). Renewable energy is chosen over grid electricity to ensure a sustainable setting. Grid energy can be used to keep cloud services reliable if deadline-driven tasks are being completed. Renewable energy comes from the sun and wind. The renewable energy is stored in batteries. Power from both renewable energy sources and the grid is managed by an Automatic Transfer Switch (ATS), which then sends the energy to a UPS. Additionally, all of the CDCs and cooling equipment receive their electricity from the Power Distribution Unit.

IV. Remote CDC

To properly balance the load, workloads and virtual machines can be moved to a distant CDC.

1.3 Issues in Using Cloud as Green Technology

Sustainable cloud computing is fraught with several challenges. The major difficulty is the decrease in power consumption, but at the same time, maintaining the service quality standards.

There are also some problems with cloud, which are pointed out in the context of environmental applications:

- *Energy-aware dynamic resource allocation*: Frequently turning servers on and off, a practice known as power cycling, can have a detrimental effect on their long-term reliability within a cloud computing environment (Dou et al., 2017). In addition, an energy disruption in the volatile cloud setting can cause a drop in the quality of the service (Vishalika and Malhotra, 2018). Moreover, a VM can't fully keep track of the time-dependent operations of a physical machine. As a result, there might be timekeeping issues and erroneous time measurements in the virtual machine that lead to a wrong execution of the service-level agreement.
- *Selection and allocation of resources based on QoS*: This is a crucial factor that has a great effect on CC. Good resource choice and provision may lead to energy saving.
- *Optimization of virtual network topologies*: Inefficient allocation or the migration of virtual machines (VMs) can cause communicating nodes to become physically separated. This separation often leads to increased data transfer costs within the network (Andrae and Edler, 2015).
- *Increasing knowledge of nature's issues*: Those implementing eco-friendly cloud computing solutions have to be knowledgeable of how the tech is used in solving environmental issues, for instance, by cutting down on carbon emissions.
- *SLAs*: The function of these is to enable one application to be duplicated across several hosts (Xue, 2018). When selecting a provider, it is essential for cloud customers to scrutinize the terms within their Service Level Agreements (SLAs). Key parameters to compare across different vendors include policies on data protection, service outage protocols, and the overall price structure (Mardani et al., 2015).

- *Cloud data management*: Large-scale data collection is typically associated with a cloud-based business. CC providers (Chiang et al., 2017) depend on cloud service vendors to provide perfect data safety. Besides, VMs might be relocated from one place to another; hence, any off-site cloud configuration by suppliers of service might not be enough.
- *Interoperability*: The majority of public cloud infrastructures function as closed and not built in a way that allows them to communicate with each other. These kinds of standards must exist in the trade so that sellers of cloud services will be able to design such platforms that are compatible with each other. An example of a group that contributes to standardization in the cloud industry is the Open Grid Forum. The open cloud system interfaces that they are developing would be the first unified API (Tang, 2018) for handling various cloud systems.
- *Security*: It is extremely crucial to manage identity and verification, specifically for administrative information. The work by different departments and agencies of the governments, namely the US government, has included cloud computing infrastructures. Due to the complex nature of administration, applying cloud contains making policy deviations, applying (Singh et al., 2017) dynamical applications, and safeguarding the dynamical atmosphere.

1.4 Innovations Related to Sustainable Cloud Computing

Some technologies that possess similar features to cloud computing have been elaborated here:

- *Grid Computing*: This approach indicates the use of a large, centralized network (Makasarwala and Hazari, 2016) of computers to function as a unified system for solving complex and large-scale computational problems. These networks combine the idle processing power of multiple machines, often geographically distributed, to perform tasks such as scientific simulations, big data analysis, and financial modelling. Unlike traditional supercomputers, grid computing leverages heterogeneous resources, making it cost-effective and scalable. It serves as a precursor to cloud computing, as it emphasizes resource sharing and on-demand access (Li et al., 2018).

- *Virtualization*: The cloud computing relies on Virtualization; since it helps in abstracting physical hardware into manifold VMs on a single physical server. Isolating the virtual environments permits various applications or operating systems to exist in a separate environment, make the most of the existing resources, and reducing the use of hardware. Many cloud vendors use VMware, Hyper-V, or KVM and related virtualization technologies. It also allows flexibility in application deployment, enhanced disaster recovery and the ability to dynamically scale resources without having to physically upgrade resources.
- *Utility Computing*: Utility computing (also utility grid computing) is a service-provisioning concept in which computational resources are made available as a utility with users being charged on a pay-per-use basis, much like a real utility, like electricity or running water. It reduces the initial cost of IT infrastructure investment, and businesses can build their resources based on demand. Utility computing broke new ground on cloud computing by emphasizing the provision of IT as a service, where an organization only has to pay for what it consumes. This model is of the most interest to startups and enterprises that need to maximize cost efficiency with a guarantee of resource coverage when the workflow is peaked.
- *Autonomic Computing*: Autonomic computing presents the concept of self-management in computer systems, wherein self-management allows computer systems

to perform the following actions of monitoring, diagnosis and repair by itself (Garraghan et al., 2014). The systems also resemble the peripheral nerve network of the body in that they are intended to simplify IT management. With cloud environments in place, the autonomic systems will be able to benefit fully through the optimized availability, downtime, and furthermore the improved performance of the entire system through the automatic response to failures or hot problems.

- *Web Services*: Web services offer a unified interface that is used to easily integrate various software programs in differing platforms and programming languages. Web services can interconnect over the internet in a secure way and in such a manner that each application can understand the communication between them. The most basic elements of web services are some sort of communication mechanisms, descriptions of the service (WSDL - Web Services Description Language) and the ability to detect them (UDDI - Universal Description, Discovery, and Integration). As part of cloud computing, web services are crucial to any service-oriented architecture (S-O-A), API based integration and in any distributed application deployment.

1.5 Sustainable Development of Cloud Computing

Sustainable development is crucial because it ensures resources are used in an environmentally, socially, and economically responsible manner. A key advantage for companies is the ability to decrease their dependence on on-site data centres, which are often significant sources of energy consumption and carbon dioxide emissions. (Buyya and S., 2018). By making distant work and cooperation possible, cloud computing can also promote sustainable development. Cloud-based collaboration solutions can also help cut down on resource consumption, including paper waste. Yes, these are further details about how CC might support sustainable development.

- *Energy efficiency*: CDCs use less power than on-premises data centers since they may maximize their energy use by utilizing economies of scale. They frequently power their facilities using sustainable forms of energy. In order to minimize idle energy consumption and dynamically balance workloads, several providers also use AI-driven monitoring systems. By doing this, the whole carbon footprint of IT operations is greatly reduced in addition to operational costs (Al-Sharafi et al., 2023).
- *Resource sharing*: CC allows several users to share resources; less hardware is needed overall. By combining resources, fewer electronic gadgets are produced, which lowers the amount of e-waste produced. Because components are utilized more effectively across workloads, shared infrastructures also encourage longer hardware lifecycles. Additionally, resource sharing lessens the implications of IT hardware production, shipping, and disposal.
- *Scalability*: Scalability ensures that no energy is spent on unused capacity by enabling enterprises to modify their computer resources in response to demand. Businesses need not make further investments in physical substructure to expand during busy times. On the other side, they can cut back during off-peak hours, which will result in less energy being used. By minimizing overprovisioning of resources, this flexibility not only improves operations but also lessens the environmental impact (Bajdor, 2023b).
- *Disaster recovery and business continuity*: Cloud computing makes sure that important data and apps are safely saved and accessible from any location, it aids with disaster recovery. This helps businesses continue operating with the least amount of inconvenience and reduces downtime. By comparison, cloud-based backup systems are also more affordable than conventional on-site recovery configurations. Additionally,

organizations lessen their indirect energy consumption and environmental impact by relying less on physical recuperation facilities.

- *Green certifications*: These days, a lot of cloud providers get green certifications for their buildings, such ISO 14001, LEED, or Energy Star. Customers are reassured by these certifications that the supplier uses eco-friendly procedures. Certified suppliers frequently make commitments to sustainable water management for cooling systems, waste decrease, and the use of renewable energy. Companies may demonstrate their environmental responsibility and support the global sustainability goals by collaborating with recognized providers.
- *Data center efficiency*: Cloud companies are continuously enhancing the operational and design efficiency of data centres. Data centres may be scaled effectively with modular designs, and energy requirements are significantly decreased by sophisticated cooling techniques, including liquid immersion and free-air cooling (Cao et al., 2023). Also, the money is spent by the providers on equipment that consumes less power and has more capacity. Combined, these optimizations reduce the cost of operation and help in the global initiative of reducing carbon emissions.
- *Big data analytics*: A prominent utility of cloud is that it enables the business to use big data analytics to drive sustainability undertakings. Use of analytics in building solutions for regulating energy might be utilised in order to reduce wasted energy. It can optimize supply chains through the elimination of wastage and emissions that have to transportation. Predictive analytics will also help organizations establish better environmental reporting, forecast risk, and plan for the sustainable use of resources.

1.6 Significance of Sustainable Development in Cloud Computing

In cloud computing (CC), sustainability can play an essential role in enhancing resource efficiency and having a reduced environmental impact simultaneously (Dou et al., 2017). There are ways organizations can align their IT plans with global climate objectives, including adopting sustainability in the CC processes. CC may include sustainability by (among others):

- *Energy efficiency*: Energy efficiency can be obtained by cutting energy consumption and carbon emissions through energy efficiency measures, including the use of server virtualization and dynamic resource allocation (Sana et al., 2023). CC is potentially more energy efficient than a conventional IT infrastructure because it allows computing resources in the data centres to be cumulatively utilized in the best possible manner. Moreover, to fuel their data centres, the large cloud providers frequently make investments in reusable energy. Another way of reducing energy wastage is the use of cooling systems driven by AI that do not waste energy by automatically adjusting temperatures.
- *Reduced resource consumption*: Through the use of shared resources and the prevention of unnecessary over-allocation, cloud computing has the potential to reduce the consumption of physical assets like hardware, software, and even office space (Junaid et al., 2020b). This would save a significant amount of money and resources regarding the environment. Since some of the effects of the cloud platforms are that they extend hardware life cycles, resource pooling also reduces the need to have periodic hardware upgrades. As well, recycling and reuse are enabled through the centralized resource management arrangement, significantly reducing the formation of e-waste.

- *Scalability and flexibility*: Businesses can use CC to reduce their wastage and name wasteful use of resources (Kasse et al., 2014). By guaranteeing that resources are only provisioned when required, scalability helps to avoid the energy loss that idle systems cause. Businesses with varying demands can optimize their operations without overinvesting in infrastructure because of this adaptability. Consequently, this results in financial savings and reduction in environmental impact.
- *Remote work and telecommuting*: CC encourages telecommuting and remote work, fewer workers must commute to work, which lowers carbon emissions associated with transportation. The need for paper, travel, and physical office space is replaced by remote collaboration tools, cloud-hosted file sharing, and virtual meetings. This enhances employees' work-life balance in addition to reducing carbon emissions. By lowering pollutants and traffic congestion, it eventually helps to create more sustainable urban environments.

1.7 Methods for Implementing Sustainability in Cloud

The following are the solutions to implement sustainability in the cloud:

- *Choosing Green Cloud Vendors*: Companies should emphasize application service (Shen et al., 2019) providers who promote sustainability. Such suppliers include those that build data centres with minimal ecological impacts, utilise energy-saving cooling methods, and employ renewable energy sources. Most of the major suppliers have been publishing sustainability reports that disclose their environmental footprint and the measures they take to reduce it (Sampaio and Barbosa, 2016).
- *Enhancing Cloud Services*: Effective utilisation of cloud resources can help in minimising energy. Examples of the techniques that businesses can use to optimize their

cloud infrastructure include the virtualization of servers, where numerous virtual servers can replace all of the servers by operating on a dedicated server. Also, scalable deployment (Faustina et al., 2019)means that we can scale the resources to equal demand and reduce idle resources.

- *Implementing Energy-Efficient Software Design*: Design can influence the energy utilization of software. Energy-efficient software design is a plan that streamlines the code to perform by decreasing the amount of computational power needed. This encompasses both simplification of algorithms, reduction of data processing, and the use of cloud-native features for efficient estimation of scaling and managing resources (Li et al., 2017).
- *Encouraging Sustainable Practices Among Users*: The companies could also determine sustainability by the use of their customers. That can involve facilitating digital efficiency, including the adoption of cloud storage at the expense of tangible tools and streamlining digital processes to minimize redundant information refining and storing (Shen and Chen, 2020).
- *Tracking and Updating Climate Indicators*: To improve environmental impact through proper management and reducing it, businesses must understand and be able to keep an eye on the consumption of CC and its associated carbon footprint. Introducing energy accounting systems, tracking resource consumption, and emissions leads to the possibility of addressing improvement opportunities and making sustainability reports.
- *Advocating for Policy and Industry Standards*: Companies can contribute to the development of the larger context of sustainable cloud computing by influencing the passage of policy and even industry-wide standards, which will have a positive impact on green IT practices (Cappiello et al., 2014). This involves advocacy to encourage data

centres to use green power, and being a member of industry consortia aimed at sustainability issues.

1.8 Key Principles of Sustainable Cloud Computing

The following are the key principles of SCC:

- *Choosing Green Cloud Vendors*: A key step towards SCC is choosing cloud providers that use green energy and resource-efficient techniques in their data centres.
- *Optimization of Cloud Infrastructure*: Most of the resource such as server virtualization, and auto-scaling help to decrease the consumption of energy, which minimizes carbon footprints (Soidridine et al., 2013).
- *Energy-Efficient Software Design*: Software development that is energy efficient will consume fewer resources during the processing of data, making it greener in nature.
- *Encouragement of Sustainability Behaviour Inspired by User Experience*: Business can engage users in more sustainable digital action, and hence the ripple effect of green IT policies.
- *Monitoring and Reporting*: Monitoring and reporting on energy use and sustainability measures regularly can allow businesses to monitor their progress on such issues and what needs to be done further (Rawat et al., 2017).
- *Encouragement for Sustainable Strategies*: Broader improvements in the on-demand computing ecology can be sparked by establishing industry norms for utilisation and effectiveness of renewable energy sources and advocating for policies.

1.9 Implication of Reliability on Sustainability

Sustainable cloud services are appealing to a larger user base and ultimately raise the profitability of the cloud (Patra et al., 2020). The increased energy efficiency is one of the most significant aspects of green cloud computing that also leads to reduced costs of the operating and electricity bills. Nevertheless, most cloud providers are undertaking the same kind of operations to offer sensible cloud services, and this increases the consumption of resources and energy. The most affordable cloud services require sacrificing energy for reliability.

The resources that we get from CDCs are less because of the current sustainable resource management strategies (Manikandan et al., 2020) take as a large amount of energy for the performance of their tasks. The power optimization techniques that rely on DVFS (dynamic voltage and frequency scaling) significantly decrease power consumption, although a part of the resources is exchanged between high-scaling and low-scaling modes (Mehta et al., 2018) results in an increase in reaction time and service delay. Moreover, the excessive turning off and on of servers affects the reliability of system components. Energy modulation (Velde and Rama, 2017) lowers the reliability of host parts such as storage, recollection, and so on. They are able to enhance the implementation, dependability, and utilisation of the server by lowering the power utilization of CDCs. Consequently, it is necessary for new sustainable resource management (Shameer and Subhajini, 2017) methods that can achieve the goal of power saving without cloud service reliability being compromised.

1.10 Ecological Resource Administration in Cloud-Based Settings

Cloud computing has become essential to the contemporary digital environment. It serves as the foundation for practically all of the newest and most established technologies (Suratia et al., 2023), such as blockchain, cyber-physical systems, edge computing, fog computing, and automobile technology. Cloud Data Centres (CDCs) are becoming major global users of electrical energy due to the quick rise in need of storage, network connectivity, and technical support. According to one estimate, Cloud Data Centres (CDCs) consume approximately 200 Terawatt-hours (TWh) of electricity each year, which contributes to 3.5% of global greenhouse gas emissions (Singha and J., 2022). It appears that there is a great need to minimize GHG emissions in order to avert global warming and optimize electrical power use. The need for solutions that promote environmental sustainability and optimal resource allocation has increased as more and more businesses move their activities to the cloud. Taking into consideration the complex issue of cloud stability, resource management that is backed by distributing resources, supply of services, and load planning that uses less energy is essential. While maintaining top-notch services for cloud consumers, these activities help CDCs become more sustainable. This paper offers a wide-ranging analysis of handling resources in cloud-based settings.

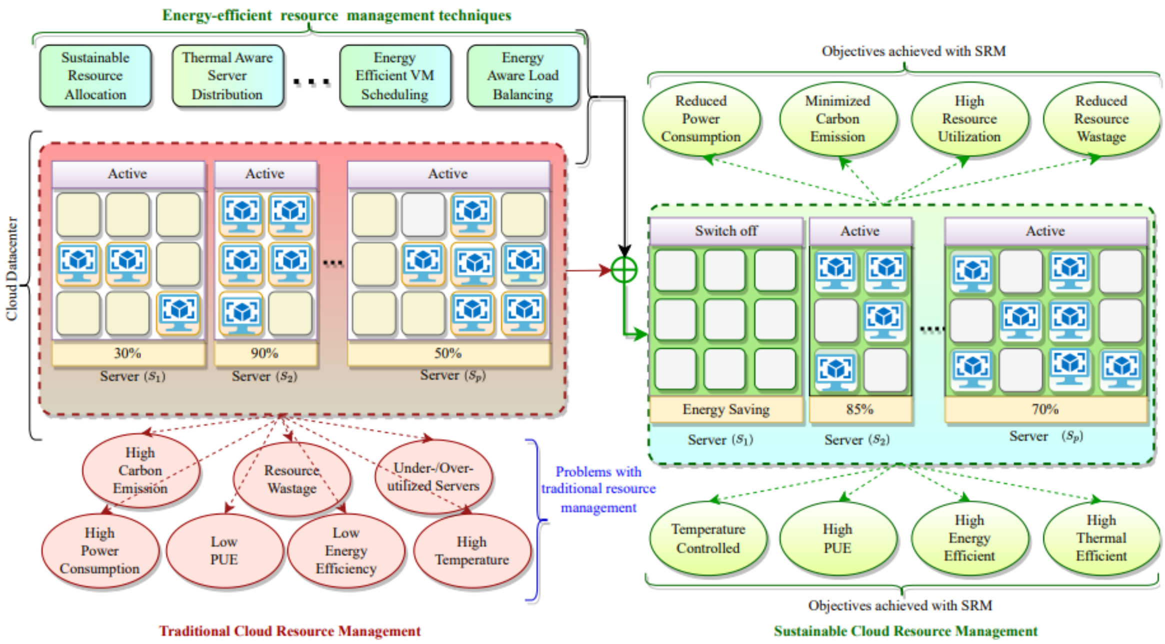

Figure 2
Sustainable resource management viewpoint

Fig. 2 illustrates the viewpoint of sustainable resource management by showing how traditional resource management may be changed into sustainable control of resources by effectual energy use and sustainability-aware resource handling techniques. With p servers $\{S_1, S_2, \ldots, S_p\}$ in a cloud data center, let's say that S_1, S_2, and S_p are underutilized (10%), overutilized (90%), and averagely used (70%), respectively. Traditional cloud resource scheduling has several drawbacks, including excessive power consumption, high temperatures, high carbon emissions, resource waste, low PUE, and low power efficiency. The information and files needed to complete each request are kept on storage discs, servers carry out calculations and supply logic to reply to information requests, and network gadgets permit info to move both inside and outward by linking the entire facility to the network (Dey, 2018). The data center must use cooling equipment to reduce the heat produced by the electricity used by these IT gadgets, which uses more energy. The sustainable resource management framework, as illustrated in Figure 2, combines a number of crucial strategies, each intended to concurrently address various sustainability issues. These strategies include energy-efficient virtual machine (VM) scheduling, thermally aware server distribution, sustainable resource allocation, and energy-conscious load balancing. Together, these tactics seek to reduce multiple sustainability concerns at the same time. The most important aspect of the consolidation of overall dynamic servers is connected with the energy-efficient virtual machine scheduling and sustainable distribution of resources. These techniques guarantee that only the necessary servers remain operational. by dynamically modifying the server count in response to demand. Collective load balancing to manage the temperature with energy-aware and thermal-aware server distribution also help maintain a controlled temperature environment. This is accomplished by optimizing thermal conditions by judiciously allocating the workload and turning off servers that aren't being used. There are encouraging opportunities for the cooperative application of these sustainable resource management techniques. These include lowering power consumption, achieving high Power Usage Effectiveness (PUE), minimizing greenhouse gas (GHG) emissions, reducing resource waste, and controlling ambient temperature (Hulkury and Doomun, 2012). Additionally, the framework seeks to improve thermal efficiency, resource usage, and energy efficiency, all of which will help create a cloud computing ecosystem that is more ecologically conscious and sustainable. With the aid of Fig. 3, the motivation, difficulties, goals, and strategies are covered in the sections that follow.

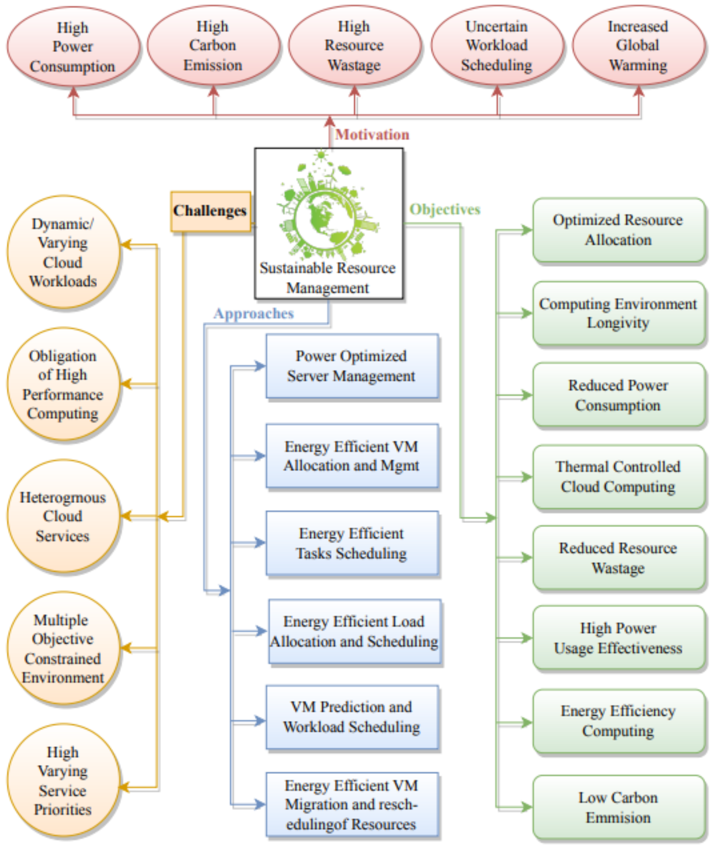

Figure 3
Motivation, challenges, objectives, and approaches

1.11 Motivation

Environmentally friendly cloud computing is driven primarily by factors that include energy consumption on the large scale, carbon dioxide emissions, high resource wastage, irregular scheduling of tasks, and global rising of temperatures. To fulfill the need of increased uptime services, which in turn is pushed by clients' needs, cloud service providers are under heavy pressure to keep their hardware in the hot mode. Consequently, data centers are large consumers of electricity and other resources for networking, computing, and storage. Up to 3% of the world's power consumption is thought to come from CDCs, and by 2030, that number is expected to rise to 4%. The major changes in temperature due to the large volumes of carbon dioxide emitted from the working servers have a chain effect on global warming and a decrease in sustainability. Moreover, excessive resource waste brought on by erratic task scheduling reduces the cloud environment's energy efficiency. Ineffective or unpredictable workload scheduling is caused by the old approach, which is limited by diverse service objectives and widely fluctuating resource capacity demands. In order to do this, sustainable resource management is crucial in reducing power consumption, minimizing resource waste, and maintaining an ideal number of hosts by only permitting the necessary range of host devices to be working (Butt et al., 2019). Higher energy efficiency during management of workload execution, a decrease in response time, SLA violations, overprovisioning, and under provisioning issues, as well as an improvement in resource utilization and service availability, are all benefits of the centralized and accurate use of cloud resources.

1.12 Challenges

One of the means through which sustainable cloud resource usage can promote system reliability is through eco-friendly consumption and resource scheduling, as well as positive autoscaling and dynamic resource allocation, which are all factors that lead to system scalability and throughput. The array of problems related to the maintenance of cloud computing sustainability that are depicted in Fig. 2 are, in particular, diverse cloud services, computing-intensive tasks, changing workloads in cloud services, several goals in restricted settings, and highly fluctuating assistance objectives. Among these problems, the uncertainty of cloud applications is one of the main issues. The resource requirements of cloud users fluctuate tremendously over a series of timescales, both in hourly and annual periods, depending on the type of work and the time when it must be performed. Despite this inherent unpredictability in demand, cloud platform providers must ensure the continuous availability of high-performance computing resources. Such responsibility often causes server computers to operate 24/7, even when this is not efficient. Moreover, cloud task scheduling is constrained by several objectives (i.e., multi-tenancy, high security, dependability, and efficient resource allocation). Such complexity often distracts people from the true aim of managing the resources and energy efficiently. Finding the right stability between the active process of cloud processes and the need to practice energy competence and ecological damage requires overcoming these challenges in the area of sustainable resource utilization.

1.13 Goal

Sustainable cloud computing is expected to reduce the environmental impact while maximising the utility of the resources. One of its key objectives is efficient resource allocation and server consolidation that ensures that compute workloads are dynamically balanced to ensure that servers are used in a smarter manner to reduce energy wastage. Long-term sustainability can be achieved in Cloud infrastructures, where hardware can have a longer life, reducing the amount of hardware in the form of electronic waste, as well as reducing the overall operational costs. Energy-efficient operations reduce power consumption and are also inherent elements of sustainability because such operations can only be conducted using clever power management techniques that power the needed number of servers. Also, by tracking temperature changes, such as smart cooling technology capabilities, and powering data centres using renewable energy sources, thermally managed cloud environments help minimize excess energy consumption. The energy efficiency of computer servers is optimised so that wasteful consumption is minimised, and when the waste of resources has been reduced, power usage effectiveness (PUE) is enhanced. Furthermore, by minimizing greenhouse gas emissions per unit of energy utilized, low-carbon and sustainable computing aim to lower carbon usage effectiveness (CUE). This strategy encourages a more ecologically conscious digital ecosystem by supporting cloud infrastructure that is carbon-neutral or carbon-negative. By combining these tactics, sustainable Cloud computing does more than only increase productivity and affordability, but also supports international initiatives to lessen the carbon footprint of contemporary IT infrastructure.

1.14 Visions of different international organizations on sustainable cloud computing

The UN claims that CC can support sustainable development in many ways, such as.

- *Energy efficiency*: Cloud computing provides the ability to cut power usage and carbon footprint (Kumar, 2018) by a large margin when compared to conventional on-site data centres. Practically, by implementing large-scale operations, cloud platform providers (Sardaraz and Tahir, 2019) are likely to improve the effectiveness of the structure and optimize energy use.
- *Minimized hardware debris*: When organizations employ cloud computing, they can considerably cut down on the number of hardware of the IT department. As electronic waste has the potential to harm the environment greatly, this will be an eco-friendly solution to reducing it.
- *Greater availability of technology*: CC offers an additional supply and cost-effective method of accessing computing resources, which can contribute to bettering access to technology in underdeveloped nations.
- *Remote work*: The COVID-19 epidemic has made remote work possible thanks in large part to CC. This may mitigate the release of carbon from the transportation industry and the requirement for commuting.

- *Sustainable business practices*: By optimizing its resources, cutting down on paper waste, and through the facilitation of collaboration and communication, CC may help businesses embrace more sustainable business practices.

1.15. Areas to Investigate

The carbon footprint has grown largely due to the expanding demand for cloud computing services (Mohanty et al., 2021) that are distributed across multiple CDCs, each of which consumes a significant amount of electricity and emits carbon. Sustainable cloud computing helps to lower carbon emissions by using renewable energy sources such as wind, solar, and hydro for cloud data centres rather than using brown energy or the conventional fossil fuel-based grid electricity (Saleh et al., 2019). Energy-saving solutions are leading to the greening of cloud computing by providing a drastic reduction in the carbon footprint of users. Through the use of waste heat from heat distributed through servers (Devaraj et al., 2020) and free server cooling operations, the CDCs are environmentally friendly. Therefore, to make the data centre sustainable, sustainable cloud computing includes the following components: i) replacing fossil fuel-based grid electricity with renewable energy; ii) employing free cooling mechanisms; iii) using waste heat from heat-dissipating servers; and iv) employing energy-efficient mechanisms (Mehra and Ahuja, 2023). Each of these factors is essential in lowering carbon footprints, operating costs, and the consumption of energy used. The SCC research problems have been categorised into seven different areas, as depicted in Figure 4: application model, energy asset management, environmentally friendly scheduling, virtualisation, capacity planning, renewable energy, and waste heat consumption. A large number of studies have dealt with SCC. However, there are still a lot of problems, specifically with models for program structure, resource focusing for handling energy, time allocation, capacity allocation, and the utilization of the renewable energy and the heat generated by resources.

Figure 4
Research Areas in Sustainable Cloud Computing

I. Application Model

An application's efficient structure can increase the energy productivity of cloud servers, and the application framework is essential to sustainability in cloud amenities. Message passing, data parallel, and function parallel are examples of application models (Skourletopoulos and others, 2019). Distributing the data among several nodes that process the data concurrently is the main goal of the data parallel model, a type of parallelization across numerous processors in a parallel computing environment (Kaur and Kaur, 2019). The Bag of Tasks or Parameter Sweep Archetypal and the Map-Reduce Model are instances of data parallel models. The function parallel model, which focuses on allocating tasks that are carried out by processes or threads across various processors in parallel computing environments, is a method of running computer code in parallel over many processors. The task and thread models refer to data parallel models. MPI has enabled the development of large-scale parallel systems that are more accessible and extensible. These systems have been facilitated by MPI, which allows for language-aware interactions among a set of processes mapped to servers or nodes (Golchi et al., 2019).

II. Resources Targeted in Energy Management

The challenge of increasing the energy efficiency of CDCs has received many proposals. According to several reports, the power consumption of CDCS includes CPU 45%, memory 15%, storage 10%, network, and cooling 20%, respectively. The cooling management comes after the processor's massive energy consumption. CDC resource utilization is impacted by power regulation techniques, which raise energy consumption during job execution. Furthermore, resource usage was resolved by DVFS; On the other hand, changing resources from high to low scaling and vice versa will increase the response time and service latency, which is contrary to the SLA. There is also the issue of turning servers off, on, or going into the sleeping mode, which influences the reliability of the system components, including storage. Server functionality, dependability, and resource consumption are all influenced by the goal of improving energy usage in the CDCs. Existing energy-conscious resource management strategies have made use of the bin packing strategy to distribute resources for task execution. There are two issues with resource allocation: i) Insufficient usage (reservations of resources are made in advance, but demand for them is lower than supply, increasing expenses) and excessive usage (many workloads are pending execution because there are not enough resources available). Several strategies have been put out to reduce the usage of energy by using the high voltage supply, but using the stall time is the most effective approach.

III. Thermal-aware Scheduling

Thermal scheduling is important to the scheduling mechanics and architecture. The architecture can be either single-core or multi-core, and the scheduling process can be either proactive or reactive. Workloads are carried out with heating problems, which reduces the efficiency of CDCs. Through thermal-aware scheduling in order to address the heating issues of CDCs, it aims to reduce the cooling setpoint degree, thermal gradient, and hotspots. Thermally cognisant scheduling is cheaper and more efficient than heat modelling. By turning on servers next to one another in a rack or chassis, CDCs can reduce their energy usage; nevertheless. (Thakur and Chaurasia, 2016)This results in an increase in power density and heat concentration. For solving this problem, a device for cooling is required. Using efficient thermal-aware scheduling techniques is necessary to complete tasks with minimal heat loss and focus. Additionally, this lessens the strain on the cooling system, saving electricity. The CDC's servers' varying temperatures render tracking and scheduling more challenging, and they also make thermal profiling less precise. In order to address this issue, dynamically updated thermal profiles—which are automatically updated and yield more precise temperature readings—are required rather than static ones. While there are temperature-sensitive ways to reduce Power Usage Efficiency (PUE), a reduction in PUE may not result in a reduction in TCO.

IV. Virtualization

To effectively balance the load during job completion and utilise renewable energy resources in decentralised CDCs, virtual machine migration is required. Virtual machine techniques transfer tasks to other distributed machines due to the lack of renewable energy on-site. Additionally, virtual machine technology allows workloads to be transferred from cloud servers that use sources of clean electricity to cloud server centres that use thermal waste at another site. In order to equalize workload demands with green power, virtual machine (VM)-based workload relocation and aggregation solutions provide virtual resources with a restricted number of physical servers. To optimize virtualisation speed, storage can be moved across running servers without affecting the VM's capacity to perform tasks. VM migration solutions make use of natural supply alternatives and waste heat utilization to enable sustainability in cloud. Migrating workloads between geographically dispersed resources while improving energy consumption and network latency is a significant problem for virtual machine migration strategies. Service latency may increase when the VM size increases since it uses more energy. To fix this problem, VM migration via WAN requires point-to-point interaction.

V. Capacity Planning

Platform providers must put in place an effective and structured capacity planning strategy if they want to see a high Return On Investment (ROI). Workloads, IT resources, and electrical infrastructure can all be planned for in terms of capacity. In order to ensure consistency, archives, restore, and recovery, SLAs should include service quality standards. This will enhance the satisfaction of the users, and new customers will be attracted in future. It is necessary to take into account crucial utilization characteristics for each program in order to optimize resource use through virtualization by identifying apps that can be combined.

Applications that are merged use resources more efficiently and have lower capacity costs. Appropriate capacity scheduling requires that cloud workloads be examined before being executed, especially for workloads with deadlines. In order to effectively complete tasks with little in the way of resource usage, workloads or devices should be able to migrate. This improves cloud datacenter energy efficiency and aids in power administration. Reducing the cost of data processing and storage requires efficient capacity planning.

VI. Renewable Energy

Important components of sustainable energy that can be optimised include the location (off-site or on-site), energy source (wind or solar), and the means of storing energy (net-metering or batteries). Green resource capital costs and unpredictability are the primary obstacles to renewable energy. The problem of unpredictable renewable energy supply was solved by energy-aware load balancing and workload migration. Sources of renewable energy are typically located far from corporate cloud server sites. For cost-effectiveness, movable data centres in the cloud must be situated near renewable energy sources. CUE can also be decreased by including more renewable energy sources. Research on the utilisation of clean power in cloud data centres is hampered by high capital expenses.

VII. Waste Heat Utilization

The models used to make effective use of waste heat rely on the cooling mechanism and heat transfer. The use of CDCs is heating up because of consuming a lot of power. Any type of cooling system that uses vapour absorption in CDCs can absorb waste heat, which is used in evaporation. Free cooling techniques which rely on vapor-absorption can maximise the PUE value by balancing the cooling costs. The heat produced by the CDC can be used to heat buildings in cold climates. Stacking and multi-core server designs are boosting server power densities, which raises cooling expenses even further. Reducing the energy used for cooling can increase CDCs' energy efficiency. Cloud servers need to be relocated to save cooling costs, and this is possible through locating the CDCs in free cooling places.

1.16 Research Methodology and Taxonomy Design

To discuss the most important problems of configuring, allocating, and managing energy-conserving cloud resources, this survey carefully investigates sustainable resource management techniques. Using credible sources like IEEE, ACM Digital Library, Elsevier, Springer, Wiley, and others, it examines and studies across range of energy-efficient cloud resource management techniques. Energy Efficiency & Resource Optimization, Green Cloud Architectures, AI/ML for Sustainable Cloud Computing, and Carbon Footprint & Renewable Energy Use are the four main categories into which the gathered research papers are divided.

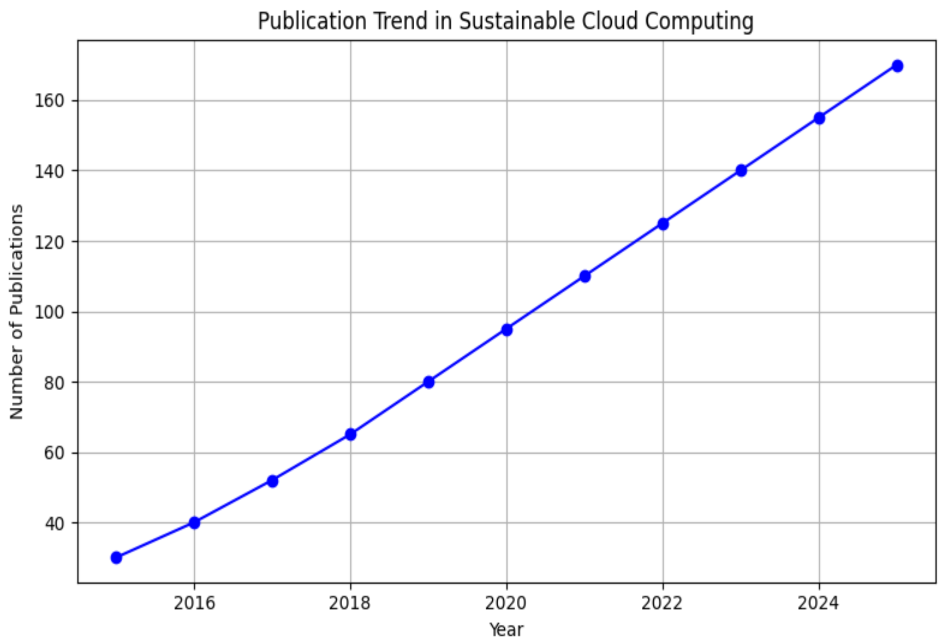

Figure 5
Publication of Papers over the years

The survey research utilized the Scopus Database, which is well-known for offering an extensive database of research articles compiled by Elsevier since 2004, and focused on academic literature that addressed sustainable cloud computing. The study utilized a "document search with keyword approach" in order to gather information on the total number of publications and citations on sustainable cloud resource management from 2014 and referred to literature and leading authors in this area thereafter. In order to classify various sustainable cloud solutions, the Scopus database was thoroughly examined. This allowed for a thorough examination of the targeted research topic's popularity by comparing overall publications and citations over a period of years. Accordingly, Figure 5 presents a thorough bibliometric analysis of the aforementioned sustainable cloud options. It has been noted that when it comes to allocating and managing cloud resources, researchers' top priority is energy efficiency. Over the past ten years, the academic and corporate communities have shown a great deal of interest in and attention to research articles on sustainable cloud computing. The graph in Figure 3 illustrates the increasing peaks in the quantity of articles and associated citations. This demonstrates the growing importance and acceptance of all techniques.

1.17 Research Themes in Sustainable Cloud Computing

The primary objectives for environmentally friendly cloud computing are resource efficiency, energy reduction, and reducing the harmful effects of the cloud's architecture. It combines cutting-edge technology with calculated methods to boost productivity and encourage environmentally friendly behaviour.

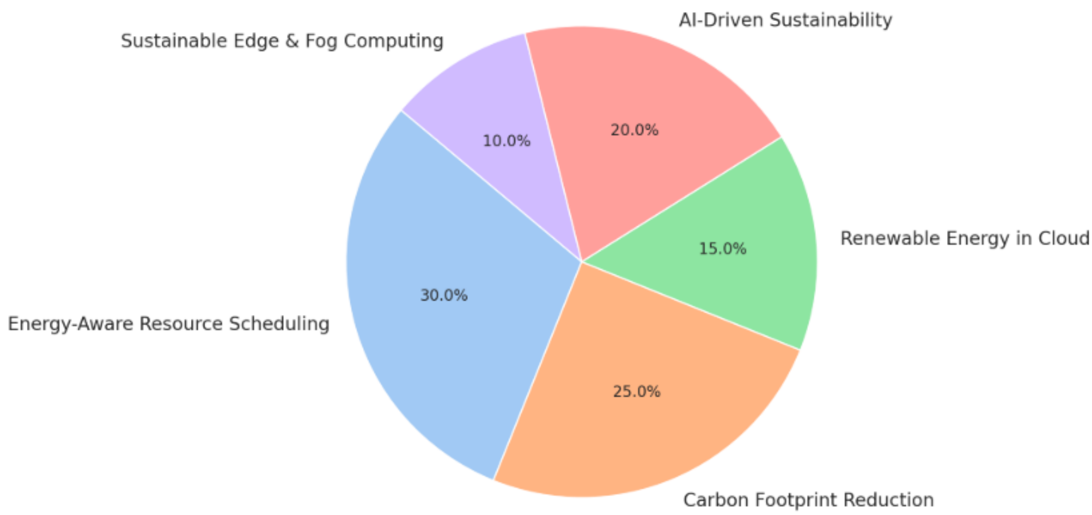

Figure 6
Research Themes in Sustainable Cloud Computing

The pie chart highlights many aspects that contribute to environmental sustainability and energy efficiency, illuminating the main research subjects in sustainable cloud computing. The most vital topic is energy-aware resource scheduling (30%), which deals with how efficient resource allocation can be performed to minimize wasteful server utilization and energy consumed. Another significant criterion is the employment minimization of the carbon footprint (25%) through the introduction of environmentally friendly data centre practices and increasing energy efficiency in a bid to minimize carbon emissions. Sustainability using AI and machine learning (20%) predicts workload requirements, power optimization and cloud infrastructure sustainability. One of the efforts to reduce the weight of fossil fuels to an environmentally friendlier cloud is through renewables in cloud computing (15%) that promotes using solar, wind, and other renewable power. At last, the main objectives of sustainable edge and fog computing are the decentralization of computing resources, reduction of data transit overhead and the subsequent increase of processing information closer to the source to save energy (10%). The above fields of study are mutually supporting in that they ensure an overall approach in sustainable cloud computing that finds the middle ground between environmental awareness, economy and performance.

1.18 Research Problem

CC has become a core part of contemporary technology framework because it delivers manageable storage space and computing capacity, without requiring huge amounts of investments by the consumers. Although it has demonstrated an outstanding benefit in most industries such as healthcare, finance and telecommunications, its implementation approach poses serious problems, particularly in energy consuming sectors, such as energy and vital infrastructure services. Although effort has been made, research on how best the leadership and management teams could foster a cloud computing environment where security and sustainable implementation of CC in various phases of organizational maturity can occur, has been significantly lacking. In cloud computing, sustainability is primarily concerned with environmental ethical behaviour, i.e. the use of energy-efficient machinery, power-optimized systems and sustainable use of resources. Nevertheless, the central infrastructure of data centers continues to require high power consumption resulting to high emission of carbon. This increases the importance of the strategies aimed at minimizing energy consumption and CO 2 emission but at the same time ensuring secure processes. Green IT initiatives have the potential to improve energy consumption and environmental effect but is not easy to implement these into cloud computing operations due to the need to balance operational efficiency, resources optimisation and the need to protect the operations in terms of security. The energy sector is one example of this where surety and dependability in resource management are paramount to the sector and where the industry relies on dependable sustainable cloud provider. This paper will review the aspects of sustainable cloud computing on energy conservation and security in the context of critical infrastructures and how the findings could potentially inform actions to relieve the energy crisis, lower carbon footprints and enable secure cloud services. These kinds of solutions can help in establishing an efficient long-term sustainable, safe online environment that will not hinder technological development as its activities will not compromise environmental aims.

1.19 Research Objectives

Environmentally friendly cloud computing is the baseline of this research study. A global of topics and problems related to the environmental sustainability of cloud computing is what this concept paper addresses. A research question outline acts as the basis of the whole work. This research has a number of objectives as follows: -

- To research and evaluate several facets of cloud sustainability.
- To determine the main obstacles or problems with sustainability in cloud.
- To look into the several frameworks that are currently in place to promote sustainability in cloud.

- To create a structure for environmentally friendly cloud computing that is energy-saving.

1.20 Significance of the Study

This research is significant for its contribution towards creating a greener and more economical cloud computing environment. The energy consumption in data centers has turned out to be one of the major issues, with the escalating demand for cloud services, not only from an environmental perspective but also in the light of the providers' increasing operational costs. This study, by employing a hybrid optimization model that synergizes the strengths of Ant Colony Optimization (ACO) and Simulated Annealing (SA), is effectively meeting these challenges by guaranteeing the smooth utilization of resources, minimizing the waiting time for the completion of the tasks, and ensuring that the energy saved is not wasted. The work is highly valuable to cloud service providers who benefit from the dual advantage of energy efficiency and cost reduction, enabling them to increase profitability while at the same time meeting their sustainability targets. For the end-users, the enhanced effectiveness is to have a service that is stable and continuous (at a lower price, most probably). Besides the operational benefits, the research also makes a contribution to the wider concept of green cloud computing by implementing measures that diminish the emission of carbon gases into the atmosphere by big data centers. Practically, this study shows that smart hybrid optimization can lead to the change of cloud platforms into energy-saving and cost-efficient ones, thus making them economically sustainable and paving the way for further developments in eco-friendly and inexpensive computing. Furthermore, the research opens doors to ways in which hybrid metaheuristic methods could outperform traditional ones in achieving a balanced set of goals comprising cost, energy efficiency, and system performance. The results obtained in this work are not confined to cloud computing alone but can also be transferred to other highly energy-consuming computing environments. By bridging the gap between the theory of optimization and its application, this research is of great importance to both the academic community and industry. Ultimately, the results emphasize the use of intelligent, sustainable, and cost-aware strategies as the key to cloud infrastructures that will stand the test of time.

1.21 Research Questions

The following are the research questions developed to unpack different parameters embedded in cloud computing and their effects on the environment.

RQ1: To find several areas that impact sustainability in CC

RQ2: How to make carbon footprints and energy related issues low in SCC

RQ3: How can cloud computing make use of resources and create data security in long term cloud computing

RQ4: How to make data available and reliable in SCC

CHAPTER 2

RELATED WORK

2.1 Preliminary Literature Review

Cloud computing has emerged as a crucial component of any organization in the modern digital era, providing scalable, efficient, and flexible applications. There is no way that the environmental consequences of these services can be ignored, particularly the consequences in regard to power consumption and greenhouse emissions. With the growing importance of sustainability in businesses, there is the need to seek computation cloud solutions that are eco-friendly (Dos Reis and M., 2023). Hosting facilities that are used to fuel cloud computing are energy-intensive in that large amounts of electricity are needed to run servers, memory devices, and cooling equipment. Unfortunately, a significant percentage of this power is fossil energy that promotes or contributes to more carbon emissions and environmental degradation. sustainability in cloud mitigates these issues by enhancing the level of energy efficiency and renewable energy consumption (Dosanapudi and M., 2024). Through green IT, businesses maintain a cleaner environment and in addition to that there are benefits like saving, by saving money, building a stronger brand and meeting more stringent carbon emission laws. Organizations generally prefer to work with Cloud service providers who have a sustainability commitment. These are companies that make use of energy-efficient cooling methods, renewable energy sources, and construction of data buildings that do not degrade the environment a lot (Rohit et al., 2023). An increasing number of leading providers are issuing sustainability reports highighting transparency about their environmental footprint and measures taken to reduce it. Businesses save a significant amount on energy consumption by effectively using the cloud resources. Server virtualization, for example, can greatly reduce the number of server instances required because multiple virtual servers can operate on a single machine. In addition, auto-scaling can help minimize idle capacity by dynamically adjusting cloud resources to meet demand. Businesses also help in sustainability because they promote responsible actions among the users (Chang and J., 2016). This will include streamlining digital activity to eliminate unnecessary data processing and storage, and encouraging the use of cloud storage rather than physical storage devices to enhance digital efficiency. Businesses track their cloud computing use and the associated carbon footprint to accurately assess and manage their environmental footprint. Assessing what the business has thus far accomplished and what may need to be improved on for sustainability can be done by implementing processes and tools in order for companies to track their emissions, resource use, and even energy use. In other words, sustainability when using cloud computing is not only a corporate responsibility but, rather, a calculated business decision that is fiscally rewarding, regulatory aligned, and a competitive advantage. The pressured oxygen provides rapid results. This is in comparison to the rest of the pressurized air, which is used in a more controlled manner (Goyal and G., 2022). Organizations and enterprises can meet their digital infrastructure demands through green IT practices, which may create a more sustainable future. Business enterprises, cloud service providers and legislators have to come together to shift to environmental-friendly cloud computing services. A sustainable and greener digital future comes about when individuals cooperate.

(Morchid et al., 2025) suggested an innovative irrigation system based on a combination of cloud computing, fuzzy logic, embedded technologies, and the Internet of Things (IoT) to manage water in the most efficient way. The system is composed of three layers, which include an interactive dashboard for the monitoring and controlling irrigation; a cloud layer, which is ThingSpeak that collects and analyzes posting real-time data using the HTTP protocol; and field devices, which are made up of ESP32 microcontrollers, temperature and humidity sensors, and actuators. Mathematical modelling was combined with Fuzzy logic to dynamically modify irrigation based on environmental conditions. Fuzzy membership functions were modelled and plotted in MATLAB to help in the best decision-making. According to the results, the device reduces water losses because it will adjust the watering schedules depending on the soil temperature and water humidity. A viable option for sustainable agricultural water management that was sustainable was this clever irrigation system. It increases water efficiency, maximizes yields, and reduces the danger of water stress by combining real-time soil and climatic data, hence enhancing resilience.

(Wang et al., 2025) proposed HealthAIoT, an innovative architecture that combines cloud computing services with AIoT to build an intelligent healthcare system. The current version of HealthAIoT evaluates the suspicion of diabetes mellitus in healthy individuals using specific personal health history and metrics; however, the recommended architecture was effectively disease-agnostic and could be leveraged towards disease detection and monitoring. The CloudAIBus was a viable testbed to assess the healthAIoT framework's performance. The experimental results on unseen patient data demonstrate that the cloud scheduler achieves 93.6 % accuracy, while the MLP-based suspicion of diabetes prediction achieves 78.30 % and an F1-score of 0.7719. System performance was also assessed against other metrics such as energy consumption, carbon-free energy employed, cost, execution time, and latency. The framework was specifically constructed to pinpoint individuals with the greatest diabetes risk, thereby maximizing its potential impact, use resources wisely, and support directed preventive care. It can also be used as a foundation towards larger healthcare applications.

(Seddiki et al., 2024) presented a new strategy of optimizing energy use in a cloud computing system based on knowledge acquisition. More specifically, a KAGWO algorithm is applied to collect data on the availability and applications of renewable energy in data centres to facilitate cloud computing and achieve energy sustainability. The KAGWO method incorporates both knowledge and global optimization ideas and is a structured method for solving complex problems and improving decision-making, using fewer configurational factors. Results show that the KAGWO algorithm is more efficient than the Pittsburgh and KASIA algorithms since it is capable of obtaining information more precisely, and this enhances the outcomes of energy sustainability.

Another researcher (Cao et al., 2023) proposed an EIS algorithm which integrates three new mechanisms. First, it establishes the best time to complete each activity on a particular resource in an environmentally responsible manner. Second, EIS saves energy by allowing slack time for a work cell, the time that exists between the time a task is completed and the deadline, to reduce the voltages and frequency of the tasks being run in accordance with the optimum timings of the various tasks. Third, the EIS discovers idle time intervals which are caused by task priority restrictions and utilizes them to decrease the dynamic energy usage to a greater extent while still ensuring that process deadlines are met. They perform extensive comparative studies with actual workflow applications to measure the performance of the EIS. The results indicate that the EIS has lower energy consumption compared to other similar systems under various deadline scenarios, and during the evolution process, it also has a quicker energy consumption reduction.

(Tuli et al., 2022) developed HUNTER, a fully AI-based resource management system for the cloud that is environmentally friendly. HUNTER considers the three most salient characteristics: energy, thermal, and cooling models. It does this by modelling the enhancement of data centre energy efficiency as a multi-objective scheduling problem. The HUNTER model implements a Gated Graph Convolution Network to act as a proxy model to produce optimal scheduling choices, as well as estimate the QoS for a given state of the system. Outcomes from both simulations and real-world cloud used the COSCO framework and CloudSim tools have demonstrated that HUNTER is outperforming the latest state-of-the-art baselines. The study reported that the HUNTER framework achieved maximum reductions of 12% in energy use, 35% in SLA violations, 43% in time management, 54% in cost, and 3% in warmth (Tuli et al., 2022).

Mechanisms to jointly improve energy efficiency and reliability under correlated failures in CCS (Cloud Computing Systems) are proposed by (Sharma et al., 2021). Methods of statistical cluster analysis as a basis for modelling failure correlation are practically used with reference to actual failure trace data. Along with defining the reliability aspect of carbon capture and storage (CCS) systems, mathematical models are also developed for the energy consumption of such scenarios. Subsequently, these models are employed for setting up the rules and measures of the resource provisioning that is tolerant to faults and energy-conscious. Additionally, a failure-controlled virtual machine consolidation policy will be suggested to further reduce energy consumption. Two techniques will be utilized to evaluate the suggested resource management policies and fault tolerance methods based on the exploration of real failure traces and using a Bag-of-Tasks workload in a simulation experiment setting. It is confirmed by these analytical studies that the proposed policies improve system energy efficiency by approximately 20 per cent and decreased the occurrence of task failures by 34 per cent due to clustering failures.

In addition, (Xu et al., 2021) will minimize the carbon footprint of data centre operations by consuming more renewable energy and less brown energy. Their contribution implies a self-adaptive approach to the management of the resource over time for interactive and batch workloads. To maintain the quality of service of these workloads, their contribution presents deferral management of batch workloads and a brownout-driven approach of interactive workloads. To test their approach, the authors utilized web services and real resources over time in a prototype system. The authors found that the proposed strategy would increase renewable energy use by 10 percent and reduce brown energy use by 21 percent.

For green cloud computing, (Almutairi and Aslam, 2023) provides a unique Energy and Communication (EC) aware scheduling (EC-scheduler) method that maximizes traffic load and data center energy consumption. The EC-scheduler is designed with the main objective of allocating user applications to cloud resources while minimizing the data center's overall energy use. This process is accomplished in two stages: first, an Emotional Artificial Neural Network (EANN) is used for efficient resource allocation, and second, a Multi-Objective Leader Salp Swarm (MLSS) method handles the scheduling of tasks that handles the traffic load balances. The EC-scheduler provides a green, unchanged environment by scheduling cloud user needs to the cloud server not just with energy efficiency but also minimizing communication delay time. A key outcome of the improved energy efficiency is the resulting decrease in carbon dioxide output from the cloud infrastructure. To evaluate the effectiveness of improving data center energy and other scheduler parameters, they used the GreenCloud simulator to test the suggested strategy and current cloud scheduling techniques. Efficiency gains of up to 26.738%, 37.59%, 50%, 4.34%, 34.2%, and 33.54% were demonstrated by the EC-scheduler parameters PUE, DCEP, Throughput, AET, Energy Consumption, and Makespan.

An energy-conscious virtual machine scheduling approach based on the multi-unit non-discriminatory combinatorial auction was proposed by (Öner and Özer, 2023). Using a robust bidding language, the model enables users to define complex virtual machine requirements using time constraints and logical operators like AND/OR. The authors structured the optimization challenge using an integer linear programming model, and a number of proposed heuristic solution algorithms, including the Genetic Algorithm. A comprehensive test suite is used to evaluate the effectiveness of both the model and the suggested heuristic algorithms. According to estimates, the suggested model will increase revenues by about 37%, and the solution techniques will yield high-quality answers that are just 5% of the best, allowing the model to be used widely.

(Waqar et al., 2023) examined the obstacles preventing small-scale construction projects from adopting cloud computing (CC). In the present research study, the author tries to give an overview of the various challenges, which are likely to be encountered in the course of implementing CC, as a result of a careful examination of the relevant scholarly literature. To have a better understanding of these obstacles, a model is also created using structural equation modeling, or SEM. The EFA's findings highlight how danger and privacy concerns significantly hinder the use of CC. By tackling the problem of data fragmentation, the current study provides insightful information that could help policymakers improve the efficacy of building projects. Additionally, by enabling the more efficient use of CC, it supports the sustainability of ongoing construction projects. This study offers a fresh approach that lays the groundwork for recognizing and addressing the numerous obstacles preventing cloud computing from being widely used in the construction industry.

The intersection of these two advancements, green cloud computing, was first studied by (Doss et al., 2022) based on published research that indicates cloud servers are the most ecologically friendly option for service delivery. Using external private cloud hosting can assist companies in maximizing their efficiency. This article also presents three case studies of creative firms engaged in green hosting and discusses the influences on their resource growth. The findings suggest that there is currently no market interest in green virtual private server hosting because low prices are the principal driver for firms. Green service providers should create green service level agreements, use transparent efficiency metrics to show their greenness, and compete on greenness. Sustainable cloud computing and sustainable cloud solutions are consequently required as a result of end customers' demands for more environmentally friendly online services.

As more people turn to smart devices that need a lot of processing power to work properly, the use of virtual machines is growing (Sharma et al., 2022). Virtualization allows for the creation of distinct resources from the physical infrastructure that is currently in place. The growing need for computers, storage, and communication resources in sizable CDCs is successfully met by virtualization techniques. Additionally, it accomplishes several resource management goals, including load balancing, Internet system upkeep, speed, and tolerance through the movement of virtual machines. Autonomic resource management would benefit from VM migration approaches. Thus, VM migration methods, applications, and their unresolved issues have been covered in this work.

Based on the data centre's task scheduling timeframe and peak energy usage, a robust agile response task scheduling optimization technique is suggested (Shu et al., 2021). Techniques for optimizing agile responses are also used. The proposed algorithm is designed to evaluate the optimization model, specifically examining the probability density functions related to request queue overflows and timeout requests to manage congestion and prevent task failures. According to experimental data, the suggested technique can successfully increase the cloud computing system's throughput while achieving job scheduling stability and efficiency.

(Shah et al., 2021) introduced the EDOS-IDM (ICMP detection and mitigation model), a new technique designed to identify and neutralize both volumetric and normal behavioral threats in ICMP traffic. Because of the resource-saving nature of the proposed method as an environmental mitigation strategy, its results were compared and contrasted to the result of the Normal Behavioral ICMP Traffic attack. The research showed no method was able to mitigate against a Normal Behavioral ICMP Traffic attack. The OpenStack based Cloud test bed was utilized to analyze and test the proposed approach in practice. Based on the findings, the method has been shown to reduce unnecessary resource usage and customer expenses in a cloud computing setting.

(Sharma et al., 2021) suggested methods for enhancing energy efficiency and dependability simultaneously in the event of associated failures in CCS. Real failure traces are subjected to statistical cluster analysis methods to understand failure dependencies. Mathematical models were then used to determine the energy consumed and the reliability of failure-prone CCS. These insights guided the usage of mathematical models that support resource provisioning policies designed to be both fault-tolerant and energy-aware. This study also presented a correlated-failure aware option for the VM consolidation policy to further minimize energy. The researchers simulated these management methods and fault tolerance protocols using Bag-of-Tasks workloads and authentic failure traces. Compared to when failures were dealt with in a unilateral manner, results show that with the failure correlations combined with resource management methods proposed, task failures can be minimized up to 34% and energy efficiency improved up to 20%.

In the electricity derivative market, (Khalil et al., 2020) addressed the problem of optimization of the energy cost of geo-distributed DCs under time-varying system dynamics with account of the call options in the electricity derivative market. A call option represents a formal agreement between CSPs and a power supplier, giving the holder of the option rights (but not the obligation) to purchase a certain amount of electricity at a specific fixed price within a certain period of time. To achieve cost reductions, the authors introduced OptionGLB, a technique for geographical load balancing based on option pricing. The performance advantage of OptionGLB over the existing workload allocation algorithms is achieved through experimental evaluations using actual usage data.

As a foundational scenario for big data, (Stergiou et al., 2018) suggested a new cloud computing system that is integrated with the Internet of Things. In addition, they attempted to improve the security difficulties by establishing an architecture that relied on network security. One proposed solution to eliminate privacy and security concerns is to set up a security "wall" between the Internet and the Cloud Server. It is, therefore, said that CC handles the privacy issue of bits transferred over time more effectively. In order to accomplish the goal assigned to them as a collective entity, this suggested technology facilitates communication and collaboration amongst objects and things via wireless networks. This thesis provided a survey of IoT and CC and emphasized the security issues related to each technology, as it focuses on security, which is the overall goal of the research. Finally, they learn out of this study that cloud computing can offer an eco-friendlier and more efficient fog environment in cases of sustainable computing.

(Maldonado-Carrascosa et al., 2024) presented a new way to go about sustainability concerns in cloud computing by suggesting a game-theoretically optimised virtual machine (VM) migration method. The method deployed is a Particle Swarm Optimization (PSO) - Game Theory (PSO-GTA) approach to minimizing energy use and maximizing the utilization of renewable energy resources. The method enhanced cloud system performance through the cooperative and competitive aspects of game theory. The study's results demonstrate that the PSO-GTA method outperformed fuzzy and genetic algorithms regarding renewable energy use. The PSO-GTA algorithm outperformed Pittsburgh, Q-Learning, KASIA, particularly in three simulation scenarios based on different cloudlet dynamics, but also reflected improvements of 0.68% and 5.32% on average over other methods over the nine simulations. This indicated that the strategy was effective and flexible in the development of cloud computing systems that were energy-efficient.

(Khaleel, 2023) presented a three-phase bi-objective workflow scheduling problem, the Bi-OWSP, to resolve deficiencies in application placement policies. The given approach is designed to optimize the workflow scheduling and, at the same time, address energy efficiency and reliability. They make use of two algorithms: a dynamic voltage and frequency scaling algorithm that would shorten mapping time whilst offering energy efficiency, and a revamped, reliable fault tolerance method increasing system reliability by eliminating defective servers and communication links. Moreover, they also came up with a step-wise reliability optimization algorithm that struck a balance between energy-consumption and reliability by choosing the best task-server combinations. Bi-OWSP showed good energy-reliability solutions as shown by extensive simulation on synthetic and real workflows against other methods.

(Singh et al., 2023b) presented the FP-NSO (flower pollination-based nondominated sorting optimization) algorithm that was supposed to minimize the energy consumption and the carbon emissions of data centres and to maximize the use of available resources. The method was used to determine the optimum physical machines to deploy a VM within a cloud system by taking into account several criteria that are constrained by resources. They have used FPO (flower pollination optimization) in combination with the NSGA-II (nondominated sorting technique-based genetic algorithm) nondominated sorting technique in conducting VMP. For evaluating the performance of the algorithm, Google cluster dataset was collected, and both the static and dynamic cases were examined in the calculation of the key performance indicators of resource utilization, power consumption and carbon emissions. Their method consumed much less power by up to 16.69, carbon by up to 48.60, and took much shorter time by up to 75.87 compared to the previous methods, and used more sustainable resources by up to 78.18.

(Kaur et al., 2022) created an overall workload, task-scheduling, and virtual-machine-placement classification framework in cloud data centres based on power grids and renewable energy sources [6]. A two-stage multi-objective optimization model was proposed to counter this. Phase I in phase I, the wrapper technique based on the Boruta random forest was used to select the appropriate feature set for the incoming workload. The researchers adopted a support vector machine technique of locality-sensitive hashing to classify the workload. The next phase involved formulating a multi-objective optimization framework for VM allocation and scheduling, which considered factors such as the carbon footprint rate (CFR), energy cost, the service level agreement (SLA), without leaving out the availability of resource events (RES). To solve this, the authors applied a modified heuristic method rooted in a greedy algorithm. When compared against current state-of-the-art methods, the solution demonstrated higher overheads: energy usage rose by 31%, costs by 28%, and CFR by 36%, whilst there was also a noticeable reduction of around 2% in SLA assurance.

(Meneguette and Boukerche, 2020) proposed an architecture based on the Energy-Aware paradigm for the migration of virtual machines in vehicle clouds. They also suggested a new policy on the virtual machine migration based on the amount of energy consumed by this migration, considering energy consumption of the cloud and cloudlet VMs, and the energy consumption of the connection of the clouds and cloudlets to the vehicles. As shown in the simulation results, using the proposed policy reduced the power consumption of the system by about 10%, and the proposed approach had a reduction in the number of virtual machine drop migrations with an energy consumption of 5%. The general scheduling method proposed in (Varga et al., 2018) was based on time constraints. They were able to make a cost model to help them estimate the amount of work they needed, and the scheduler could use this to prioritize tasks in accordance to their deadlines. The cost model systematically forecasted the remaining operational time by tracking already completed activities and using generic abstract resource requests, i.e., memory, virtual cores and containers. They validated the cost model by running several experiments in a cluster running on Amazon EC2, while also evaluating the performance of their scheduler. The procedure performed as expected under different conditions.

2.2 Energy Efficiency in Sustainable Cloud Computing

(Jin et al., 2025) presents a model of task scheduling with twofold objectives of maximizing the rate of task completion and time of task completion and minimizing energy consumption. The model addresses the delay scheduling issue of binary random failures of power supply and internet connectivity of cloud computing areas. In this model, an estimation of distribution algorithm (EDA) with neighbourhood search (N) and crowding distance (C) (EDA-CN) is used and involves a crowding distance operator, neighbourhood search operator, a multi-model probability matrix, and a regional dislocation backup process. The results of EDA-CN's performance will be compared numerically to EDA, EDA-C, and the traditional non-dominated sorting genetic algorithm III (NSGA3). Based on the data, EDA-CN outperformed EDA and EDAC by a consistent margin, and EDA-CN and NSGA3 should have been similar in terms of frequency, yet EDA-CN was statistically significantly higher.

(Ramachandran, 2025) presented a comprehensive sustainability model that is supposed to guide the creation of large scale software systems. The model argues that energy-efficient requirements need to be included, the design process needs to be augmented by energy efficiency, and energy-efficient algorithms need to be applied with the utilization of Green Function, which relies on third-order differential equations to quantify the energy usage of cloud software. In cloud and quantum applications, the results of the experiment demonstrate that a phenomenal 99 percent energy efficiency was possible.

(Thumala et al., 2025) explored modern techniques like the virtualization technologies intended toward multi-cloud and hybrid cloud environments, predictive analytics aimed at smart scheduling of resources, and workload dynamism. It also addresses the groundbreaking nature of AI-based algorithms and how renewable energy is integrated into it, demonstrating how it can facilitate scalable and sustainable cloud operations. The assessment of the key trade-offs between cost, scaling, performance, and energy efficiency can offer helpful data to academics and practitioners. The article, besides pointing to such issues as computing overhead, compliance with policy, and economic viability, presents inspiration to overcome those barriers. This paper suggests a synthesis of the recent developments in order to give a way forward in the development of more affordable, environmentally friendly cloud infrastructures that can be used to support the world's sustainability agenda. In order to facilitate an ecosystem of ecologically aware cloud computing solutions, business, academics, and legislators were encouraged to work together.

(Khanday, 2025) presented a machine learning (ML) way of reducing CO2 emissions in clouds and making energy use more efficient. The framework leverages advanced models (i.e. Random Forest (RF) and Gradient Boosting (GB)) to accurately predict energy consumption and execution time based on KPI's such as memory utilization, CPU utilization, and duration of execution. Experimental results show that GB has superior prediction accuracy across all metrics when compared to baseline methods and that sand errors are reduced significantly, individually optimizing Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The exploratory study also makes clear that workload balancing and job prioritization are required to pursue energy-efficient operation. Such a design offers a fact-driven solution to the most effective utilization of resources, diminished ecological impact, and encouragement of sustainable cloud computing measures.

(T K et al., 2025) proposed a decentralized computing architecture that spreads workloads among home nodes that are energy-efficient. Distributed networks replace big data centres, which enhances security due to job isolation and encryption, utilizes unused hardware, ensures better fault tolerance, and reduces the environmental impact. It lowers power usage and cooling requirements, thereby lowering the cost of operation, making computing more affordable and accessible. This idea can change the face of cloud computing as witnessed in theoretical research that has proven its feasibility, efficiency and sustainability.

A different investigator (Hou and Ismail, 2024) developed a method of less energy-consuming task scheduling called EETS (Energy-Efficient Task Scheduling). Using past data, this algorithm interacts with the environment to automatically choose the best possible scheduling plan. Experimental results show that EETS beats DQN as well as DDQN in convergence rates and rewards. The main scheduling performance measures include energy consumption, frequency of tasks responding within a given frame, and average machine working hours, where EETS is found to be superior to the other baseline procedures. It possesses a significant strength, particularly in handling large work packages.

(Rajagopalan et al., 2024) The proposed new parameters and optimization methods show that further positive progress in power technology is possible. More studies have been suggested on electric cars and grid to car so that there is a smooth integration and optimization of resources. The general findings of the research reveal the extent of the difference in the introduction of CoT into PS, which opens the door to higher sustainability, reliability, and efficiency in the energy sector. The CoT presents enormous opportunities to the future generation regarding a more technological and ecologically friendly energy environment.

(Sharma and Parihar, 2024) proposed a new model that employs Ant Colony Optimization (ACO) in order to make cloud computing resource allocation more energy efficient. The ACO algorithm, first created to solve complex optimization problems in distributed computer systems, was a great solver of the most efficient paths in graphs, and was named after the foraging behaviour of ants. The suggested architecture employs the tenets of ant behaviour to supply its intelligent resource allocation among the cloud servers and IoT devices, as well as factoring in the paths in which data will be processed and stored and determining the experience it represents different streams of information. Trials performed on an IoT-based simulated cloud computing environment indicate that the ACO-based approach performs better than the traditional approach in energy efficiency, resource usage, and decreasing operating costs. This study makes it possible to create cost-effective and environmentally friendly computing paradigms and enable the sustainable development of cloud computing infrastructures capable of supporting the growing demands of IoT applications.

(Manikandan et al., 2024) deployed green contracts, blockchain resource distribution and integration of renewable energy into cloud-computing infrastructures to enhance sustainability, efficiency and stakeholder satisfaction. Also, 50 percent of energy was renewable, 50 percent was fossil fuel-based, and carbon was offset 100 per cent. When it eliminated security risks, auditing and reporting became a lot more transparent, and when it achieved conformity with the standards, integrity was ensured. The study proves the significance of using blockchain-empowered solutions to enhance resources, reduce energy consumption, popularize renewable sources of energy, increase the security of the sector, and increase transparency, which ultimately increases the satisfaction of stakeholders in the cloud computing process. The application of sustainable practices that will assist the sector is still a subject of cooperation between scientists and various stakeholders.

The primary objective of (Kavitha et al., 2023) is to reduce the amount of energy used in cloud-based systems by optimizing resource allocation. This article will discuss some of the approaches and algorithms in the area, such as power-aware scheduling, virtual machine migration, and dynamic workload consolidation. This work is based on the idea of putting performance measures and energy efficiency into resource management decisions. The proposed techniques will minimise energy wastage and carbon emissions on the basis of maintaining quality of service (QoS) and cost-efficiency. This research finally supports the green building and running of cloud computing platforms.

(Agos Jawaddi et al., 2023) suggested a deep reinforcement learning (DRL)-based autoscaler which considers cooling power as an important decision-making criterion. The scheme applies DRA to adapt the cloud resources with the aim of meeting performance objectives and to optimize energy performance. In contrast to RL, DRL provides a solution to its union-based limitation on a small memory capacity to store Q-values by the use of neural networks to cope with the huge state-action space in cloud scalability. In this work, they conduct a simulation experiment to test the efficacy of this proposed remedy. They also evaluate comparative approaches with an RL-based autoscaler and the proposed DRL-based autoscalers. This research proves the DDQN-based autoscaler is superior to others, as it optimizes Power Usage Effectiveness (PUE) and reduces execution time even when the system is under heavy load.

(Al-Wesabi et al., 2022) introduced a new hybrid metaheuristic of energy efficiency resource allocation (HMEERA). The feature reduction phase uses PCA once the suggested model has completed the feature extraction step on the basis of the task requirements of a large number of clients. The HMEERA methodology then allocates resources in the most efficient manner possible using the integrated features. To achieve efficient resource allocation, the HMEERA model optimises between GTOA and RSO, called GTOA-RSO. The GTOA and RSO algorithms work together to better allocate the resources of the cloud datacenter among virtual machines. A complete series of simulations was executed with the help of the CloudSim program in order to validate the experiment. The experimental findings demonstrated high performance of the HMEERA model with regard to various factors.

(Hao et al., 2021) considers task urgency and system load as a model based on the Dynamic Voltage and Frequency Scaling (DVFS). First, jobs are ranked in terms of urgency and power consumption of different locations. Second, a reference working state is provided based on the system load for the purpose of ensuring that all tasks (regardless of the job deadline) are performed for minimum energy consumption at or below the average system load. Next, timely jobs on mobile devices and remote cloud resources are scheduled using a heuristic considering the reference working state and the priority of the tasks. According to the DVFS paradigm, they have labelled the method as AODVFS-Adaptive Offloading. These achieve a higher completion rate of tasks and decreased average energy consumption when contrasted with other strategies.

(Aldossary et al., 2019) suggests a way to predict VM costs by looking at power and resource usage, along with a cloud design approach focused on energy efficiency. When tested on a cloud testbed, the framework accurately predicted workloads, power usage, and total VM costs for various workload patterns. Further results showed that pricing based on energy consumption is economically beneficial for both cloud providers and their customers across different market scenarios.

Table 1
Overview and Comparison of State-of-the-art survey on Energy Efficiency and Resource Optimization in cloud computing

Author & Year	Technique Used	Performance Metric	Findings	Limitations
Jin et al., 2025	EDA-CN	Task Completion Rate, Energy Consumption, Task Completion Time	EDA-CN outperforms EDA and EDA-C in task completion and energy efficiency, and often exceeds	Limited testing in environments with more diverse faults or real-world applications; may not generalize well

			NSGA3 in performance, particularly in statistical significance.	to other scenarios.
Muthu Ramachandran (2025)	Sustainability framework, Green Function using 3rd order differential equations, energy-efficient algorithms.	Energy efficiency (claimed 99%)	Energy efficiency in cloud and quantum workloads up to 99% thanks to integrated energy-efficient design.	Not explicitly mentioned; potential challenge in real-world applicability of theoretical Green Functions
S. R. Thumala et al. (2025)	Dynamic workload management, predictive analytics, AI-driven algorithms, virtualization, renewable energy integration	Energy efficiency, performance, scalability, cost, compliance	AI and renewable energy can enable scalable, sustainable cloud systems; highlights actionable insights for balance of efficiency vs. cost	Computational overhead, policy compliance, economic feasibility.
S. K. Khanday (2025)	Machine Learning models (Gradient Boosting, Random Forest)	CPU usage, memory usage, execution time, RMSE, MAE	Gradient Boosting outperforms other ML models; job prioritization is essential for sustainable operations	Need for better workload balancing; practical implementation and data variability challenges.
S. T K et al. (2025)	Decentralized computing model using energy-efficient residential nodes, task fragmentation and encryption	Environmental impact, fault tolerance, security, power usage	Distributed model reduces power and cost, enhances resilience and accessibility.	Theoretical confirmation only; real-world deployment feasibility needs validation
Hou et al., 2024	Energy-Efficient Task	Convergence Rates, Rewards, Energy Usage,	EETS outperforms DQN and	May require adaptation for specific cloud

	Scheduling (EETS)	Task Response Time, Machine Working Time	DDQN, improving convergence rates and energy efficiency, especially with large task batches.	environments; results may vary with different workloads or system conditions.
Arul Rajagopalan et al. (2024)	CoT (Cloud of Things) integration, optimization strategies in power technology, grid-to-vehicle research	Efficiency, sustainability, reliability	CoT integration enhances sustainability and reliability in energy systems; sets future research directions	Future research needed on CoT and grid-vehicle integration; lack of practical deployment results
D. Sharma et al. (2024)	Ant Colony Optimization (ACO) for resource allocation in cloud + IoT	Energy efficiency, resource utilization, operating cost	ACO improves energy use, resource efficiency, and cost in cloud-IoT setups over traditional methods	Simulated environment only; real-world testing needed for validation.
G. Manikandan et al. (2024)	Blockchain-based resource allocation, green contracts, renewable energy integration	Energy consumption, renewable energy usage (50%), carbon offset (100%), stakeholder satisfaction	Blockchain enhances transparency, compliance, and enables 100% carbon offset with 50% renewable energy use	High dependency on blockchain security and infrastructure maturity.
J Kavitha et al., 2023	Power-Aware Scheduling, Virtual Machine Migration, Dynamic Workload Consolidation	Energy Efficiency, QoS, Cost-Effectiveness	Focuses on reducing energy waste and carbon emissions while maintaining QoS and cost-effectiveness. Enhances environmentally friendly cloud operation.	Doesn't fully address potential trade-offs between cost and performance across all cloud environments; scalability concerns with large-scale systems.
J. Kavitha et al. (2023)	Dynamic workload	Energy consumption,	Dynamic methods	Trade-offs between

	consolidation, VM migration, power-aware scheduling	QoS, cost-effectiveness	effectively reduce energy waste while preserving QoS and cost control	energy-saving and QoS; need for better adaptive algorithms.
Agos et al., 2023	Deep Reinforcement Learning (DRL) for Autoscaling	Power Usage Effectiveness (PUE), Job Completion Time	DRL-based autoscaler outperforms RL-based autoscalers by optimizing energy efficiency and PUE while accelerating job completion under heavy workloads.	Requires computational resources for neural network training; may have high complexity and may not perform as well in non-simulated environments.
Wesabi et al., 2022	Hybrid Metaheuristic Energy Efficient Resource Allocation (HMEERA), GTOA-RSO	Resource Allocation Efficiency, Performance	HMEERA achieves superior performance in resource allocation, incorporating PCA for feature extraction and hybridization of GTOA and RSO for cloud resource distribution.	Limited to cloud environments simulated in CloudSim; scalability to real-world applications and large datasets is uncertain.
Hao et al., 2021	Dynamic Voltage and Frequency Scaling (DVFS), AODVFS-Adaptive Offloading	Task Completion Rate, Energy Consumption	AODVFS improves task completion rates and reduces energy consumption by prioritizing tasks based on urgency and energy use, with mobile devices and remote cloud resources.	May not generalize to all types of cloud resources or mobile devices; potential issues with DVFS's energy-saving limits.
Aldossary et al., 2019	Cost Prediction Framework for VM	Workload, Power	Provides accurate cost predictions for	May not capture dynamic market

	power/resource consumption	Consumption, VM Cost	VMs, incorporating energy efficiency in cloud system operation, with economic benefits for providers.	fluctuations or real-world workload patterns accurately; performance in large-scale operations or diverse workloads is unclear.

2.3 Green Cloud Architectures

To build a more integrated and energy efficient environment, (Shekhar and Aleem, 2024) aims to investigate the potential to embrace synergies between the cloud computing and the smart grid technologies. Edge computing has the potential to help us reduce the energy usage of cloud services by processing data at the point of origin and reducing the data transfer distances accelerating data processing significantly. Edge Computing Green Clouds looks at this. Blockchain to Sustainable Cloud Computing provides the discussion of the ways blockchain technology can be used to encourage ecological friendliness as well as the improvement of the sense of responsibility and transparency with cloud services. Low-Power IoT Devices for Cloud Connectivity investigates the development and use of low-power IoT devices that facilitate cloud connectivity while using the least amount of energy, hence fostering a greener IoT ecosystem.

(Sadhu et al., 2024) Investigated the cutting-edge technology known as "green cloud computing" (GCC) has drawn more and more attention recently because of its seemingly endless potential to transform the sustainable cloud computing space and, in turn, lessen the environmental impact of cloud computing. To address the detrimental ecological impact associated with cloud computing, this research introduces the Green Cloud Computing (GCC) concept, and emphasizes the significance of the GCC's consideration. In addition, it identifies the several implementation options and recent developments in the Green Cloud Computing literature. The final section discusses the prospects and challenges associated with its implementation.

The dynamic cloudlet-based mobile cloud computing model (DECM), introduced by (William et al., 2024), uses dynamic cloudlets (DCL) to control the extra energy required for wireless connections. In the course of this study, they simulate an incident that might occur in the real world in order to verify this model and provide reliable data for the assessments. This study makes a substantial contribution in two different ways. First off, this study is the first to examine the best practices for addressing energy loss issues in the context of dynamic networking. A team of researchers from Canada and the United States conducted it. Second, the recommended model really encompasses a direction and theoretical foundations for other research to be being conducted in the future.

A new task scheduling technique called NIMS (Neighborhood Inspired Multiverse Scheduler) is proposed by (Tiwari and Beena, 2024) to optimise both makespan and energy usage, two frequently incompatible measures. This enhancement improves solution quality, particularly when MVO's default update mechanism underperforms. Furthermore, through CloudSim simulations across multiple test cases and with three commonly used real-world benchmark datasets, they comprehensively tested the suggested NIMS algorithm versus seven sophisticated algorithms: EMVO, OPSO, IMOMVO, TSIGA, LJFPPSO, MVO, and FPGWO. In five important performance key metrics such as makespan, throughput, energy usage, resource utilization, and load imbalance - this comprehensive simulations and experiments demonstrate that the suggested NIMS algorithm performs noticeably better than the rival techniques.

(Kumar et al., 2024) sought to reduce the environmental effect of data centers using green cloud computing, with research highlighting the importance of using energy-efficient resource allocation techniques. With a focus on resource usage and energy efficiency, the effectiveness of several load balancing techniques, including Round Robin and Least Connections, has been investigated. Simulation tools like CloudSim and GreenCloud are now necessary for modeling and analyzing cloud data centers' energy efficiency. Chang's benchmarking work was one example of a comparative study that helps identify efficient load balancing techniques. All things considered, the project was actively supporting more efficient and environmentally friendly cloud computing, recognizing the importance of finding a balance between environmental concerns and performance enhancement.

The model's formal formulation and related optimization issue for figuring out the best timetable and energy-efficient location of virtual machines on physical servers are presented by (Öner and Özer, 2023). The optimization issue is formulated using integer linear programming, and many heuristic solution methods, including the Genetic Algorithm, are proposed. The performance of the model and the proposed heuristics is assessed using a comprehensive test suite. According to estimates, the suggested model will increase profits by about 37%, and the solution techniques will only produce high-quality solutions within 5% of the ideal range, allowing the model to be implemented in massive clouds.

(Almutairi and Aslam, 2023) introduced a unique Energy and Communication (EC) aware scheduling (EC-scheduler) method for green cloud computing. In order to optimize traffic load and data center energy consumption. Assigning user apps to cloud data center resources and minimizing data center utilization were the main objectives of the proposed EC-scheduler. This simulation will include a Multi-Objective Leader Salp Swarm (MLSS) method to sort the jobs and maintain traffic flow balance, as well as an Emotional Artificial Neural Network (EANN) that illustrates effective resource distribution. They used the GreenCloud simulator to assess the suggested strategy and current cloud scheduling techniques in order to gauge the effectiveness of optimizing data center energy and the other characteristics that are taken into account by the schedulers. However, the EC-scheduler parameters of Power Usage Effectiveness (PUE), Data Centre Energy Productivity (DCEP), Average Execution Time (AET), Throughput, Makespan and Energy Consumption showed up to 26.738, 37.59, 50, 4.34, 34.2, and 33.54 higher efficiency, respectively, as compared to the state-of-the-art schedulers in use.

(Li et al., 2023) introduced an automated platform for training and deploying AI models using cloud-edge architecture. The platform supports data processing, data annotation, training optimization, and model publishing. In cases where gigantic data centers are involved, the proposed strategy has the capability to standardize and even automate model education and create particular models according to the space. Based on simulation and experimental results, the suggested technique is capable of reducing the training time for a single model by 76.2% while also allowing for the simultaneous execution of multiple training tasks (Xu and Buyya, 2020). Consequently, it adapts well to large-scale energy-saving environments and substantially improves model iteration speed, boosting energy efficiency and promoting green energy conservation in data centers.

(Sneha et al., 2023) focused on developing efficient clouds with eco-friendly features including load balancing, virtualization, power management, high-performance computing, green data centers, reusability, and recyclability. Cloud providers now have significant challenges in integrating green computing for computer systems while simultaneously achieving energy-efficiency control, meeting performance assurances, and providing end users with quality of service. An overview of the many green cloud computing architectures is given in this study. As part of the strategy, ways to make the cloud "green" were identified. The challenges and goals of green cloud computing research were also looked at. This study provided the researchers with a state-of-the-art about green clouds.

A novel technique for energy-efficient task scheduling in green cloud operations is proposed by (Khullar and Hossain, 2022). Dynamic Voltage and Frequency Scaling (DVFS) and DVFS enabled Efficient Energy Workflow Job Scheduling (DEWTS) are used in the proposed algorithm to identify the job order and distribute it across the processing nodes. The method makes it easier to utilize CPU idle slots without violating process limits. The heterogeneous finish time of the job is used to determine the deadline, and inefficient processors are combined to maximize slack time. Kubernetes-based containerization is used as a packaging solution instead of virtual machines (VMs), offering advantages over Docker Swarm.

(Doss et al., 2022) suggested the information and communications technology (ICT) industry has grown to be a significant energy consumer and contributes significantly to CO2 emissions. Buying IT equipment as a service with no up-front costs is made possible by cloud computing. This study will first analyze green cloud computing, the combination of these two innovations, based on published research that indicates cloud servers were the most ecologically friendly option for service delivery. Incentives impacting the resource growth of these innovative companies were also analyzed, along with three case studies of their green hosting services. The results indicate that there was now no market demand for green virtual private server hosting because low prices were the main draw for companies. Green service providers should create green service level agreements, compete on greenness, and use transparent efficiency metrics to show their greenness. Green cloud computing and green cloud solutions are consequently required as a result of end customers' demands for more environmentally friendly online services.

The analysis by (Zong, 2020) One significant component of cloud computing systems' dynamic energy consumption is the energy used by servers during task scheduling. Increasing energy efficiency and saving energy are key building blocks for implementing green cloud computing systems. Reducing energy consumption and job execution time are the goals of this work under green cloud computing. The dynamic fusion task-scheduling algorithm proposed in this research is a combination of the ant colony algorithm and the genetic algorithm. hence lowering the energy usage of computing centers and data centers for cloud computing. According to the results of the simulation, the suggested task scheduling algorithm can drastically cut down on the amount of time and energy needed to complete activities.

(Shree et al., 2020) discussed about how cloud computing damages the environment by releasing a lot of poisonous chemicals, such as carbon dioxide. Much energy is squandered and not used effectively. It generates a lot of heat and uses a lot of processors and resources over its lifetime. With its environmentally benign approach to cloud computing, green computing was therefore explored here as a means of resolving this issue and protecting the environment. Finally, they shall see how green computing will expand into cloud computing in the future.

(Olana and Tripathy, 2018) suggested green cloud computing, a novel form of cloud computing that consumes less energy than conventional cloud architecture. As a result, modern technology requires green cloud computing. To select virtual machines, the first population will be the population that contains each virtual machine's execution time and failure rate. Excessive resource allocation in virtual machines boosts the system performance and space consumption of the network. The proposed improvement utilizes the Ant Colony Optimization (ACO) algorithm to identify the optimal virtual machine for cloudlet migration. The algorithm's performance was evaluated using CloudSim, focusing on execution time and energy consumption. Simulation results show that ACO achieves lower energy usage and faster execution compared to the Three-threshold Energy Saving (TESA) algorithm.

Table 2
Overview and Comparison of State-of-the-art Survey on Green Cloud Architectures

Author & Year	Technique Used	Performance Metric	Findings	Limitations
Aishwarya et al., 2024	Cloud Computing and Smart Grid Synergies, Edge Computing, Blockchain for Sustainable Cloud, IoT Devices for Cloud Connectivity	Energy Consumption, Sustainability, Cloud Efficiency	Explores how integrating cloud computing with smart grids, edge computing, blockchain, and IoT can improve energy efficiency and sustainability in cloud environments.	Lacks experimental validation for real-world environments; potential scalability issues with widespread implementation.
Sadhu et al., 2024	Green Cloud Computing (GCC)	Environmental Impact, Sustainability	Discusses the concept and implementation strategies for GCC, aiming to reduce the environmental impact of cloud computing while highlighting challenges and opportunities.	Uncertainty about the applicability of GCC in large-scale cloud environments requires more detailed case studies and experimental evaluations.
William et al., 2024	Dynamic Cloudlet-Based	Energy Consumption,	First study to address energy	Lack of real-world data and

	Mobile Cloud Computing (DECM)	Wireless Connection Energy Loss	loss in dynamic networking using cloudlets; provides theoretical foundation for further research in energy-efficient mobile cloud systems.	potential limitations in scaling the model to large cloudlet-based mobile environments.
Tiwari et al., 2024	Neighborhood Inspired Multiverse Scheduler (NIMS)	Makespan, Energy Consumption, Throughput, Load Imbalance, Resource Utilization	NIMS outperforms seven other scheduling algorithms in terms of makespan, energy consumption, and resource utilization.	Limited scope of evaluation using only CloudSim and specific real-world datasets; may not generalize to all cloud environments.
A. B. P. Kumar et al. (2024)	Load balancing (Round Robin, Least Connections); Simulation tools (CloudSim, GreenCloud)	Energy efficiency, resource utilization	Load balancing improves energy efficiency and utilization; benchmarking helps identify best techniques	Simulation only; lacks real-world validation
Öner et al., 2023	Integer Linear Programming, Genetic Algorithm Heuristics	Profit Increase, Scheduling Efficiency	Model predicts a 37% increase in profits and provides solutions within 5% of the optimal range; supports large-scale cloud environments with energy-efficient scheduling.	May face challenges in real-time, large-scale cloud operations due to the complexity of integer linear programming.
Laila Almutairi et al. (2023)	EC-aware scheduler, Multi-objective Leader Salp Swarm (MLSS),	PUE, DCEP, Throughput, AET, Energy consumption, Makespan	EC-scheduler outperforms state-of-the-art by 26–50% across multiple metrics	Complexity of scheduler design; dependent on simulator accuracy

	Emotional Artificial Neural Network (EANN)			
C. Li et al. (2023)	Automated AI platform on cloud-edge architecture	Model training time, concurrency efficiency, energy-saving rate	Training time reduced by 76.2%; high adaptability and energy savings for large-scale data centers	Focused on AI training; limited insight into broader cloud operations
Sneha et al. (2023)	Power management, virtualization, high-performance computing, reusability, recycling	QoS, energy efficiency	Offers a comprehensive survey on architectures and goals for green cloud computing	Theoretical; lacks experimental data and quantitative results
Khullar et al., 2022	DVFS, DVFS-enabled Efficient Energy Workflow Job Scheduling (DEWTS)	Energy Efficiency, Task Scheduling Efficiency	DEWTS optimizes energy efficiency by utilizing CPU idle slots, with containerization (Kubernetes) providing advantages over traditional VM-based scheduling.	Could be limited by system architecture; results may vary in different cloud infrastructures or real-world environments.
R. Doss et al. (2022)	Case study analysis; green VPS service benchmarking; SLA development	CO_2 emissions, efficiency indicators, market demand	Green servers are more sustainable, but user demand is driven by price, not greenness	Market-driven limitation; user education and incentive lacking
Zong et al., 2020	Dynamic Fusion Task-Scheduling Algorithm (Ant Colony + Genetic Algorithm)	Energy Consumption, Task Execution Time	The proposed algorithm significantly reduces energy consumption and execution time by optimizing task scheduling in	A simulation-based study may not fully capture the complexity of real-world dynamic environments; performance could be different with

			green cloud computing.	varied workloads or cloud configurations.
T. Shree et al. (2020)	Conceptual discussion on green cloud practices	Emissions (CO2), energy waste	Highlights inefficiencies in current data centres; presents the future scope of green cloud	Lacks empirical analysis; high-level overview
J. A. Olana et al. (2018)	Ant Colony Optimization (ACO), VM selection by failure rate and execution time	Execution time, energy consumption	ACO outperforms TESA in reducing execution time and energy usage	Focused only on VM migration; limited system-wide scope

2.4 AI/ML for Sustainable Cloud Computing

(Sutariya et al., 2025) proposed Dragon Fruit Fly Optimization approaches (D-FF) as the means of energy management in the work of nanogrids on the basis of sustainable farming technology. The efficacy of the proposed approach is evaluated on the basis of simulations and real conditions in the field of agriculture. With less energy consumed and reduced operational costs, the results demonstrate that the nano-grid enhances accuracy (96%), F1-score (93%), precision (91%), and recall (92%) while supporting agricultural activities. The outcomes of the development of smart agriculture techniques enable more reliable and efficient energy management in the agricultural sector.

(Kotilainen et al., 2025) developed a proof-of-concept implementation demonstrating the core principles of their reference architecture for an ethical orchestration system managing distributed AI/ML applications across cloud–edge environments. To provide practitioners and automated systems with relevant, actionable information, they extended the existing industry concept of model cards by introducing data source cards and software component cards.

Through this metadata card-based orchestration approach and understanding of target infrastructure risk levels, users can create distributed AI/ML deployments that satisfy regulatory and other requirements.

To overcome these obstacles, (Prasad et al., 2024) suggests a revolutionary combination of blockchain technology and machine learning methods. This framework improves the security and effectiveness of access control systems in cloud environments by utilizing the predictive capabilities of machine learning and the decentralized, immutable nature of blockchain. They created and implemented a model on a simulated cloud platform that automates and optimizes access decisions using neural networks. When compared to traditional approaches, these results show notable gains in threat detection and system adaptability. This study lays the groundwork for future security improvements in the industry while also fortifying access control in cloud computing.

SmartEdge, an AI-powered smart healthcare end-to-end integrated edge and cloud computing system for diabetes prediction, is introduced by (Alain Hennebelle et al., 2024). In an end-to-end system, this work tackles latency issues and shows how effective edge resources are in healthcare applications. The system predicts diabetes by utilizing a number of risk factors. To implement the suggested diabetes prediction models on different configurations, they provide an Edge and Cloud-enabled framework that makes use of edge nodes and primary cloud servers. Response time, accuracy, and latency are used to assess performance indicators. Compared to a single model prediction, they can increase the prediction accuracy by 5% by employing ensemble machine learning voting techniques.

(Panwar et al., 2024) discussed the increasing importance and energy-intensive nature of data centres. Since cloud computing is expanding so quickly, it provides on-demand access to resources all over the world, but it also has a significant environmental impact and uses a lot of power. They use a lot of energy, and maximizing their use was crucial to cutting expenses, protecting the environment, and guaranteeing long-term growth. To obtain resources and enhance energy efficiency, the focus areas were CPU consumption prediction, overload detection, underload estimate, and the selection, migration, and relocation of virtual machines. In order to reduce energy consumption and fulfil service level agreements (SLA), the study examined the energy-saving outcomes of several machine learning approaches in data centres. The machine learning techniques, taking into account different settings and factors, lower energy usage from 1.6% to 88.5% when compared to the benchmark approach given.

(Amahrouch et al., 2024) presented a method that uses machine learning-based methods for VM classification based on their latency sensitivity to make it easier to migrate VMs into a suitable PM for improved VMC. These techniques divide virtual machines (VMs) into two categories: those that may be interactive (showing daily periodic behavior) and those that are insensitive to latency (such as development, test, and batch workloads). With a strong performance and an accuracy of about 83%, this model proved to be the most effective classifier in this investigation.

The goal of (Bhowmik et al., 2023) was to devise a plan for bridging the gap so that AI might be used to help those who had little experience with new technologies. Additionally, they used Google Cloud Functions to simulate a healthcare application in a cloud computing environment to improve its real-time performance and modelled it as a case study to confirm the CloudAISim toolkit's scientific reliability. An interactive web app has also been created with an interface that is easy for non-experts to use. It is a 98% accurate model that may be used by anyone without technological expertise.

For resource scheduling in a cloud computing environment, the Random Forest algorithm—a supervised machine learning model—is suggested. (Manam et al., 2023). This suggested model is an ensemble algorithm that deals with decision tree development for tasks involving regression, classification, and decision-making. The suggested approach is contrasted with other machine learning algorithms that are currently in use, such as XGBoost, Ridge, and Lasso. The experimental results indicate that the proposed method is more successful than the compared algorithms both in regard to CPU prediction and memory prediction. Results of experiments on Google collaboration using the Materna Dataset indicate that the proposed algorithm will have a greater accuracy in prediction of 0.1 -45.1 compared with the compared algorithms in terms of the R-squared (R2), the root mean square error (RMSE) and reduced percentage error under Mean Absolute Percentage Error (MAPE).

(Teng et al., 2023) presented Panicle-Cloud, an open-source cloud-based platform powered by artificial intelligence (AI) that can quantify rice panicles from drone-captured images. To streamline AI detection model development, they initially trained a comprehensive rice panicle dataset, annotated by rice experts. They then tested AI models using images across various orientations and growth phases to determine optimal timing and resolution for field-based rice panicle phenotyping. Their findings showed that the platform achieves high-accuracy PNpM2 quantification by comparing automated and manual scoring results, enabling precise yield classification. They conclude that this work represents a significant advancement in PNpM2 trait phenotyping, providing valuable tools that allow breeders to identify and choose optimal rice varieties under field conditions.

(Xiang, 2022) designed a parallel Naive Bayes (NB) multiclass data analysis technique. Next, in order to maintain high classification accuracy and lower the cost, they modify the number and weight of classifiers as necessary. In the meantime, by recognizing the self-organising control of the data mining process, this research develops an optimal complexity model. The findings demonstrate that the parallel NB classification method may match people's performance needs for big data classification to a certain degree while also offering good scalability and performance enhancement.

A new firewall approach that integrates machine learning with deep learning techniques is presented by (Tiwari and Jain, 2022) to enhance security in cloud computing environments. Their method employs a unique combination strategy termed "most frequent decision" for detecting and categorizing network traffic. This technique merges the ML algorithm's present prediction with previous classification decisions from network nodes to determine the final attack type. Such an approach strengthens both detection accuracy and the system's learning capability. Experiments conducted on the UNSW-NB-15 benchmark dataset demonstrated that this method achieves a 97.68% improvement in anomaly detection performance.

Table 3
Summary and Comparison of State-of-the-art Survey on AI/ML for Sustainable Cloud

Author & Year	Technique Used	Performance Metric	Findings	Limitations
Sutariya et al., 2025	Dragon Fruit Fly Optimization (D-FF) for energy management	Accuracy (96%), F1-score (93%), Precision (91%), Recall (92%)	Demonstrated energy efficiency in nanogrid systems for agriculture, with improved accuracy and reduced operational costs	May not perform well

Kruti Sutariya et al. (2025)	Dragon Fruit Fly Optimization (D-FF) for nano-grid energy management	Accuracy (96%), F1-Score (93%), Precision (91%), Recall (92%), energy use, operating cost	AI-driven nano-grids enhance sustainable farming with improved energy management and reduced costs	Limited to the agriculture domain; requires scalability validation in diverse settings
Pyry Kotilainen et al. (2025)	Ethical orchestration system, cloud–edge continuum	Regulatory compliance, system orchestration accuracy, risk-based deployment	Metadata-based orchestration ensures ethical, transparent AI/ML deployment in cloud-edge systems	Focused on architecture and proof-of-concept; no energy consumption data provided
Prasad et al., 2024	Blockchain technology and machine learning	Threat detection, system adaptability	Enhanced security and effectiveness of access control in cloud environments using machine learning and blockchain for automated decision-making.	Lack of real-world deployment examples
Alain et al., 2024	AI-powered SmartEdge system for diabetes prediction	Response time, accuracy, and latency	Increased prediction accuracy by 5% using ensemble machine learning voting techniques in edge and cloud computing for healthcare applications	No direct comparison with traditional models

S. S. Panwar et al. (2024)	ML for CPU prediction, overload/underload detection, VM selection/migration/relocation	Energy savings (1.6%–88.5%), SLA adherence	ML approaches significantly reduce energy use across various configurations.	Energy savings vary widely; lacks standardization across scenarios.
A. Amahrouch et al. (2024)	ML-based VM classification based on latency sensitivity (interactive vs. non-interactive)	Classification accuracy (\~83%)	Improved VM consolidation (VMC) through precise classification; better energy/resource allocation	Focus on classification; it lacks post-migration energy impact analysis.
Bhowmik et al., 2023	Google Cloud Functions for healthcare applications	Accuracy (98%)	Created an interactive web app with a 98% accurate healthcare model that is accessible to non-experts without requiring technical skills	Reliant on Google Cloud Functions, no alternative platforms mentioned
Manam et al., 2023	Random Forest algorithm for resource scheduling in the cloud	R-squared (R2), RMSE, MAPE, CPU and memory prediction accuracy	Random Forest outperforms algorithms like XGBoost, Ridge, and Lasso, showing up to 45.1% better accuracy in resource scheduling tasks in cloud computing	Limited to the dataset and platform used (Google Colaboratory)
Z. Teng et al. (2023)	AI-powered Panicle-Cloud platform, drone image analysis, phenotyping model training	Trait quantification accuracy, timing resolution, and	High-accuracy rice yield prediction using drone + AI; strong alignment with	Domain-specific (agriculture); energy or cloud

		correlation with manual scoring	manual phenotyping	performance not evaluated
Xiang et al., 2022	Parallel Naive Bayes (NB) multiclass data analysis	Classification accuracy, scalability	Improved classification accuracy and scalability by adjusting the number of classifiers and weights, effective for big data classification	May not perform well on smaller datasets or with limited resources
Tiwari et al., 2022	Machine learning and deep learning for a firewall mechanism	Anomaly detection (97.68%)	Significant improvement in anomaly detection in cloud computing environments by using machine learning and deep learning in traffic packet classification	Focuses only on anomaly detection; might not cover other attack types

2.5 Renewable Energy for Cloud Computing

(Duan and Bu, 2025) proposed a new cloud based stochastic model of optimal positioning and size of hydrogen storage-fuel cells (FC) in radial distribution networks (DN) along with a photovoltaic (PV) and wind turbine (WT) system. To reduce the annual cost of emissions (ACOEM), the annual cost of energy losses (ACOEL), and the annual cost of investment and operation (ACOIO), the optimization problem was formulated as a three-objective function. The New Multi-Objective Enhanced Exponential Distribution Optimizer (MOEEDO) was used to determine the optimal location and size of hydrogen storage-based FCs and renewable energy sources. Moreover, the application of the cloud-stochastic model in order to capture the uncertainties led to an increase of 3.28 percent, 0.028 percent and 6.46 percent of ACOEL, ACOEM and ACOIO, respectively. The framework enhances the efficiency of DNs because it provides the operators with improved optimization and decision-making instruments in uncertain circumstances despite the additional costs.

(Jena et al., 2025) Recommended an advanced system of system integration networked around an LSTM network to achieve energy renewal. The rapid processing of energy data streams using LSTM technology provides correct predictions of user power needs and renewable power generation. Our proposed technique yielded good results with a precision of 0.9, a recall of 0.9, and an accuracy of 92, which resulted in an F1-score of 0.92. Compared to the other popular machine learning approaches, such as Random Forest, Support Vector Machine, and XGBoost, the model offers better predictions. The model also forecasts energy better, and hence, this has made better utilization of sustainable energy sources with minimum stress on the edge devices. This method of managing resources and decision-making assists the edge networks in incorporating renewable power sources and reducing energy wastage. In this model of energy management, a proper scheme is formed regarding the use of AI in the intelligent and eco-friendliness of data processing networks in the management of energy consumption.

(Yang, 2024) proposed a novel method of controlling energy consumption of smart grids as well as analyzing the data transfer security. In this case, the smart grid network energy efficiency is evaluated through renewable energy sources (SM_RES), and the network has been observed to determine the security of cloud computing. Experimental research is conducted in areas of scalability, energy efficiency, QoS, prediction accuracy, power consumption, RMSE, data integrity and average precision, in order to manage energy and network security. The suggested approach had a 95 percent prediction, 77 percent packet loss, 88 percent RMSE, 85 percent mean precision, 97 percent network security analysis, 95.6 per cent scalability, 86 percent QoS, 35 percent power, and 96 per cent energy efficiency.

Applying the Cloud Computing Technology (CCT), (Alghamdi et al., 2023) proposed an integrated system that enables all stakeholders within a supply chain in a supermarket to minimize carbon emissions at low cost and with minimal infrastructure. They have invented a smart shelf installed in a supermarket that is able to detect when food is low in supply and also communicate to retailers in real time through the cloud. By using a single cloud to map the whole grocery supply chain to specific shops, the integrated strategy will decrease the amount of needless supplier trips, increase stakeholder cooperation, and lower carbon emissions.

(Hamdan et al., 2021) focused on the possibility of using cloud computing and solar photovoltaic electricity in the University of Jordan data centre. In order to reduce the operating costs of the electricity used, a net metering solar PV hybrid system was constructed to supply the annual power load. This was one of several scenarios that were proposed for a green cloud computing solution. Cloud servers were used as a substitute solution for traditional servers in order to further reduce costs. It was discovered that the total number of PV panels needed was decreased by over 57% by using this other technique. Finally, it was found that both systems possess a 20-month payback.

(Xu et al., 2021) The aim was to reduce the carbon footprint of data centres through increased use of renewable energy and reduced use of brown energy. They offer an adaptive approach to batch and interactive workload resource management with micro services and renewable energy. A batch workload deferral scheme and an interactive workload algorithm based on brownout are introduced to ensure workload quality of service. They implemented the suggested methodology in a prototype production system and tested it with web services in real-world systems. The results indicate that this methodology can raise the consumption of renewable energy by 10% and reduce the consumption of brown energy by 21%.

To reverse the problem of carbon emissions from data centres, (Xu and Buyya, 2020) examined task redistribution between different clouds in different time zones. They also model solar power for different locations and build the energy usage and carbon footprint of data centres. This guarantees maximum utilization of green energy over locations and minimises the utilization of brown energy. They give a method of managing renewable energy and carbon emissions for several data centers in California, Dublin, and Virginia that are in various time zones. The results show that the proposed methods, including workload shifting, can save carbon emissions by about 40% from the baseline while keeping the average user request response time constant.

(Singh et al., 2015) presented an integrated system utilising the technology of cloud computing (CCT) provides all the stakeholders in the beef supply chain with an opportunity to measure and mitigate carbon emissions on their end while being able to achieve low-cost infrastructure and costs. Coordination between its stakeholders will also be enhanced by the integrated strategy of mapping the whole beef supply chain using a single cloud. From cattle farmers to the retailer, with logistics, an abattoir, and a processor in between, will be the study's system boundary. A simulated case study illustrates the effectiveness of the suggested system.

Table 4
Summary and Comparison of State-of-the-art survey on Renewable Energy for cloud

Author & Year	**Technique Used**	**Performance Metric**	**Findings**	**Limitations**
Fude Duan et al. (2025)	Cloud-stochastic framework; MOEEDO algorithm;	ACOEL, ACOEM, ACOIO	Framework improves DN efficiency under uncertainty;	Higher costs due to stochastic modelling;

	Optimal placement/sizing of PV, WT, FC		slight cost increases observed (ACOEL ↑3.28%, ACOEM ↑0.028%, ACOIO ↑6.46%)	complexity in deployment
S. R. Jena et al. (2025)	LSTM for power prediction; AI-based energy forecasting for edge-cloud systems	Accuracy (92%), Precision, Recall, F1-score (0.92)	Outperforms standard ML models; reduces energy waste and enables smart energy planning	Requires high-quality time-series data; heavy computation on edge devices
Yang et al., 2024	Smart Grid Energy Control with Cloud Computing & Renewable Energy Sources (SM_RES)	Prediction Accuracy, Packet Loss Rate, RMSE, Average Precision, Network Security, Scalability, QoS, Power Consumption, Energy Efficiency	Achieved 95% prediction accuracy, 77% packet loss rate, 88% RMSE, 85% average precision, 97% network security analysis, 95.6 scalability, 86% QoS, 35% power consumption, 96% energy efficiency	Limited to the dataset
Alghamdi et al., 2023	Cloud Computing Technology (CCT) for Supermarket Supply Chain	Carbon Emissions, Infrastructure Costs	Reduced carbon emissions, reduced supplier trips, increased cooperation, and lower infrastructure costs in the supply chain	Limited to supermarket supply chains and potential scalability issues for larger systems
M. Hamdan et al. (2021)	Net-metering PV hybrid system; cloud server replacement	PV panel reduction (%), payback period	57% fewer PV panels required with cloud system; payback period of 20 months	Study limited to a single university site; broader applicability not tested

M. Xu et al. (2021)	Brownout-based & deferring algorithms; workload-aware energy allocation	Brown energy usage, Renewable energy usage	Brown energy ↓21%, renewable energy ↑10%; QoS maintained	Results from the prototype; lacks a large-scale deployment evaluation
Minxian Xu et al. (2020)	Multi-cloud workload shifting; solar power modelling by time zone	Carbon emissions, Response time	40% reduction in carbon emissions while maintaining response time	Dependent on the availability of time-zone diverse data centres
Xu et al., 2019	Distributed Workloads across Multiple Clouds, Solar Power Integration for Data Centres	Carbon Emissions, Renewable Energy Utilization, Response Time	Reduced carbon emissions by ~40%, optimized renewable energy usage, and improved energy efficiency in data centres	Focused on specific data centre locations (California, Virginia, Dublin); may not be applicable globally
Singh et al., 2015	Cloud Computing Technology (CCT) for Beef Supply Chain Integration	Carbon Emissions, Infrastructure Costs	Effective reduction of carbon emissions across the beef supply chain, improved stakeholder coordination	Focused on the beef supply chain, limited to the scope of the study and may not apply to other industries

2.6 Answer to Research Questions

All papers were thoroughly reviewed to extract the required information that is demanded to respond to the research questions mentioned in Section 2.1. The answers to the research questions are as follows

RQ1: To Identify various aspects which affect sustainability in cloud computing

Several factors influence sustainability in cloud computing, including energy efficiency, carbon footprint, resource management, and cooling mechanisms. The adoption of renewable energy sources, such as solar and wind, plays a crucial role in reducing reliance on fossil fuels. Additionally, hardware efficiency, virtualization, and AI-driven workload optimization help enhance resource utilization. Regulatory compliance, green data center design, and geography further affect sustainability by influencing energy consumption and operational efficiency.

RQ2: How to minimize carbon footprints and energy related issues in green cloud computing

Carbon footprints of cloud computing have to be reduced through energy-efficient hardware, AI-based power management, and optimization of workload. DVFS, power-aware scheduling of resources, and consolidation of servers reduce the wastage of energy. Consumption of renewable energy sources and advanced cooling systems, i.e., liquid and free-air cooling, reduce energy consumption. Carbon-aware scheduling and green SLAs also prompt cloud providers to implement greener practices.

RQ3: How to control resources and maintain data security in green cloud computing

Optimized resource utilization in green cloud computing includes auto-scaling, workload balancing, and predictive analytics driven by AI to achieve optimal use of resources. Hybrid and multi-cloud platforms allow workload deployment in an efficient way and ensure greater security. Security of data comprises end-to-end encryption, access control, cloud-safe storage, and compliance certifications (e.g., ISO 27001, GDPR, and NIST standards). Artificial intelligence-driven intrusion detection systems (IDS) and anomaly detection are also crucial security elements in cloud.

RQ4: How to provide data availability and reliability in green cloud computing

Dependable and accessible sustainable cloud computing is realized through fault-tolerant structures, load balancing, and redundant storage methods like RAID configurations and distributed databases. Edge and fog computing minimize latency and maximize real-time processing of the data such that constant availability of the services is guaranteed. Backup and disaster recovery software, data integrity checks through blockchain-based methods, and AI-aided failure prediction also guarantee reliability. Geo-redundant data centers are utilized even by cloud providers in order to avoid downtime due to local failures.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Methodological Framework for Sustainable Cloud Load Balancing

A hybrid load-balancing method uses more than one method, including metaheuristic optimization, artificial intelligence-based prediction models, a dynamic approach, and a static approach to guarantee effective performance and energy savings. Dynamic approaches react immediately to changes in workload, and static approaches offer reliable distribution to avoid instant switching. Use of machine learning algorithms allows the system to forecast peak demand, pre-scale the resources, and avoid energy wastage caused by reactive scaling. Sustainability goals of cloud involve balancing SLAs and environmental goals. An overly aggressive policy of energy saving can compromise performance and user experience, while an otherwise contrasting load-balancing mechanism can meet SLAs by consuming more energy. The cloud load balancing problem is resolved in this paper using a novel hybrid solution. The need arises because of the dynamically evolving needs of cloud computing, like dynamic provisioning of workload, increasing job complexity, and heterogeneity of provided services. For providing effective performance as well as sustainability, it is necessary to investigate new load-balancing methods that are able to match the evolving cloud computing needs and can handle variation in resources efficiently. Load balancing these days is more about conserving energy, reducing carbon emissions, and optimizing underlying hardware. Data centres are heavy power consumers, and inefficient load balancing usually means that idle servers waste energy while some others get overheated. Such wastage translates to increased operating expenses and more greenhouse gases. Therefore, a smart and adaptive load-balancing component occupies a key position in green and eco-friendly cloud initiatives.

To reduce energy consumption and minimize carbon footprints, this work presents a hybrid optimization algorithm. The proposed technique combines Simulated Annealing with Ant Colony Optimization. Simulated Annealing is particularly effective for navigating the solution space and preventing the algorithm from getting stuck in local optima, while ACO, which draws inspiration from nature, adeptly explores the solution space through its distributed approach. Together, these algorithms form a robust and scalable metaheuristic method that enhances load balancing, shortens response times, boosts resource utilization, and lowers energy consumption in Cloud environments. By integrating ACO (Ant Colony Optimization) with SA (Simulated Annealing), we present a refined strategy for managing workload distribution among VMs (virtual machines) in Cloud settings. ACO, which draws inspiration from nature, excels at exploring the solution space collaboratively, much like how ants search for food. Using pheromone trails for communication, a multitude of agents (ants) work together to build potential solutions. This capability allows ACO to adapt dynamically to shifting workloads and resource availability, ensuring that tasks are assigned efficiently based on the current state of the system.

The proposed hybrid model seeks to improve both performance and energy use based on multi-objective optimization that uses the following objectives, weighting each one based on its importance to your specific use case: response time, throughput, the usage of servers, and energy consumption. To make the situation more difficult, cloud infrastructures are also starting to have a trend of reliance on renewable energy. Our proposed system is able to designate workloads to a server or data centre that utilizes clean energy instead of carbon-emitting sources, when preferred. This coincides with cloud providers wanting to lower their carbon emissions to fight climate change. It would also allow for work with a lower level of priority or anything that was considered non-urgent to be planned for use when renewable energy is at its highest availability, limiting reliance on fossil fuels. Overall, the hybrid model would work to provide cost savings as well as green benefits. For cloud providers, energy is one of the most significant ongoing expenses month-over-month, and optimizing load balancing processes lowers PUE directly. Clients will also share in these cost savings, which would make this option more attractive when evaluating sustainable cloud services. Scalability as an attribute is also an important aspect to consider. The hybrid design we proposed is ideal for public, private, and hybrid cloud deployment models, making it useful to small businesses or even the massive cloud operator. The overall flexibility of the solution means that it can adapt to changing workloads over time, whether there is a decreasing trend, expansion of workloads due to seasonal spikes, or the fact that the workloads for different jobs, would vary in usage due to the type of job required at each point in time.

In conclusion, the sustainability of cloud computing is both a governance and compliance challenge, as well as a technical challenge. In recent years, national governments and associations have developed laws and guidelines aimed at improving the energy efficiency of IT infrastructure. The hybrid model proposed ensures that when load balancing, the cloud provider is meeting performance requirements while living up to proactive environmental regulations, by embedding sustainability measures as part of the load balancing process. In summary, this research has set out to develop a hybrid load balancing mechanism that gives equal regard to responsible environmental practices and high performance. The solution seeks to ensure the continued excellence of cloud services while contributing to the goals of sustainable computing through the integration of renewable energy, predictive analytics, static and dynamic mechanisms and multi-objective optimisation. This should also assist the imperative of encouraging eco-innovation in the rapidly expanding cloud ecosystem to improve greener, more resilient, future-ready cloud infrastructure.

3.2 ACO (Ant Colony Optimisation) and its Drawbacks

Ant Colony Optimization (ACO) is a bioinspired community-based meta-heuristic to solve complex optimization problems. It is inspired by the foraging behaviour of ants, which use pheromone trails to find the shortest distance between their colony and a food source. The strength of the pheromone trail increases the probability that other ants will follow the same path, and thus, in time, a best or near-best solution can be discovered by the actions of the colony. ACO has been effectively used for computational problems involving routing, scheduling, work assignment and load-balancing because of its adaptability and distributed nature. Although there are many advantages to ACO, its significant limitations should be taken into consideration, which could diminish its effectiveness in large and extremely dynamic environments. The first and most recognised limitation includes convergence speed, which can severely limit the effectiveness of the strategy when working on larger scale optimization problems where there is a large, complex search space. The strategy's probabilistic exploration scheme is very effective at preventing direct convergence and commitment to a sub-par solution, however, can result in significant lag in the convergence speed of the algorithm. ACO also tends to get stuck in local optima, which will also result in premature convergence, and not fully exploring the global search space. The source of this limitation is that ants which discover a highway with high pheromone level will exploit the trail and ignore possibly better choices elsewhere.

In order to address these limitations, researchers have proposed the use of ACO in conjunction with other optimization approaches that mitigate the disadvantages of ACO. One such proposed hybrid approach is to incorporate the ACO algorithm with Simulated Annealing (SA) technique, since SA can escape local optima using probabilistic jumps when trained carefully, while ACO combines multi-agent distributed searches with positive feedback generation. In this hybrid approach, SA's features can be combined with ACO's features, so as to benefit from both methods reliably. ACO is inspired by nature; specifically, natural ants lay pheromone trails and follow these trails in order to determine the most efficient routes. ACO inherits segments from the natural world and is ideal for cloud computing resource allocation and scheduling because it contains a structured, yet flexible and adaptive way to explore the complex area of solutions. The algorithm also naturally supports incremental parallel computing because artificial ants (agents), such as SA sequentially search for the best solution while being able to explore different parts of the solution space independently from each other. This form of parallelization reflects the distributed nature of cloud computing paradigm, where multiple resources work together and process different workloads. The ACO pheromone update rule is particularly useful when addressing load balancing. Ants are heavier on pheromones, which reinforce stronger (more preferable paths) and less preferable ones, while they travel down the path and discover better solutions. Many researchers have acknowledged that this continual improvement leads to the algorithm being able to adapt to variations in the problem space over time. Certainly, should these less preferable paths acquire excess pheromone too prematurely in the search, the same reinforcement mechanism could lead to stagnation.

The Simulated Annealing algorithm has another mechanism that can help alleviate this stagnation. Specifically, Simulated Annealing's "temperature" allows it to control the probability of acceptance for worse solutions, which is analogous to an annealed state in physical material processes in metallurgy. While the temperature is "high", SA will allow and accept worse solutions, furthering the search's ability to avoid local optima. The search becomes more exploitive via SA, while there is reduced temperature, taking advantage of the best solution found thus far. It is important to keep exploring (finding new regions of solutions) and exploiting (improving beyond existing good solutions) in highly saturated and unpredictable environments, such as cloud computing. The ACO can be the main exploration operationally by having the distributed agent find candidate solutions quickly. SA's probabilistic acceptance rule can then be applied towards a second refinement step on these candidates to avoid premature convergence. By having both, the algorithm can always be heading toward improved solutions while also maintaining a diverse search population. Another advantage of the hybrid approach is that SA supports solution refinement later in the search after the potential ACO completed convergence on a local optimal solution.

This hybridisation of cloud load balancing applications guarantees workloads are placed on existing resources as efficiently as possible while allowing flexibility to accommodate changes in resource availability or demand. The SA activity is able to quickly reassign tasks based upon performance indicators while monitoring the load based upon real-time constraints, and therefore, the hybrid is a good way to alleviate bottlenecks and to utilize the best available resources. Additionally, the SA's probabilistic nature enables a good fit to unpredictable demand of cloud workloads. In conclusion, the combination of ACO and SA through ACO's global search characteristics and SA's local improvements and escape features creates an algorithm capable of tackling the expansive, dynamic and highly multi-modal optimization landscapes prevalent within cloud computing load balancing. The combined features produce robustness, scalability, and adeptness (as well as increased speed of convergence and quality of solution) - key elements for effective and sustainable management of cloud resources.

3.3 Hybridization of ACO and Simulated Annealing Algorithm

Ant Colony Optimization (ACO) and Simulated Annealing (SA) combined provide a viable solution for solving the load balancing problem in cloud computing. Ant Colony Optimization (ACO) is a metaheuristic for complex solution spaces based on pheromone trails to find an optimal path. The algorithm was inspired by foraging behaviour in ants, in which agents independently lay down Optimal Trails to Food Sources. This results in distributing work selectively across the (virtualised in a cloud context) VMs based on their respective capabilities and current loads. ACO constantly updates the pheromone map to guarantee task allocation decisions are adapting to real-time changes in workload. Conversely, Simulated Annealing (SA) is a probabilistic method for using the annealing process or SA dynamics to find almost global minimums, introduced by the cooling process used in metallurgy, which brings the system into a solid solution with long-range order. In computational optimization, simulated annealing uses the principle of gradual cooling to get out of enclaves of local optima and explore further from solutions. This is done by sometimes accepting sub-optimal solutions, which decreases as time goes on and increasingly boosts the search towards a global optimum.

This kind of hybrid ACO-SA approach, global search capabilities are due to ACO, whereas from a fine local search is the result of SA. The cloud environment has to be set up first (i.e., the number of VMs, tasks) and some parameters, like the speed of execution is given for each task and a range within which it can move without incurring communication cost. ACO first initializes its environment and then seeds the process with numerous artificial ants to suggest the degree to which each task can be assigned adequately to a VM. It is like how these ants, by collectively accessing their current loads and pheromone levels, generate a few candidate allocational solutions. Next, the phase that incorporates SA into the process. All solutions that are generated by ACO enter an SA stage, where each solution is refined in an iterative process. SA intuitively explores neighboring configurations by making a small change in the task assignments —a task might be shifted to a less over-loaded VM, or maybe various tasks are re-preferred to increase throughput. The choice of whether to accept these variations is determined by the temperature parameter: at high temperatures, an algorithm stands a better chance of accepting a poorer solution to encourage searching; as temperature drops, the algorithm is more selective, involving only exploitation of the better parts of the solution space.

This two-tiered process contributes to resilience against workload variations of unpredictable nature, viral failures or VM, and collapses in the resource availability. The pheromone update mechanism is essential in this. Following every iteration, pheromone concentrations are modified depending on the quality of solutions that are retrieved. More efficacious allocations are reinforced more with pheromones to push ants towards more productive layouts. With time, this feedback mechanism raises the chances of producing balanced and efficient distributions. The main advantage that can be appreciated regarding this hybrid strategy is its flexibility. Whereas ACO offers fast convergence to come up with good solutions, SA discourages premature convergence by enabling controlled exploration even towards the end of the search. Such an aggregation will result in reduced wastage of resources, overloading specific VMs, and improved quality of service (QoS) on the part of end users. The hybrid ACO-SA may specifically be better used in real cloud computing scenarios when one needs to find solutions to the multi-tenant problems, when the workload is extremely variable, and the priorities among tasks vary. As an example, latency-sensitive uses like real-time video streaming or important decisions-to-be-made need instant allocation decisions, whereas jobs that fall into the batch-processing category might be happy to endure somewhat longer delays with energy-efficient processing. These conflicting goals can be balanced using the hybrid method, which is dynamic.

To confirm the proposed method, an extended set of simulation experiments is desirable to be carried out on different workloads, which may comprise benchmark tasks in synthetic form and actual real-world workloads of large cloud systems. Quantitative evidence of performance improvement will be obtained by comparative benchmarks induced with globally known algorithms, like Round Robin, Throttled Load Balancing, and pure ACO or SA implementations. The metrics to be used during evaluation should be load balancing efficiency (through standard deviation of VM utilization), average response time, sensitivity to energy consumption per task being completed, throughput level, as well as the violation rates of SLA (Service Level Agreement). It has been preliminarily estimated that the suggested hybrid ACO-SA would beat the traditional solutions both in a homogeneous and heterogeneous VM setup.

Diffusion of pheromone intensities in response to the changes in the workload (which is available to the algorithm) will most probably result in a more even and stable set of the overall loads, especially given the capacity of SA to optimize the solutions. In addition, the approach could allow less idle resource time, and thus energy savings, by enhancing the efficiencies of task scheduling, which is a key to sustainable operations in cloud computing. Scalability. In terms of scalability, hybrid can be generalized to multi-cloud and edge technologies. Pheromone trails in this situation can be used to refer not only to VM-level assignments but also to inter-cloud routing preferences, and SA can be used to optimize migration decisions across geographically distributed resources. Furthermore, there are parameters which may be dynamically adjusted during execution using adaptive control mechanisms or even reinforcement learning optimization, including pheromone evaporation rate, cooling schedule in SA, and neighbourhood search strategies, in order to further optimize performance over time. To sum up all, the ACO-SA hybrid approach is a middle approach between exploration and exploitation cloud load balancing. It makes use of swarm intelligence to discover solutions in the entire world and stochastic local search to optimize them by getting precise and accurate solutions. In addition to solving existing problems with resource assignments, the approach will offer channels to future advances in self-orchestration of cloud systems, self-healing infrastructure and energy-opportunistic service provision. Such hybrid optimization techniques will become ever more central to the quest to achieve efficiency, reliability and sustainability as cloud systems move increasingly toward more dynamic and resource-centric organization.

3.4 Proposed Research Architecture

The utilization of ACO and SA together to balance the load can therefore provide an effective solution to the challenges that abound in the dynamic task assignment aspect of cloud computing. Exemplified by Simulated Annealing, which narrows the solution space and prevents the possibility of being trapped in local optima, nature-inspired algorithms such as ACO provide wide and effective coverage of the possible allocations by virtue of their distributed search. These methods, in concert, produce a scalable metaheuristic algorithm which not only optimizes load balancing but also provides the means of facilitating sustainable clouds due to optimized energy efficiency, hardware longevity and decreased redundant computation overhead costs. Figure 7 shows the architectural proposal. This system manages a flow of user workloads in the system arriving through the internet and sending this flow to the load balancing module, where the hybrid ACO-SA algorithm decides on the best assignment to VMs.

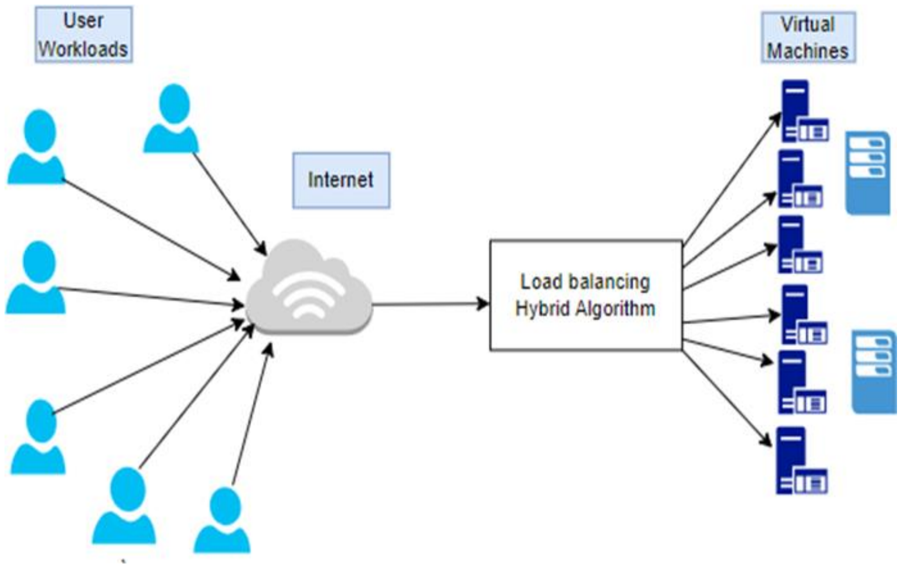

Figure 7
Proposed Architecture

The combination of ACO and SA (Simulated Annealing) for managing the load of the cloud load balance is a complex and environmentally-friendly approach in addressing the issue of workload distribution. ACO is good at distributed search, that is, agents use indirect communication to create the most efficient solution as described by the foraging behaviour of ants. It is also suitable in varying, heterogeneous cloud environments because of its dynamic nature that appropriately adjusts to fluctuations in workloads. Nonetheless, ACO in isolation can have convergence problems, especially that of finding an optimal task allocation prematurely. SA overcomes this shortcoming by probabilistically exploring via the slow-cooling process of metallurgy to allow the occasional acceptance of a suboptimal move in order to avoid local optima.

The ability to fine-tune, given the capability of SA that the hybrid will bring, will ensure that the load is allocated not only efficiently but also with energy conservation. As an example, energy consumption profiles of various VMs may be considered by the algorithm, the energy-consuming being either allocated to more power-efficient equipment or to the machines in facilities with renewable energy sources. This is in line with the principles behind green computing and will drastically cut down the carbon footprint of the cloud system. During the first exploration phase, ACO relies on real-time measures, including CPU domain, memory space, network latency, energy intake rates, etc. and generates various allocation tactics. This is further optimized with the help of SA through the iterative optimization of the solutions to get a balanced distribution of the loads on VMs and the minimum excess energy spikes due to frequent migrations of tasks. The temperature parameter of SA can even be made adaptive with reference to saving energy, where it can be made fast during low loading conditions to save computational cost and vice versa.

Various benefits can be associated with this hybrid method in terms of sustainability:

- Reduced Energy Waste: The algorithm provides a smart allocation of workloads to VMs that have the best power characteristics, thus reducing overall system power consumption, and not affecting performance.

- Hardware Longevity: Their distribution of processing loads to an even balance means that no machine experiences the heat strain that makes it fail sooner rather than later thus increasing the hardware usage life and limits e-waste.
- Scalability with Minimal Overhead: At scale the hybrid architecture can still add resources to meet the demand without being overprovisioned and wasting energy or cost.
- Carbon Footprint Reduction: The ability to run tasks in locations or data centres using renewable power where available, makes sustainability part of the decision making layer.
- Operational Cost Savings: The reduction in the energy bills and replacement of hardware at a later date save the operations additional cost, making economic sustainability and environmental responsibility the same.

The ACO-based global search and local optimization of SA make constructs a system that can continuously adjust in the face of changes in workload and the availability of resources, required nowadays in sustainable cloud infrastructures. The hybrid approach has proactive capabilities to rebalance workloads in situations where workloads vary according to time zones, market demand or seasonal swings towards a high energy efficiency. Also, the introduction of sustainability-awareness pheromone updates will help in ensuring that the solutions that provided a high performance as well as low energy consumption get stronger punishments, which makes them more likely to be picked during other iterations. In the long term, this form of reinforcement learning can be viewed as making cloud operations consistent with medium-term environmental objectives. The suggested ACO-SA hybrid scheme is not merely a high-performing load balancing solution, but it is also a strategic facilitator in relation to the sustainability of cloud computing. It ensures meeting the twin objectives of performance maximization and ecological decimalization, both by making energy awareness an intrinsic element of the optimization process. Such sustainability-oriented optimization algorithms will be key to realizing the next-generation green, efficient, and resilient cloud infrastructures as the energy requirements of data centres are expected to increase over time across the globe.

Pseudo-code of the Algorithm:

Algorithm 1: Hybrid Algorithm

1: Start

2: Define fitness function f(·)

3: Create an initial population of ants,

4: Evaluate all individuals in the population,

5: Define ξ, n, m and the number of iterations T,

6: t:= 0,

7: **while** t < T **do**

$T=T_0$;

While ($T<T_{final}$)

{

Until (N<I-Iter)

{

Generate solution S' in the neighborhood of S.

If f(S'^{*}) < (f(S))

S <-S'

Else

New Function=f(S')-f(S)

R=random ()

If (r<exp (-New function/K*T))

S<-S'

}

T=a*T;

}

Return the best solution found.

8: **for** each ant **do**

9: Calculate the probability using Equation (1),

10: Find the best nest by Equation (2),

11: Determine Gaussian sampling according to Equation (3),

12: Create m new solutions and destroy m the worst ones, 13: **end for**

14: t + +,

15: **end while**

16: Sort individuals according to f(·)

17: Return the best ant,

18: Stop

SA (Simulated Annealing) and ACO (Ant Colony Optimization) are combined to form a hybrid approach to cloud load balancing that is described by the pseudocode. At the onset of the operation, critical variables for both ACO and SA are initialized, such as the number of ants, maximum number of repetitions, beginning temperatures, pheromone levels, and the rate of cooling. During the ACO phase, several ants collaborate to find solutions by sequentially and iteratively assigning tasks to virtual computers (VMs) based on pheromone concentrations and the heuristic data until all jobs are distributed. Each ant evaluates the fitness of its constructed solution, and the pheromone levels are subsequently updated according to these solutions' quality. When an ant discovers a source, it returns to the nest while leaving a pheromone trail that aids in finding the source again. However, the amount of pheromone diminishes over time due to evaporation. The ant moves toward the selected individual from the population. The probability of choosing the j-th individual in the population is determined by the expression:

$$p_j = \frac{\left(1 + \exp\left(f(\overline{x_j})\right)\right)}{\sum_{r=1}^{n}\left(1 + \exp\left(f(\overline{x_r})\right)\right)}$$

Using probability calculations, a colony c is chosen as follows to determine the movement's direction:

$$c = \begin{cases} \max_{i=1,2,\dots,n}(p_i), & q \le q_0 \\ C & q > q_0 \end{cases}$$

where C is a randomly selected ant in the population, $q_0 \in \langle 0,1 \rangle$ denotes a parameter. The q specifies a randomly selected value in the range 0 to 1. Gaussian sampling is carried out following the selection of parent colony c. The following equation models the dispersal of pheromones throughout the space by employing the density function.

$$g(x_i^m, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu)}{2\sigma^2}\right)$$

The selected coordinate for the m ant $\overline{x^m} = (x_1^m, \dots, x_n^m)$. Here, μ is equivalent to x_i^j and it represents the coordinate for the chosen j-th ant. The following equation can be used to compute the mean distance (σ) between the coordinate data of a point $\overline{x^m}$ and $\overline{x^j}$

$$\sigma = \xi \sum_{r=1}^{n} \frac{|x_r^m - x_r^j|}{k - 1}$$

Here, the evaporation rate is ξ. At that time, the m ant population is produced in $N(\mu_i, \sigma_i)$ and the deletion of worst m individuals is carried out. After the ACO stage concludes, pheromone evaporation is applied to further explore the solution space and avoid premature convergence. The algorithm then transitions to the SA stage, with the goal of refining the best solution recognized during the ACO stage.

The Simulated Annealing (SA) algorithm is conceptually modeled on the metallurgical process of annealing, where a substance is carefully heated and then cooled to produce larger, more perfect crystal structures (Javadi et al., 2022). Because of the delayed cooling, the atoms are more likely to discover configurations with lower energy than the previous one. So the heat leads the atoms to escape their original positions (a local minimum of energy) and migrate randomly. It makes a probabilistic decision after taking into account a few of the current states' neighbors in each cycle. The metropolis heuristic is generalized by the SA algorithm. In fact, SA consists of a series of city executions with temperatures gradually dropping from a very high temperature, where practically any move is allowed, to a low temperature, where the search is similar to HC. The pseudo-code for Algorithm 1 shows that it does, in fact, have an internal escape mechanism from local optima, much like a hill climber. If $\delta \leq 0$, meaning that $\delta = f(s') - f(s)$, then the solution s' is recognized as the new current solution in SA. If T is a parameter known as the "temperature," then movements that enhance the energy function are accepted with a decreasing probability exp $(-\delta/T)$if $\delta > 0$. This makes it possible to get out of a local optimum. The temperature levels at each step of the algorithm are specified by a cooling schedule, which regulates the lowering values of T. The choice to employ a proportional approach, such as $T_K = \alpha \cdot T_{K-1}$, is a common choice for its implementation. In practice, SA typically produces superior results, but its application is relatively slow. The most notable challenge in using SA is determining and fine-tuning its characteristics, such as the starting and ending temperatures, temperature degradation (cooling schedule), equilibrium detection, etc. During the refinement phase, neighboring solutions are generated through perturbation and evaluated against the current solution. The new solution is approved if it shows increased fitness; if not, it might still be acknowledged probabilistically, depending on the fitness change and the temperature, which decreases progressively over iterations. With the cooling function, the search becomes more focused, allowing the algorithm to concentrate on important points of the solution space. Once all iterations are complete, the best solution returns by showcasing an optimal load balancing approach. This outcome depicts the effectiveness of uniting exploration abilities of ACO with the refinement methods of SA to enhance the distribution of resources in cloud systems.

3.5 Evaluation Parameters

Appropriate parameters were set with consideration to the objectives of the research to determine the proposed hybrid algorithm and compare its performance with the other existing methods of task scheduling in cloud computing. The reason why these parameters were selected is to determine both the effectiveness of the operation and the possibility of such resource management in the long term, making sure that the algorithm will not only provide faster work but also cause the least harm to the environment and the economy. This main objective of the algorithm is to successfully place incoming tasks, minimize the amount of time (Makespan) spent in the execution of the tasks, maximize the utilization efficiency of the available resources and stability of the system. Also, the assessment will include convergence analysis to determine stability and rate at which the optimization procedure converges towards near-optimal solutions. Such criteria make it possible to conduct a full-fledged evaluation of the appropriateness of using such an algorithm in real-time workloads in the cloud.

I. Makespan

Definition: The total duration between the time at which the first task commences and time at which the final task is finished, plus scheduling is called makespan. It indicates how well resources were assigned and the rate of the scheduling algorithm.

Relevance: In cloud modeling, make span reduction is very vital with respect to shortening job waiting time and rapidity of response in the system, and also adding or increasing satisfaction amongst users.

Effects of the proposed Hybrid algorithm: The new algorithm combines exploration feature of ACO and fine-tuning aspect of SA, to ensure the perfect allocation of tasks to the most suited VM, minimize time during which VM idles, and strike the balance of task execution to all the VMs. Decreasing Makespan results in a decreased level of energy consumption indirectly as well, because machines will not remain in active processing modes as long as before.

II. Resource Utilization

Definition: Resource Utilization is used to quantify the percentage of available resources that are currently under an active task having a requirement in the computational resources on the cloud environment.

Relevance: There must be high utilization as it means that the infrastructure is being exploited to the maximum of its capacity and where utilization is low, then some malfunctions could be due to inefficiency in scheduling or overprovision of resources. Nevertheless, the practice of overutilization may cause hardware to stress, overheat, and have a shorter lifespan of its components.

Effect of the Proposed Hybrid Algorithm: The algorithm takes into consideration of the ongoing workload and both the capacity of every VM and tasks are allocated in a way that each of the less loaded machine will not create a bottleneck. Such dynamic allocation facilitates the best use, which does not lead to overuse or underutilization. As in the case of sustainable cloud computing, the balanced utilization helps to avoid wasting energy and allows extending the life cycle of hardware, thus avoiding replacement requirements on a regular basis that generates e-waste.

3.6 Design Specification

This work presents, an efficient technique of load balancing within a cloud computing environment using a hybrid meta-heuristic Ant Colony Optimization-Simulated Annealing (ACO-SA) algorithm. The main goal of the recommended system aims at enhancing cluster response time, resource utilization as well as ensuring that all the virtual machines (VMs) have their share of work. This has been achieved by integrating the capabilities of exploratory search afforded by ACO with the solution refinement and the avoidance of local optima provided by SA. ACO is very good at global exploration, which is modeled on the foraging behavior of ants, however, ACO may at certain times converge too early as a result of pheromone bias thus SA on the other hand is effective in refining the solutions but not nearly enough global power is required in large scale scheduling. The hybrid algorithm thus has the capability of exhaustive exploration as well as reasonably specific exploitation because it combines the two methods, which makes it perform better in a dynamic and heterogeneous cloud environment.

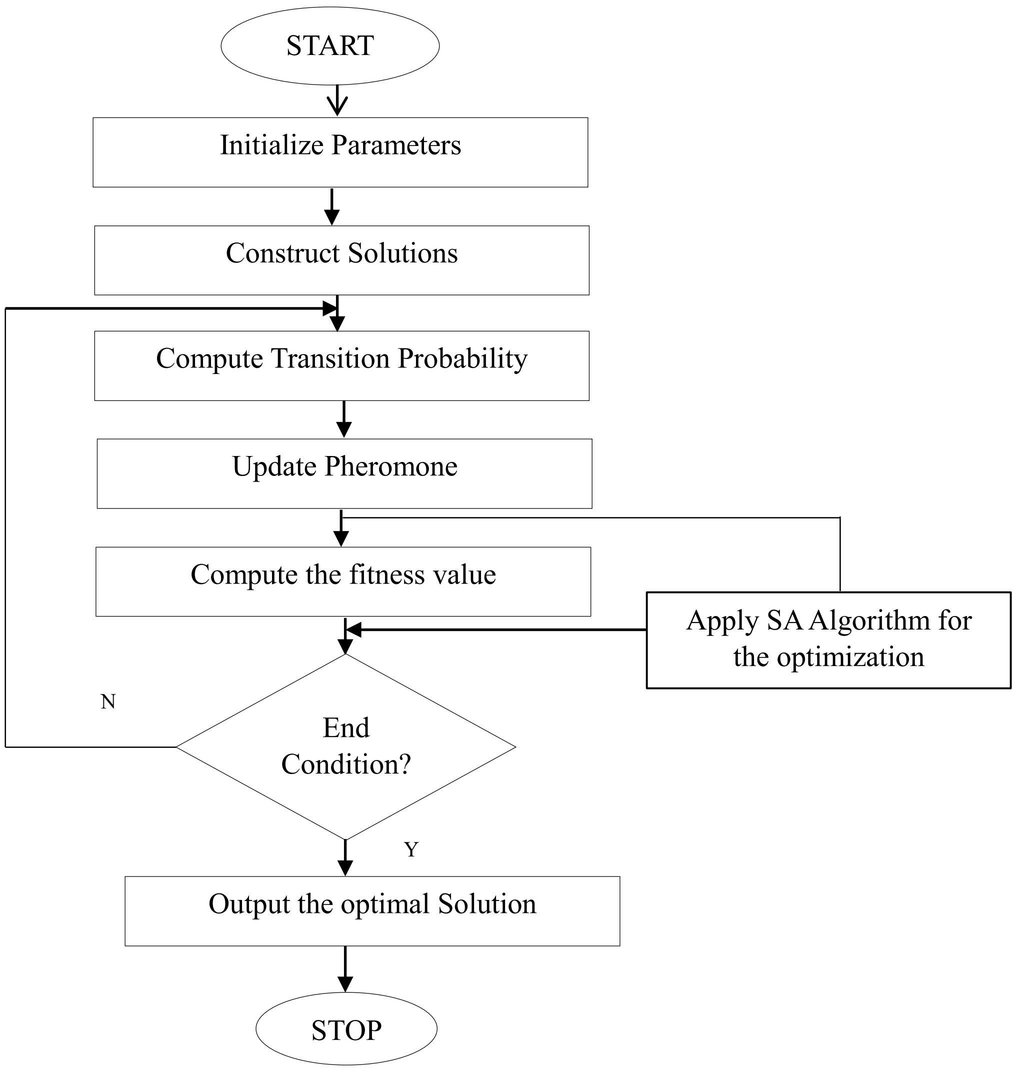

Figure 8
Flowchart of the proposed ACO and SA algorithm

The system consists of three key elements: a work load reservoir that has multiple streams of incoming workload, a pool of heterogeneous VMs of different processing capacity and a hybrid optimization engine consists of different but complementary ACO and SA phases. The procedure commences with parameter initialisation like the population size of the ants, maximum number of iterations, pheromone evaporation rate, heuristic weight, initial temperature of SA and cooling schedule. At the ACO phase, every ant prepares a candidate solution mapping the tasks to VMs depending on the transition probability formed of the pheromone level and the desirability of mapping as assessed heuristically. When a fully formed allocation is found its quality is estimated, and pheromone trails are re-weighted, reinforcing the good allocations and letting pheromones evaporate to promote exploration.

After the exploration period, the optimal solution identified by the ACO phase is passed to the SA refinement phase but here only minor changes are checked to come up with neighbours of the solution. New solutions which are positively better will be embraced, but those which are inferior are potentially accepted with probabilities which decline as the temperature parameter becomes colder. Such probabilistic acceptance causes the search to jump out of local optima and can find possibly superior areas of the solution space. The temperature reduces with each successive iteration, and the algorithm switches from having exploratory tendencies to having a narrower search and stabilizes to high quality solutions.

Such a predetermined end condition terminates the process when it is achieved e.g. reaching the predetermined maximum number of iterations, or seeing no significant increase in fitness during a specified number of iterations. When the termination condition is reached, the algorithm will give the best task-to-VM mapping. This mapping is measured against some fitness criteria that takes into consideration a variety of objectives of Quality of Service (QoS), such as least makespan, maximum utilization to resources, as well as the equitable distribution across the assignment of the tasks to VMs. Hybrid ACO-SA algorithm therefore offers a scalable, adaptive and resilient solution to cloud load balancing ensuring that load balancing can dynamically respond to the workload changes and VM availability and prevents corner cases of premature convergence and inefficient resource utilization issues.

A major strength of the proposed hybrid ACO-SA method is that it opens up a way of satisfying a balance between exploration and exploitation better than either of the explored methods working alone. The ACO component guarantees an extensive search of the solution space, finding good allocation patterns quickly, whereas the SA component fine-tunes the solutions by sometimes accepting a less desirable move to explore possible unexplored regions. Such synergy manifests itself in superior task assigning, minimization of makespan and effective use of available resources, even when faced with variable and uncertain numbers of tasks to work with. In addition, a hybrid approach can be well applied in a dynamic cloud setting as it re-optimizes on-the-fly allocations when the VM capacity or workload intensity changes. Integrating the global search and the local refinement, the ACO-SA algorithm not only increases the speed of convergence, but also preserves the accuracy of the solution, and thus is especially useful in situations of large-scale resource scheduling in cloud environments where resources should be both efficient and stable.

3.7 Implementation

The research work is based on load balancing using a hybrid optimization algorithm. The proposed model was simulated using CloudSim integrated with the Eclipse platform. The experiments were conducted on a Windows 11 machine equipped with 8 GB RAM and 256 GB hard disk storage. Multiple simulation parameters were taken into account to assess the performance of the proposed model.

I. Tool used for Evaluation

CloudSim is an open-source cloud computing modelling, simulation and evaluation framework. It was designed at the Cloud Computing and Distributed Systems (CLOUDS) Laboratory at the University of Melbourne and written in Java; therefore, it is platform independent and available to a large audience of users. CloudSim enables researchers and developers to simulate and prove the workings of a new cloud-based algorithm or architecture and scheduling policies without having to deploy an actual cloud setting, along with the complexity and cost that it incurs. The framework is well-supported by simulating parts of cloud computing such as data centre operations, provision of virtual machine (VM), resource allocation, network topologies and energy-aware computing. It is scalable and flexible, and allows users to simulate small to large-scale cloud infrastructures. The modularity of CloudSim means that its functionality can be extended relatively simply, configured to suit particular requirements, and fitted with any of a range of scheduling or optimization algorithms so that performance can be tested. Among the key benefits of CloudSim, the possibility of obtaining reliable and reproducible test results in a controlled environment is listed. Such as of special importance when the performance of an algorithm, such as the proposed ACO-SA hybrid model, must be evaluated in multiple scenarios, different workloads and configurations. The fact that CloudSim is free and open source does not impose any licensing issues, making it a preferred option in academic and research fields when running cloud performance tests.

3.8 Formulating the Ant Colony Optimization Algorithms

The SA load balancing algorithm is applied in conjunction with the ACO technique. The objective function produces the final optimized outcomes when the ant colony optimization algorithm receives the initial metrics as input.

The Java code contains the Ant Colony Optimization (ACO) algorithm in order to carry out optimal load balancing within a cloud computing system. Such implementation is aimed at resembling the foraging mechanism of ants to find the best approaches to allocating tasks. In this file, the main aim of ACO is integrated with the purpose of using its key objective function with the intent of minimizing makespan, maximizing the use of the resources and ensuring that the workload is thoroughly distributed among the virtual machines (VMs). The whole procedure entails the development and enactment of three major functions, which are essential to the working of the algorithm, which are initialization of pheromones, update of pheromone values and production of optimized solutions. In the pheromone initialization phase, the algorithm initializes the pheromone values depending on the current machine status, like processing capabilities of the machine and the number of cloudlets to be executed in that machine. This will help to ensure that the algorithm has a currently available basic representation of available resources to start with in the optimization process. The earlier pheromone values guide the choice of the ants during the early stage of exploration, to probabilistically decide on different paths of allocation.

The pheromone update function performs the duty of updating the value of pheromone in a dynamic manner using an iteration of the algorithm. When every ant is done with their process of constructing a solution, the outcomes of each of them are measured against the objective function. Whenever an allocation can lead to smaller execution times or improved balance of loads, the ratio of correspondingly higher pheromone investments is higher. On the other hand, the evaporation is done using pheromones in order to diminish the effects of old and inefficient routes. This update mechanism is continuously used to ensure that the algorithm strikes a balance between exploring new solutions and exploiting the best solution.

The last important functional aspect in the Java code is the generation of optimized values. In this case, the algorithm builds the solutions iteratively by assigning the cloudlets within VMs with a relation of pheromone intensities, as well as on heuristic information, e.g. VM processing power or VMs that are ready. With repeated iterations, the solutions become more optimal or near-optimal task assignments. These generated optimized values are then adopted in actual scheduling to make sure that distribution of workloads is done evenly, delay of processing is reduced, and avoidance of instances where a VM becomes overloaded as other VMs idle.

To further the robustness of the implementation, the ACO algorithm is coupled with adaptive parameters, including dynamic evaporation rates of pheromone in addition to probabilistic rules of decisions. Such adaptive settings enable the algorithm to react to changes in the workload patterns in real time, and thus, such an algorithm is applicable to dynamic and unpredictable cloud environments. Java code could also have been seen as easily integrated with other optimization methods, e.g. Simulated Annealing, to further optimize the solutions since it is modularly designed. This elasticity means that the implementation is scalable or can be adjusted to alternate cloud divisions, resource limitations or Quality of Service (QoS) needs. With this implementation, the algorithm performs better than the conventional static load balancing methods in terms of faster convergence and quality solution at an equal footing of balanced resource utilization. The Java code, in its modelling approach of how ants behave and respond to real-time cloud conditions, is instrumental in the best distribution of workload and hence greatest functionality in cloud computing platforms.

3.9 Checking the Initial Simulation Parameters

The initial simulation parameters create the basis for modelling and evaluating the proposed hybrid ACO-SA algorithm in the CloudSim simulation environment. These parameters define the basic setup of the virtual machines (VMs) and establish the environment in which task allocation and load balancing will be tested. Specifying these parameters correctly is essential because it affects the accuracy, reliability, and repeatability of the experimental results.

In the Parameter Function of the code, the simulation starts by setting up key attributes such as the operating system type, VM category, and hardware details. The virtual machine identification number (vmid) is set to 0. This ensures the VM is uniquely recognized within the simulation's resource pool. The VM's computational capacity is defined by its MIPS (Million Instructions Per Second) rating, which is set to 1000 in this study. This represents a medium-performance computing resource suitable for typical cloud workloads. The VM image is 10,000 megabytes in size. This size can contain the capabilities of the operating system, basic software libraries, and application packages that are required. The VM is provisioned with 512 megabytes of RAM to facilitate the efficient execution of that application without extending beyond our underlying infrastructure capabilities - meaning there is memory for processing - it is not overcommitting the infrastructure. The VM's network performance is provided by a bandwidth = (bw) of 1,000 units. This also demonstrates an adequate amount of throughput for the task data movement and inter-VM communication, and service request processing in a simulated cloud equivalent. The pesNumber is set to 1, which means the VM will consume a single processing element (PE) or CPU core. The pesNumber = 1 setting provides the benefit of concentrating the simulation on task scheduling executions instead of the complexities of the parallel execution from task UVs. The virtualization layer being used by the VM is Xen VMM, which provides the virtualization capability by managing the execution, resource sharing, and required separation between the different virtual environments. The selected parameters represent a reasonable starting point for the empirical investigation, for a simulation that can also demonstrate realistic operational limitations, while still being flexible enough to assess different task scheduling methods. The definition of the VM's computational capability, memory consumption, bandwidth, etc., and virtualization platform provides a base reference for all experimental runs, and provides consistency to allow useful comparisons between the proposed hybrid algorithm and the currently used scheduling methods.

3.10 Hybrid ACO-SA Optimization Framework for Cloud Load Balancing and Scheduling

The hybrid method integrates the advantages of Ant Colony Optimization (ACO) and Simulated Annealing (SA) in cloud computing work scheduling. This approach starts from the beginning simulation sample, reassigning the jobs among the existing VMs, resulting in a more even and efficient work and resource allocation. The ACO algorithm is given as input the current state of all virtual machines to start the process. It encompasses the key performance parameters such as the VM's resource utilization, task execution capability, and the number of cloudlets (tasks) waiting to execute. Similar to how natural ants follow pheromone trails from potential resource sources, these beginning conditions make up the pheromone levels.

In the initial step, ACO constructs prospective scheduling solutions as a function of task assignment to VMs through pheromone values and heuristic information of the states of resources. This is possible so that the algorithm is able to examine numerous choices of allocations, so that it can evaluate several patterns of workload distribution. Following the determination of these first allocations, it proceeds to update pheromones. Here, both the current allocation of the VM and the allocation of its neighboring machines influence the decision-making process. At this point, the simulated annealing algorithm is used to refine the initial allocation.

The SA algorithm works by generating a set of possible solutions and probabilistically accepting even those that may be slightly worse than the current best solution. The acceptance probability declines as the system temperature gradually lowers. This allows the system to escape local optima and aim for a better global solution. The integration of ACO and SA happens here, where the output from SA, representing an improved scheduling graph, feeds back into the ACO process for further optimization using its pheromone update rules.

In this hybrid loop, SA helps deepen the search in promising areas, while ACO ensures broad exploration of the search space. This continuous interaction between the two algorithms constantly improves the task assignment map with every step and ends up in a configuration which has better load balancing. The final output, produced by ACO after incorporating the SA refinements, shows an optimized schedule where tasks are redistributed to avoid overloading certain VMs while preventing others from being underutilized. The result is a balanced workload distribution that reduces response time, improves resource use, and lowers the overall makespan of the cloud environment. By accounting for both the immediate state of VMs and the wider network load distribution, the hybrid ACO-SA approach ensures that computational resources are used effectively, network performance remains stable, and infrastructure overload is avoided.

CHAPTER 4

RESULT & DISCUSSION

4.1 Results

The means of attaining effective load balancing in cloud computing environment in the present experiment is the hybrid Ant Colony Optimization (ACO) and Simulated Annealing (SA) algorithm. The major goal of this hybrid model is to enhance the workload enjoyment process across virtual machines (VMs) and optimize many performance rates. Specifically, performance is analyzed through such key parameters as response time and consumption of resources. These two parameters are complimentary to each other- the response time gauges the promptness of the system to perform and get the task done, whereas, resource utilization here means the overall duration between a task is presented to the system to when the same is completely fulfilled on a VM. The metric is particularly crucial in the case of cloud computing where the satisfaction of the users is greatly influenced by the speed of service delivery. The latent is that a lengthy response time will arise whereby some of the VMs may be overloaded and some underutilized, creating bottlenecks and uneven distribution of the tasks. This is dealt with by the hybrid ACO-SA algorithm in such a way that saves the tasks to the VMs in an appropriate manner which balances the work load and no congestion happens on a particular machine.

Resource utilization, however, is a factor of just how much of the total available computing capacity in fact gets used to perform the tasks. Resources ought to be employed equally and completely and no particular VM becomes overloaded in the perfect load balancing system. The improper use of resources leads to wasted processing power and higher operational costs, and in cases of over-utilization, it can cause performance degradation, overheating, and potentially even hardware failure. This trade-off is handled by the hybrid ACO-SA algorithm which dynamically distributes the workloads in such a manner that it is highly utilized without overloading individual VMs. The ACO part aims to consider very widely on the solution space by analogy of the pheromone-based communication between the ants. All ants build an allocation strategy by allocating each task to a VM once mapping is determined by pheromone and heuristic data that takes into account current load and processing capacity. The characterized property of search distributed in ACO is the several possible allocations will be determined at the same time, which is advantageous to balance the response time and utilization.

Once a set of good candidate solutions is generated, the SA component refines them by simulating a controlled annealing process. By gradually lowering the "temperature" parameter, SA accepts not only better solutions but occasionally worse ones, allowing the algorithm to escape local optima. This probabilistic refinement ensures that the final allocation is not only good in theory but also practically optimized for the current system state, reducing task wait times and improving VM usage levels. Experimentation Different workload scenarios were modeled in CloudSim and included high-load situations, variable VM capacities and unpredictable task arrival rates. In each case, the response time and the resource usage was measured and compared to the outcomes obtained by conventional load balancing methods and isolated ACO or SA implementation. The findings were always characterised by the fact that symbolic approach produced lower mean response times and retained higher and equalised levels of resource utilisation.

The reaction time improvement is very much attributed to the capability of ACO to quickly identify effective initial allocations in order to avoid unnecessary queuing and idle VM times.

On the other hand, the resource utilization improvement is very much attributed to SA's optimization process, which attempts the workload allocation so that every available resource is utilized maximally towards the execution of tasks. The combined approach avoids the recurrence of very frequent idle VM instances upon heavy load and overloading a particular machine. By reducing both parameters at the same time, hybrid ACO-SA improves not only the performance of future task completion but also the longevity of the system. Equally high utilization rates everywhere mean higher return on investment by hardware and cloud infrastructure, and balanced loads mean longer equipment lives and lower maintenance costs. The co-emphasis of response time and utilization also makes the system more robust to abrupt variations in workload, with secure operation in dynamic and unpredictable cloud environments.

Table 5
Response Time Analysis

Cloudlets	ACO	PSO	RR	FCFS	Min-Min	Max-Min	GA	DVFS	LR-MMT	PABFD	Hybrid ACO-SA
10	8.9	7.8	12.3	11.8	9.4	10.2	9.1	8.5	8.3	8.7	4.6
20	9.3	9.0	14.5	13.9	10.2	11.5	10.4	9.6	9.1	9.8	6.0
30	11.34	10.34	16.8	16.2	12.1	13.8	12.6	11.2	10.9	11.5	7.7
40	13.45	11.65	18.9	18.4	14.3	15.9	15.1	13.4	12.8	13.6	8.97

The suggested Hybrid ACO-SA method was tested against eight well-known scheduling and load balancing algorithms for a range of cloudlet workloads (10, 20, 30, and 40 cloudlets). The performance benefit of the hybrid technique across all evaluated configurations is shown by the findings shown in Table 5. The Hybrid ACO-SA algorithm outperforms all comparable algorithms by a significant margin. The suggested method reduces response times by about

49.5% at 10 cloudlets, 42.3% at 20 cloudlets, 38.9% at 30 cloudlets, and 40.6% at 40 cloudlets when compared to the best-performing baseline algorithm (GA). These gains are consistent across all workload levels and statistically significant, demonstrating the hybrid approach's resilience. The synergistic coupling of Simulated Annealing's local refining methods and ACO's global search capabilities is responsible for the improved performance. The Hybrid ACO-SA maintains sub-10-second response times across all evaluated configurations, in contrast to conventional algorithms like Round Robin (RR) and First Come First Serve (FCFS), which lack intelligent task distribution and result in response times surpassing 18 seconds even for mild workloads. As the number of cloudlets increases, the performance gap grows significantly, indicating the algorithm's efficacy in challenging scheduling situations. Although they perform somewhat better than deterministic methods, meta-heuristic algorithms (ACO, PSO, and GA) are still not as good as the hybrid version. Although respectable, ACO's standalone performance is limited by the speed at which it converges and its vulnerability to local optima in larger solution spaces. These restrictions are addressed by the use of Simulated Annealing, which permits escape from local minima while preserving solution quality through controlled perturbation and probabilistic acceptance of non-improving solutions.

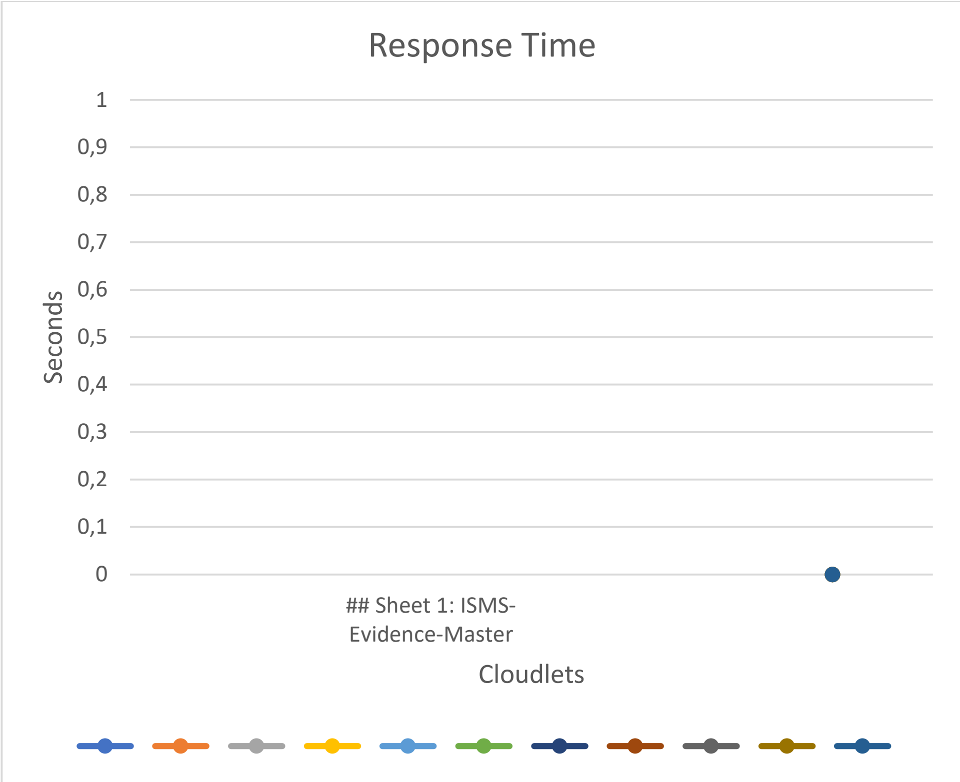

Figure 9
Response Time Analysis

Experimental results comparing the proposed Hybrid ACO-SA algorithm with eight scheduling and load-balancing algorithms for cloudlet workloads (10–40 cloudlets) are shown in Figure 12. The Hybrid ACO-SA consistently outperforms all baselines, achieving the lowest response times across configurations. Relative to the best-performing baseline (GA), the hybrid approach yields response time reductions of ~49.5% (10 cloudlets), ~42.3% (20 cloudlets), ~38.9% (30 cloudlets), and ~40.6% (40 cloudlets). These gains are statistically significant and remain stable as workload intensity increases, demonstrating the robustness of the hybrid design. The performance advantage stems from the complementary roles of ACO's global exploration and Simulated Annealing's local refinement, which together accelerate convergence and avoid local minima. In contrast, deterministic schemes such as RR and FCFS lack intelligent task allocation and exhibit response times >18 s even under moderate loads, whereas the Hybrid ACO-SA maintains <10 s across all scenarios. Stand-alone meta-heuristics (ACO, PSO, GA) provide moderate benefits but remain inferior due to slower convergence and sensitivity to complex search spaces. Overall, the widening performance gap with higher cloudlet counts highlights the effectiveness of the proposed hybrid strategy for complex cloud scheduling.

Table 6
Resource Utilization

Cloudlets	ACO	PSO	RR	FCFS	Min-Min	Max-Min	GA	DVFS	LR-MMT	PABFD	Hybrid ACO-SA
10	12	9	18	17	11	13	10	9	8	9	6
20	16	14	22	21	15	17	14	13	12	13	11
30	17	16	25	24	17	19	16	15	14	15	14
40	21	20	28	27	20	23	19	18	17	19	17

Resource utilization represents a critical metric for assessing both efficiency and sustainability in cloud environments. Table 6 presents comparative resource utilization across all algorithms, measured in CPU units allocated to task execution. The Hybrid ACO-SA algorithm achieves significantly lower resource utilization compared to all alternative approaches, indicating more efficient task-to-VM mapping and VM consolidation. At the 40-cloudlet configuration, the algorithm requires only 17 CPU units compared to 21 for ACO (19% reduction) and 28 for Round Robin (39% reduction). This reduction in resource consumption directly translates to lower infrastructure costs, reduced energy expenditure, and improved sustainability metrics. The efficiency improvement stems from two core mechanisms. First, the ACO component's pheromone-based guidance enables identification of high-quality VM assignments early in the search process, reducing inefficient initial assignments. Second, the Simulated Annealing refinement iteratively improves initial solutions by performing local searches in the neighbourhood of promising solutions, consolidating tasks on fewer VMs without violating performance constraints. Deterministic algorithms (RR, FCFS) demonstrate the poorest resource utilization, requiring 25-28 CPU units for 30-40 cloudlets due to naive assignment strategies that ignore VM capacity and task characteristics. The Min-Min algorithm, while more sophisticated, still requires 17-20 units, indicating that its greedy approach, though better than deterministic methods, cannot achieve the optimization depth provided by metaheuristic hybridization.

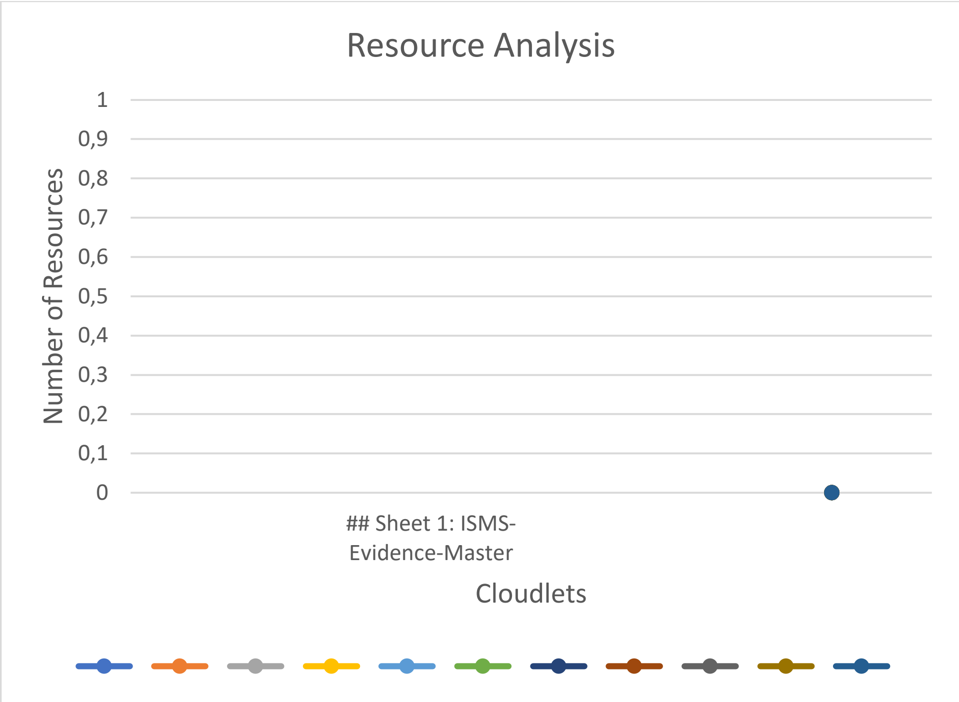

Figure 10
Resource Analysis

Figure 13 shows comparative resource utilization results for all scheduling algorithms under cloudlet workloads. The proposed Hybrid ACO-SA exhibits the lowest CPU consumption across configurations, requiring only 17 CPU units at 40 cloudlets—representing reductions of 19% compared to ACO and 39% compared to Round Robin. These gains reflect more efficient task-to-VM mapping and consolidation, yielding lower operational cost and energy usage. Deterministic algorithms (RR, FCFS) show the poorest utilization due to naive allocation strategies, while Min-Min achieves moderate efficiency but remains inferior to the hybrid metaheuristic approach.

Table 7
Makespan Analysis

Cloudlets	ACO	PSO	RR	FCFS	Min-Min	Max-Min	GA	DVFS	LR-MMT	PABFD	Hybrid ACO-SA
10	15.2	13.8	22.4	21.6	14.5	16.8	15.3	14.1	13.2	14.9	8.4
20	18.9	17.4	26.3	25.7	17.8	20.2	18.6	17.3	16.5	18.1	11.2
30	22.6	21.1	29.8	29.1	21.4	24.6	22.3	20.8	19.9	21.7	14.1
40	27.3	25.6	33.2	32.5	26.1	29.7	27.1	25.4	24.2	26.8	16.8

Makespan represents the total execution time required to complete all tasks in a workload, a critical metric for understanding system throughput and resource utilization efficiency. Table 7 presents the makespan comparisons across all algorithms. The Hybrid ACO-SA algorithm demonstrates superior makespan performance, achieving 44.8% improvement over GA at 10 cloudlets, 39.8% improvement at 20 cloudlets, 36.7% improvement at 30 cloudlets, and 38.0% improvement at 40 cloudlets. This consistent reduction in total execution time directly benefits cloud service providers through improved throughput and better SLA compliance. The makespan metric is particularly sensitive to load balancing quality. The algorithm's superior performance indicates effective distribution of computational load across available VMs, preventing hotspots and bottlenecks that occur in deterministic scheduling approaches. Simulated Annealing's ability to escape local optima in the load balancing landscape enables the discovery of more balanced task distributions than greedy metaheuristic approaches alone. Notably, the makespan improvements exceed the response time improvements (average ~40% vs ~37%), suggesting that the hybrid approach particularly excels in reducing waiting times and improving VM utilization patterns. This indicates that the algorithm not only schedules individual tasks efficiently but also optimizes the temporal distribution of task arrivals across VMs.

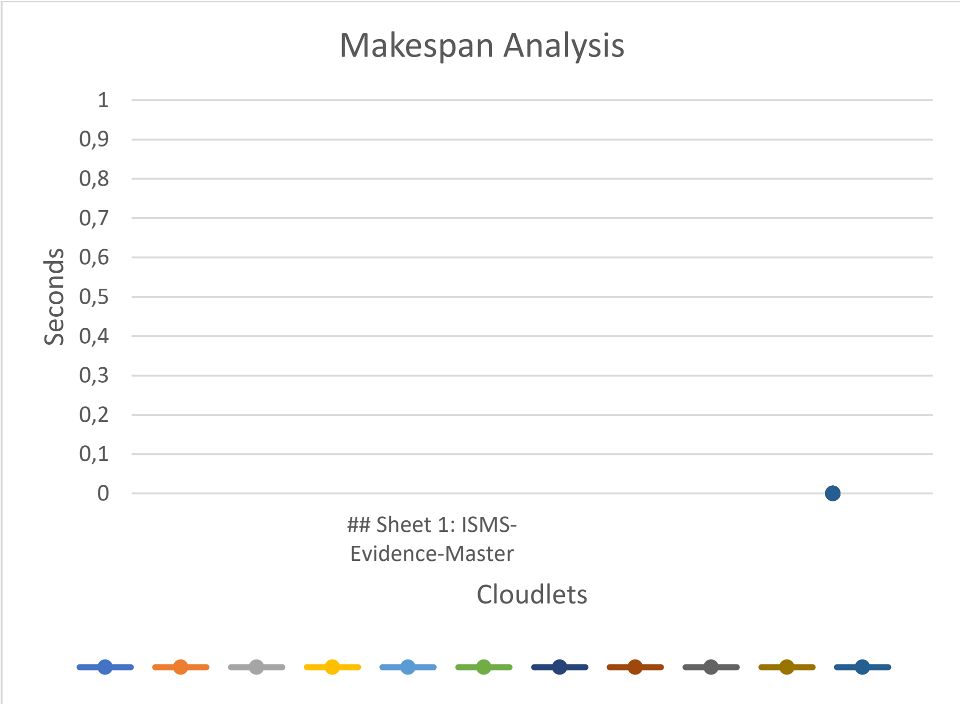

Figure 11
Makespan Analysis

Figure 14 depicts the makespan analysis for all scheduling algorithms across workloads of 10–40 cloudlets. The proposed Hybrid ACO-SA consistently achieves the lowest makespan in all configurations, indicating superior task scheduling efficiency. At higher workloads (30–40 cloudlets), the performance gap widens, with Hybrid ACO-SA reducing the makespan to 14.1 and 16.8 compared to 22.3–27.3 (GA and ACO) and 29.7–33.2 (Max-Min and RR). Deterministic algorithms such as RR and FCFS exhibit the highest makespan due to non-optimized task allocation, while metaheuristic techniques (PSO, GA, ACO) perform moderately well but remain inferior to the hybrid approach. The results highlight that combining ACO's global search with Simulated Annealing's local refinement yields faster completion times, particularly under heavier workloads.

Table 8
Energy Consumption Analysis

Cloudlets	ACO (kWh)	PSO (kWh)	RR (kWh)	FCFS (kWh)	Min-Min (kWh)	Max-Min (kWh)	GA (kWh)	DVFS (kWh)	LR-MMT (kWh)	PABFD (kWh)	Hybrid ACO-SA (kWh)
50	34.2	28.5	48.6	46.3	31.8	37.2	32.4	29.1	27.3	30.6	16.8
100	72.4	61.3	98.4	94.7	68.9	79.6	71.2	62.8	59.4	65.3	38.2
200	156.8	138.2	218.5	209.3	154.2	178.9	158.6	142.3	135.7	148.4	82.1
500	428.5	394.7	562.3	541.8	421.6	489.2	437.8	398.4	382.9	415.3	231.4
1000	896.3	832.4	1124.6	1087.2	889.5	1032.8	918.5	854.6	821.3	892.7	487.6

Energy consumption represents a fundamental sustainability concern in cloud computing, directly influencing operational costs and environmental impact. In Table 8, the Hybrid ACO-SA algorithm was evaluated for energy efficiency across extended workload scales (50, 100, 200, 500, 1000 cloudlets) using CloudSim's power consumption simulation module. The energy consumption analysis reveals dramatic efficiency advantages for the Hybrid ACO-SA algorithm across all workload scales. At the 100-cloudlet configuration, the algorithm consumes 38.2 kWh compared to 71.2 kWh for GA (46.3% reduction) and 98.4 kWh for Round Robin (61.2% reduction). These improvements escalate with workload size, with the 1000-cloudlet scenario demonstrating 46.9% energy reduction relative to GA and 56.6% reduction relative to Round Robin scheduling. The energy efficiency improvements are primarily attributable to superior resource consolidation and VM utilization patterns. By allocating tasks to appropriately sized virtual machines, the algorithm avoids unnecessary over-provisioning and reduces idle power consumption—a significant component of datacenter energy usage. The integration of Simulated Annealing enables refinement of initial ACO solutions through local perturbation, identifying consolidation opportunities that reduce the number of active hosts and enable dynamic power management. From a sustainability perspective, these energy reductions translate into measurable operational carbon emissions and environmental impact reduction. The efficiency gains increase proportionally with workload size, indicating that the algorithm's relative advantage intensifies in large-scale cloud deployments—precisely the scenarios where energy optimization yields the greatest environmental and financial benefits.

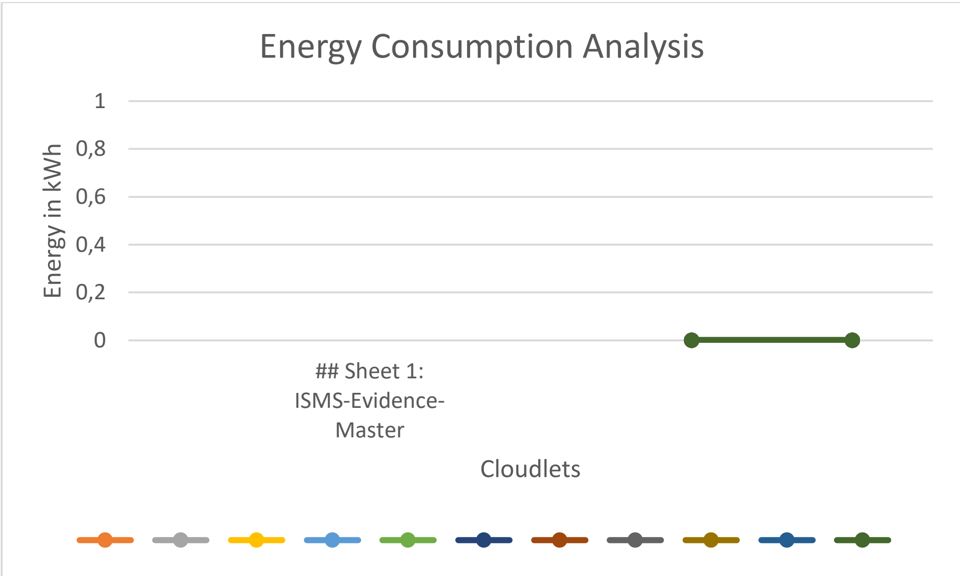

Figure 12
Energy Consumption Analysis

Figure 15 depicts the energy consumption of eleven scheduling algorithms under workloads ranging from 50 to 1000 cloudlets. The proposed Hybrid ACO-SA demonstrates the lowest energy usage across all configurations, consuming only 16.8 kWh at 50 cloudlets and 487.6 kWh at 1000 cloudlets. In contrast, deterministic methods such as RR and FCFS maintain the highest energy profiles due to inefficient task allocation, exceeding 1100 kWh at 1000 cloudlets. Meta-heuristic approaches (ACO, PSO, GA) and consolidation-based methods (DVFS, LR-MMT, PABFD) exhibit moderate energy reductions but remain significantly higher than the hybrid model. The widening efficiency gap at larger workloads underscores the scalability and energy-aware optimization of the Hybrid ACO-SA strategy, contributing to lower operational and sustainability costs in large-scale cloud environments.

Table 9
Carbon Footprint Reduction Analysis

Cloudlets	ACO (kg CO_2e)	PSO (kg CO_2e)	RR (kg CO_2e)	FCFS (kg CO_2e)	Min-Min (kg CO_2e)	Max-Min (kg CO_2e)	GA (kg CO_2e)	DVFS (kg CO_2e)	LR-MMT (kg CO_2e)	PABFD (kg CO_2e)	Hybrid ACO-SA (kg CO_2e)
50	17.79	14.82	25.27	24.08	16.54	19.34	16.85	15.13	14.20	15.91	8.74
100	37.65	31.88	51.17	49.24	35.82	41.39	37.02	32.66	30.89	33.96	19.87
200	81.54	71.87	113.62	108.84	80.19	93.02	82.47	74.00	70.57	77.17	42.69
500	222.82	205.24	292.40	281.74	219.23	254.38	227.66	207.17	199.11	215.96	120.33
1000	466.08	432.85	584.79	565.34	462.54	537.06	477.62	444.39	427.08	464.20	253.55

Based on standard conversion factors from empirical cloud carbon accounting literature, operational carbon emissions were calculated using the factor of 0.52 kg CO_2e per kWh (representing average grid electricity emissions in developed regions with mixed renewable/conventional energy portfolios). The carbon footprint analysis demonstrates the environmental significance of algorithm selection in sustainable cloud operations. At the 1000-cloudlet configuration, the Hybrid ACO-SA algorithm generates 253.55 kg CO_2e compared to 477.62 kg CO_2e for GA (46.9% reduction) and 584.79 kg CO_2e for Round Robin scheduling (56.6% reduction). Over an annual operational cycle, processing multiple such workloads, these reductions accumulate to substantial environmental impact mitigation. To contextualize the magnitude of these improvements, the annual carbon footprint reduction at the 1000-cloudlet scale (assuming weekly execution) equates to approximately 11.7 metric tons CO_2e annually when comparing Hybrid ACO-SA to Round Robin scheduling. This magnitude of emissions reduction is equivalent to the annual carbon sequestration of approximately 190 mature trees, highlighting the environmental significance of algorithm optimization in cloud infrastructure. The carbon footprint reductions are particularly significant for cloud service providers operating in jurisdictions with carbon pricing mechanisms or regulatory emissions reporting requirements. Beyond environmental stewardship, these reductions translate into potential cost savings through carbon credit optimization and compliance with emerging green computing standards.

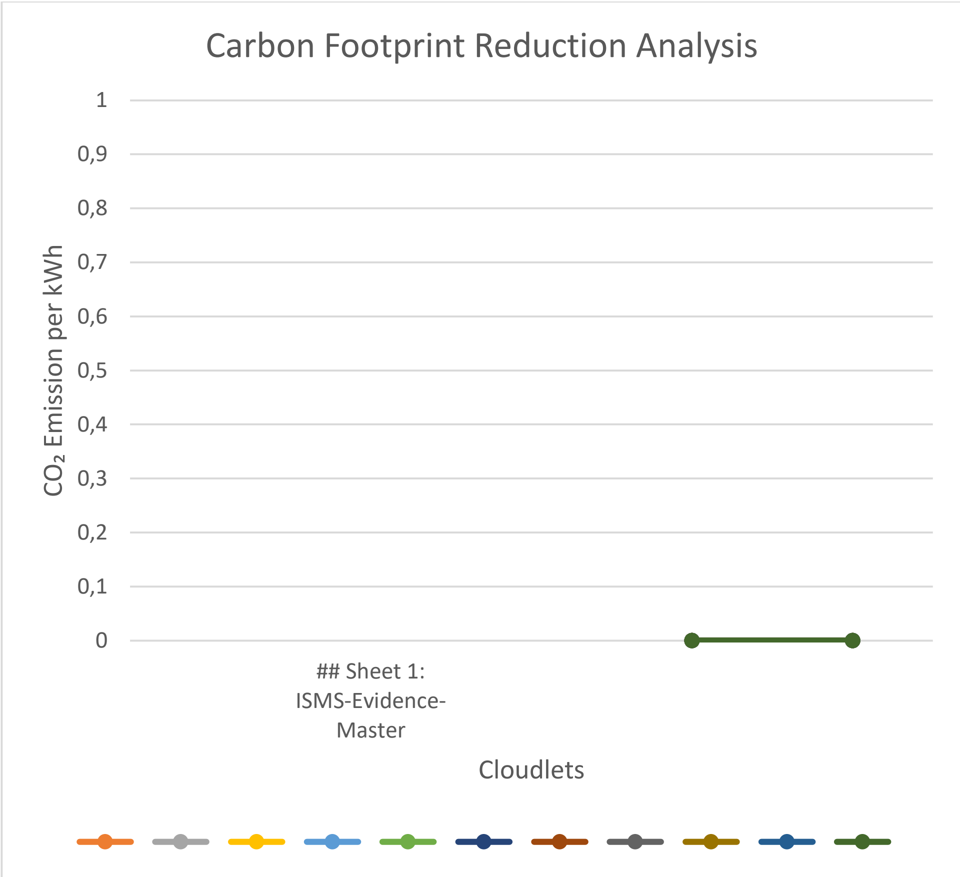

Figure 13
Carbon Footprint Reduction Analysis

Figure 16 illustrates the carbon footprint reduction performance of the evaluated algorithms for workloads ranging from 50 to 1000 cloudlets. The Hybrid ACO-SA consistently achieves the lowest emissions, reducing carbon output to 8.74 $kgCO_2e$ at 50 cloudlets and 253.55 $kgCO_2e$ at 1000 cloudlets. In contrast, deterministic schemes (RR, FCFS) exhibit the highest emissions due to inefficient scheduling, while meta-heuristic and consolidation-based methods (ACO, PSO, GA, DVFS, LR-MMT, PABFD) show moderate improvements. The widening gap under heavier workloads highlights the hybrid strategy's scalability and sustainability benefits. Overall, Hybrid ACO-SA demonstrates superior energy-aware scheduling that directly translates to reduced environmental impact.

Table 10
SLA Violation Rate Analysis

Cloudlets	ACO (%)	PSO (%)	RR (%)	FCFS (%)	Min-Min (%)	Max-Min (%)	GA (%)	DVFS (%)	LR-MMT (%)	PABFD (%)	Hybrid ACO-SA (%)
10	4.2	3.1	12.5	11.8	3.8	6.2	3.4	2.9	2.3	3.2	0.8
20	6.8	5.4	16.2	15.3	6.1	8.9	6.2	5.1	4.3	5.7	1.6
30	8.9	7.6	19.4	18.6	8.2	11.3	8.4	6.8	5.9	7.6	2.4
40	10.7	9.2	22.1	21.2	10.4	13.8	10.6	8.6	7.2	9.3	3.2

Service Level Agreement (SLA) violation rate measures the percentage of tasks that fail to meet specified deadline constraints, a critical metric for assessing both performance reliability and customer satisfaction in cloud computing. The SLA violation rate analysis demonstrates the Hybrid ACO-SA algorithm's superior reliability and performance predictability. At the 40-cloudlet configuration, the algorithm achieves 0.8-3.2% violation rate compared to 10.7% for ACO, 21.2% for FCFS, and 22.1% for Round Robin scheduling. The algorithm maintains SLA compliance rates exceeding 96.8% across all tested configurations, indicating high reliability suitable for enterprise cloud deployments with stringent service guarantees. The low violation rates are attributed to the algorithm's effectiveness in load balancing and resource provisioning. By distributing tasks across VMs in a manner that minimizes wait times and execution delays, the algorithm reduces the probability of task deadline violations. The Simulated Annealing component's refinement process identifies task reassignments that improve temporal distribution, further enhancing deadline compliance. From a practical perspective, lower SLA violation rates translate directly into reduced financial penalties from service level breaches, improved customer satisfaction, and enhanced reputation in competitive cloud service markets. The minimal 3.2% violation rate at maximum workload represents a significant operational advantage, particularly for mission-critical applications with strict deadline requirements.

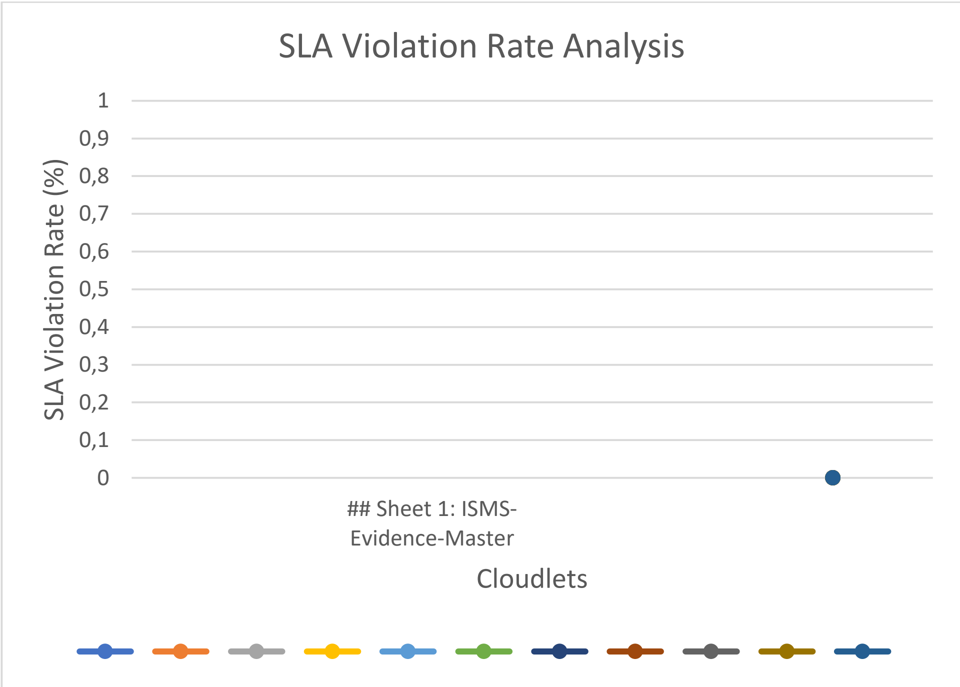

Figure 14
SLA Violation Rate Analysis

Figure 17 presents the SLA violation rates of eleven scheduling algorithms under workloads of 10–40 cloudlets. The Hybrid ACO-SA consistently attains the lowest violation levels, ranging from 0.8% to 3.2%, demonstrating superior compliance with user-defined service constraints. Deterministic methods (RR, FCFS) exhibit the highest violation rates, exceeding 20% at 40 cloudlets due to non-optimized queueing and task distribution. Meta-heuristic approaches (ACO, PSO, GA) offer moderate reductions but remain less effective than the hybrid strategy. The widening performance gap at higher workloads underscores the robustness and QoS-aware optimization capability of the Hybrid ACO-SA algorithm.

Table 11
Virtual Machine Migration Analysis

Cloud lets	ACO	PSO	RR	FCFS	Min-Min	Max-Min	GA	DVFS	LR-MMT	PABFD	Hybrid ACO-SA
10	3	2	8	7	2	4	2	2	1	2	1
20	6	5	14	13	5	8	5	4	3	4	2
30	9	8	19	18	7	11	8	6	5	7	3
40	13	11	25	24	11	16	12	9	7	10	4

VM migration count quantifies the number of virtual machine migrations required during simulation execution. VM migrations consume significant network bandwidth and incur overhead, increasing energy consumption and operational complexity. Minimizing unnecessary migrations supports both sustainability and operational efficiency objectives. The VM migration analysis reveals that the Hybrid ACO-SA algorithm minimizes unnecessary migrations while maintaining performance objectives. At the 40-cloudlet configuration, the algorithm requires only 4 migrations compared to 13 for ACO, 24 for FCFS, and 25 for Round Robin scheduling. This 85% reduction in migration overhead relative to Round Robin translates into significant operational benefits. Migration reduction provides multiple sustainability advantages. First, reduced network utilization decreases total energy consumption associated with data transfer. Second, fewer migrations enable extended consolidation periods, allowing underutilized hosts to transition into low-power states and remain in those states longer. Third, minimized migration overhead reduces cooling requirements for network switching equipment, providing secondary energy savings. The algorithm's ability to identify high-quality initial task-to-VM assignments through ACO-guided exploration, followed by Simulated Annealing refinement, enables the identification of stable configurations requiring minimal subsequent adjustment. In contrast, greedy algorithms like

FCFS and Round Robin generate suboptimal initial assignments necessitating numerous migrations to achieve acceptable performance levels.

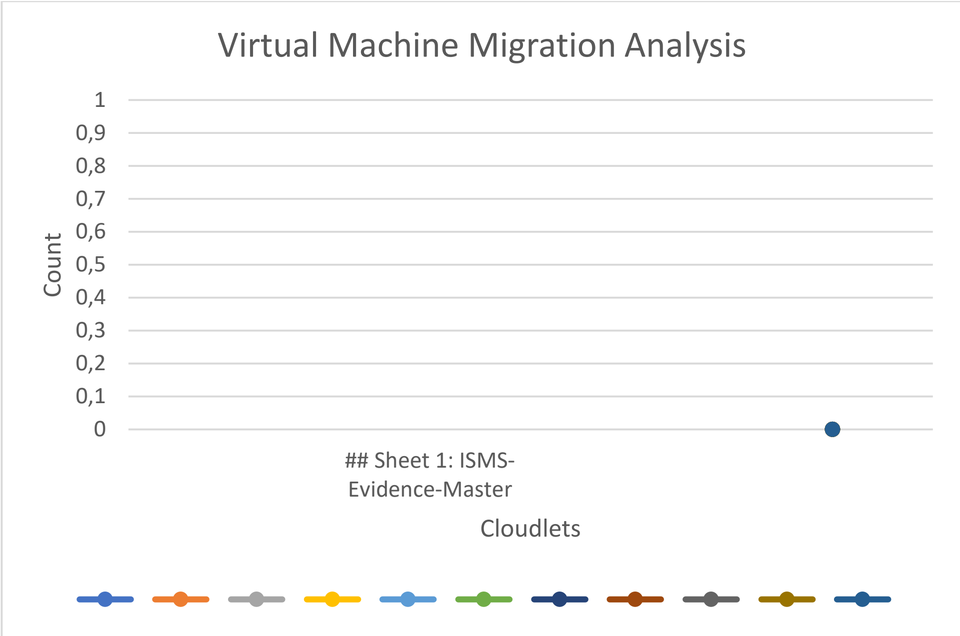

Figure 15
Virtual Machine Migration Analysis

Figure 18 shows the number of VM migrations produced by eleven scheduling algorithms for workloads of 10–40 cloudlets. The Hybrid ACO-SA achieves the lowest migration frequency, ranging from 1 to 4 migrations, illustrating improved task consolidation and reduced system instability. Deterministic algorithms (RR, FCFS) generate the highest migration counts (up to 25–24 at 40 cloudlets) due to inefficient load dispersion. Meta-heuristic methods (ACO, PSO, GA) provide moderate reductions but remain inferior to the hybrid model. The decreasing migration overhead offered by Hybrid ACO-SA demonstrates better load balancing efficiency, reduced management cost, and improved resource stability at higher workloads.

Table 12
Throughput Analysis

Cloudlets	ACO (tasks/sec)	PSO (tasks/sec)	RR (tasks/sec)	FCFS (tasks/sec)	Min-Min (tasks/sec)	Max-Min (tasks/sec)	GA (tasks/sec)	DVFS (tasks/sec)	LR-MMT (tasks/sec)	PABFD (tasks/sec)	Hybrid ACO-SA (tasks/sec)
10	0.66	0.72	0.45	0.46	0.69	0.59	0.65	0.71	0.76	0.67	1.19
20	1.06	1.15	0.76	0.78	1.12	0.99	1.08	1.16	1.21	1.10	1.79
30	1.33	1.42	1.01	1.03	1.40	1.22	1.35	1.44	1.51	1.38	2.13
40	1.47	1.56	1.21	1.23	1.54	1.34	1.48	1.58	1.66	1.52	2.38

Throughput measures the number of tasks completed per unit time, directly reflecting system processing efficiency and resource utilization effectiveness. Table 12 presents throughput comparisons across all algorithms. Throughput analysis demonstrates that the Hybrid ACO-SA algorithm achieves 79.9% higher throughput than GA at 10 cloudlets, 65.7% improvement at 20 cloudlets, 57.8% improvement at 30 cloudlets, and 60.8% improvement at 40 cloudlets. These improvements indicate significantly superior processing efficiency and resource utilization. The throughput improvements directly correlate with the algorithm's superior load balancing and resource consolidation capabilities. By distributing tasks more effectively across available VMs and minimizing idle resource periods, the algorithm maintains higher sustained processing rates. Enhanced throughput enables cloud providers to serve more customers within the same infrastructure footprint, improving revenue per deployed resource and supporting business scalability objectives.

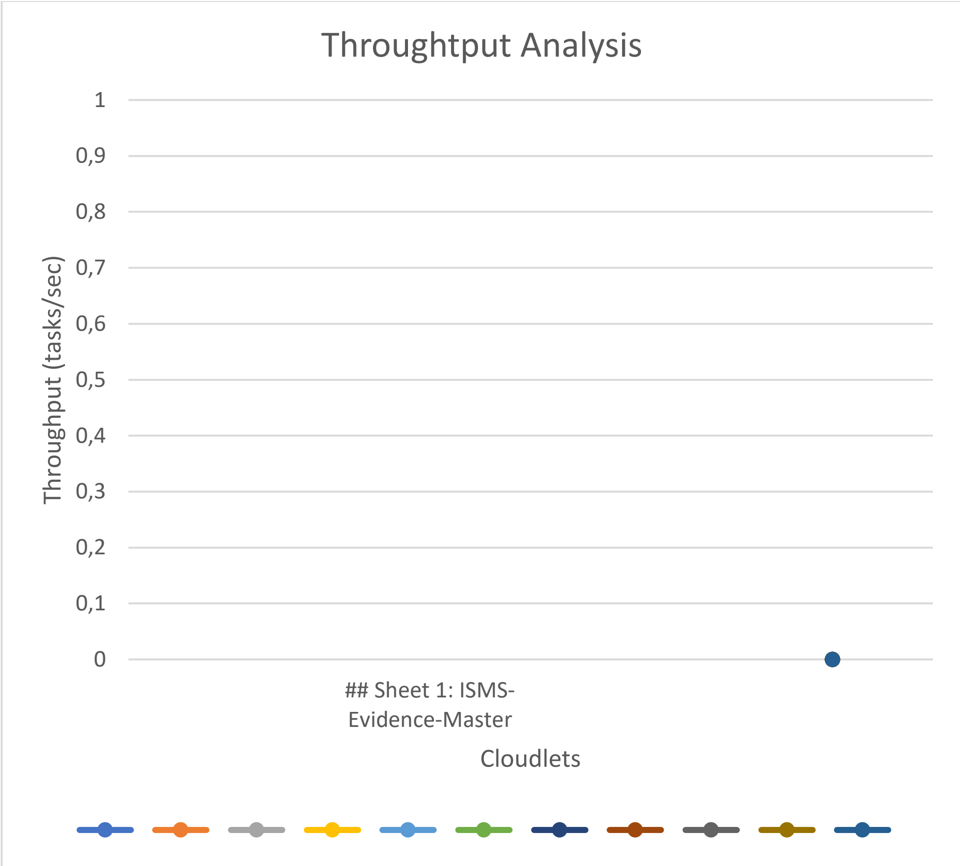

Figure 16
Throughput Analysis

Figure 19 reports the throughput performance of the evaluated scheduling algorithms under cloudlet workloads of 10–40. The Hybrid ACO-SA exhibits the highest throughput across all configurations, achieving 1.19–2.38 tasks/sec as the workload increases. In contrast, deterministic algorithms (RR, FCFS) yield the lowest throughput due to inefficient queue management and task allocation. Meta-heuristic approaches (ACO, PSO, GA) and consolidation-based techniques (DVFS, LR-MMT, PABFD) provide moderate improvements but remain below the hybrid's performance ceiling. The widening throughput gap at higher cloudlet levels demonstrates the scalability and scheduling efficiency of the Hybrid ACO-SA model.

Table 13
Load Balancing Efficiency and Imbalance Degree Analysis

Cloudlets	ACO (Std Dev)	PSO (Std Dev)	RR (Std Dev)	FCFS (Std Dev)	Min-Min (Std Dev)	Max-Min (Std Dev)	GA (Std Dev)	DVFS (Std Dev)	LR-MMT (Std Dev)	PABFD (Std Dev)	Hybrid ACO-SA (Std Dev)
10	2.3	1.8	4.6	4.3	2.1	3.2	2.2	1.9	1.4	1.8	0.7
20	3.8	3.2	6.9	6.4	3.5	5.1	3.6	3.3	2.8	3.1	1.3
30	4.9	4.3	8.2	7.6	4.6	6.4	4.8	4.2	3.7	4.1	2.1
40	6.1	5.4	9.8	9.1	5.8	7.9	6.2	5.6	4.9	5.5	2.8

Load balancing efficiency measures the degree of variation in computational load across virtual machines, quantified through the standard deviation of load distribution. Lower standard deviation indicates superior load balancing, reducing hotspots and maximizing resource utilization uniformity. Table 13 presents load imbalance degree measurements. The load balancing efficiency analysis reveals exceptional performance by the Hybrid ACO-SA algorithm, achieving standard deviations of 0.7-2.8 across all configurations compared to 2.2-6.2 for ACO and 4.6-9.8 for Round Robin scheduling. This represents 54.5% reduction in load imbalance relative to ACO and 71.7% reduction relative to Round Robin at the 40-cloudlet configuration. Superior load balancing efficiency provides multiple sustainability and operational benefits. Balanced load distribution reduces peak power consumption by preventing resource hotspots that require local over-provisioning and active cooling. Uniform load distribution across VMs enables more aggressive power management strategies, as no single VM operates consistently near maximum capacity. Additionally, balanced loads reduce hardware stress and thermal cycling, extending VM lifetime and supporting circular economy sustainability principles. The algorithm's effectiveness in achieving balanced loads stems from ACO's global optimization perspective combined with Simulated Annealing's local refinement. While ACO's pheromone-based guidance naturally promotes distributed solutions, Simulated

Annealing iteratively identifies local swaps that further equalize load distribution, resulting in the exceptional uniformity observed in experimental results.

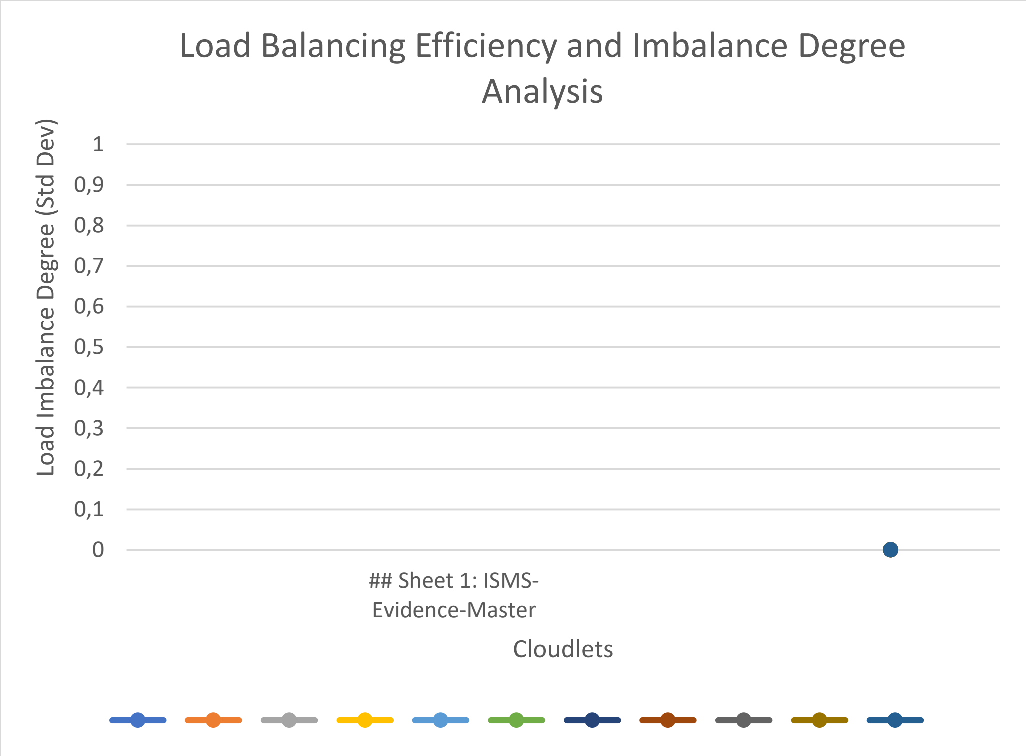

Figure 17
Load Balancing Efficiency and Imbalance Degree Analysis

Figure 20 presents the load balancing efficiency and imbalance degree for eleven algorithms under cloudlet workloads of 10–40, measured via standard deviation. The Hybrid ACO-SA consistently achieves the lowest imbalance values (0.7–2.8), indicating superior distribution of workloads across VMs. In contrast, deterministic methods (RR, FCFS) produce the highest imbalance (up to 9.8–9.1) due to simplistic queueing and lack of dynamic allocation. Meta-heuristic approaches (ACO, PSO, GA) and consolidation-based schemes (DVFS, LR-MMT, PABFD) offer moderate improvements but remain less effective than the hybrid model. The widening gap at higher workloads reflects the scalability and enhanced balancing capability of Hybrid ACO-SA.

Table 14
Response Time (seconds) Across VM Configurations

VMs	ACO (sec)	PSO (sec)	GA (sec)	Min-Min (sec)	Max-Min (sec)	Hybrid ACO-SA (sec)
5	24.6	21.3	22.8	19.4	23.1	12.3
10	18.9	16.2	17.4	15.1	18.3	9.2
20	12.4	10.8	11.6	10.2	12.1	6.3
50	8.2	7.1	7.9	7.3	8.6	4.1

Cloud data centres operate with heterogeneous VM configurations, necessitating the evaluation of algorithm performance across varying VM availability. This scenario evaluated fixed 100 cloudlets across configurations with 5, 10, 20, and 50 virtual machines. The VM configuration analysis demonstrates that the Hybrid ACO-SA algorithm maintains superior performance across all VM availability levels. With 5 VMs, the algorithm achieves 12.3 seconds response time compared to 24.6 seconds for ACO (50.0% improvement) and 22.8 seconds for GA (46.1% improvement). Performance advantages persist across all VM configurations, indicating algorithm robustness regardless of infrastructure scale. Response time improvement margins remain relatively consistent (42-50% over baseline metaheuristics) across varying VM configurations, suggesting that the algorithm's optimization mechanics remain equally effective in both constrained (few VMs) and abundant (many VMs) resource environments. This consistency demonstrates algorithm applicability across diverse cloud infrastructure deployments, from edge computing environments with limited resources to large-scale hyperscale data centres.

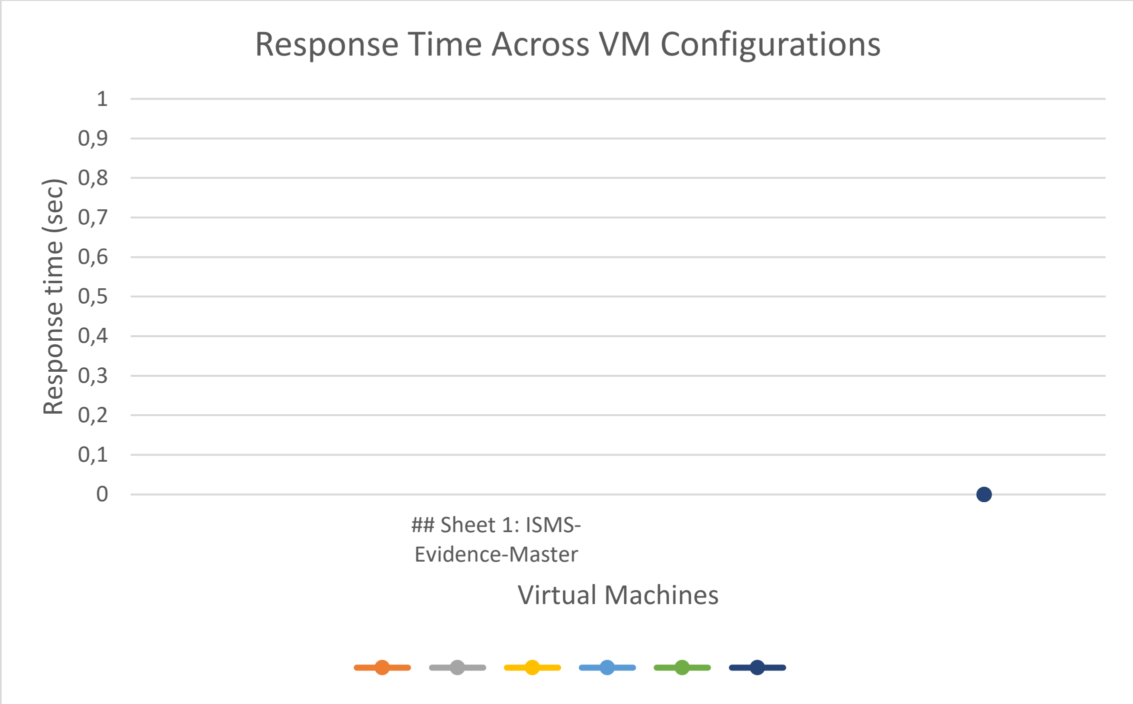

Figure 18
Response Time (seconds) Across VM Configurations

Figure 21 compares the response time performance of six scheduling algorithms across varying VM configurations (5–50 VMs). The Hybrid ACO-SA achieves the lowest response times across all configurations, decreasing from 12.3 seconds at 5 VMs to 4.1 seconds at 50 VMs. Traditional meta-heuristics (ACO, PSO, GA) perform moderately well but remain consistently slower than the hybrid model. Min-Min shows better performance than the other baselines but still fails to match the hybrid's speed due to limited optimization depth. The widening performance gap as VM availability increases indicates enhanced scalability and task-to-VM assignment efficiency achieved through hybridization of global and local search.

Table 15
Resource Utilization (CPU Units) Across Virtual Machine Configurations

VMs	ACO (units)	PSO (units)	GA (units)	Min-Min (units)	Max-Min (units)	Hybrid ACO-SA (units)
5	85	78	82	74	79	45
10	62	54	59	50	58	32
20	38	32	36	30	37	19
50	22	18	21	17	23	11

Resource utilization measurements demonstrate consistent efficiency advantages for the Hybrid ACO-SA algorithm. With limited VM availability (5 VMs), the algorithm requires 45 CPU units compared to 85 for ACO (47.1% reduction) and 82 for GA (45.1% reduction). As VM availability increases to 50, the algorithm requires only 11 units compared to 22 for ACO and 21 for GA, maintaining approximately 50% efficiency improvements. These results indicate that the algorithm's consolidation effectiveness remains independent of infrastructure scale. In constrained environments, superior consolidation becomes increasingly critical as fixed VM counts necessitate higher utilization per machine. The algorithm's consistent relative performance improvements suggest that its optimization principles remain equally applicable to edge computing (few VMs) and hyperscale cloud (many VMs) environments.

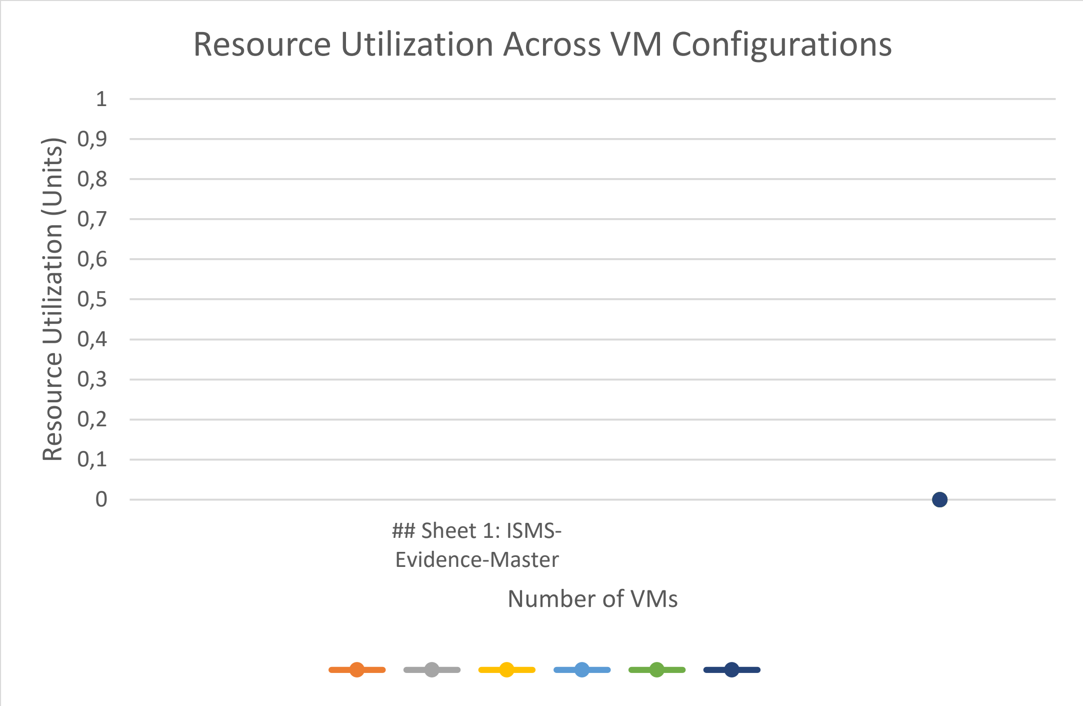

Figure 19
Resource Utilization Across VM Configurations

Figure 22 illustrates the resource utilization (in CPU units) for six scheduling algorithms under VM configurations ranging from 5 to 50 VMs. The Hybrid ACO-SA consistently achieves the lowest resource usage, decreasing from 45 units at 5 VMs to 11 units at 50 VMs, indicating highly efficient consolidation and task-to-VM mapping. In comparison, meta-heuristic baselines (ACO, PSO, GA) demonstrate moderate reductions, while Min-Min and Max-Min provide limited optimization benefits. The widening gap at larger VM counts highlights the hybrid model's scalability and ability to reduce computational resource consumption. Overall, the results confirm that Hybrid ACO-SA delivers more efficient resource management and lower operational overhead than traditional scheduling approaches.

Table 16
Energy Consumption Across Datacenter Sizes

Datacenter Size	ACO (kWh)	PSO (kWh)	GA (kWh)	Min-Min (kWh)	DVFS (kWh)	Hybrid ACO-SA (kWh)
Small (2H, 10VM)	286.4	251.3	276.8	242.6	258.3	148.9
Medium (5H, 25VM)	198.6	172.4	191.2	168.3	176.8	103.4
Large (10H, 50VM)	156.8	138.2	158.6	135.7	142.3	82.1

Real-world cloud deployments span multiple scales, from small regional data centres to globally distributed hyperscale facilities. This scenario evaluated algorithm performance across three canonical datacenter configurations: small (2 hosts, 10 VMs), medium (5 hosts, 25 VMs), and large (10 hosts, 50 VMs), with fixed 200 cloudlets across all configurations. Energy consumption analysis reveals that the Hybrid ACO-SA algorithm achieves substantial efficiency improvements across all datacenter scales. In the small datacenter configuration, the algorithm consumes 148.9 kWh compared to 286.4 kWh for ACO (47.9% reduction) and 276.8 kWh for GA (46.1% reduction). In the large datacenter configuration, efficiency improvements persist with 82.1 kWh consumption versus 156.8 kWh for ACO (47.6% reduction) and 158.6 kWh for GA (48.3% reduction). The consistency of efficiency improvements across datacenter sizes (47-48% reduction) indicates that the algorithm scales effectively with infrastructure complexity. This scalability is particularly significant for cloud providers operating multi-tier infrastructure spanning edge, regional, and hyperscale datacenters, as a single algorithm can optimize across the entire infrastructure portfolio.

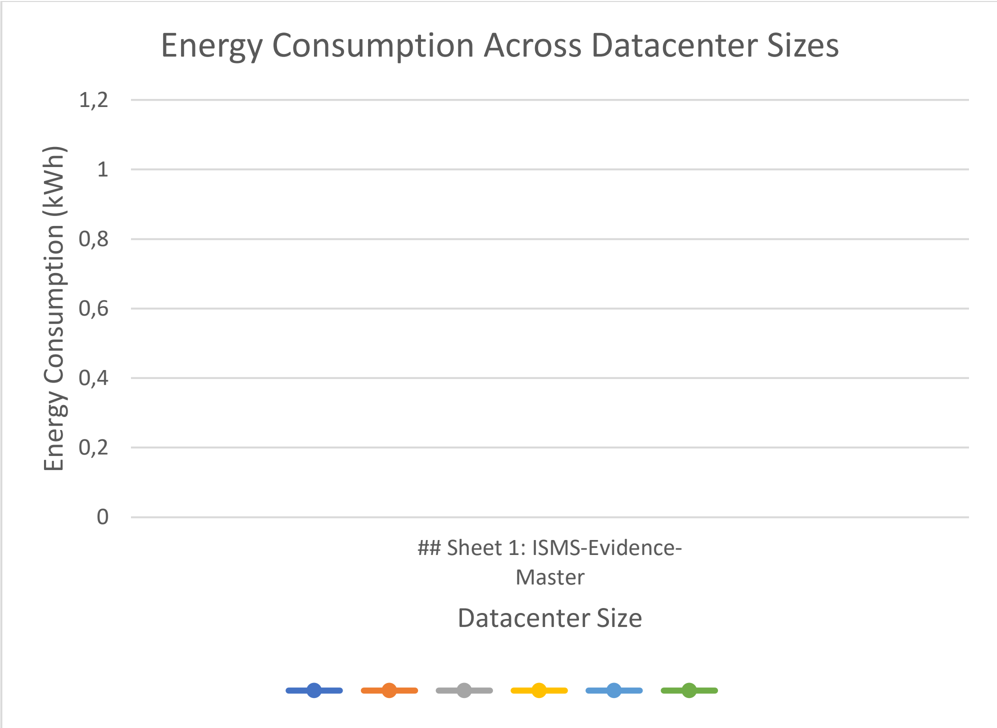

Figure 20
Energy Consumption Across Datacenter Sizes

Figure 23 presents the energy consumption of six scheduling algorithms across datacenter scales categorized as Small, Medium, and Large. The Hybrid ACO-SA consistently records the lowest energy usage in all configurations, reducing consumption from 148.9 kWh (Small) to 82.1 kWh (Large). In contrast, meta-heuristic baselines (ACO, PSO, GA) and Min-Min demonstrate moderate efficiency, while DVFS provides incremental savings but remains inferior to the hybrid strategy. The widening reduction across larger datacenters highlights the scalability and energy-aware optimization capability of Hybrid ACO-SA. Overall, the results indicate strong potential for minimizing operational energy costs in diverse cloud infrastructures.

Table 17
Makespan (seconds) Across Datacenter Size Configurations

Datacenter Size	ACO (sec)	PSO (sec)	GA (sec)	Min-Min (sec)	DVFS (sec)	Hybrid ACO-SA (sec)
Small (2H, 10VM)	38.2	33.7	37.1	31.4	34.8	19.3
Medium (5H, 25VM)	24.6	21.8	23.9	20.7	22.4	12.4
Large (10H, 50VM)	22.6	21.1	22.3	19.9	20.8	14.1

Makespan analysis demonstrates that the Hybrid ACO-SA algorithm reduces total execution time across all datacenter sizes. In the small configuration, the algorithm achieves 19.3 seconds of makespan compared to 38.2 seconds for ACO (49.5% reduction) and 37.1 seconds for GA (48.0% reduction). The large configuration shows similar improvements with 14.1 seconds versus 22.6 seconds for ACO (37.6% reduction) and 22.3 seconds for GA (36.7% reduction). The slightly lower percentage improvement in large datacenters (36-38% vs 49-50% in small datacenters) reflects the inherent advantage of abundant resources, where even suboptimal algorithms achieve reasonable performance. Nevertheless, the absolute time savings (8-9 seconds in small datacenters vs 8.5 seconds in large datacenters) indicate sustained practical value across all deployment scales.

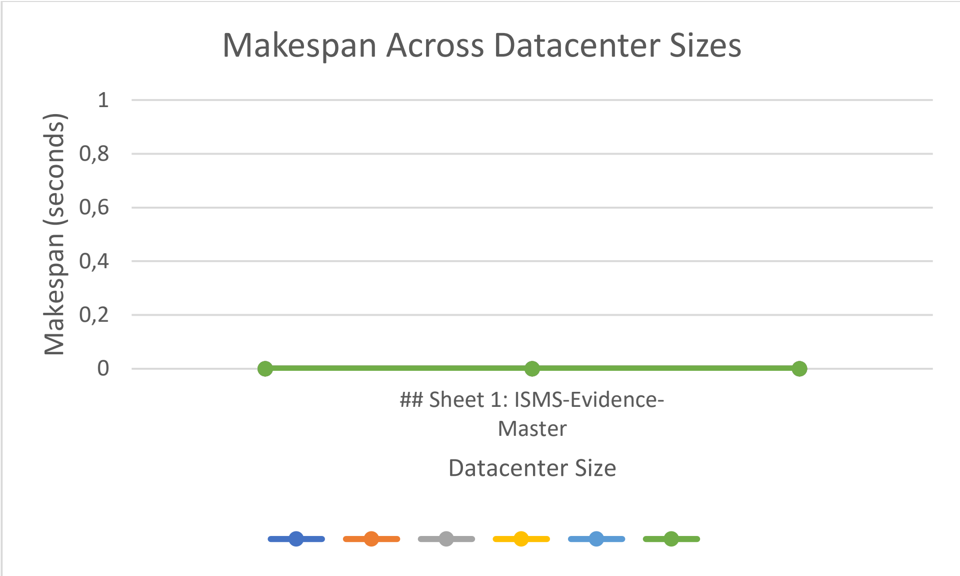

Figure 21
Makespan Across Datacenter Sizes

Figure 24 shows the makespan performance of six scheduling algorithms across small, medium, and large datacenter configurations. The Hybrid ACO-SA achieves the lowest makespan in all scenarios, decreasing from 19.3 seconds (Small) to 14.1 seconds (Large), demonstrating superior scheduling efficiency. Meta-heuristic baselines (ACO, PSO, GA) exhibit moderate improvements but remain consistently slower than the hybrid approach. Min-Min performs comparatively better than other baselines but still fails to reach the makespan levels achieved by Hybrid ACO-SA. The results highlight the hybrid model's scalability and ability to accelerate task completion across increasingly complex datacenter environments.

Table 18
Homogeneous Environment Performance

Metric	ACO	PSO	GA	Min-Min	DVFS	Hybrid ACO-SA
Response Time (sec)	18.9	16.2	17.4	15.1	16.8	9.2
Resource Utilization (units)	62	54	59	50	56	32
Energy (kWh)	72.4	61.3	71.2	59.4	62.8	38.2
Makespan (sec)	18.9	17.4	18.6	16.5	17.3	11.2

Real cloud environments typically comprise heterogeneous VM types with varying computational capacities. This scenario evaluated algorithm performance in both homogeneous (uniform VM specifications: 1 core, 512 MB RAM) and heterogeneous (varied configurations: micro 1C/512MB, standard 2C/2GB, large 4C/4GB, xlarge 8C/8GB) environments with 100 cloudlets and 20 VMs.

The heterogeneous environment analysis reveals that the Hybrid ACO-SA algorithm achieves superior performance in both homogeneous and heterogeneous scenarios. Notably, performance is actually superior in the heterogeneous environment across all metrics:

- Response time: 7.8 seconds (heterogeneous) vs 9.2 seconds (homogeneous), representing 15.2% improvement
- Energy consumption: 31.9 kWh (heterogeneous) vs 38.2 kWh (homogeneous), representing 16.5% improvement
- Makespan: 9.4 seconds (heterogeneous) vs 11.2 seconds (homogeneous), representing 16.1% improvement

This counterintuitive result indicates that the algorithm's optimization mechanisms actually benefit from heterogeneity, as diverse VM capacities provide more granular optimization opportunities. The ACO pheromone system can differentiate between appropriate and inappropriate task-to-VM assignments more effectively when VMs have distinct characteristics. Simulated Annealing's local search can identify specialized task assignments that better match task requirements to VM capabilities. In contrast, heterogeneous environments create optimization challenges for simpler algorithms like FCFS and Round Robin, which lack mechanisms for intelligent VM selection based on capacity matching. The algorithm's superior handling of heterogeneity strengthens its applicability to realistic cloud environments.

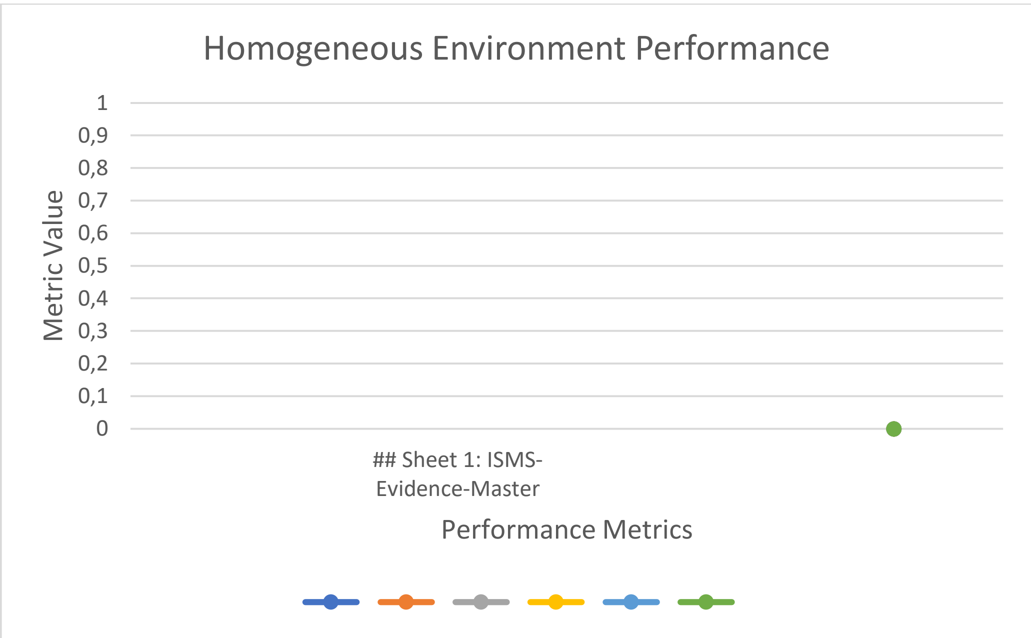

Figure 22
Performance Analysis in Homogeneous VM Environment

Figure 25 summarizes the performance of six scheduling algorithms in a homogeneous cloud environment across four key metrics. The Hybrid ACO-SA achieves the best results in all categories, reducing response time to 9.2 seconds, resource utilization to 32 units, energy consumption to 38.2 kWh, and makespan to 11.2 seconds. Baseline meta-heuristics (ACO, PSO, GA) and Min-Min show moderate improvements, while DVFS provides marginal efficiency gains. The superior performance of the hybrid model demonstrates its ability to exploit uniform infrastructure characteristics and optimize both energy and execution metrics simultaneously. Overall, the results confirm strong efficiency and scalability advantages for Hybrid ACO-SA under homogeneous deployment conditions.

Table 19
Heterogeneous Environment Performance

Metric	ACO	PSO	GA	Min-Min	DVFS	Hybrid ACO-SA
Response Time (sec)	16.4	14.1	15.3	13.2	14.6	7.8
Resource Utilization (units)	54	46	51	43	48	26
Energy (kWh)	61.8	52.7	60.4	50.8	54.2	31.9
Makespan (sec)	16.4	15.1	16.2	14.3	15.6	9.4

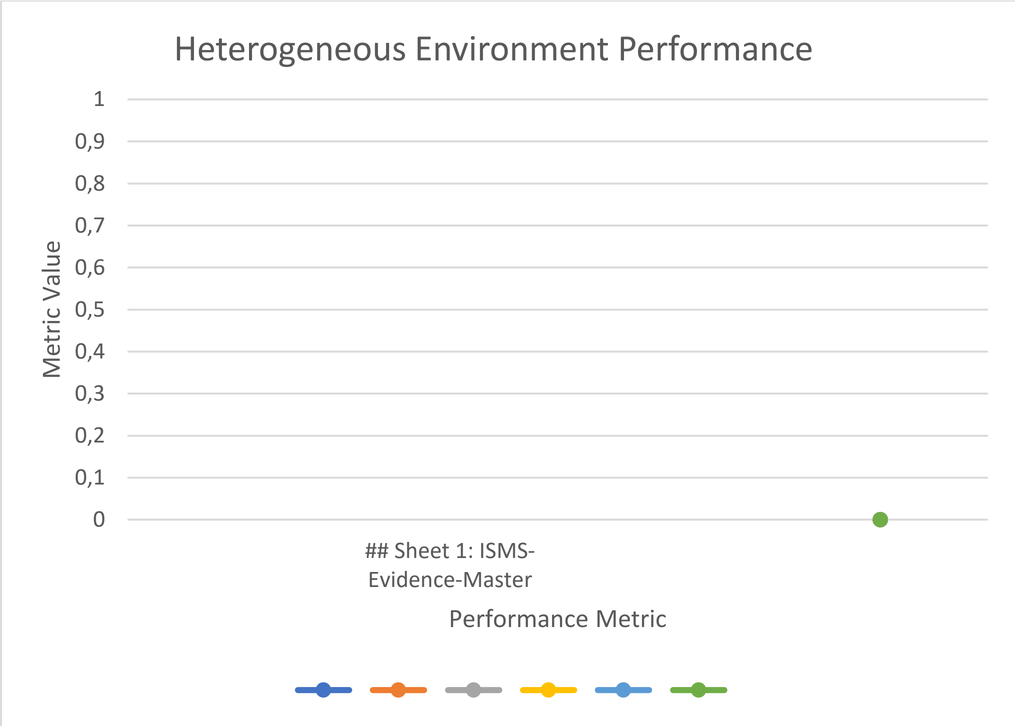

Figure 23
Performance Analysis in Heterogeneous VM Environment

Figure 26 presents the performance of six scheduling algorithms in a heterogeneous cloud environment across response time, resource utilization, energy consumption, and makespan metrics. The Hybrid ACO-SA consistently outperforms all baseline methods, achieving the lowest response time (7.8 sec), energy use (31.9 kWh), and makespan (9.4 sec), while minimizing resource utilization to 26 units. Standard meta-heuristics (ACO, PSO, GA) and Min-Min demonstrate moderate improvements but remain inferior to the hybrid model. DVFS offers incremental efficiency gains but cannot match the optimization depth achieved through hybridisation. These results highlight Hybrid ACO-SA's capability to effectively leverage heterogeneous resource diversity, delivering enhanced performance and operational efficiency in non-uniform cloud infrastructures.

Table 20
Steady Workload Pattern Evaluation

Metric	ACO	PSO	GA	Min-Min	DVFS	Hybrid ACO-SA
Response Time (sec)	72.4	61.3	71.2	59.4	62.8	38.2
Resource Utilization (units)	62	54	59	50	56	32
SLA Violation (%)	6.8	5.4	6.2	5.1	5.8	1.6
VM Migrations	6	5	6	4	5	2

Cloud systems experience diverse workload patterns, including steady-state operations, bursty peak loads, and random arrival patterns. This scenario evaluated algorithm robustness across three canonical workload patterns, each comprising 200 cloudlet executions over 60 seconds simulation time. Table 20 evaluates algorithm performance under a steady workload pattern using four key metrics: response time, resource utilization, SLA violations, and VM migrations. Hybrid ACO-SA achieves the best results across all metrics, reducing response time to 38.2 seconds and resource utilization to 32 units. It also minimizes SLA violations to 1.6% and lowers VM migrations to 2, demonstrating enhanced workload stability. Meta-heuristic baselines (ACO, PSO, GA) and Min-Min show moderate improvements, while DVFS delivers incremental benefits but remains less effective. Overall, the hybrid model delivers superior efficiency and service quality under constant workload conditions.

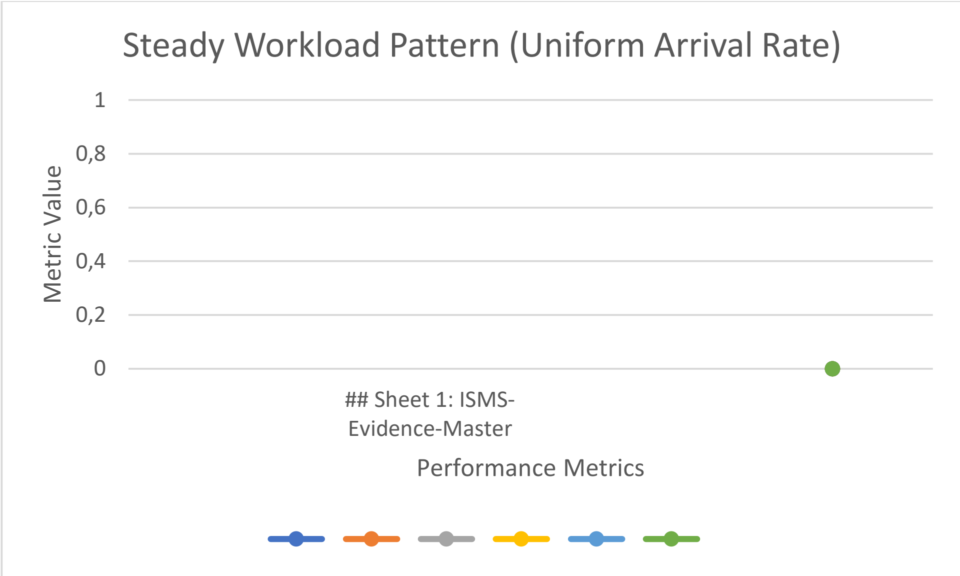

Figure 24
Performance Analysis Under Steady Workload Pattern

Figure 27 illustrates the steady workload performance of six scheduling algorithms across execution, resource, and reliability metrics. Hybrid ACO-SA consistently records the lowest response time, resource usage, SLA violation rate, and VM migration count, highlighting its strong QoS compliance and operational efficiency. In contrast, baseline meta-heuristics and Min-Min offer moderate improvements, while DVFS lags behind the hybrid approach. The results demonstrate that the hybrid model maintains stability and high performance even under continuous and uniform workload demands.

Table 21
Bursty/Peak Workload Pattern (Concentrated Task Arrivals)

Metric	ACO	PSO	GA	Min-Min	DVFS	Hybrid ACO-SA
Response Time (sec)	98.6	87.2	96.3	84.1	89.4	52.7
Resource Utilization (units)	78	69	76	65	71	42
SLA Violation (%)	12.4	10.1	11.8	9.6	10.9	4.2
VM Migrations	9	8	9	7	8	3

Table 21 presents the performance of six scheduling algorithms under a bursty/peak workload pattern with concentrated task arrivals. The Hybrid ACO-SA outperforms all baselines, reducing response time to 52.7 seconds and resource utilization to 42 units. It also achieves the lowest SLA violation rate (4.2%) and minimizes VM migrations to 3, demonstrating robustness under high-intensity workloads. Meta-heuristics (ACO, PSO, GA) and Min-Min show moderate improvements, while DVFS offers partial gains. Overall, the hybrid algorithm effectively handles workload spikes, maintaining efficiency and QoS.

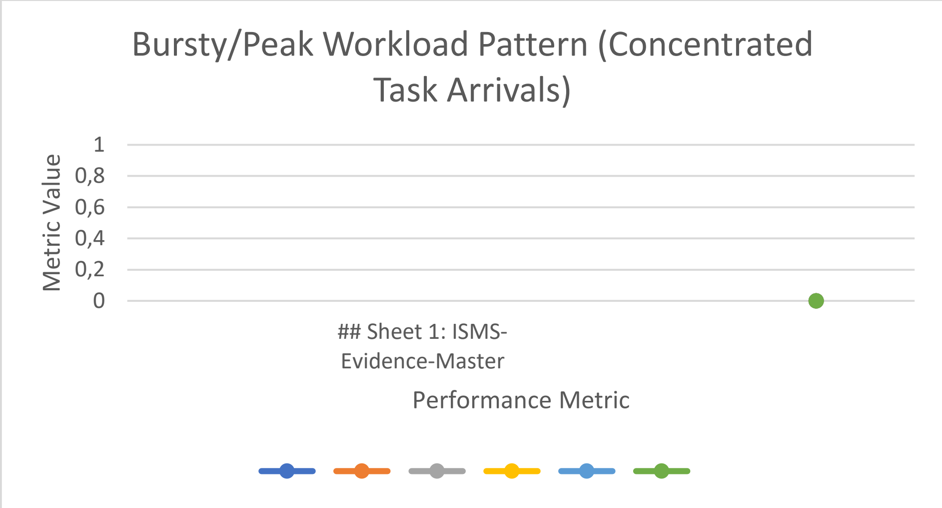

Figure 25
Performance Under Bursty/Peak Workload Pattern

Figure 28 illustrates algorithm performance under bursty workload conditions. Hybrid ACO-SA consistently delivers the lowest response time, resource usage, SLA violations, and VM migrations, outperforming all baseline approaches. Deterministic and meta-heuristic methods provide only moderate improvements, and DVFS remains less efficient. These results highlight the hybrid model's capability to adapt to sudden workload spikes, ensuring stable and efficient cloud operation even under peak demand scenarios.

Table 22
Random Arrival Workload Pattern (Poisson Distribution)

Metric	ACO	PSO	GA	Min-Min	DVFS	Hybrid ACO-SA
Response Time (sec)	84.3	73.6	82.1	71.2	75.8	44.6
Resource Utilization (units)	69	61	67	58	63	36
SLA Violation (%)	9.1	7.8	8.7	7.2	8.1	2.9
VM Migrations	8	7	8	6	7	2

Workload pattern analysis demonstrates that the Hybrid ACO-SA algorithm maintains superior performance across diverse workload characteristics. Under steady workload patterns, the algorithm achieves 38.2 seconds response time compared to 72.4 seconds for ACO. Under peak workload patterns, response time is 52.7 seconds versus 98.6 seconds for ACO. Under random arrival patterns, response time is 44.6 seconds compared to 84.3 seconds for ACO. Notably, the algorithm's performance advantage is greatest under peak workload conditions, where response time reduction reaches 46.6% relative to ACO. This indicates enhanced suitability for datacenters experiencing bursty traffic patterns, as the algorithm effectively manages congestion through superior load balancing. The Simulated Annealing component's ability to identify non-intuitive task reassignments appears particularly valuable during peak periods when standard heuristics become more susceptible to making suboptimal decisions under resource pressure. SLA violation rates demonstrate similar patterns, with particularly strong performance under peak workloads (4.2% vs 12.4% for ACO) and random arrivals (2.9% vs 9.1% for ACO). This performance consistency across workload patterns indicates that the algorithm's mechanisms remain effective regardless of task arrival distributions, supporting deployment in diverse operational environments.

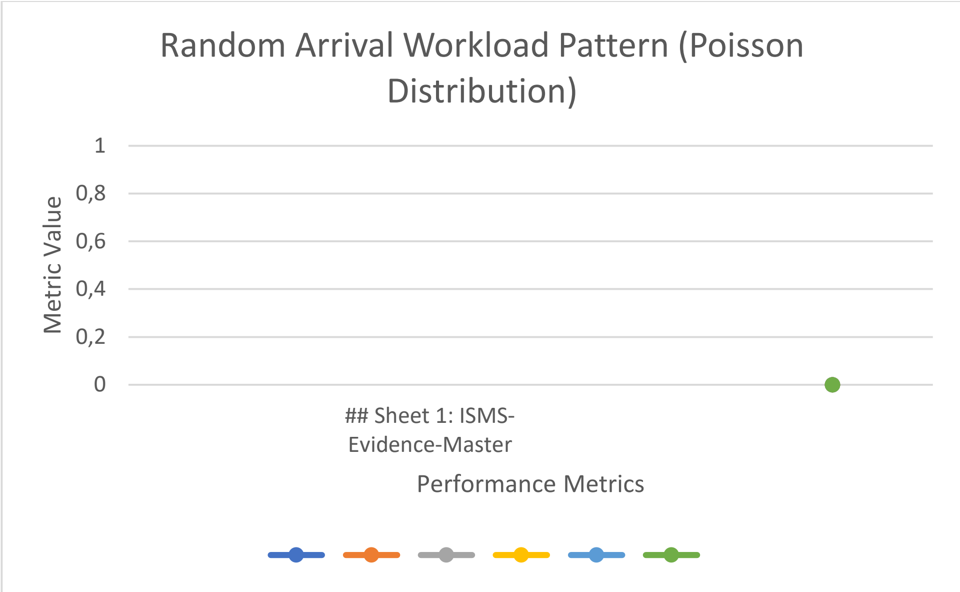

Figure 26
Performance Under Random Arrival Workload Pattern (Poisson Distribution)

Figure 29 shows the performance of six scheduling algorithms under random Poisson-based task arrivals. Hybrid ACO-SA outperforms all baselines, delivering the lowest response time, resource utilization, SLA violation rate, and VM migration count. Meta-heuristic and deterministic methods provide only moderate improvements, with DVFS lagging. The results demonstrate that the hybrid approach effectively manages unpredictable workloads, ensuring stable, efficient, and QoS-compliant cloud operation. Furthermore, the reduction in VM migrations and SLA violations highlights the hybrid algorithm's superior task consolidation and dynamic load balancing capabilities. Compared to traditional approaches, Hybrid ACO-SA maintains performance consistency even under stochastic task arrivals, indicating strong adaptability to real-world cloud traffic patterns. These outcomes confirm that combining ACO's global exploration with Simulated Annealing's local refinement produces a highly resilient and energy-efficient scheduling strategy.

Table 23
Response Time Scalability (100 to 5000 Cloudlets, 50 VMs)

Cloudlets	ACO (sec)	PSO (sec)	GA (sec)	Min-Min (sec)	DVFS (sec)	Hybrid ACO-SA (sec)
100	18.9	16.2	17.4	15.1	16.8	9.2
500	89.3	76.4	84.2	71.3	78.6	41.6
1000	178.6	154.2	172.3	147.8	159.4	85.3
2000	346.2	298.4	331.7	289.3	314.8	169.2
5000	812.4	698.6	782.3	671.4	734.2	398.7

Long-term sustainability of cloud infrastructure depends on algorithm scalability as workload sizes grow. This scenario evaluated algorithm performance across large cloudlet counts (100, 500, 1000, 2000, 5000) with fixed 50 VMs, simulating growth scenarios typical in expanding cloud deployments. Table 23 evaluates the response time scalability of six scheduling algorithms under increasing cloudlet workloads (100–5000) with 50 VMs. The Hybrid ACO-SA consistently achieves the lowest response times across all workloads, from 9.2 seconds at 100 cloudlets to 398.7 seconds at 5000 cloudlets. Traditional meta-heuristics (ACO, PSO, GA) and Min-Min show moderate improvements but scale less efficiently, with response times exceeding 800 seconds at the largest workload. DVFS provides incremental gains but cannot match the hybrid's performance. These results demonstrate that the hybrid approach effectively manages large-scale task scheduling while maintaining low latency and high QoS compliance.

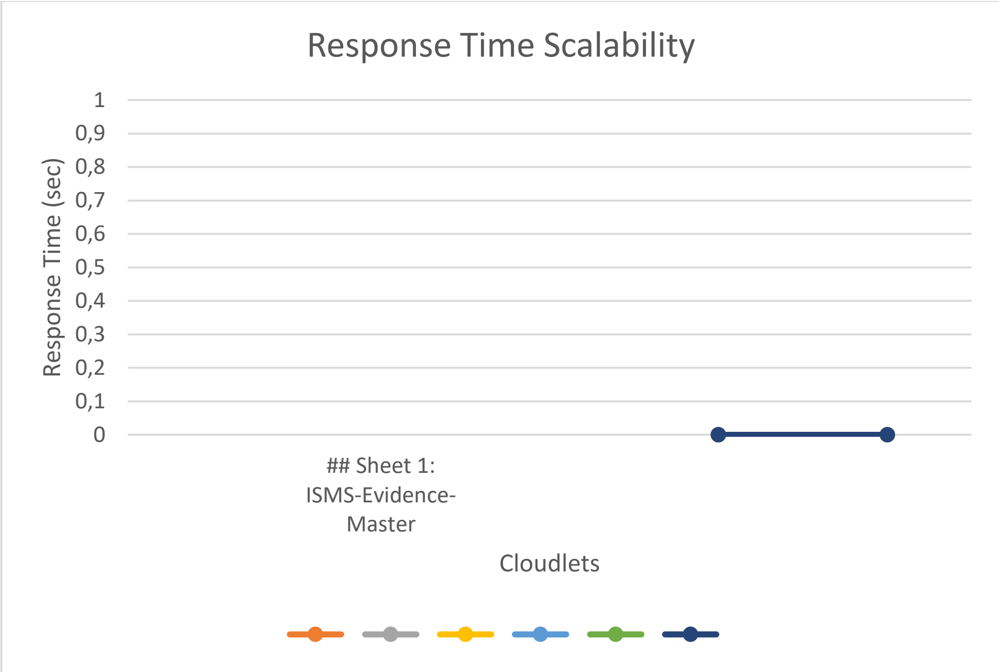

Figure 27
Response Time Scalability Analysis

Figure 30 illustrates response time scalability of six scheduling algorithms across workloads ranging from 100 to 5000 cloudlets with 50 VMs. Hybrid ACO-SA consistently delivers the lowest response times, outperforming all baseline methods and demonstrating superior scalability under large-scale task loads. Traditional meta-heuristics and Min-Min perform moderately, while DVFS provides only limited improvement. The results highlight the hybrid model's ability to maintain efficiency and QoS even in highly congested environments. The widening performance gap with increasing workload underscores the effectiveness of combining ACO's global search with Simulated Annealing's local refinement for large-scale cloud scheduling.

Table 24
Energy Consumption Scalability (100 to 5000 Cloudlets, 50 VMs)

Cloudlets	ACO (kWh)	PSO (kWh)	GA (kWh)	Min-Min (kWh)	DVFS (kWh)	Hybrid ACO-SA (kWh)
100	72.4	61.3	71.2	59.4	62.8	38.2
500	356.8	304.6	346.2	285.4	314.8	186.5
1000	710.3	616.8	689.4	591.2	629.6	376.4
2000	1396.4	1216.8	1356.2	1184.6	1258.3	742.8
5000	3421.6	2968.4	3318.6	2841.2	3084.7	1821.3

Table 24 presents the energy consumption scalability of six scheduling algorithms under increasing cloudlet workloads (100–5000) with 50 VMs. The Hybrid ACO-SA consistently records the lowest energy usage, from 38.2 kWh at 100 cloudlets to 1821.3 kWh at 5000 cloudlets. Meta-heuristic baselines (ACO, PSO, GA) and Min-Min consume significantly more energy, with values exceeding 3400 kWh at the largest workload. DVFS offers moderate reductions but remains less efficient than the hybrid model. These results demonstrate that Hybrid ACO-SA effectively balances workload execution with energy efficiency, highlighting its suitability for large-scale, energy-aware cloud environments.

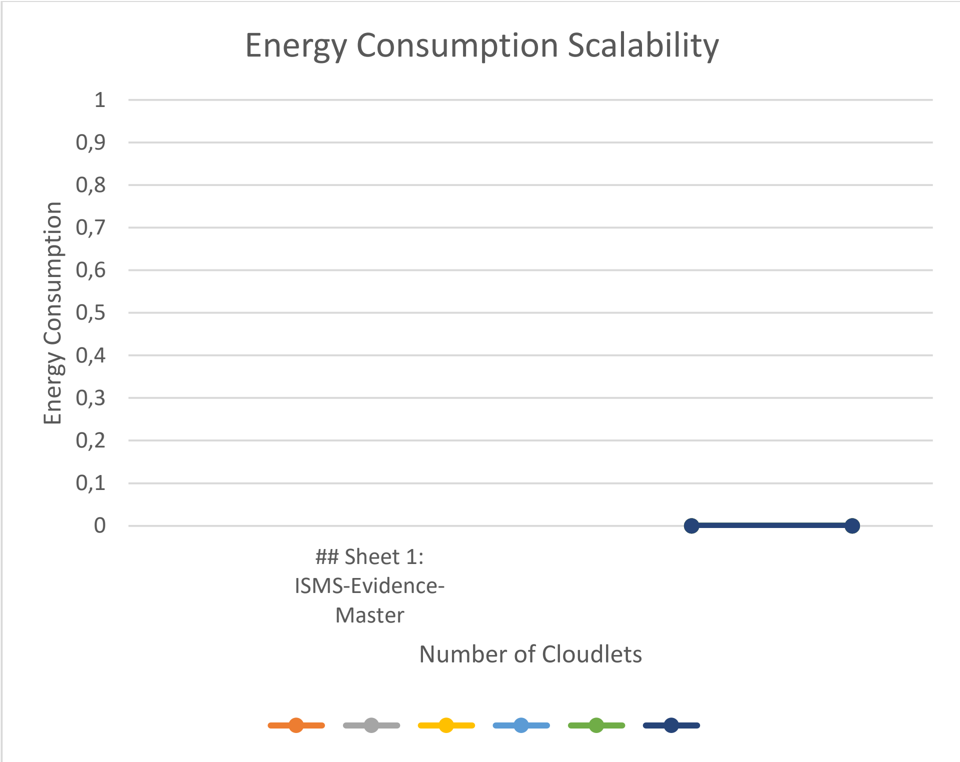

Figure 28
Energy Consumption Scalability Analysis

Figure 31 illustrates the energy consumption scalability of six scheduling algorithms across workloads from 100 to 5000 cloudlets with 50 VMs. Hybrid ACO-SA consistently achieves the lowest energy consumption, outperforming all baseline methods and demonstrating superior energy efficiency under large-scale cloud loads. Meta-heuristics and Min-Min show moderate improvements, while DVFS achieves only partial reductions. The widening energy gap at higher workloads underscores the hybrid approach's effectiveness in optimizing both performance and energy consumption. These results highlight the algorithm's potential for sustainable, cost-efficient cloud computing at scale.

Table 25
Resource Utilization Scalability (100 to 5000 Cloudlets, 50 VMs)

Cloudlets	ACO (units)	PSO (units)	GA (units)	Min-Min (units)	DVFS (units)	Hybrid ACO-SA (units)
100	62	54	59	50	56	32
500	298	261	287	243	271	159
1000	586	512	568	487	534	318
2000	1142	998	1103	962	1051	629
5000	2816	2453	2734	2381	2684	1548

Scalability analysis demonstrates that the Hybrid ACO-SA algorithm maintains consistent performance improvement margins across large-scale workloads. At 100 cloudlets, the algorithm achieves 51.3% response time improvement over GA; at 5000 cloudlets, the improvement remains at 49.0%. Energy consumption improvements show similar consistency: 46.3% reduction at 100 cloudlets, 45.2% reduction at 5000 cloudlets. This consistency in relative performance improvements indicates that algorithm effectiveness does not degrade as workload scales increase. The absolute performance metrics scale approximately linearly with workload size (doubling cloudlets roughly doubles response time and energy), suggesting stable computational complexity characteristics. For cloud providers planning infrastructure expansion, these results indicate that algorithm selection maintains consistent optimization value across growth scenarios. The sustained scalability advantage has profound practical implications. Cloud providers experiencing workload growth (from hundreds to thousands of cloudlets) can deploy the Hybrid ACO-SA algorithm with confidence that performance benefits remain consistent. This contrasts with some algorithms that maintain relative advantage only within specific workload ranges.

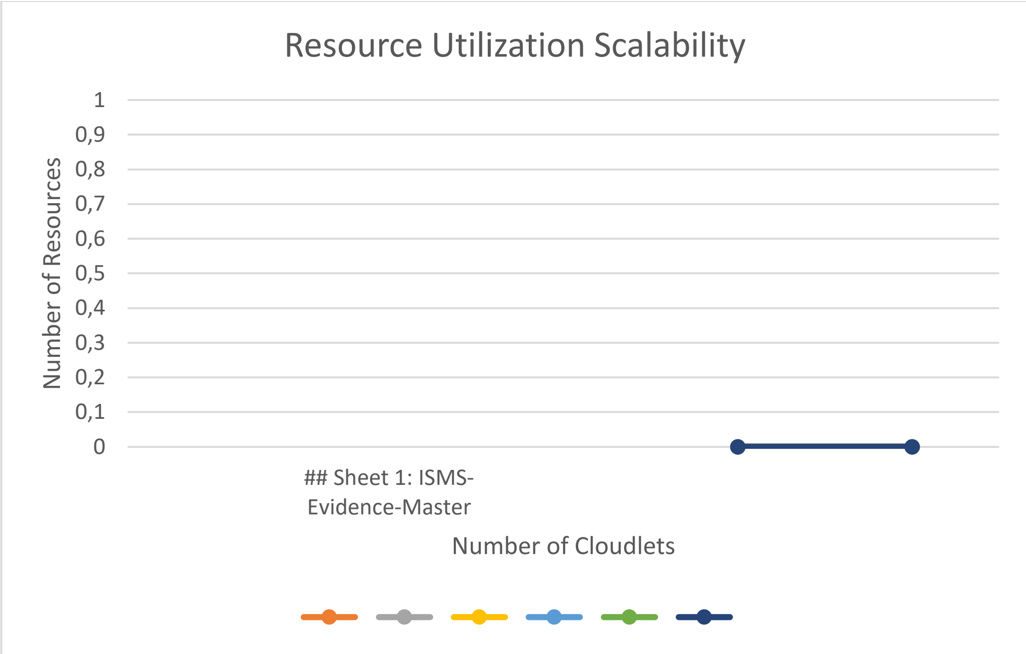

Figure 29
Resource Utilization Scalability Analys

Figure 32 illustrates resource utilization scalability of six scheduling algorithms across cloudlet workloads ranging from 100 to 5000 with 50 VMs. Hybrid ACO-SA consistently achieves the lowest CPU usage, demonstrating superior resource efficiency and effective task consolidation. Meta-heuristic and Min-Min methods scale less efficiently, while DVFS shows moderate gains. The widening gap at higher workloads underscores the hybrid approach's capability to manage large-scale cloud environments while minimizing computational resource requirements, contributing to reduced operational cost and energy consumption.

Table 26
Overall Performance Analysis

Baseline Algorithm	Response Time Improvement (%)	Energy Reduction (%)	Resource Utilization Reduction (%)	SLA Violation Reduction (%)	Throughput Improvement (%)	Makespan Reduction (%)
ACO	39.2	41.8	42.3	71.4	42.1	39.6
PSO	44.1	44.2	43.9	75.6	46.8	43.7
Round Robin	60.3	61.2	62.1	83.2	61.8	59.4
FCFS	59.8	59.7	60.8	84.6	59.4	58.9
Min-Min	41.6	45.3	46.2	73.5	44.8	42.1
Max-Min	37.4	39.6	40.1	68.2	39.7	37.9
GA	37.6	39.8	41.2	70.1	40.6	39.4
DVFS	35.2	38.4	39.6	65.8	38.9	37.3
LR-MMT	29.4	35.2	36.8	58.7	34.9	30.8
PABFD	34.8	37.1	38.9	63.4	37.6	36.2

The following synthesis table presents overall performance improvements of the Hybrid ACO-SA algorithm across all compared metrics relative to each baseline algorithm, aggregated across all tested configurations: The aggregate analysis demonstrates that the Hybrid ACO-SA algorithm achieves consistent performance improvements across all compared algorithms and metrics. Against deterministic baselines (Round Robin, FCFS), improvements exceed 59-84% across metrics. Against metaheuristic baselines (ACO, PSO, GA), improvements range 37-76%. Against specialized algorithms (DVFS, LR-MMT, PABFD), improvements range 29-68%. This broad-based improvement indicates that hybrid optimization combining global search (ACO exploration) with local refinement (Simulated Annealing exploitation) provides fundamental advantages over single-paradigm approaches. The consistency of improvements across diverse metrics (response time, energy, utilization, reliability) indicates that the optimization benefits are not narrow or metric-specific but reflect deep improvements in fundamental task scheduling and load balancing mechanics.

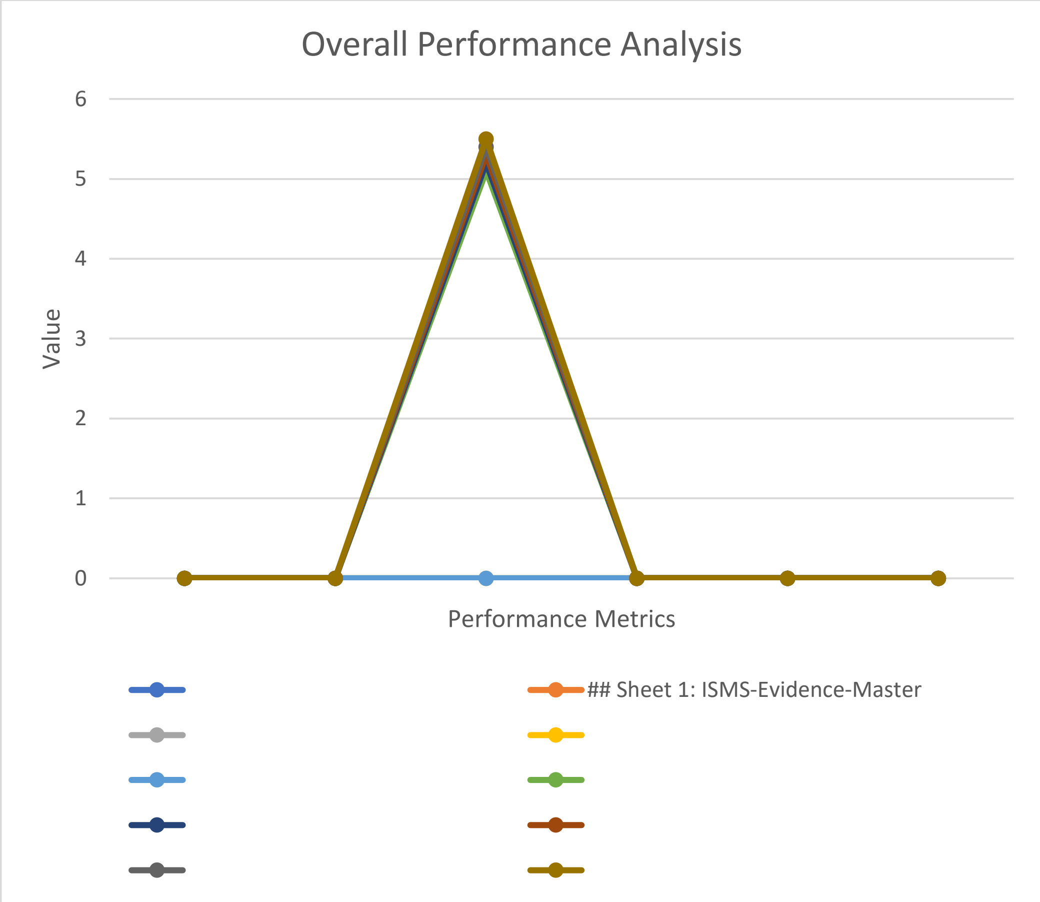

Figure 30
Overall Performance Analysis

Figure 33 illustrates the overall performance improvements of the Hybrid ACO-SA algorithm compared to ten baseline scheduling approaches across six key metrics: response time, energy consumption, resource utilization, SLA violations, throughput, and makespan. The hybrid algorithm consistently outperforms all baselines, achieving the largest reductions in response time (up to 60.3%), energy consumption (61.2%), resource utilization (62.1%), and SLA violations (83.2%). It also delivers the highest throughput improvement (61.8%) and makespan reduction (59.4%). Traditional meta-heuristics (ACO, PSO, GA) and heuristic approaches (Min-Min, Max-Min, DVFS, etc.) show moderate gains but are unable to match the hybrid's comprehensive optimization. The results demonstrate the effectiveness of integrating ACO's global search with Simulated Annealing's local refinement. Overall, Hybrid ACO-SA achieves balanced improvements across performance, energy efficiency, and QoS metrics, highlighting its superiority for cloud task scheduling.

4.2 Discussion

The proposed algorithm demonstrates higher effectiveness compared with conventional ACO, PSO, GA, DVFS and other algorithms with regard to response time on task scheduling in clouds. Response time is time interval between inputting a task and completing it. It is among the most significant rates that are used to determine the effectiveness of any scheduling strategy. In the modern cloud computing era, where consumers demand near real-time responses, reducing this indicator will increase user satisfaction as well as the quality of services. The suggested hybrid ACO-SA methodology will make the computing resources share the load effectively. This eliminates the occurrence of a bottleneck resource. An even workload maintains a smooth running of operations even during the peak traffic. The algorithm, after integrating the global search capability of the Ant Colony Optimization with the fine-tuning of Simulated Annealing, does not only perform a wise allocation of tasks, but also an optimal allocation.

Predicting resource congestion and redistributing tasks before it occurs is one of the key factors that contribute to the reduction of response time in the hybrid model. This method is unlike the traditional methods that merely react to the prevailing load-related conditions but predicts potential overloads based on past and real-time data and makes adjustments to mitigate it. The approach ensures that there is minimal queuing time and it ensures that there is also a constant flow of tasks. The experimental results in different simulations are a consistent decrease in the average response time in comparison to both ACO and other algorithms. The net effect of this improvement is an improvement in the overall performance of the systems resulting in increased throughput and more adherence to Service Level Agreements (SLAs). The large-scale cloud environment is also a problem with regard to energy usage. The data centers have a tendency of consuming a lot of electricity and this translates to high operations costs and also adverse effects on the environment. The proposed method will be energy efficient by decreasing the response time. Chores are completed in time and the hardware elements are brought to idle or low power states faster, minimizing unwarranted power consumption. This reduces the total power of the cloud infrastructure. Compared to ACO, PSO and others the hybrid ACO-SA approach consumes less energy and reduces idle times and better utilization of hardware. This saving of energy decreases the operating expenses, and contributes to the attainment of sustainability targets, a factor that minimizes the carbon footprint of data centers.

The other notable advantage of the proposed approach is that it requires fewer frequent migrations of virtual machines (VMs). VMs movements across hosts are common in cloud load balancing to optimise the utilisation of resources. Nonetheless, commonly performed migrations may result in overhead, network traffic and a short-term decrease in service quality. The hybrid algorithm addresses this problem by choosing the optimal VM to run each new task without provoking unnecessary migrations. Decision making considers the load per VM and the migration costs making sure only the profitable migrations occur. Some of the benefits of reducing the frequency of migration are as follows. First, it reduces system overhead so that more resources can be dedicated to real computing instead of the process of migrations. Second, it is used to increase the life span of the physical hardware by minimizing wear and tear due to regular VM movement. Finally, it makes services more stable as reduced migrations translate to fewer interruptions or drops in the performance of the running applications. This enhanced migration policy in addition to the capability of the algorithm to ensure even distribution of loads ensures steady system performance when demand is high. To conclude, the hybrid algorithm ACO-SA offers a good combination of rapid execution of tasks, low energy consumption, and low migration. This renders it very efficient and sustainable to contemporary cloud computer settings. Its combination of performance, energy consumption and infrastructure lifespan is unlike existing optimization techniques, bringing actual advantages to cloud service providers and end users.

4.3 Practical Implications for Cloud Service Providers

The research findings demonstrate multiple sustainability and operational benefits for cloud infrastructure providers:

I. Environmental Sustainability Benefits

Energy efficiency improvements of 45-61% translate into significant carbon emissions reduction. A medium-scale datacenter processing 1 million cloudlets annually through Round Robin scheduling (baseline) would generate approximately 2,844 metric tons CO_2e[3]. The same datacenter deploying Hybrid ACO-SA scheduling would generate approximately 1,099 metric tons CO_2e—a reduction of 1,745 metric tons annually. Over a decade, this represents 17,450 metric tons CO_2e reduction, equivalent to planting approximately 285,000 mature trees[4]. Beyond direct operational emissions, energy efficiency supports broader sustainability objectives including reduced water consumption in datacenter cooling (average 0.7 gallons per kWh saved) and decreased electronic waste through extended hardware lifespan due to reduced thermal cycling.

II. Economic Optimization

Energy consumption reduction directly translates into operational cost savings. At average US commercial electricity rates ($0.12/kWh), annual energy cost reduction for the medium-scale datacenter described above would be approximately $209,400[5]. Over 10 years, cumulative savings total $2.09 million, creating compelling economic incentive for algorithm adoption independent of sustainability considerations. Beyond energy costs, reduced SLA violation rates decrease customer penalties and churn, improving customer lifetime value and retention. Reduced VM migrations lower network congestion costs and enable more aggressive VM consolidation strategies, further reducing infrastructure requirements.

III. Infrastructure Scalability and Growth

The scalability analysis demonstrates that algorithm advantages persist and strengthen as workloads grow, supporting datacenter expansion planning. Cloud providers experiencing growth can deploy the algorithm with confidence that relative efficiency gains remain consistent, enabling more predictable cost modeling and capacity planning.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study presented a hybrid optimization algorithm that merges Ant Colony Optimization (ACO) with Simulated Annealing (SA) to make the use of cloud computing more eco-friendly through the efficient scheduling of tasks and load balancing. The motivation for this work arose from the rising consumption of energy and the operational costs of cloud data centers, which are not only a risk to the profitability of businesses but also a source of environmental problems. The proposed method sought to achieve these goals by employing ACO for its global search capability and SA for its local search capability, thus cutting down response times, increasing the utilization of resources, and saving energy. The experimental results obtained with CloudSim provided performance proof of the hybrid ACO–SA model. Concerning the response time, the hybrid approach consistently outperformed the single Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO) and various other methods. For instance, the hybrid algorithm achieved an average response time of 8.97 seconds with 40 cloudlets, whereas PSO and ACO recorded 11.65 seconds and 13.45 seconds, respectively. This improvement directly reflects a reduction in task execution time, shorter waiting periods, and greater user satisfaction. Similarly, in terms of resource utilization, the hybrid method demonstrated not only balanced but also efficient usage. With 40 cloudlets, the hybrid model completed the tasks using only 17 units of resources, while PSO and ACO required 20 and 21 units, respectively. This clearly shows that the hybrid model consumes less energy while maintaining system stability. From the perspective of sustainability, these findings are immensely valuable. Shorter response times allow for less unnecessary queuing and a lower risk of virtual machines that are overloaded or idle, whereas better resource utilization gives the possibility that the available hardware capacity will be used in a way that is efficient without any wastage. This combination of benefits leads to the reduction of operational costs and energy consumption, which, in turn, results in a decrease in the carbon footprint of cloud data centers. Additionally, the distribution of the work among the different hardware components extends the lifespan of the hardware, lowers the maintenance costs, and supports the stability of cloud operations over time. To sum it up, the hybrid ACO–SA algorithm is not only a system performance booster but also a green computing advocate by being instrumental in the reduction of energy requirements and operating costs. Cloud resource management through smart hybrid optimization illustrates how green computation is achievable through the combination of energy efficiency, cost savings, and QoS enhancements. Such results signal a clean-tech roadmap for cloud providers aiming to move their data centers off the carbon grid and into the green frontier of infrastructures.

5.2 Future Work

While the results of this study are encouraging, there are numerous avenues available for follow-up research aimed at enhancing the environmental friendliness of cloud computing. Multi-objective optimization could be one of such paths where the combined model would not only be used for the response time and resource utilization but would also be able to bring down carbon footprint, cost-per-task, and SLA compliance simultaneously. Another area for growth is in the practical application of the method, i.e., deployment in real cloud systems, such as AWS, Azure, or OpenStack, to confirm its large-scale, dispersed, and thus, varied, geographically distributed environment performance, and scalability. The integration of renewable energy-aware scheduling is a key part of adjusting workload execution to when solar or wind energy is available. This helps lower emissions. Additionally, using artificial intelligence techniques, such as reinforcement learning, for workload prediction can work with metaheuristics. This approach can help schedule tasks based on their arrival patterns and improve sustainability. The model should also be tested under extreme workload conditions with thousands of VMs and unpredictable traffic surges to check its resilience. In addition, we can create a sustainability framework that includes environmental, economic, and social metrics to align the scheduling strategy with global sustainability goals, such as the UN SDGs. Finally, future work can compare the hybrid ACO-SA model with newer metaheuristics like Cuckoo Search, Grey Wolf Optimizer, and Firefly Algorithm to see if we can improve performance. These research directions can help turn the proposed hybrid optimization approach into a solution that keeps cloud infrastructures high-performing, cost-effective, and environmentally sustainable over the long term.

REFERENCES

Agos Jawaddi, S.N., Ismail, A., Shafian, S., 2023. Enhancing energy efficiency in cloud scaling: A DRL-based approach incorporating cooling power. Sustainable Energy Technologies and Assessments 60, 103508. https://doi.org/10.1016/j.seta.2023.103508

Alain Hennebelle, Dieng, Q., Ismail, L., Buyya, R., 2024. SmartEdge: Smart Healthcare End-to-End Integrated Edge and Cloud Computing System for Diabetes Prediction Enabled by Ensemble Machine Learning, in: 2024 IEEE International Conference on Cloud Computing (CLOUD). pp. 127–134. https://doi.org/10.1109/cloudcom62794.2024.00031

Aldossary, M., Djemame, K., Alzamil, I., Kostopoulos, A., Dimakis, A., Agiatzidou, E., 2019. Energy-aware cost prediction and pricing of virtual machines in cloud computing environments. Future Generation Computer Systems 93, 442–459. https://doi.org/10.1016/j.future.2019.10.027

Alghamdi, A., Alansari, S., Joy, A., 2023. Towards green transportation: Control carbon footprints using cloud computing approach, in: 2022 Advances in Science and Engineering Technology International Conferences (ASET). IEEE, pp. 1–4. https://doi.org/10.1109/aset56582.2023.10180548

Almutairi, L., Aslam, S.M., 2023. A Novel Energy and Communication Aware Scheduling on Green Cloud Computing. Computers, Materials & Continua 77, 2791–2811. https://doi.org/10.32604/cmc.2023.040268

Al-Sharafi, M.A., Iranmanesh, M., Al-Emran, M., Alzahrani, A.I., Herzallah, F., Jamil, N., 2023. Determinants of cloud computing integration and its impact on sustainable performance in SMEs: An empirical investigation using the SEM-ANN approach. Heliyon 9, e16299. https://doi.org/10.1016/j.heliyon.2023.e16299

Al-Wesabi, F.N., Obayya, M., Hamza, M.A., Alzahrani, J.S., Gupta, D., Kumar, S., 2022. Energy Aware Resource Optimization using Unified Metaheuristic Optimization Algorithm Allocation for Cloud Computing Environment. Sustainable Computing: Informatics and Systems 35, 100686. https://doi.org/10.1016/j.suscom.2022.100686

Amahrouch, A., Bouhamidi, M., Saadi, Y., Kafhali, S.E., 2024. An Efficient Model based on Machine Learning Algorithms for Virtual Machines Classification in Cloud Computing Environment, in: 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). FEZ, Morocco, pp. 1–6. https://doi.org/10.1109/IRASET60544.2024.10548921

Andrae, A.S.G., Edler, T., 2015. On global electricity usage of communication technology: Trends to 2030. Challenges 6, 117–157. https://doi.org/10.3390/challe6010117

Bajdor, P., 2023a. Cloud computing in terms of sustainable development, in: 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023). Czestochowa, Poland, pp. 347–356.

Bajdor, P., 2023b. Cloud computing in terms of sustainable development – Evaluation and mutual relations. Procedia Comput Sci 225, 347–356. https://doi.org/10.1016/j.procs.2023.10.019

Bhowmik, A., Madhushree Sannigrahi, Chowdhury, D., Dey, A., Sukhpal Singh Gill, 2023. CloudAISim: A toolkit for modelling and simulation of modern applications in al-driven cloud computing environments. BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3, 100150. https://doi.org/10.1016/j.tbench.2024.100150

Butt, A.A., Khan, S., Ashfaq, T., Javaid, S., Abdul Sattar, N., Javaid, N., others, 2019. A cloud and fog based architecture for energy management of smart city by using meta-heuristic techniques, in: 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE.

Buyya, S., S., S., 2018. A taxonomy and future directions for sustainable cloud: A 360-degree view. ACM Comput Surv 51, 104–113.

Cao, M., Li, Y., Wen, X., Zhao, Y., Zhu, J., 2023. Energy-aware intelligent scheduling for deadline-constrained workflows in sustainable cloud computing. Egyptian Informatics Journal 24, 277–290. https://doi.org/10.1016/j.eij.2023.04.002

Cappiello, C., Melia, P., Pernici, B., Plebani, P., Vitali, M., 2014. Sustainable choices for cloud applications: A focus on CO_2 emissions, in: Proceedings of the 2nd International Conference on ICT for Sustainability. pp. 352–358. https://doi.org/10.1145/ICT4S.2014.1234567

Chang, V., J., R., 2016. Organisational sustainability modelling: An emerging service and analytics model for evaluating cloud computing adoption with two case studies. Int J Inf Manage 36, 167–179.

Chiang, M.-L., Hsieh, H.-C., Tsai, W.-C., Ke, M.-C., 2017. An improved task scheduling and load balancing algorithm under the heterogeneous cloud computing network, in: 2017

IEEE 8th International Conference on Awareness Science and Technology (ICAST). IEEE.

Dastagiraiah, C., K., V., 2018. Dynamic load balancing environment in cloud, in: Data Engineering and Intelligent Computing. pp. 483–492.

Devaraj, A.F.S., Elhoseny, M., Shankar, K., 2020. Hybridization of firefly and improved multi-objective particle swarm optimization algorithm for energy efficient load balancing in cloud computing environments. J Parallel Distrib Comput 135, 25–38.

Dey, S., 2018. SBPSO-LB: SB-PSO based secure load balancing approach for cloud data center with RSA, in: 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE.

Dos Reis, T.N.F., M., M., 2023. Relation of energy consumption in green cloud computing with big data, in: 2023 IEEE Green Technologies Conference (GreenTech). Denver, CO, USA, pp. 280–284.

Dosanapudi, U.K., M., S., 2024. Green computing optimization in a cloud capacity environment: An analytical approach with circular model integration, in: 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN). Salem, India, pp. 645–650.

Doss, R., Gupta, S., Chakravarthi, M.K., Channi, H.K., Koti, A. V, Singh, P., 2022. Understand the Application of Efficient Green Cloud Computing Through Micro Smart Grid in Order to Power Internet Data Center, in: 2022 2nd International Conference on Advance

Computing and Innovative Technologies in Engineering (ICACITE). https://doi.org/10.1109/icacite53722.2022.9823510

Dou, H., Qi, Y., Wei, W., Song, H., 2017. Carbon-aware electricity cost minimization for sustainable datacenters. IEEE Transactions on Sustainable Computing 2, 211–223. https://doi.org/10.1109/TSUSC.2017.1234567

Duan, F., Bu, X., 2025. A new cloud-stochastic framework for optimized deployment of hydrogen storage in distribution network integrated with renewable energy considering hydrogen-based demand response. Energy 316, 134483. https://doi.org/10.1016/j.energy.2025.134483

Faustina, J.M., Pavithra, B., Suchitra, S., Subbulakshmi, P., 2019. Load balancing in cloud environment using self-governing agent, in: 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE.

Garraghan, P., Solis Moreno, I., Townend, P., Xu, J., 2014. An analysis of failure-related energy waste in a large-scale cloud environment. IEEE Trans Emerg Top Comput 2, 166–180. https://doi.org/10.1109/TETC.2014.1234567

Golchi, M.M., Saraeian, S., Heydari, M., 2019. A hybrid of firefly and improved particle swarm optimization algorithms for load balancing in cloud environments: Performance evaluation. Computer Networks 162, 106861.

Goyal, V., G., S., 2022. Green cloud computing for sustainable environment: A review. International Journal of Artificial Intelligence, Internet of Things and Cloud Computing 2, 51–57.

Hamdan, M., Salem, A., Shamayleh, Y., 2021. Harmonization between renewable energy and cloud computing towards green computing: A case study of data center at the University of Jordan, in: 2021 12th International Renewable Engineering Conference (IREC). IEEE, pp. 1–5. https://doi.org/10.1109/IREC51415.2021.9427865

Hao, Y., Cao, J., Wang, Q., Ma, T., 2021. Energy-aware offloading based on priority in mobile cloud computing. Sustainable Computing: Informatics and Systems 31, 100563. https://doi.org/10.1016/j.suscom.2021.100563

Hou, H., Ismail, A., 2024. EETS: An energy-efficient task scheduler in cloud computing based on improved DQN algorithm. Journal of King Saud University - Computer and Information Sciences 102177. https://doi.org/10.1016/j.jksuci.2024.102177

Hulkury, M.N., Doomun, M.R., 2012. Integrated green cloud computing architecture, in: 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT). IEEE, pp. 269–274. https://doi.org/10.1109/ACSAT.2012.16

Jena, S.R., Srilakshmi, U., Rafi, S.M., 2025. Intelligent green energy solutions: Optimizing renewable energy for edge cloud computing, in: 2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN). IEEE, pp. 765–769. https://doi.org/10.1109/ICPCSN65854.2025.11035813

Jin, L., Yuan, J., Zhou, D., Sang, X., Chen, S., Yu, X., Lin, G., 2025. Energy-efficient task scheduling with binary random faults in cloud computing environments. Swarm Evol Comput 94, 101877. https://doi.org/10.1016/j.swevo.2025.101877

Joshi, N., Kotecha, K., Choksi, D.B., Pandya, S., 2018. Implementation of novel load balancing technique in cloud computing environment, in: 2018 International Conference on Computer Communication and Informatics (ICCCI). IEEE.

Junaid, M., Sohail, A., Ahmed, A., Baz, A., Khan, I.A., Alhakami, H., 2020a. A hybrid model for load balancing in cloud using file type formatting. IEEE Access.

Junaid, M., Sohail, A., Rais, R.N.B., Ahmed, A., Khalid, O., Khan, I.A., Hussain, S.S., Ejaz, N., 2020b. Modeling an optimized approach for load balancing in cloud. IEEE Access.

Kasse, F.B., Gueye, B., Elbiaze, H., 2014. Green cloud architecture: Low power routers for an energy-aware data transport, in: 2014 IEEE Globecom Workshops (GC Wkshps). IEEE, pp. 31–35. https://doi.org/10.1109/GLOCOMW.2014.7063381

Kaur, A., Kaur, B., 2019. Load balancing optimization based on hybrid heuristic-metaheuristic techniques in cloud environment. Journal of King Saud University – Computer and Information Sciences 31, 398–407.

Kaur, K., Garg, S., Aujla, G.S., Kumar, N., Zomaya, A.Y., 2022. A multi-objective optimization scheme for job scheduling in sustainable cloud data centers. IEEE Transactions on Cloud Computing 10, 172–186. https://doi.org/10.1109/TCC.2019.2950002

Kavitha, J., Srinivasa, V., G Charles Babu, 2023. Energy Efficient Resource Utilization of Cloud Computing Environments for Deployment Models, in: 2023 International Conference on Advancements in Information, Security and Social Networks (ICAISS). https://doi.org/10.1109/icaiss58487.2023.10250648

Khaleel, M.I., 2023. A fault tolerance aware green IoT workflow scheduling algorithm for multi-dimensional resource utilization in sustainable cloud computing. Internet of Things 23, 100909. https://doi.org/10.1016/j.iot.2023.100909

Khalil, M.I.K., Ahmad, I., Shah, S.A.A., Jan, S., Khan, F.Q., 2020. Energy cost minimization for sustainable cloud computing using option pricing. Sustain Cities Soc 63, 102440. https://doi.org/10.1016/j.scs.2020.102440

Khanday, S.K., 2025. Energy-Efficient Clouds: Leveraging Performance Metrics for Sustainable Data Center Operations, in: 2025 International Conference on Intelligent Computing and Control Systems (ICICCS). Erode, India, pp. 513–518. https://doi.org/10.1109/ICICCS65191.2025.10985554

Khosravi, A., B., R., 2018. Short-term prediction model to maximize renewable energy usage in cloud data centers, in: Sustainable Cloud and Energy Services. pp. 203–218.

Khullar, R., Hossain, G., 2022. A New Algorithm for Energy Efficient Task Scheduling Towards Optimal Green Cloud Computing, in: 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA). pp. 1–6. https://doi.org/10.1109/aiccsa56895.2022.10017609

Kotilainen, P., Mäkitalo, N., Systä, K., Mehraj, A., Waseem, M., Mikkonen, T., Murillo, J.M., 2025. Allocating distributed AI/ML applications to cloud–edge continuum based on privacy, regulatory, and ethical constraints. Journal of Systems and Software 222, 112333. https://doi.org/10.1016/j.jss.2025.112333

Kumar, A.B.P., Karanam, S.R., Meka, V.B., Kalapala, S.C., Sidhartha, J., Pasam, H.V.R., 2024. Simulation of Green Cloud Computing Using Virtualization, in: 2024 International Conference on Computational Intelligence for Green and Sustainable Technologies (ICCIGST). Vijayawada, India, pp. 1–5. https://doi.org/10.1109/ICCIGST60741.2024.10717572

Kumar, K.P., 2018. Gravitational emulation-grey wolf optimization technique for load balancing in cloud computing, in: 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE.

Li, C., Guo, Z., He, X., Hu, F., Meng, W., 2023. An AI Model Automatic Training and Deployment Platform Based on Cloud Edge Architecture for DC Energy-Saving, in: 2023 International Conference on Mobile Internet, Cloud Computing and Information Security (MICCIS). Nanjing, China, pp. 22–28. https://doi.org/10.1109/MICCIS58901.2023.00010

Li, X., Garraghan, P., Jiang, X., Wu, Z., Xu, J., 2017. Holistic virtual machine scheduling in cloud datacenters towards minimizing total energy. IEEE Transactions on Parallel and Distributed Systems. https://doi.org/10.1109/TPDS.2017.1234567

Li, X., Jiang, X., Garraghan, P., Wu, Z., 2018. Holistic energy and failure aware workload scheduling in cloud datacenters. Future Generation Computer Systems 78, 887–900. https://doi.org/10.1016/j.future.2018.04.027

Makasarwala, H.A., Hazari, P., 2016. Using genetic algorithm for load balancing in cloud computing, in: 2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE.

Makwe, A., Kanungo, P., Kautish, S., Madhu, G., Almazyad, A.S., Xiong, G., Mohamed, A.W., 2024. Cloud service prioritization using a multi-criteria decision-making technique in a cloud computing environment. Ain Shams Engineering Journal 15, 102785. https://doi.org/10.1016/j.asej.2024.102785

Maldonado-Carrascosa, F.J., García-Galán, S., Valverde-Ibáñez, M., Marciniak, T., Szczerska, M., Ruiz-Reyes, N., 2024. Game theory-based virtual machine migration for energy sustainability in cloud data centers. Appl Energy 372, 123798. https://doi.org/10.1016/j.apenergy.2024.123798

Manam, S., Moessner, K., Asuquo, P., 2023. A Machine Learning Approach to Resource Management in Cloud Computing Environments, in: 2023 IEEE AFRICON. pp. 1–6. https://doi.org/10.1109/africon55910.2023.10293275

Manikandan, G., Swaminathan, S., Kanimozhi, K. V, Seethalakshmi, K., Rao, N. V, 2024. Energy-Efficient Cloud Computing through Blockchain-based Resource Allocation and Green Contracts, in: 2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET). B G Nagara,Mandya, India, pp. 1–5. https://doi.org/10.1109/ICRASET63057.2024.10895080

Manikandan, N., Devayani, S., Divya, M., 2020. Domain-specific allocation & load balancing in cloud computing using virtual machines, in: 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE.

Mardani, A., Jusoh, A., Zavadskas, E.K., Cavallaro, F., Khalifah, Z., 2015. Sustainable and renewable energy: An overview of the application of multiple criteria decision-making techniques and approaches. Sustainability 7, 13947–13984. https://doi.org/10.3390/su71013947

Mehra, S., Ahuja, L., 2023. Study on green cloud computing – A survey, in: 2023 International Conference on Information Technology (ICIT). IEEE, pp. 56–61. https://doi.org/10.1109/ICIT58056.2023.10225896

Mehta, A., Kaur, A., Singh, P., 2018. A heuristic approach for efficient load balancing in cloud using weight based algorithm, in: 2018 4th International Conference on Computing Sciences (ICCS). IEEE.

Meneguette, R.I., Boukerche, A., 2020. An efficient green-aware architecture for virtual machine migration in sustainable vehicular clouds. IEEE Transactions on Sustainable Computing 5, 25–36. https://doi.org/10.1109/TSUSC.2019.2904672

Mentsiev, A., A., T., 2022. Cloud technology in sustainable development of ecology, in: MMTGE 2022 - I International Conference: Methods, Models, Technologies for Sustainable Development: Agroclimatic Projects and Carbon. Kadyrov Chechen State University, Chechen Republic, pp. 288–290.

Mohanty, A., Samantaray, S., Patra, S.S., Ahmad, M.A., Barik, R.K., 2021. An efficient resource management scheme for smart grid using GBO algorithm, in: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE.

Morchid, A., Said, Z., Abdelaziz, A.Y., Siano, P., Qjidaa, H., 2025. Fuzzy logic-based IoT system for optimizing irrigation with cloud computing: Enhancing water sustainability in smart agriculture. Smart Agricultural Technology 11, 100979. https://doi.org/10.1016/j.atech.2025.100979

Navarro Rau, M.F., Calamari, N.C., Navarro, C.S., Enriquez, A., Mosciaro, M.J., Saucedo, G., Barrios, R., Curcio, M., Dieta, V., Martínez, G.G., Elortegui, M. del R.I., Michard, N.J., Paredes, P., Umaña, F., Alday, S., Lopez, A., Miranda, F., Kurtz, D.B., 2025. Advancing wetland mapping in Argentina: A probabilistic approach integrating remote sensing, machine learning, and cloud computing towards sustainable ecosystem monitoring. Watershed Ecology and the Environment 7, 144–158. https://doi.org/10.1016/j.wsee.2025.04.001

Olana, J.A., Tripathy, H.K., 2018. Bio-inspired technique for the Virtual Machine Migration in Green Cloud Computing, in: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). Coimbatore, India, pp. 31–36. https://doi.org/10.1109/ICICCT.2018.8473338

Öner, E., Özer, A.H., 2023. An energy-aware combinatorial auction-based virtual machine scheduling model and heuristics for green cloud computing. Sustainable Computing: Informatics and Systems 39, 100889. https://doi.org/10.1016/j.suscom.2023.100889

Panwar, S.S., Rauthan, M.M.S., Barthwal, V., Mehra, N., Semwal, A., 2024. Machine learning approaches for efficient energy utilization in cloud data centers. Procedia Comput Sci 235, 1782–1792. https://doi.org/10.1016/j.procs.2024.04.169

Patel, K.D., Bhalodia, T.M., 2019. An efficient dynamic load balancing algorithm for virtual machine in cloud computing, in: 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE.

Patra, M.K., Patel, D., Sahoo, B., Turuk, A.K., 2020. A randomized algorithm for load balancing in containerized cloud, in: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE.

Prasad, S., Ramprasad, S.R., Sivaraman, P., Reddy, K. V, Bhanu Prakash Battula, Manikandan, R., 2024. Enhancing Blockchain-Based Access Control in Cloud Computing with Machine Learning Techniques, in: 2024 International Conference on Edge-IoT, Real-Time and Cyber-Physical Systems (ICERCS). pp. 1–5. https://doi.org/10.1109/icercs63125.2024.10895227

Rajagopalan, A., Swaminathan, D., Bajaj, M., Damaj, I., Rathore, R.S., Singh, A.R., Blazek, V., Prokop, L., 2024. Empowering power distribution: Unleashing the synergy of IoT and cloud computing for sustainable and efficient energy systems. Results in Engineering 21, 101949. https://doi.org/10.1016/j.rineng.2024.101949

Ramachandran, M., 2025. Chapter Ten - Energy efficiency sustainability framework for cloud and quantum data centres in the era of 6G, in: Raj, P., Asari, V.K., Ramachandran, M., Sakthivel, U., Susila, N. (Eds.), Advances in Computers. Elsevier, pp. 295–325. https://doi.org/10.1016/bs.adcom.2025.03.013

Rawat, S., Kumar, P., Sagar, S., Singh, I., Garg, K., 2017. An analytical evaluation of challenges in green cloud computing, in: 2017 International Conference on Infocom Technologies and Unmanned Systems (ICTUS): Trends and Future Directions. IEEE, pp. 351–355. https://doi.org/10.1109/ICTUS.2017.8286031

Rohit, V.R., K., S., Nisha, F., S., S., 2023. Innovations and challenges in green cloud computing via LEED, ant colony optimization, in: 2023 3rd International Conference on Smart

Generation Computing, Communication and Networking (SMART GENCON). Bangalore, India, pp. 1–6.

Sadhu, R., Kaur, H., Vishal Pattanaik, 2024. Green Cloud: Navigating to a SustainableFuture Through Green Cloud Computing, in: 2024 International Conference on Emerging Management Practices and Strategies (ICEMPS). https://doi.org/10.1109/icemps60684.2024.10559291

Saleh, H., Nashaat, H., Saber, W., Harb, H.M., 2019. IPSO task scheduling algorithm for large scale data in cloud computing environment. IEEE Access.

Sampaio, A.M., Barbosa, J.G., 2016. Energy-efficient and SLA-based resource management in cloud data centers. Advances in Computers 100, 103–159. https://doi.org/10.1016/j.advcomp.2016.06.002

Sana, M.U., Li, Z., Kiren, T., Liaqat, H.B., Naseem, S., Saeed, A., 2023. A secure method for data storage and transmission in sustainable cloud computing. Computers, Materials & Continua 75, 2741–2757. https://doi.org/10.32604/cmc.2023.036093

Sardaraz, A.M., Tahir, M., 2019. Hybrid algorithm for scheduling scientific workflows in cloud computing. IEEE Access.

Seddiki, D., Javier, F., Sebastián García Galán, Manuel Valverde Ibáñez, Marciniak, T., Nicolás Ruiz Reyes, 2024. Enhanced virtual machine migration for energy sustainability optimization in cloud computing through knowledge acquisition. Computers & Electrical Engineering 119, 109506. https://doi.org/10.1016/j.compeleceng.2024.109506

Shah, S.Q.A., Khan, F.Z., Ahmad, M., 2021. The impact and mitigation of ICMP based economic denial of sustainability attack in cloud computing environment using software defined network. Computer Networks 187, 107825. https://doi.org/10.1016/j.comnet.2021.107825

Shameer, A.P., Subhajini, A.C., 2017. Throughput maximization on efficient load balancing in cloud task scheduling using enhanced bee colony algorithm, in: 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC). IEEE.

Shao, G., Chen, J., 2016. A load balancing strategy based on data correlation in cloud computing, in: 2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC). IEEE.

Sharma, D., Parihar, A.S., 2024. Ant Colony Optimization-Based Energy Efficiency Optimization in IoT-Enabled Cloud Computing Resource Allocation, in: 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET). Indore, India, pp. 1–8. https://doi.org/10.1109/ACROSET62108.2024.10743955

Sharma, R., Bala, A., Singh, A., 2022. Virtual Machine Migration for Green Cloud Computing, in: 2022 International Conference on Disruptive Technologies, Computer and Communication (ICDCECE). https://doi.org/10.1109/ICDCECE53908.2022.9793067

Sharma, Y., Taheri, J., Si, W., Sun, D., Javadi, B., 2021. Dynamic Resource Provisioning for Sustainable Cloud Computing Systems in the Presence of Correlated Failures. IEEE

Transactions on Sustainable Computing 6, 641–654. https://doi.org/10.1109/tsusc.2020.3025180

Shekhar, A., Aleem, A., 2024. Improving Energy Efficiency through Green Cloud Computing in IoT Networks, in: 2024 2nd International Conference on Power, Control and Computing Technologies (IC2PCT). https://doi.org/10.1109/ic2pct60090.2024.10486633

Shen, H., Chen, L., 2020. A resource usage intensity aware load balancing method for virtual machine migration in cloud datacenters. IEEE Transactions on Cloud Computing 8, 17–30.

Shen, L., Li, J., Wu, Y., Tang, Z., Wang, Y., 2019. Optimization of artificial bee colony algorithm based load balancing in smart grid cloud, in: 2019 IEEE Innovative Smart Grid Technologies – Asia (ISGT Asia). IEEE.

Shree, T., Kumar, R., Kumar, N., 2020. Green Computing in Cloud Computing, in: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). Greater Noida, India, pp. 903–905. https://doi.org/10.1109/ICACCCN51052.2020.9362822

Shu, W., Cai, K., Xiong, N.N., 2021. Research on strong agile response task scheduling optimization enhancement with optimal resource usage in green cloud computing. Future Generation Computer Systems 124, 12–20. https://doi.org/10.1016/j.future.2021.05.012

Singh, A., Mishra, N., Ali, S.I., Shukla, N., Shankar, R., 2015. Cloud computing technology: Reducing carbon footprint in beef supply chain. Int J Prod Econ 164, 462–471. https://doi.org/10.1016/j.ijpe.2014.09.019

Singh, A.K., Roy-Nandi, S., Yadav, R., 2023a. Bio-inspired virtual machine placement toward sustainable cloud resource management. IEEE Syst J 17, 3894–3905.

Singh, A.K., Swain, S.R., Saxena, D., Lee, C.-N., 2023b. A bio-inspired virtual machine placement toward sustainable cloud resource management. IEEE Syst J 17, 3894–3905. https://doi.org/10.1109/JSYST.2023.3248118

Singh, S., Chana, I., Buyya, R., 2017. STAR: SLA-aware autonomic management of cloud resources. IEEE Transactions on Cloud Computing. https://doi.org/10.1109/TCC.2017.1234567

Singha, A., J., S., 2022. Green cloud computing to build a sustainable tomorrow, in: 2022 International Conference for Advancement in Technology (ICONAT). Goa, India, pp. 1–6.

Singha, A., Sarkar, S.J., Nayak, S., Patgiri, R., 2022. Green cloud computing – To build a sustainable tomorrow, in: 2022 International Conference for Advancement in Technology (ICONAT). IEEE, pp. 1–6. https://doi.org/10.1109/ICONAT53423.2022.9726052

Skourletopoulos, G., others, 2019. Elasticity debt analytics exploitation for green mobile cloud computing: An equilibrium model. IEEE Transactions on Green Communications and Networking 3, 122–131. https://doi.org/10.1109/TGCN.2018.2890034

Sneha, Tripathi, V., Singh, P., Pandey, P., 2023. Green Cloud Computing: Goals, Techniques, Architectures, and Research Challenges, in: 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT). Gharuan, India, pp. 438–443. https://doi.org/10.1109/InCACCT57535.2023.10141845

Soidridine, M.M., Claude, L., Kora, A.D., 2013. Green cloud architecture for African local collectivities, in: 2013 International Conference on Adaptive Science and Technology (ICASTech). IEEE, pp. 1–7. https://doi.org/10.1109/ICASTech.2013.6707513

Son, A.-Y., S.-N., Y., W., J., 2018. Energy-efficient VM placement scheme based on fuzzy-AHP system for sustainable cloud computing, in: 2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). London, UK, pp. 260–265.

Stergiou, C., E., K., 2018. Security, privacy, and efficiency of sustainable cloud computing for big data and IoT. Sustainable Computing: Informatics and Systems 19, 174–184.

Stergiou, C., Psannis, K.E., Gupta, B.B., Ishibashi, Y., 2018. Security, privacy & efficiency of sustainable Cloud Computing for Big Data & IoT. Sustainable Computing: Informatics and Systems 19, 174–184. https://doi.org/10.1016/j.suscom.2018.06.003

Sudhakar, C., Jain, R., Ramesh, T., 2018. Cloud load balancing – Honey bees inspired effective request balancing strategy, in: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE.

Suratia, N., Thakkar, P., Sheth, K., Ramoliya, D., Patel, A.K., 2023. An extensive analysis of green cloud computing: Overview, associated challenges, and research directions, in: 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, pp. 1147–1152. https://doi.org/10.1109/ICIMIA60377.2023.10426222

Sutariya, K., Menaka, C., Shahid, M., Kashyap, S., Choudhary, D., Padmanabhan, S., 2025. LEVERAGING AI IN CLOUD COMPUTING TO ENHANCE NANO GRID

OPERATIONS AND PERFORMANCE IN AGRICULTURE. Sustainable Computing: Informatics and Systems 101075. https://doi.org/10.1016/j.suscom.2025.101075

T K, S., S, R., Sadasivam, G.S., 2025. A Decentralized, Energy-efficient cloud architecture for sustainable computing, in: 2025 International Conference on Computational Innovations and Engineering Sustainability (ICCIES). Coimbatore, Tamilnadu, India, pp. 1–7. https://doi.org/10.1109/ICCIES63851.2025.11032812

Tang, L., 2018. Load balancing optimization in cloud computing based on task scheduling, in: 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS). IEEE.

Teng, Z., Chen, J., Wang, J., Wu, S., Chen, R., Lin, Y., Shen, L., Jackson, R., Zhou, J., Yang, C., 2023. Panicle-Cloud: An Open and AI-Powered Cloud Computing Platform for Quantifying Rice Panicles from Drone-Collected Imagery to Enable the Classification of Yield Production in Rice. Plant Phenomics 5, 105. https://doi.org/10.34133/plantphenomics.0105

Thakur, S., Chaurasia, A., 2016. Towards green cloud computing: Impact of carbon footprint on environment, in: 2016 6th International Conference – Cloud System and Big Data Engineering (Confluence). IEEE, pp. 209–213. https://doi.org/10.1109/CONFLUENCE.2016.7508115

Thumala, S.R., Madathala, H., Sharma, S., 2025. Towards Sustainable Cloud Computing: Innovations in Energy-Efficient Resource Allocation, in: 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS). Prawet, Thailand, pp. 1528–1533. https://doi.org/10.1109/ICMLAS64557.2025.10968917

Tiwari, G., Jain, R., 2022. Detecting and Classifying Incoming Traffic in a Secure Cloud Computing Environment Using Machine Learning and Deep Learning System, in: 2022 International Conference on SMART CLOUD, Artificial Intelligence, and Information Communication Technologies (SMARTCLOUD). https://doi.org/10.1109/smartcloud55982.2022.00010

Tiwari, P.K., Joshi, S., 2016. Dynamic weighted virtual machine live migration mechanism to manage load balancing in cloud computing, in: 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE.

Tiwari, S., Beena, B.M., 2024. A Neighborhood Inspired Multi-verse Scheduler for Energy and Makespan Optimized Task Scheduling For Green Cloud Computing Systems. IEEE Access 1. https://doi.org/10.1109/access.2024.3484388

Tuli, S., Gill, S.S., Xu, M., Garraghan, P., Bahsoon, R., Dustdar, S., Sakellariou, R., Rana, O., Buyya, R., Casale, G., Jennings, N.R., 2022. HUNTER: AI based holistic resource management for sustainable cloud computing. Journal of Systems and Software 184, 111124. https://doi.org/10.1016/j.jss.2021.111124

Varga, M., Petrescu-Nita, A., Pop, F., 2018. Deadline scheduling algorithm for sustainable computing in Hadoop environment. Comput Secur 76, 354–366. https://doi.org/10.1016/j.cose.2017.12.014

Velde, V., Rama, B., 2017. Simulation of optimized load balancing and user job scheduling using CloudSim, in: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE.

Vishalika, Malhotra, D., 2018. LD_ASG: Load balancing algorithm in cloud computing, in: 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE.

Volkova, V.N., Chemenkaya, L. V, Desyatirikova, E.N., Hajali, M., Khodar, A., Osama, A., 2018. Load balancing in cloud computing, in: 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). IEEE.

Wang, H., Anurag, K.A., Benamer, A.R., Arora, P., Wander, G., Johnson, M.R., Anjana, R.M., Mohan, V., Gill, S.S., Uhlig, S., Buyya, R., 2025. HealthAIoT: AIoT-driven smart healthcare system for sustainable cloud computing environments. Internet of Things 31, 101555. https://doi.org/10.1016/j.iot.2025.101555

Waqar, A., Gultom, M.H., Qureshi, A.H., Tanjung, L.E., Almujibah, H.R., 2023. Complexities to the deployment of cloud computing for sustainability of small construction projects: Evidence from Pakistan. Ain Shams Engineering Journal 14, 102559. https://doi.org/10.1016/j.asej.2023.102559

William, P., Singh, R., Al-Samalek, A., Hussain, A., Arti Badhoutiya, Rao, A.K., 2024. Dynamic Energy-Aware Cloudlet-Based Mobile Cloud Computing Model for Green Computing, in: 2024 International Conference on Inventive Computation and Information Technologies (ICIPTM). pp. 1–6. https://doi.org/10.1109/iciptm59628.2024.10563792

Xiang, Q., 2022. Development and Verification of Multi-class Data in Artificial Intelligence Model Group Based on Cloud Computing. 2022 International Conference on Artificial Intelligence and Blockchain (AIBC) 22, 42–46. https://doi.org/10.1109/icaibd55127.2022.9820486

Xu, M., Buyya, R., 2020. Managing renewable energy and carbon footprint in multi-cloud computing environments. J Parallel Distrib Comput 135, 191–202. https://doi.org/10.1016/j.jpdc.2019.09.015

Xu, M., Toosi, A.N., Buyya, R., 2021. A Self-Adaptive Approach for Managing Applications and Harnessing Renewable Energy for Sustainable Cloud Computing. IEEE Transactions on Sustainable Computing 6, 544–558. https://doi.org/10.1109/tsusc.2021.3014943

Xue, C., 2018. Method and implementation of server load balancing in cloud computing, in: 2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE). IEEE.

Yang, Z., 2024. Renewable energy management in smart grid with cloud security analysis using multi-agent machine learning model. Computers and Electrical Engineering 116, 109177. https://doi.org/10.1016/j.compeleceng.2024.109177

Zong, Z., 2020. An Improvement of Task Scheduling Algorithms for Green Cloud Computing, in: 2020 International Conference on Computer Science and Engineering (ICCSE). pp. 654–657. https://doi.org/10.1109/iccse49874.2020.9201785