Optimizing Energy Efficient Cloud Architectures for Edge Computing: A Comprehensive Review

TA Gamage*, Indika Perera
Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka

***Abstract*—Now-a-days, edge computing and cloud computing are considered for collaborating together to produce computing solutions that are more effective, scalable and adaptable. The proliferation of cloud infrastructures has drastically increased energy consumption leading to the need for more research in optimizing energy efficiency for sustainable and efficient systems with reduced operational costs. In addition, the edge computing paradigm has gained wide attention during the last few decades due to the rise of the Internet of Things (IoT) devices, the emergence of applications that require low latency, and the widespread demand for environmentally friendly computing. Moreover, lowering cloud-edge systems' energy footprints is essential for fostering sustainability in light of growing concerns about environmental effects. This research presents a comprehensive review of strategies aimed at optimizing energy efficiency in cloud architectures designed for edge computing environments. Various strategies, including workload optimization, resource allocation, virtualization technologies, and adaptive scaling methods, have been identified as techniques that are widely utilized by contemporary research in reducing energy consumption while maintaining high performance. Furthermore, the paper investigates how advancements in machine learning and AI can be leveraged to dynamically manage resource distribution and energy-efficient enhancements in cloud-edge systems. In addition, challenges to the approaches for energy optimization have been discussed in detail to further provide insights for future research. The conducted comprehensive review provides valuable insights for future research in the edge computing paradigm, particularly emphasizing the critical importance of enhancing energy efficiency in these systems.**

***Keywords*—Cloud computing; edge computing; energy efficiency; sustainability**

I. INTRODUCTION

Cloud computing is a concept for providing computer resources such as servers, storage, databases, networking, software, and analytics over the internet. Users can obtain physical infrastructure and data centers on-demand from cloud service providers, negating the need to own and manage them and providing more flexibility, scalability, and cost-efficiency. Without the upfront expenses and hassles of maintaining traditional IT systems, cloud computing allows businesses and individuals to scale their IT needs as needed, covering everything from data storage to sophisticated application development [1] [2] [3].

The adaptability of cloud technology is one of its most intriguing features. Businesses can simply reallocate their computer resources to meet changing demands, allowing for ongoing scaling in response to changing business requirements. Global accessibility is yet another fundamental benefit of cloud computing. Moreover, cloud services reduce geographical barriers, enabling real-time collaboration across several regions and supporting the delivery of services to a global clientele with less idle time [2].

Edge computing is a distributed computing paradigm that moves data storage and computation closer to the point of demand, which is generally at the network's edge, close to the data generating source. Edge computing reduces latency, bandwidth utilization, and reaction times by processing data locally on IoT sensors, gateways, or edge servers rather than depending on a centralized cloud [4], [5]. This method is more efficient for time-sensitive applications where rapid decisions are essential, like real-time analytics, industrial automation, and driverless cars. Edge computing improves performance, security, and dependability in a variety of applications with the use of decentralized computing [6], [7].

Cloud computing has revolutionized the way businesses and organizations operate, offering scalable, on-demand computing resources. Nevertheless, the growth of cloud infrastructures has increased their energy consumption that has been a major concern. Therefore, the demand for energy-efficient cloud architectures has become a critical area of research due to environmental concerns, operational costs, and the need for sustainability. In addition, the proliferation of edge computing where data is processed closer to the data source to reduce latency and bandwidth usage adds another dimension to cloud energy optimization [8], [9].

The rapid growth of energy consumption by edge cloud computing infrastructures has been a significant focus of contemporary research considering the operational costs and environmental concerns aiming at the sustainability of cloud computing architectures. Among the approaches for a sustainable edge computing paradigm, dynamic resource allocation, AI driven energy optimization, energy-aware scheduling and load balancing, green data centres, virtualizations and containerizations are acquiring significant focus by the researchers [10], [11], [12], [13].

Nevertheless, the exponential growth of cloud computing infrastructures has resulted in inflated demand for further research on optimizing energy efficient cloud architectures in order to move towards a sustainable edge computing paradigm. In addition, maintaining performance and scalability while reducing energy consumption has also been a controversial topic that is further researched [14] [15]. Therefore, further research that enhance the green cloud architectures for edge

The open-access publication fee of this paper is supported by the SRC publication fee scheme of the University of Moratuwa

computing are required for the improved sustainability of the cloud computing paradigm.

The rest of the paper is laid out as follows. Section II discusses approaches that have been widely concerned to achieve sustainability in the edge computing paradigm focusing on energy efficiency while Section III broadly explores recent related work that makes use of different strategies for energy efficiency in the edge computing paradigm. Section IV provides a discussion of the conducted review work along with challenges and further insights into the energy-efficient approaches in the edge computing domain. Finally, Section V concludes the research findings along with future directions for the conducted research work.

II. Strategies for Energy Optimization

The following strategies were identified as the main approaches that are widely focused in research that aim at energy optimizations in edge computing. Nevertheless, the techniques that are utilized for edge computing and even in the cloud computing paradigm in contemporary research are a combination of the below-discussed approaches.

A. Energy Aware Scheduling and Load balancing

Energy-aware scheduling and Load Balancing in edge computing aims to optimize resource allocation while reducing energy consumption. In the context of an edge environment, resources are distributed across a number of small, geographically separated devices, requiring efficient scheduling algorithms to manage tasks without overloading nodes or consuming excessive amounts of energy [16], [17]. Energy-aware scheduling approaches prioritize tasks based on their energy demands and resource requirements, ensuring that high-priority tasks are executed on energy-efficient nodes while low-priority tasks are deferred across less critical resources. By optimizing task scheduling, systems can minimize energy wastage, extend device lifespans, and ensure seamless service delivery [18] [19].

Load balancing complements energy-aware scheduling by distributing computational workloads evenly across available nodes to prevent resource overuse and reduce energy consumption. In edge computing, load balancing techniques must consider not only the computational capacity of each node but also its current energy state. Dynamic load balancing ensures that tasks are allocated to nodes with sufficient energy reserves, minimizing the risk of node failure due to energy depletion. These approaches improve the efficiency and sustainability of edge computing setups, enabling them to support more applications and services without overwhelming the energy resources of distributed nodes [20] [21].

B. AI-Driven Energy Optimization

AI-driven energy optimization approaches that are utilized in Edge computing incorporate artificial intelligence strategies to improve the energy efficiency of edge networks and devices, which are frequently decentralized and resource-constrained. AI algorithms, in particular deep learning and machine learning, are able to track and forecast edge device energy consumption and task trends [22]. AI systems are able to make decisions about how best to assign jobs to nodes, modify power settings, and dynamically manage resource utilization based on needs by evaluating real-time data. With this real-time optimization, energy conservation is guaranteed without compromising the necessary performance and service levels.

AI can also provide intelligent load balancing and scheduling, which will optimize the energy consumption of edge computing systems. AI models, for instance, can anticipate high load periods and move workloads to nodes that use less energy or have unused resources. AI algorithms can also allow edge devices to transition into low-power modes while they are inactive or when processing demands drop, which will cut down unnecessary energy consumption. AI-driven energy optimization may dramatically increase the lifespan of edge devices, slash operating costs, and lessen the environmental effect of edge computing infrastructures by continually learning from and reacting to the system [23] [24].

C. Virtualization and Containerization

Virtualization and Containerization in edge computing play a significant role in improving resource usage, flexibility, and scalability of edge networks. Several Virtual Machines (VMs) can operate on a single physical server owing to virtualization, which makes it possible to abstract hardware resources. This makes it possible to consolidate edge resources and execute various applications or services in separate environments, making the best use of the hardware that is available. Virtualization assists in the management of various workloads and offers the flexibility to dynamically scale resources in response to variations in demand in edge computing [25], [26]. In addition, it makes it possible to handle software upgrades and system maintenance effectively without interfering with ongoing services that in turn increase system reliability.

Nevertheless, containerization, which bundles apps and their dependencies into containers, is a less complex substitute for virtualization. Since containers share the host operating system kernel, containers are more efficient than VMs in terms of startup times and overhead. Moreover, containerization facilitates the quick deployment and scalability of applications among dispersed nodes in edge computing. Due to its great degree of mobility, moving apps between various edge devices or contexts is much simpler. This is especially critical in edge environments, which are dynamic and heterogeneous and have a wide range of device capabilities [27], [28], [29]. Therefore, it can be stated that containerization is a perfect solution for the changing needs of edge computing since it allows for quicker application development cycles, better system responsiveness, and more effective resource management.

D. Network Optimization

Network optimizations in edge computing primarily aims to reduce latency, enhance bandwidth efficiency, and improve overall performance by bringing data processing and storage closer to end devices, such as IoT sensors or mobile users. By decentralizing computing power, edge computing reduces the amount of data that needs to be sent to central cloud data centers, minimizing delays in data transmission [30], [31]. Techniques like adaptive routing allow the network to choose the optimal path for data, considering real-time conditions such as congestion, which improves response times [32].

Moreover, Dynamic Voltage and Frequency Scaling (DVFS) is another approach in network optimization for energy efficiency that reduces energy consumption when full performance is not needed by adjusting the power consumption of networking devices depending on real-time demands [33], [34]. Network Function Virtualization (NFV) is also widely researched as an efficient approach for network optimization for energy efficiency in edge computing as it enables the operation of several network services on a single physical server, which eliminates the need for specialized hardware [35]. Adaptive transmission power control [36], [37] and the application of energy-efficient communication protocols [38], such as the enhanced energy efficiency features of 5G [39] are other approaches that wireless networks can reduce overall power consumption. Therefore, these approaches can significantly lower the energy footprint of networks without sacrificing the performance by following the network optimization techniques discussed above.

E. Green Data Centers

Edge data centers are typically smaller and distributed, which makes traditional data center energy efficiency techniques less applicable. By utilizing sustainable methods like the usage of renewable energy sources, cutting-edge cooling systems, and energy-efficient hardware, green data center features aim to reduce their environmental impact for sustainable and green data centers [40], [41]. The transition towards greener operations decreases carbon footprints and operational expenses, making these data centers a crucial element of sustainable IT infrastructure.

In contemporary research, other than the shift towards renewable energy sources in data centers that assists in reducing dependency on the grid, automation techniques, low power consuming electric equipment, and devices with extended thermal limits have also been incorporated in order to improve sustainability and to move towards green data centers [42]. In addition, thermal energy harvesting, kinetic energy and hybrid systems have also been noted in data centers to improve the resilience, and contribute to more sustainable and energy efficient infrastructures.

III. Recent Work

The authors in study [43] have conducted a comprehensive analysis of energy consumption across various cloud-related architectures, including cloud, fog, and edge computing. It introduces a taxonomy that categorizes these architectures based on their characteristics, such as the number and role of data centers and their connectivity. The authors propose a generic energy model that accurately estimates and compares the energy consumption of these infrastructures, taking into account factors like cooling systems and network devices. The findings of this research work indicate that fully distributed architectures can consume significantly less energy between 14% and 25% compared to centralized and partly distributed architectures, highlighting the importance of energy efficiency in the design and deployment of modern computing solutions. In summary, the study aims to provide a foundational framework for future research in energy consumption analysis within the evolving landscape of cloud computing technologies. However, while the proposed model effectively highlights architectural differences and provides insights into energy efficiency, it may benefit from more consideration of real-world variables, such as the uneven distribution of end users and the impact of varying application workloads on energy consumption. In addition, the reliance on existing simulators, which have their limitations, could affect the accuracy of the results.

Another group of researchers in study [13] have presented a novel framework that employs Deep Reinforcement Learning (DRL) to optimize workflow scheduling in edge-cloud computing environments, specifically targeting the challenges introduced by the proliferation of IoT devices. Traditional cloud architectures often struggle with the demands of IoT applications due to issues like high latency and limited bandwidth. This study aims to address these challenges by balancing the conflicting objectives of minimizing energy consumption and execution time while ensuring that workflow deadlines are met. The proposed DRL technique demonstrates significant improvements over baseline algorithms, achieving 56% better energy efficiency and 46% faster execution times. Key innovations include a hierarchical action space that distinguishes between edge and cloud nodes, as well as a multi-actor framework that enhances the learning process by allowing separate networks to manage task allocation. The results indicate that this approach is particularly effective for latency-sensitive applications, such as video surveillance, where efficient resource management is critical. Overall, the research highlights the potential of DRL in optimizing resource allocation and scheduling in edge-cloud environments, providing valuable insights for future advancements in this rapidly evolving field.

The researchers of study [11] address the crucial issue of resource allocation in cloud computing, particularly in the context of increasing energy consumption and performance demands on data centers. The authors propose a hybrid model that combines Genetic Algorithms (GA) and Random Forest (RF) techniques to optimize the allocation of VMs to physical machines (PMs). The GA is employed to generate an optimized training dataset that maps VMs to PMs, which is then utilized by the RF for classification and allocation tasks. This approach aims to minimize power consumption while maximizing resource utilization and maintaining load balance across the data center. The effectiveness of the proposed model is evaluated using real-time workload traces from PlanetLab, showing significant improvements in energy efficiency and execution time compared to traditional methods. The study contributes to the existing body of knowledge by demonstrating the potential of hybrid optimization techniques in enhancing cloud infrastructure management. However, the authors acknowledge the need for further research to assess the model's adaptability to diverse workloads and its scalability in heterogeneous cloud environments.

The research in [10] introduces an Energy-Efficient Task Offloading Strategy (ETOS) aimed at enhancing energy efficiency in Mobile Edge Computing (MEC) environments for resource-intensive mobile applications. The study formulates the task offloading problem as a non-linear optimization challenge, proposing a hybrid approach that combines Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO) to effectively allocate resources while considering capacity and latency constraints. The proposed ETOS leverages the collaborative capabilities of MEC servers to minimize energy consumption during task execution. Extensive simulations demonstrate that ETOS outperforms existing baseline methods in terms of energy utilization, response delay, and offloading utility, particularly under limited resource conditions. Despite its promising results, the research highlights the need for real-world validation and addresses the potential complexity of implementing the hybrid optimization approach in practical scenarios, suggesting directions for future work to enhance applicability and effectiveness in real-world MEC systems.

The research in study [12] focuses on developing a novel multi-classifier algorithm aimed at optimizing energy-efficient task offloading in Fog Computing environments for IoT applications. As the number of connected devices increases, efficient resource management becomes crucial to minimize energy consumption and enhance service quality. The proposed algorithm evaluates various attributes related to tasks, network conditions, and processing capabilities of Fog nodes to determine the most suitable node for task execution. By leveraging machine learning techniques, the algorithm aims to improve decision-making processes regarding task offloading, thereby reducing execution time and energy usage. The study emphasizes the importance of balancing energy efficiency with performance metrics, demonstrating that the multi-classifier approach can significantly enhance Quality of Service (QoS) parameters. In summary, this research contributes to the ongoing efforts to optimize Fog Computing frameworks, making them more effective in handling the computational demands of IoT applications while addressing energy constraints.

The authors of study [44] propose a novel framework to optimize energy consumption and computational efficiency in IoT environments. It introduces a three-layer architecture comprising sensor, edge, and cloud layers, facilitating effective task offloading and resource management. The study employs Long Short Term Memory (LSTM) networks for accurate workload prediction, enabling the system to adapt to dynamic conditions. Additionally, it utilizes Lyapunov optimization methods to address the non-convex nature of resource allocation problems. Simulation results demonstrate significant improvements in energy efficiency and processing rates. However, the research acknowledges limitations, including concerns about scalability, the assumptions made regarding user behavior, and a lack of focus on security aspects. Therefore, the paper contributes valuable insights into mobile edge computing, highlighting the potential for enhanced performance in IoT networks while suggesting areas for further exploration and refinement.

The research of [22] presents the Edge Intelligent Energy Consumption Model (ECMS) aimed at optimizing energy usage in MEC environments. As energy consumption in edge data centers becomes increasingly critical, the ECMS model provides a framework for predicting and managing energy needs based on varying workloads, including CPU-intensive, Web transactional, and I/O-intensive tasks. The authors validate the model through experimental results, demonstrating its superior performance in accuracy and training time compared to existing models like TW BP PM and FSDL. The study categorizes energy consumption modeling into two main approaches: system resource utilization and Performance Monitor Counter (PMC)-based modeling. The ECMS model leverages a simpler network topology and lower input dimensions, resulting in reduced training time and CPU workload during execution. The findings indicate a strong correlation between energy consumption and CPU utilization, emphasizing the need for precise energy models to inform optimization algorithms. The research concludes with future directions, suggesting the extension of the ECMS model to mixed workloads and integration with advanced AI/ML techniques, such as reinforcement learning, to further enhance energy efficiency in edge computing environments, ultimately contributing to sustainable practices in the industry.

A technical analysis on service placement approaches in the context of edge computing has been focused by the authors [45] with the aim of addressing energy efficiency in edge computing paradigm in IoT systems. The main objective of this research work has been to identify the effective and efficient strategies for service placement in IoT environments along with a taxonomy to categorize the studies in this field and in terms of cloud edge service placement approaches and algorithms that have been utilized. In addition to the technical analysis, a statistical analysis has also been provided by the authors along with evaluation factors to which the research findings provide further insights on future research in this paradigm. The results suggest that the server placement approaches fall under three types of categories, namely, decentralized, centralized and hierarchical while Genetic Algorithm has been widely utilized by researchers compared to other machine learning algorithms such as Greedy, Markov, BSAP, Topsis and Polynomial algorithms. Furthermore, time, cost, and latency evaluation metrics have been identified as the most concerned evaluation metrics in the context of service placement. Finally, the authors provide insights on to open issues and challenges in the context of service placement that is vital for future research focusing on service placement.

The research in study [19] presents a heterogeneous cluster-based wireless sensor network (WSN) model aimed at optimizing task allocation to minimize energy consumption and balance load. The network consists of clusters, each with a cluster center and several sensor nodes, where the cluster center collects data from the nodes and communicates with a central processor. The study establishes a task model where complex tasks are divided into independent subtasks, each requiring specific resources, data sizes, and computation times. To address the task allocation problem, the authors propose a Fusion Algorithm (FA) that integrates Genetic Algorithm (GA) and Ant Colony Optimization (ACO) techniques. This algorithm features a novel mutation operator and a new population initialization method, enhancing its effectiveness in reducing energy consumption and balancing the load across the network. Experimental results demonstrate that the FA outperforms traditional GA, achieving 8.1% lower energy consumption and significantly reducing the load on both sensor nodes and cluster centers. The proposed approach ensures all sensors remain operational throughout task execution, thereby increasing the reliability and longevity of the WSN. Therefore, the research contributes valuable insights into efficient task allocation strategies in edge computing environments.

The researchers of study [46] investigate an RIS-assisted Non-Orthogonal Multiple Access (NOMA) Mobile Edge Computing (MEC) network, focusing on minimizing energy consumption for users. The authors propose a joint optimization approach that includes RIS phase shifts, data transmission rates, power control, and transmission time. Due to the non-convex nature of the problem, the authors decompose it into two sub-problems: firstly, utilizing a dual method for a closed-form solution with a fixed RIS phase vector, and the other employing a penalty method for suboptimal power control solutions. The optimization process alternates between these sub-problems until convergence is achieved. To demonstrate the effectiveness of their proposed NOMA-MEC scheme, the authors compare it against three benchmark schemes: TDMA-MEC partial offloading, full local computing, and full offloading. These researchers have also introduced an alternating 1-D search method for optimizing RIS phase shifts in the TDMA-MEC scheme. Numerical results indicate that the proposed scheme significantly reduces overall energy consumption and highlights the impact of user distance on performance. The paper concludes by acknowledging the potential for future work on robust transmission design to address channel state information estimation errors.

Another group of researchers in [47] have worked on a deep reinforcement learning (DRL) approach for delay-aware and energy-efficient computation offloading in dynamic MEC networks with multiple users and servers. The primary objective is to maximize the number of tasks completed before their deadlines while minimizing energy consumption. The proposed DRL model operates in an end-to-end manner, eliminating the need for post-action optimization functions, and can handle a large action space without relying on traditional optimization methods. The study formulates the offloading problem as a Markov Decision Process (MDP), capturing the complexity of the MEC system by incorporating time-varying channel conditions and various task profiles. Extensive simulations demonstrate that the proposed DRL model significantly outperforms existing DRL models and greedy algorithms in terms of task completion and energy efficiency. The results indicate that the DRL model learns optimal policies over time, effectively managing the trade-off between exploration and exploitation during training. The conducted research highlights the potential of DRL in enhancing the performance of MEC systems, making it a valuable contribution to the field of IoT and edge computing.

The research in [48] presents a novel approach to enhance task offloading in dynamic vehicular environments. It addresses the limitations of existing centralized and decentralized Deep Reinforcement Learning (DRL) algorithms, which often struggle with computational constraints and coordination issues. The proposed framework introduces a multi-layer Vehicular Edge Computing (VEC) architecture that optimizes task management across vehicles, edge servers, and cloud resources. Key contributions include the development of an energy-efficient VEC framework that considers the diverse computing capabilities of network entities and introduces a utility function to enhance energy efficiency. Additionally, a decentralized Multi-Agent Deep Reinforcement Learning (MADRL) algorithm is proposed, which effectively adapts to changing conditions while minimizing latency and maximizing task completion rates. The research also provides a comprehensive performance evaluation using simulations in a realistic environment, demonstrating the effectiveness of the proposed solution in managing varying traffic densities. The conducted research highlights the potential of decentralized approaches in improving the efficiency and responsiveness of vehicular networks, paving the way for advancements in autonomous vehicle applications and real-time data processing.

The research in [24] presents a novel cloud-edge cooperative content-delivery strategy aimed at minimizing network latency in asymmetrical Internet of Vehicles (IoV) environments. By leveraging Deep Reinforcement Learning (DRL), the authors propose a Deep Q Network (DQN) policy that optimizes content caching and request routing based on perceptive request history and current network states. The study formulates the joint allocation of heterogeneous resources as a queuing theory-based delay minimization objective, addressing the challenges of computation complexity and dynamic network conditions. Extensive simulations demonstrate that the proposed strategy significantly reduces network latency compared to existing solutions, showcasing its adaptability to varying user requirements and network states. The findings indicate that the DQN model achieves fast convergence and improved performance across different scenarios. The paper concludes with a discussion on future work, including the exploration of end-user mobility and deeper collaboration among mobile users, edge, and cloud networks to enhance overall Quality of Experience (QoE) while balancing network delay and energy consumption. This research contributes valuable insights into optimizing resource allocation in IoT-edge-cloud systems, particularly in the context of intelligent transportation systems.

Table I illustrates a summary of the recent work that were discussed in detail under objective, approach, key findings and remarks of the particular research work.

TABLE I. SUMMARY OF THE LITERATURE REVIEW

Reference	Objective	Approach	Findings	Remarks
[43]	To evaluate and compare energy consumption of Cloud, Fog, and Edge computing infrastructures	A taxonomy of different cloud architectures and a generic energy model	Fully distributed architectures consume 14%-25% less energy than centralized ones	Model may not account for real-world variables and simulator constraints
[13]	Energy-efficient resource scheduling in edge cloud environment	Deep Reinforcement Learning	• Energy consumption (56% of improvement) • Execution time (46% of improvement)	The proposed reinforcement learning framework is designed to operate in a centralized manner
[11]	Optimize virtual machine allocation in cloud infrastructure, aiming to minimize power consumption while maintaining load balance and maximizing resource utilization	Genetic Algorithm (GA) and Random Forest (RF) techniques	• Execution Time (37% of reduction) • Resource Utilization (11% Improvement)	Limited generalizability due to specific workload traces, complexity in real-world implementation, and potential oversight of other critical factors
[10]	Propose a task offloading scheme that minimizes the overall energy consumption along with satisfying capacity and delay requirements	Hybrid approach established based on Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO)	The proposed strategy considerably outperforms other baseline approaches, such as OEOS, ROA-DPH, ATO, and Local execution in terms of energy consumption, execution time, and offloading utility	Lacks real-world validation and may face implementation complexity in resource-constrained environments
[12]	Propose a novel energy-efficient task offloading method for IoT, Fog, and Cloud computing paradigms	Multi classifier-based approach	• Energy Consumption (11.36% of reduction) for Cloud-only • Energy Consumption (9.30% of reduction) for edge-ward • Network usage (67% of reduction) for Cloud-only • Network usage (96% of reduction) for edge-ward	Lacks extensive empirical validation and does not address decentralized approaches for dynamic environments
[44]	Propose a hierarchical communication and computation framework for jointly optimizing energy consumption and computation rate is proposed	The Long Short Term Memory (LSTM) network	The proposed method can greatly improve system performance by saving energy costs and achieving a high processing rate	Scalability concerns, assumptions about user behavior, limited security focus, and complexity in practical implementation
[22]	Predicting energy consumption and monitor edge servers	Intelligent energy modeling approach that combines Elman Neural Network (ENN)	The proposed ECMS outperforms the baseline power models (FSDL, CMP, TW_BP_PM, AEC, CUBIC, Power regression)	The research primarily focuses on specific workloads, limiting its applicability to broader, mixed workload scenarios in MEC environments
[45]	Identify studies related to service placement strategies to categorize the relevant studies as a knowledge source for further research	Perform a technical analysis on the cloud edge service placement approaches	• Methods and algorithms of the existing service placement approaches • Evaluation metrics for service placement approaches • Tools and environments developed for service placement approaches	Lacks any further implementation towards the service placement paradigm
[19]	Develop an efficient task allocation strategy for heterogeneous wireless sensor networks that minimizes energy consumption and balances load to extend network lifetime and enhance reliability	Fusion Algorithm combining Genetic Algorithm and Ant Colony Optimization for effective task allocation in WSNs	• Load on sensors was reduced by 58% with the FA, while the load on cluster centers decreased by 30.8%. • FA achieved an 8.1% reduction in energy consumption compared to the traditional GA.	The research may not address complex task dependencies and real-world scenarios requiring dynamic resource allocation and execution order
[46]	Minimize energy consumption in RIS-assisted NOMA-MEC networks	Jointly optimize RIS phase shifts, transmission rates, and power control	Significant energy savings compared to benchmark schemes.	Non-convex optimization and CSI estimation errors affect performance
[47]	Maximize task completion before deadlines while minimizing energy consumption in MEC systems	Deep Reinforcement Learning model	Proposed model outperforms existing methods in task completion and energy efficiency through extensive simulations	Model's performance may affect the complexity of real-world environments
[48]	Optimize task offloading and resource management in dynamic vehicular environments using a decentralized framework	A multi-layer edge computing architecture and a decentralized multi-agent deep reinforcement learning algorithm	Significant reduction in energy consumption and improved task completion rates when compared to existing algorithms in simulations	Scalability issues in highly dynamic vehicular networks
[24]	Minimize network latency in asymmetrical IoV environments through optimized resource allocation	Deep Reinforcement Learning	• Network Delay (44% reduction) • Reward per episode (39% improvement)	Future work needed on end-user mobility and deeper collaboration among network components and the tradeoff between network delay

				and energy consumption have not been focused to compromise among multiple network indicators

IV. Discussion

The conducted literature review provides insights into future research that aims at energy efficiency in the paradigms of both cloud and edge computing along with the approaches that are utilized concerning the reduction of energy consumption. Nevertheless, it also noted that the researchers utilize a combination of techniques in their work towards optimizing energy efficiency for sustainable and efficient systems with reduced operational costs. Moreover, it is noted that the techniques widely spread towards the AI-driven approaches with the advancements in machine learning.

Furthermore, it is also vital to identify the challenges that are encountered when focusing on the energy-efficient aspects in edge computing due to the proliferation of IoT devices and advancements in modern computing. Among the challenges to optimizing energy efficiency in the edge computing paradigm, distributed dynamic workloads among edge nodes with vast range of heterogeneity, distributed, and resource constrained nature is much prominent since accurate modeling would be complicated due to the fluctuating workloads based on user demands. In addition, different and unpredictable workloads have imposed a lot of issues towards the edge computing paradigm. Nevertheless, the task scheduling techniques employed by contemporary research are more towards the independent task-oriented workloads while few studies have focused on complex workloads [13].

Moreover, delayed critical applications that run on devices in the mobile edge computing paradigm also impose challenges to this arena [10] that directly leads to the insufficient quality of experience for the end users and high cost of energy and bandwidth utilization that are unfavourable. Moreover, limited power sources, processing and storage capabilities also impose challenges to energy efficiency optimization strategies for a sustainable edge computing paradigm [49], [50]. Resource management is key to optimizing energy efficiency and has also been a challenging task owing to the highly dynamic nature of IoT traffic. Furthermore, many tasks have dependencies that dictate the order of execution. Managing these dependencies while optimizing resource allocation adds another level of complexity to the task allocation process.

Ensuring that the network meets specific QoS requirements, such as latency and reliability, while performing task allocation is a critical challenge that must be addressed. In addition, striking a balance between energy efficiency and QoS awareness has also been a challenging aspect where the QoS is highly impacted by the majority of mechanisms that are utilized in distributed computing environments [51]. Furthermore, energy efficiency with application performance and user experience is crucial, as overly aggressive energy-saving measures may negatively affect user satisfaction. Addressing these challenges is essential for advancing research and developing effective energy management strategies in MEC environments.

Additionally, the need for extensive data collection for relevant energy-related parameters can introduce overhead and impact performance, while selecting the most pertinent features for modeling remains a challenge. Real-time processing requirements further complicate the development of accurate models, and integrating advanced AI/ML techniques introduces complexities in training and deployment.

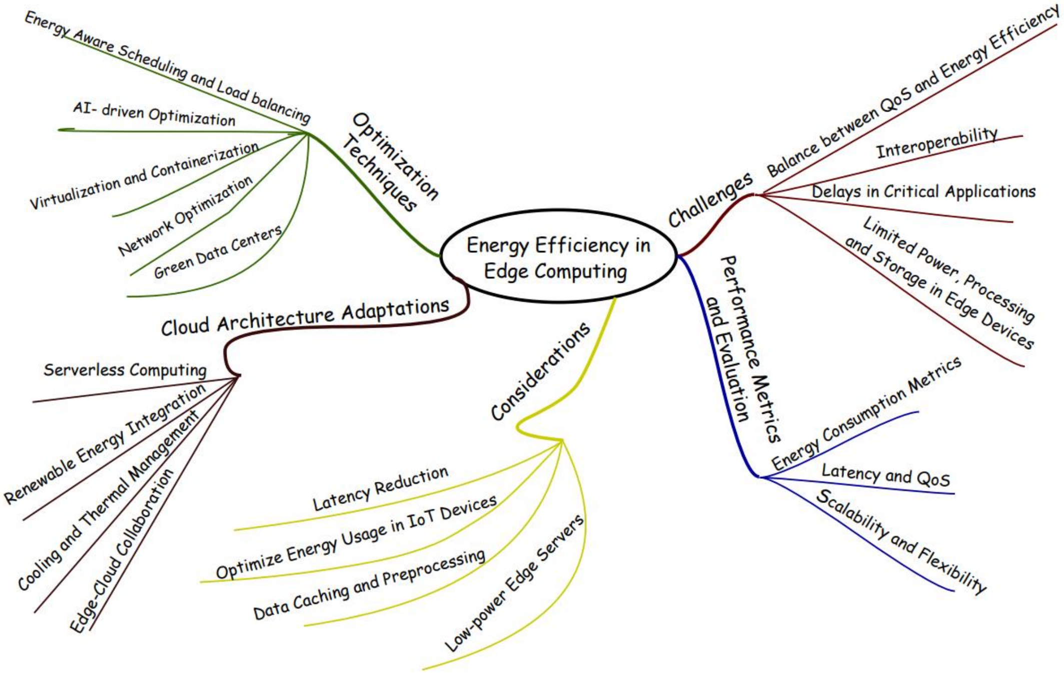
Energy Efficiency in Edge Computing
Optimization Techniques
Energy Aware Scheduling and Load balancing
AI- driven Optimization
Virtualization and Containerization
Network Optimization
Green Data Centers
Cloud Architecture Adaptations
Serverless Computing
Renewable Energy Integration
Cooling and Thermal Management
Edge-Cloud Collaboration
Considerations
Latency Reduction
Optimize Energy Usage in IoT Devices
Data Caching and Preprocessing
Low-power Edge Servers
Challenges
Balance between QoS and Energy Efficiency
Interoperability
Delays in Critical Applications
Limited Power, Processing and Storage in Edge Devices
Performance Metrics and Evaluation
Energy Consumption Metrics
Latency and QoS
Scalability and Flexibility

Fig. 1. Mind map with key findings from the conducted review.

With the scalability concerns in both edge and cloud computing, as the number of nodes and tasks increase, maintaining efficient communication and coordination among nodes becomes more difficult. Algorithms must be scalable to handle larger networks without significant performance degradation. Moreover, interoperability also could be a major challenge where a diverse range of IoT devices and platforms can lead to compatibility issues, making it difficult to implement a unified mechanism. Fig. 1 provides a mind map of the key findings from the conducted literature review comprising of energy optimization techniques, cloud architecture adaptations, major considerations and challenges to the energy efficiency aspect of the edge computing paradigm that provides further insights into research in this domain.

Therefore, when the above aspects are concerned, it is apparent that there is a lack of unified metrics and benchmarks for assessing energy efficiency across different edge computing environments, which makes it difficult to compare solutions. In addition, the use of AI/ML models for decision-making in edge computing often introduces significant energy overhead that urges need of optimized lightweight AI/ML algorithms for edge environments without compromising accuracy or efficiency. Heterogeneity of Edge devices also necessitates the scalable, device-agnostic optimization models that can adapt to heterogeneous edge environments. Moreover, real-time applications such as autonomous vehicles, healthcare monitoring systems, etc., have strict latency and reliability requirements, which complicate energy optimization efforts. Therefore, developing solutions that balance energy efficiency with real-time performance guarantees has also been a critical consideration in this research paradigm.

V. Conclusion

The research findings are indicative of the current approaches to energy efficient edge computing systems with reduced energy consumption and costs to provide further insights into future research in this arena. In addition, the different strategies that were identified as the key techniques to energy efficient systems in cloud computing further assists future research while a combination of different strategies has also been noted in contemporary research. Furthermore, AI-driven energy optimization techniques have been widely focused and researched and this approach has been able to provide better mechanisms for optimizing energy efficiency in both edge and cloud computing paradigms.

In conclusion, this comprehensive review on optimizing energy efficiency for edge computing highlights the growing importance of balancing performance and sustainability in modern cloud systems since as edge computing gains prominence in reducing latency and improving data processing efficiency, the need for energy optimization becomes critical. Moreover, various approaches, including workload distribution, resource allocation, and hardware improvements, have demonstrated the potential to significantly reduce energy consumption while balancing the quality of service. As future research, the authors are to refine these strategies and explore innovative methods to further enhance energy efficiency, ensuring sustainable cloud-edge ecosystems that meet the rising demands of modern applications.

References

[1] A. Sunyaev and A. Sunyaev, ‘Cloud computing’, Internet computing: Principles of distributed systems and emerging internet-based technologies, pp. 195–236, 2020.

[2] H. K. Mistry, C. Mavani, A. Goswami, and R. Patel, ‘The Impact Of Cloud Computing And Ai On Industry Dynamics And Competition’, Educational Administration: Theory and Practice, vol. 30, no. 7, pp. 797–804, 2024.

[3] A. K. Y. Yanamala, ‘Emerging Challenges in Cloud Computing Security: A Comprehensive Review’, International Journal of Advanced Engineering Technologies and Innovations, vol. 1, no. 4, pp. 448–479, 2024.

[4] K. Cao, Y. Liu, G. Meng, and Q. Sun, ‘An overview on edge computing research’, IEEE access, vol. 8, pp. 85714–85728, 2020.

[5] L. Kong et al., ‘Edge-computing-driven internet of things: A survey’, ACM Computing Surveys, vol. 55, no. 8, pp. 1–41, 2022.

[6] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, ‘Resource scheduling in edge computing: A survey’, IEEE Communications Surveys & Tutorials, vol. 23, no. 4, pp. 2131–2165, 2021.

[7] T. Qiu, J. Chi, X. Zhou, Z. Ning, M. Atiquzzaman, and D. O. Wu, ‘Edge computing in industrial internet of things: Architecture, advances and challenges’, IEEE Communications Surveys & Tutorials, vol. 22, no. 4, pp. 2462–2488, 2020.

[8] Y. Chen, S. Ye, J. Wu, B. Wang, H. Wang, and W. Li, “Fast multi-type resource allocation in local-edge-cloud computing for energy-efficient service provision,” Information sciences, pp. 120502–120502, Mar. 2024, doi: https://doi.org/10.1016/j.ins.2024.120502.

[9] K. Sadatdiynov, L. Cui, L. Zhang, J. Z. Huang, S. Salloum, and M. S. Mahmud, ‘A review of optimization methods for computation offloading in edge computing networks’, Digital Communications and Networks, vol. 9, no. 2, pp. 450–461, 2023.

[10] M. P. J. Mahenge, C. Li, and C. A. Sanga, “Energy-efficient task offloading strategy in mobile edge computing for resource-intensive mobile applications,” Digital Communications and Networks, Apr. 2022, doi: https://doi.org/10.1016/j.dcan.2022.04.001.

[11] M. H. S, S. Kumar T, S. M. F. D. S. Mustapha, P. Gupta, and R. P. Tripathi, “Hybrid Approach for Resource Allocation in Cloud Infrastructure Using Random Forest and Genetic Algorithm,” Scientific Programming, vol. 2021, pp. 1–10, Oct. 2021, doi: https://doi.org/10.1155/2021/4924708.

[12] M. K. Alasmari, S. S. Alwakeel, and Y. A. Alohali, “A Multi-Classifiers Based Algorithm for Energy Efficient Tasks Offloading in Fog Computing,” Sensors, vol. 23, no. 16, pp. 7209–7209, Aug. 2023, doi: https://doi.org/10.3390/s23167209.

[13] A. Jayanetti, S. Halgamuge, and R. Buyya, “Deep reinforcement learning for energy and time optimized scheduling of precedence-constrained tasks in edge-cloud computing environments,” Future Generation Computer Systems, Jun. 2022, doi: https://doi.org/10.1016/j.future.2022.06.012.

[14] D. Alsadie, “Efficient Task Offloading Strategy for Energy-Constrained Edge Computing Environments: A Hybrid Optimization Approach,” IEEE Access, vol. 12, pp. 85089–85102, 2024, doi: https://doi.org/10.1109/access.2024.3415756.

[15] J. A. Ansere et al., ‘Optimal computation resource allocation in energy-efficient edge IoT systems with deep reinforcement learning’, IEEE Transactions on Green Communications and Networking, vol. 7, no. 4, pp. 2130–2142, 2023.

[16] S. Sangeetha, J. Logeshwaran, M. Faheem, R. Kannadasan, S. Sundararaju, and L. Vijayaraja, ‘Smart performance optimization of energy-aware scheduling model for resource sharing in 5G green communication systems’, The Journal of Engineering, vol. 2024, no. 2, p. e12358, 2024.

[17] A. Asghari, H. Azgomi, A. A. Zoraghchian, and A. Barzegarinezhad, ‘Energy-aware server placement in mobile edge computing using trees social relations optimization algorithm’, The Journal of Supercomputing, vol. 80, no. 5, pp. 6382–6410, 2024.

[18] F. Ramezani Shahidani, A. Ghasemi, A. Toroghi Haghighat, and A. Keshavarzi, ‘Task scheduling in edge-fog-cloud architecture: a multi-

objective load balancing approach using reinforcement learning algorithm', Computing, vol. 105, no. 6, pp. 1337–1359, 2023.

[19] J. Wen, J. Yang, T. Wang, Y. Li, and Z. Lv, 'Energy-efficient task allocation for reliable parallel computation of cluster-based wireless sensor network in edge computing', Digital Communications and Networks, vol. 9, no. 2, pp. 473–482, 2023.

[20] M. Raeisi-Varzaneh, O. Dakkak, A. Habbal, and B.-S. Kim, 'Resource scheduling in edge computing: Architecture, taxonomy, open issues and future research directions', IEEE Access, vol. 11, pp. 25329–25350, 2023.

[21] H. Huang, W. Zhan, G. Min, Z. Duan, and K. Peng, 'Mobility-aware computation offloading with load balancing in smart city networks using MEC federation', IEEE Transactions on Mobile Computing, 2024.

[22] Z. Zhou, M. Shojafar, J. Abawajy, H. Yin, and H. Lu, 'ECMS: An Edge Intelligent Energy Efficient Model in Mobile Edge Computing', IEEE Transactions on Green Communications and Networking, vol. 6, no. 1, pp. 238–247, 2022.

[23] K. Sathupadi, 'Ai-driven energy optimization in sdn-based cloud computing for balancing cost, energy efficiency, and network performance', International Journal of Applied Machine Learning and Computational Intelligence, vol. 13, no. 7, pp. 11–37, 2023.

[24] T. Cui, R. Yang, C. Fang, and S. Yu, 'Deep reinforcement learning-based resource allocation for content distribution in IoT-edge-cloud computing environments', Symmetry, vol. 15, no. 1, p. 217, 2023.

[25] Y. Mansouri and M. A. Babar, 'A review of edge computing: Features and resource virtualization', Journal of Parallel and Distributed Computing, vol. 150, pp. 155–183, 2021.

[26] C. Jian, L. Bao, and M. Zhang, 'A high-efficiency learning model for virtual machine placement in mobile edge computing', Cluster Computing, vol. 25, no. 5, pp. 3051–3066, 2022.

[27] L. Urblik, E. Kajati, P. Papcun, and I. Zolotová, 'Containerization in Edge Intelligence: A Review', Electronics, vol. 13, no. 7, p. 1335, 2024.

[28] J. Zhang, X. Zhou, T. Ge, X. Wang, and T. Hwang, 'Joint task scheduling and containerizing for efficient edge computing', IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 8, pp. 2086–2100, 2021.

[29] S. Hu, W. Shi, and G. Li, 'CEC: A containerized edge computing framework for dynamic resource provisioning', IEEE Transactions on Mobile Computing, vol. 22, no. 7, pp. 3840–3854, 2022.

[30] F. Zhou and R. Q. Hu, 'Computation Efficiency Maximization in Wireless-Powered Mobile Edge Computing Networks', IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 3170–3184, 2020.

[31] X. Zhou, X. Yang, J. Ma, I. Kevin, and K. Wang, 'Energy-efficient smart routing based on link correlation mining for wireless edge computing in IoT', IEEE Internet of Things Journal, vol. 9, no. 16, pp. 14988–14997, 2021.

[32] T. Vaiyapuri, V. S. Parvathy, V. Manikandan, N. Krishnaraj, D. Gupta, and K. Shankar, 'A novel hybrid optimization for cluster-based routing protocol in information-centric wireless sensor networks for IoT based mobile edge computing', Wireless Personal Communications, vol. 127, no. 1, pp. 39–62, 2022.

[33] A. Javadpour et al., 'An energy-optimized embedded load balancing using DVFS computing in cloud data centers', Computer Communications, vol. 197, pp. 255–266, 2023.

[34] S. K. Panda, M. Lin, and T. Zhou, 'Energy-efficient computation offloading with DVFS using deep reinforcement learning for time-critical IoT applications in edge computing', IEEE Internet of Things Journal, vol. 10, no. 8, pp. 6611–6621, 2022.

[35] A. Cañete, M. Amor, and L. Fuentes, 'HADES: An NFV solution for energy-efficient placement and resource allocation in heterogeneous infrastructures', Journal of Network and Computer Applications, vol. 221, p. 103764, 2024.

[36] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, 'Optimized power control design for over-the-air federated edge learning', IEEE Journal on Selected Areas in Communications, vol. 40, no. 1, pp. 342–358, 2021.

[37] X. Cao, G. Zhu, J. Xu, and S. Cui, 'Transmission power control for over-the-air federated averaging at network edge', IEEE Journal on Selected Areas in Communications, vol. 40, no. 5, pp. 1571–1586, 2022.

[38] X. Mo and J. Xu, 'Energy-efficient federated edge learning with joint communication and computation design', Journal of Communications and Information Networks, vol. 6, no. 2, pp. 110–124, 2021.

[39] H. Koumaras et al., '5G-enabled UAVs with command and control software component at the edge for supporting energy efficient opportunistic networks', Energies, vol. 14, no. 5, p. 1480, 2021.

[40] X. Shao, Z. Zhang, P. Song, Y. Feng, and X. Wang, 'A review of energy efficiency evaluation metrics for data centers', Energy and Buildings, vol. 271, p. 112308, 2022.

[41] Q. Zhang et al., 'A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization', Journal of Systems Architecture, vol. 119, p. 102253, 2021.

[42] M. Manganelli, A. Soldati, L. Martirano, and S. Ramakrishna, 'Strategies for improving the sustainability of data centers via energy mix, energy conservation, and circular energy', Sustainability, vol. 13, no. 11, p. 6114, 2021.

[43] E. Ahvar, A.-C. Orgerie, and A. Lebre, 'Estimating Energy Consumption of Cloud, Fog, and Edge Computing Infrastructures', IEEE Transactions on Sustainable Computing, vol. 7, no. 2, pp. 277–288, 2022.

[44] Q. Wang, L. T. Tan, R. Q. Hu, and Y. Qian, 'Hierarchical Energy-Efficient Mobile-Edge Computing in IoT Networks', IEEE Internet of Things Journal, vol. 7, no. 12, pp. 11626–11639, 2020.

[45] L. Heng, G. Yin, and X. Zhao, 'Energy aware cloud-edge service placement approaches in the Internet of Things communications', International Journal of Communication Systems, vol. 35, no. 1, p. e4899, 2022.

[46] Z. Li et al., 'Energy Efficient Reconfigurable Intelligent Surface Enabled Mobile Edge Computing Networks With NOMA', IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 2, pp. 427–440, 2021.

[47] L. Ale, N. Zhang, X. Fang, X. Chen, S. Wu, and L. Li, 'Delay-Aware and Energy-Efficient Computation Offloading in Mobile-Edge Computing Using Deep Reinforcement Learning', IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 3, pp. 881–892, 2021.

[48] M. Fardad, G.-M. Muntean, and I. Tal, 'Decentralized vehicular edge computing framework for energy-efficient task coordination', in 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), 2024, pp. 1–7.

[49] K. Kaur, S. Garg, G. S. Aujla, N. Kumar, J. J. P. C. Rodrigues, and M. Guizani, 'Edge Computing in the Industrial Internet of Things Environment: Software-Defined-Networks-Based Edge-Cloud Interplay', IEEE Communications Magazine, vol. 56, no. 2, pp. 44–51, 2018.

[50] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, 'A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications', IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1125–1142, 2017.

[51] U. M. Malik, M. A. Javed, S. Zeadally, and S. ul Islam, 'Energy-Efficient Fog Computing for 6G-Enabled Massive IoT: Recent Trends and Future Opportunities', IEEE Internet of Things Journal, vol. 9, no. 16, pp. 14572–14594, 2022.