Optimizing Next-generation Cloud and Data Center Networks: A Review of Routing, Resource Management, and Emerging Technologies

Sultan Alanazi

Department of Computer Science, College of Computer Engineering and Sciences,
Prince Sattam bin Abdulaziz University, Alkharj, Kingdom of Saudi Arabia
Email: sa.alanazi@psau.edu.sa

***Abstract*—As cloud computing and data centers become integral to global Information Technology (IT) infrastructure, optimizing routing and resource management in these networks is critical for maintaining performance, scalability, and energy efficiency. This paper presents a structured review of optimization models in cloud and data center environments using a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guided methodology covering literature from 2016 to 2025. Key advancements include adopting Software-Defined Networking (SDN), machine learning-based routing algorithms, and energy-efficient resource allocation strategies. This paper critically analyzes the challenges posed by scalability, latency, energy consumption, security, and interoperability, alongside the opportunities presented by Artificial Intelligence (AI)-driven autonomous networks, edge computing, and the integration of 5G technologies. Comparative evaluation highlights key trade-offs between performance gains and real-world feasibility, particularly for machine learning and deep reinforcement learning approaches. Furthermore, the study examines emerging trends, including cloud-edge collaboration and multi-objective optimization frameworks. The findings reveal that the adaptive methods can improve throughput, reduce latency, and enhance energy efficiency under specific datasets, simulation settings, traffic models, and network configurations. Overall, this review provides a consolidated perspective on current approaches, open challenges, and promising research directions for next-generation cloud and data center network optimization.**

***Keywords*—cloud and data networks, routing, resource management, optimizing next-generation, emerging technologies**

I. Introduction

Cloud computing and data center networks have become indispensable components of modern information technology infrastructure, providing scalable, flexible, and on-demand computing resources to industries across the globe [1]. From enterprises running mission-critical applications to consumers enjoying entertainment services and cloud-based applications, the demand for efficient, reliable, and high-performance data centers has surged in recent years. These networks support a wide range of services, including real-time data processing, machine learning, Artificial Intelligence (AI), video streaming, and online gaming, all of which require highly optimized routing and resource management mechanisms to meet their performance requirements [2, 3]. As data centers continue to expand in scale and complexity, optimizing these mechanisms has become more challenging but essential for ensuring seamless service delivery and operational efficiency [4].

Despite rapid technical advancements, current systems frequently rely on static, hardware-centric, or manually programmed methods that are unsuitable for dynamic and large-scale situations. The literature does not yet provide a comprehensive review of emerging technologies, particularly Software-Defined Networking (SDN), AI-driven optimization, and energy-aware frameworks. This review attempts to close this gap by thoroughly examining modern routing and resource management methods, evaluating their performance, and recommending future research areas.

Routing and resource management are the backbone of cloud and data center network performance. Routing involves determining the most efficient paths for data to travel through the network, ensuring minimal latency, reduced congestion, and optimal use of available bandwidth [5]. Resource management, on the other hand, focuses on the allocation and optimization of computational, storage, and network resources within the data center. Efficient resource management ensures that these resources are utilized effectively to meet demand while minimizing waste, operational costs, and energy consumption [6]. However, traditional approaches to routing and resource management are increasingly

Manuscript received February 1, 2026; revised March 24, 2026; accepted May 7, 2026; published July 10, 2026.

inadequate to handle the demands of modern, large-scale cloud environments. Conventional routing algorithms, such as Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP), while widely used, are often too rigid to adapt to the dynamic nature of data center traffic patterns. In addition, static resource allocation models can lead to inefficiencies, particularly in the face of variable workloads and traffic spikes [7].

In recent years, significant advancements have been made in the technologies and algorithms used to optimize routing and resource management. One of the most transformative developments has been the advent of SDN, which decouples the control plane from the data plane, allowing for centralized and programmable control over the network [8]. This enables real-time adjustments to routing policies based on current network conditions, resulting in more flexible and efficient traffic management. Similarly, Network Function Virtualization (NFV) allows network services to be virtualized, enabling more efficient resource allocation and reducing the reliance on specialized hardware [9]. Machine Learning (ML) techniques are increasingly used to predict traffic patterns, identify congestion points, and optimize routing and resource allocation dynamically. Despite these advancements, numerous challenges remain in optimizing routing and resource management within cloud and data center environments [10]. Scalability is a significant concern, as data centers grow to accommodate millions of interconnected devices and servers. The complexity of managing these networks increases exponentially, and traditional methods struggle to maintain optimal performance at such a scale [11]. Latency, particularly for time-sensitive applications such as video streaming and real-time analytics, remains a critical challenge. Ensuring that data travels the shortest, least congested path is essential, but it requires sophisticated algorithms that can adapt to fluctuating network conditions in real-time [12].

Additionally, energy efficiency is becoming a pressing issue [13], as data centers are among the largest consumers of electricity globally [14]. Reducing energy consumption while maintaining performance is a balancing act that requires innovative solutions, such as dynamic voltage scaling, server consolidation, and renewable energy integration. Security is another major challenge in cloud and data center networks. As the reliance on SDN and NFV grows, so do the security risks associated with these technologies [15]. Centralized control planes and virtualized network functions are attractive targets for cyberattacks, making it imperative to develop robust security frameworks that can safeguard the integrity of the network [16]. Furthermore, as cloud environments become more complex, ensuring interoperability between different cloud providers, platforms, and architectures becomes essential for efficient resource management, especially in hybrid and multi-cloud deployments [17].

In addition to these challenges, there are numerous opportunities for innovation in routing and resource management. The rise of AI and machine learning presents new possibilities for creating intelligent, self-optimizing networks that can autonomously manage traffic and resources based on real-time analysis of network conditions [18]. Edge computing and 5G integration open up further opportunities to optimize resource distribution and reduce latency by bringing computation closer to the data source [19]. Moreover, the potential of blockchain technology to enhance the security and transparency of resource management in decentralized cloud networks is also gaining attention [20].

II. Review Methodology

This review adopts a structured literature review methodology guided by the PRISMA framework to ensure transparency, reproducibility, and systematic coverage of relevant research. Fig. 1 shows the annual publications from 2016 to 2025, with the final search conducted on 23 November, 2025.

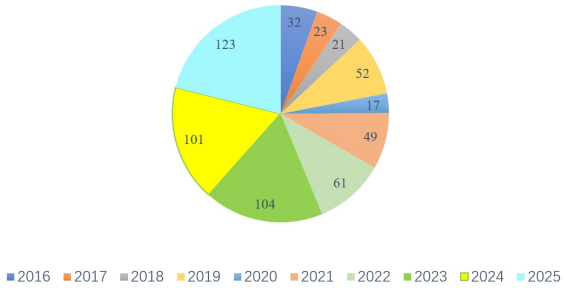

Fig. 1. A bibliometric analysis of WoS, IEEE Xplore, and Scopus (2016–2025).

Fig. 1 shows that 583 papers have been published since 2016. The search included studies published between 2016 and 2025, and the final search was conducted on 23 November 2025. The publication rate reflects the growing demand for optimizing next-generation cloud and data center networks. Across global databases, such as Web of Sciences, IEEE Xplore, and Scopus, a thorough literature search was performed to find the relevant studies. Searches were designed to capture multidisciplinary evidence related to cloud computing, data center networks, load balancing, SDN, NFV, energy efficiency, edge computing, and “interoperability. The main author conducted both the screening and eligibility evaluation to ensure the inclusion and exclusion criteria were applied consistently. Ambiguous cases were re-evaluated iteratively against the predefined criteria to minimize selection inconsistency.

Cross-sectional studies are primarily used to identify new patterns or research gaps, not to establish cause and effect. We collected these publications from various databases and applied clear selection criteria to determine which to keep and which to discard, as shown in Table I.

Table I shows that articles such as short-length publications or those lacking relevant technical content were excluded. We selected publications for further evaluation based on their quality assessment scores. Many review and survey studies have been examined, each possessing distinct advantages and disadvantages. Fig. 2 shows a PRISMA literature identification and selection workflow, describing the stages of identification, screening, eligibility assessment, and inclusion.

TABLE I. PUBLICATION SELECTION CRITERIA

Included Studies	Excluded Studies	Extracted Research Information
Research focused on cloud computing, data center networks, routing optimization, SDN, NFV, AI/ML-based optimization, and related emerging technologies.	Articles not related to cloud or data center networks, SDN, ML based optimization.	Study objective and research focus.
Peer-reviewed journal articles, conference papers, and relevant review papers with clear technical contribution.	Short-length articles Articles published in non-indexed Journals.	Type of study (survey, experimental, simulation-based, framework/proposal, comparative study).
Studies reporting conceptual, analytical, simulation-based, on routing and resource management in cloud/data center environments.	Minum details without proper references.	Network environment or architecture studied (cloud, data center, hybrid cloud, edge-cloud, SDN-enabled network, NFV environment).
Studies published in English.	Non-English studies.	Techniques/methods used (e.g., OSPF, BGP, SDN, ML, RL, DVFS, live migration, server consolidation, blockchain, edge integration).
Studies published within the selected review period, with earlier foundational studies included where necessary for background.	Duplicate studies or overlapping versions of the same work.	Evaluation metrics reported (latency, throughput, packet loss, energy efficiency, QoS, fault tolerance, scalability).

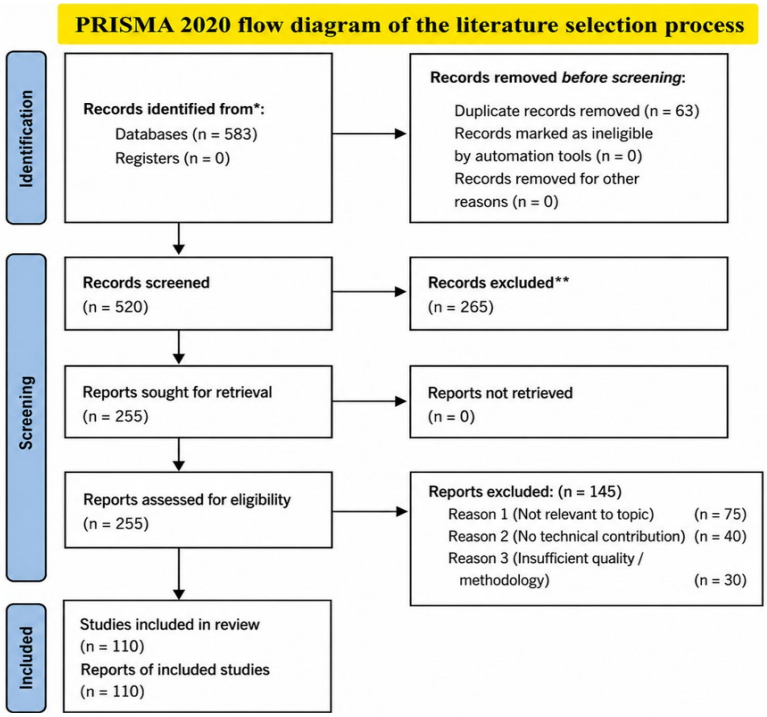

Fig. 2. PRISMA flow diagram for literature identification and selection workflow.

Fig. 2 shows how to use the rules in Table I to decide which studies to include or exclude at each step: finding studies, checking titles and summaries, reviewing full articles, and deciding which to keep.

The main contributions of this paper are as follows. First, we provide a comprehensive review of the current state-of-the-art techniques for routing and resource management in cloud and data center networks. We analyze the strengths and limitations of traditional methods while highlighting the benefits of emerging technologies such as SDN, NFV, and machine learning. Second, we identify and discuss the key challenges that cloud and data center operators face, particularly in terms of scalability, latency, energy efficiency, and security. Third, we explore the potential opportunities presented by new technologies and frameworks, including AI-driven automation, edge computing, 5G, and blockchain. Finally, we propose future research directions aimed at addressing these challenges and leveraging these opportunities to create more efficient, secure, and sustainable cloud and data center networks. Fig. 3 shows the cloud data center network optimization.

The main contributions of this paper are as follows. First, we provide a comprehensive review of the current state-of-the-art techniques for routing and resource management in cloud and data center networks. We analyze the strengths and limitations of traditional methods while highlighting the benefits of emerging technologies such as SDN, NFV, and machine learning. Second, we identify and discuss the key challenges that cloud and data center operators face, particularly in terms of scalability, latency, energy efficiency, and security. Third, we explore the potential opportunities presented by new technologies and frameworks, including AI-driven automation, edge computing, 5G, and blockchain. Finally, we propose future research directions aimed at addressing these challenges and leveraging these opportunities to create more efficient, secure, and sustainable cloud and data center networks. Fig. 3 shows the cloud data center network optimization.

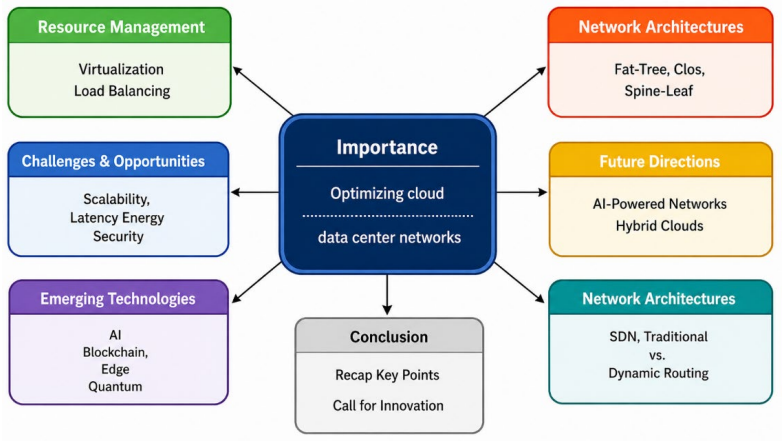

Fig. 3. Cloud data center network optimization: Issues, innovations, design, and perspectives of the future.

III. Cloud and Data Center Network Architectures

To analyze optimization approaches, the overall design of data center networks must be defined. Layers are implemented in data centers using Fat-Tree, Clos, or Spine-Leaf topologies to achieve bandwidth scalability, besides efficient traffic management [21]. Designed to handle high-volume data traffic, both within and between servers. The Fat-Tree and Clos topologies boast of having relatively high bandwidth together with the ability to handle faults, while the Spine-Leaf comes with an easier and more preferable relative scale design that can efficiently support the east-west traffic, which is more commonplace in data centers [22]. Understanding these architectures is essential for improving the network performance and resource utilization. The following subsections describe the basic components of such architectures as shown in Fig. 4.

A. Fat-Tree and Clos Topologies

Fat-Tree and Clos topologies are being adopted in large-scale data centers due to bandwidth utilization and fault tolerance features. These hierarchical network architectures are highly suitable for today's applications, having high traffic levels, cloud computing, data analytics, artificial intelligence, and others [23]. For instance, Fat-Tree topology facilitates the distribution of traffic across multiple network paths and hence will not congest when busy, as most other networks do. Likewise, Clos topology, which was designed for telephone networks, has found its niche in the data center as non-blocking, under which messages can be exchanged between switches and servers with ease. Overall, Fat-Tree and Clos topologies are highly scalable interconnect networks; thus, they are capable of handling the growing nature of demands placed on data centers as they grow [24]. However, this scalability brings the problem of large volumes of traffic that pass through the network for management. Lack of traffic control means traffic congestion, which in turn impacts data center productivity because of latency and packet droppage. To avoid this, these topologies demand mature routing protocols that can accommodate the queue and respond accurately according to the current network state [25]. Other techniques, like Equal Cost Multi-Path (ECMP) are used to load balance traffic over all the available paths. While these topologies have highly desirable characteristics in terms of fault tolerance and bandwidth utilization they are highly complex and require enhanced routing algorithms to fully exploit large-scale networks [26].

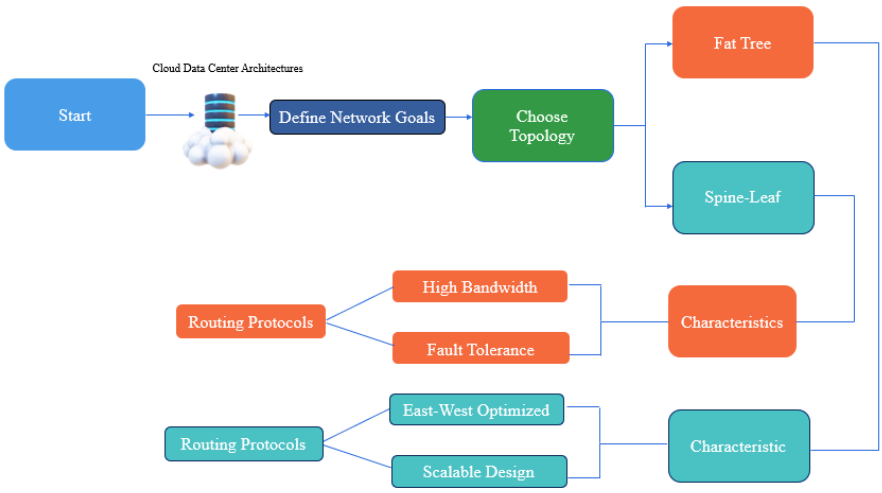

Fig. 4. Cloud data center networks design: Performance, reliability, and scalability.

B. Spine-Leaf Topologies

Spine-Leaf topologies have emerged as an optimal topology for implementing innovative cloud data centers based on their simplicity, scalability as well as efficiency. This architecture is particularly designed for east-west traffic, which means the horizontal traffic occurring within the data center, as distinct from the north-south traffic that goes in and out of the data center [27]. The ever-growing amount of east-west traffic in cloud environments, in which applications are built on top of distributed computing across multiple servers, makes traditional hierarchical networks ineffective. This is resolved by this Spine-Leaf topology, whereby each piece of data between any two servers only has to go through a single spine switch, hence reducing latency [28]. The architecture consists of two main layers: The leaf layer is connected to the servers, and the spine layer connects all the leaf switches [29]. This uniform connection of the spine to the leaf switches means that each server is only a single hop away from any other server and thus there are no bottlenecks and congestion, which are usual in most designs. Also, the design is highly scalable because new spine or leaf switches may be incorporated into the topology without affecting existing paths in the communication network. With the increasing size of cloud data centers, Spine-Leaf topologies can be easily and naturally extended at minimal latency and maximum bandwidth [30]. This makes the Spine-Leaf architecture suited for high-performance, low-latency applications that need to be scaled, without compromising the network solution, and are suited to big data analytics, artificial intelligence, and virtualization environments.

IV. Routing Optimization in Data Center Networks

Optimization of routes is crucial in an attempt to enhance the quality of the network in addition to cutting down the flow of traffic, which may slow down the system. The subsequent sections delve more into the important routing methods and strategies that have been put in place to enhance the performance of cloud and data center networks.

A. Traditional Routing Protocols

Core protocols, such as Open Shortest Path First (OSPF) and BGP, continue providing inter-routing between large networks or keeping traditional routing in networks with legacy systems [7, 31]. However, in the context of modern data centers, these protocols hit certain deficiencies. OSPF and BGP were originally developed for networks that were not rapidly growing and mainly followed the hierarchy system, but data center networks are no longer like this. In today's environment, data traffic within the modern data center tends to be heavy with east-west data flow communication between servers, and these protocols are not well-suited to optimize efficiently [32]. Table II shows the Comparative methods across different parameters and emerging technologies.

Another weakness of OSPF and BGP, which are ideal for data-center applications, is their static design. These protocols use defined routes for packet flow; therefore, they fail to consider the dynamic conditions of the network, including congestion, fluctuations in traffic patterns, and failure [33]. As a consequence, they result in unsatisfactory performance, especially when density demands start curving sharply, something not unusual in such systems. The consequence can be bottlenecks, increased latency time, and generally unutilized bandwidth. On the other hand, modern prospective methods like SDN are capable of managing traffic according to the real needs of the network measurements [34]. Although OSPF and BGP are still suitable for the specified purposes, their drawbacks prompt the search for new intelligent routing solutions for large-scale, modern data centers with high demands for flexibility and speed.

According to recent studies, conventional protocols such as OSPF and BGP have typical speeds of 20–30 ms in east-west traffic scenarios, but SDN-based routing methods reduce latency to 5–10 ms, resulting in a 60–75% increase in time-sensitive conditions [32]. Furthermore, standard methods have packet loss rates of up to 1.2% under high-load conditions, compared to less than 0.4% with dynamic SDN routing.

A 2023 investigation by Google found that traditional BGP configurations had packet retransmission rates 30% quicker during peak loads than SDN configurations. Additionally, a test by Microsoft Azure revealed that static routing systems had a 20% decline in performance during busy periods due to difficulties in handling fluctuating east-west traffic. Table III shows the comparative analysis of routing optimization techniques.

The table compares the two routing optimization techniques used in cloud and data center networking. It highlights the algorithm type, the optimization objectives, the evaluation environment, and the results obtained. It helps the readers to easily compare how each routing technique, including traditional, SDN, DRL, and hybrid approaches, copes with latency, throughput, and traffic efficiency.

B. Software-Defined Networking (SDN)

SDN has recently been studied as a solution to organize routing in a data center; it significantly differs from conventional approaches and provides greater flexibility. One of the most revolutionary elements of SDN is the separation of the control plane, which is responsible for the routing decisions, from the data plane that performs the actual transmission of the data [35]. This mechanism enables network administrators to implement routing at runtime on a case-by-case basis instead of setting up rigid routing protocols. Consequently, SDN is in a better position to adapt to the shifting traffic patterns, congestion levels within the network, or device failures to improve the delivery of data across the network [8, 32].

TABLE II. COMPARING EXISTING METHODS ACROSS DIFFERENT PARAMETERS AND EMERGING TECHNOLOGIES FOR RESOURCE MANAGEMENT AND OPTIMIZING ROUTING IN CLOUD AND DATA CENTER NETWORKS

Parameter	Traditional Methods	Emerging Technologies (SDN, NFV, ML, AI)
Scalability	Limited scalability due to static routing algorithms and resource allocation methods. Requires manual intervention and pre-configured rules.	High scalability is enabled by dynamic, centralized control (SDN) and automated decisions through machine learning and AI.
Flexibility	Rigid configurations; difficult to adapt to changing traffic patterns or workloads.	Highly flexible due to real-time control of network resources and programmable infrastructure (SDN/NFV).
Latency	Often higher, as traditional routing protocols (OSPF, BGP) may not optimize for latency.	Lower latency with real-time traffic routing adjustments and edge computing integration.
Energy Efficiency	Energy management is reactive, often leading to high power consumption. Limited use of energy-saving techniques.	Proactive energy management through AI-driven optimization, server consolidation, and energy-efficient algorithms like dynamic voltage scaling.
Traffic Engineering	Traffic routing is largely static, with limited capacity for real-time adjustments.	Advanced traffic engineering is possible via SDN and AI-based models, allowing real-time adjustments to optimize network traffic.
Resource Allocation	Static resource allocation; often leads to over-provisioning or underutilization.	Dynamic resource allocation based on real-time demand using AI and ML, improving efficiency.
Load Balancing	Simple algorithms (e.g., round-robin or least connection) with limited adaptability.	Advanced, dynamic load balancing using AI and SDN to optimize workloads across multiple servers in real-time.
Security	Security relies on traditional firewalls and VPNs, vulnerable to attacks on static configurations.	Enhanced security via centralized SDN control, automated threat detection, and blockchain-enabled resource management.
Cost Efficiency	Higher operational costs due to over-provisioning, inefficient energy usage, and manual configuration.	Reduced operational costs through automation, energy efficiency, and optimal resource usage enabled by machine learning and AI.
Adaptability to Emerging Tech	Limited adaptability to integrate emerging technologies such as edge computing, 5G, or AI.	Seamless integration with emerging technologies like edge computing, 5G, AI, and blockchain, enabling future-proof solutions.
Management Complexity	High complexity in managing and configuring the network manually, requiring significant expertise.	Simplified management with centralized control (SDN), AI-driven automation, and real-time analytics reducing manual intervention.

TABLE III. COMPARATIVE ANALYSIS OF ROUTING OPTIMIZATION APPROACHES

Ref.	Algorithm Category	Technique / Method	Primary Objective(S)	Evaluation Environment	Workload / Traffic Setting	Reported Performance Outcome
[32]	SDN optimization	Routing optimization	Delay, load balancing	SDN environment	Traffic scenario	Reduced delay
[35]	SDN	Proactive SDN Route	Routing optimization	Experimental/simulation study	Data center traffic flows	Improved route efficiency
[36]	SDN load balancing	Dynamic load balancing	Throughput, delay	SDN test environment	Dynamic traffic	Lower end-to-end delay,
[37]	Segment routing	Energy-aware segment routing	Energy, path efficiency	Software-defined data center setting	Controlled traffic	Improved routing efficiency
[38]	DRL	Deep reinforcement-learning-based routing	Adaptive routing	Data center network simulation	Dynamic traffic conditions	Better adaptive routing

The primary advantage of creating a centralized control system in SDN is that it offers an overview of an entire network, allowing the traffic that flows through it to be rerouted according to latency and bandwidth. For instance, with SDN, the routing decisions are made by a central controller that can determine network conditions and change traffic paths holistically to areas that experience high congestion to achieve balance [39]. Such flexibility is critical in today's data centers, where traffic patterns can shift drastically in a shorter span of time, and conventional routing protocols can be slow to adapt. Furthermore, the architecture of SDN-based solutions is inherently programmable and has policies that can be tuned with high precision, thus enabling the ability to prioritize traffic and resources [40].

1) Centralized control

SDN adds the advantage of having an overall view of the network, therefore allowing for routing decisions that are way more central than the default ones that are in the distributed architecture [41]. In existing network models, it is quite a norm that each piece of equipment or router independently initiates routing tasks using minimum local information. This divorce results in routinization, often because routers do not possess the holistic view they require to make the best routing decisions, given other networks and specifically within rapidly evolving, complex, and large networks such as modern data centers [42]. In SDN, the controller has a periscope vision of the whole current state of the network, the amount of traffic, congestion, and available path information.

This enables the SDN controller to employ the correct evaluation of possible paths for data traffic, given the situation within the network, without reliance on old or partial information [36]. In that sense, SDN involves centralizing routing decisions so that wasted areas and underlying bottlenecks are prevented concerning the overall bandwidth asset of the network. Also, such a view of the network allows for a quicker response to changes in network load or hardware problems, for instance.

Centralized control makes it easy to dynamically divert traffic, eliminating traffic congestion and improving network performance and reliability [43].

This capability is critical in large-scale high-performance data center layers where typical distributed routing protocols would not manage to address the dynamic and dispersed traffic design.

2) Dynamic routing adjustments

One of the most notable benefits of the SDN controllers is that they can dynamically redirect traffic, thereby avoiding a buildup of traffic at any particular point in the network and, at the same time, decreasing the delay in real-time. SDN is distinct from conventional protocols that depend on statically configured paths; instead, the centralized control of SDN lets the administrator keep track of activity comprehensively, permitting changes based on current status to be made instantly [44]. In the case of congestion, or when congestion is anticipated, the SDN controller is fully capable of noticing a congestion issue and immediately finding other non-congested paths to choose, thus keeping the communication stream fluid and untangled from any congestion issues. This real-time traffic control is especially helpful in large Hew, high-performance networking facilities, including data centers, where the network traffic may peak because of the numerous flows of east-west traffic between the servers [45]. Because of managing data paths, SDN is not only able to prevent congestion but also to reduce the delays due to high traffic loads in improved network performance. Another advantage of the capability of manipulating routes in real-time terms also improves the quality of the reaction from the applicability and services, hence making SDN efficient for iterative applications and workloads such as video streaming, online gaming, and real-time analytics [46]. Last but not least, through the power of SDN, traffic can be rerouted whenever passing through a link that has already experienced heavy traffic flow in a certain period of time, which maintains the high quality of data delivery to the last mile [47]. These strengths make the network architecture more flexible and provide network administrators with better control over their infrastructures, and also create more dependable networks.

SDN-enabled systems with dynamic routing reduce average end-to-end delay by 45%, boost throughput by 30–40%, and retain higher fault tolerance, with up to 99.9% network availability during node or link failures due to centralized rerouting features [36, 44]. These gains are particularly relevant in large-scale data centers that handle varying workloads.

Although SDN is flexible, it has some major drawbacks. In a large network, a centralized controller may become a single point of failure. In a highly trafficked network, scalability and latency issues arise when decisions are made by the controller. Centralized systems are also difficult to implement due to scalability issues.

C. Traffic Engineering with Segment Routing (SR)

SR is an appealing traffic engineering solution that enhances network administration by addressing complex data plane designs and optimizing routing mechanisms. Unlike conventional routing protocols, whereby routing tables prescribe the route that a packet will follow in a network, SR imprints the whole path inside a packet [37]. This provides high control over the flow that the packet should take; there is no need to manage the state of every router in the route. Table IV illustrates the comparison of different routing methods.

SR methods improve network stability, with recent evaluations demonstrating 40% faster convergence after a failure and 35% lower average path setup latency than older routing alternatives. Furthermore, SR-integrated SDN systems have 50% fewer missed packets during rerouting events, which increases fault resilience.

Specifically, in a data center environment where traffic flows are unpredictable and the workloads are distributed among thousands of hosts, traffic control at the per-flow level offered by SR can be highly useful. The combination of exact path information into the packet header facilitates connectivity that SR creates for the network operators, allowing more careful steering of data based on the performance criteria to avoid areas that may be congested in the network [48]. This level of control is especially important for east-west traffic, which is the focus of many current data center designs and is the key to achieving high throughput in applications running on cloud computing systems [49].

In addition, variable state, deployed in SC, reduces the quantity of data plane information necessary for intermediate routers, facilitating network design and scalability. SR works with an ability to change the traffic in accordance with the network conditions, enhancing the flexibility of data center networks and making the networks more ideal as they increase in efficiency, use less latency, and are more ideal in terms of the network resources they are likely to use [50]. It makes it an important tool for traffic engineering within heterogeneous, high-performance, large-scale data center environments.

Segment routing provides it more power over traffic, but it also adds more header overhead and makes configuration more complicated. Managing path encoding at scale can be highly expensive in terms of computing power. Also, compatibility with current routing protocols and hardware limitations could make deployment more difficult. The requirement for precise knowledge of the network state makes it less useful in situations where traffic is highly dynamic or difficult to predict.

TABLE IV. COMPARISON OF DIFFERENT ROUTING METHODS IN TERMS OF DIFFERENT PARAMETERS

Parameter	Traditional Routing (OSPF, BGP)	Emerging Routing (SDN, ML)
Control Plane	Distributed, static control	Centralized, programmable control (SDN)
Routing Flexibility	Limited, predefined routes	Highly flexible, dynamic route adjustments
Traffic Awareness	Low, no real-time traffic awareness	High, real-time traffic monitoring and response (SDN/ML)
Adaptability	Slow to adapt to changes in network conditions	Fast adaptation to real-time conditions (AI/ML predictions)
Latency Optimization	Suboptimal, may not prioritize latency	Low latency, with optimized routes through real-time updates
Scalability	Limited scalability for large-scale data centers	Scalable with centralized SDN control

V. Resource Management in Data Centers

Resource optimization facilitates that information technology computing, storage, and networking resources meet demand in a manner that optimizes cost. Dynamic scaling, live migration, and DVFS are all resource management methods that require trade-offs between performance and efficiency. Aggressive energy-saving strategies may make real-time apps slower and lower their Quality of Service (QoS). In large-scale cloud environments, it remains a problem that resource allocation algorithms don't perform well with a large number of users. This section covers the central resource management approaches applied in the cloud and data center contexts.

A. Virtualization and Resource Allocation

In Virtual Machines (VM), several virtual machines can share one or more physical computing resources, hence improving the space utilization and resource capacity in datacenters. However, working in a shared environment implies that resources should be well managed to achieve the best results possible [51]. The manner in which the individual VMs are distributed over the physical servers affects the efficiency of resource usage, such as CPU, memory, and storage. The improper VM assignment causes resource fragmentation in that physical servers are often left with small, isolated resources that cannot be deployed to other VMs. Consequently, this leads to an underutilization of the hardware used and may lead to higher operational costs due to needless utilization of multiple physical servers [52]. Table V shows the comparison of different resource management methods.

How to prevent wastage of resources is therefore a very critical success factor, especially in relation to VM placement. The problem is how to distribute system load between physical servers on the fly so that some machines are not overloaded and, at the same time, the load is not fragmented among too many machines either. This must be done with the use of sophisticated algorithms that consider information such as current and anticipated consumption of resources on the cloud, and the indicated parameters of individual VMs [53]. Furthermore, live VM migration, in which VMs are migrated from one physical host to another as needed in real-time, is applied to adjust VM placements and server loads that are dynamically shifting. Effective allocation of organizational resources combined with effective assignment of VMs can enhance total system performance while decreasing energy consumption; huge servers can be turned off during low-traffic periods [54]. With the growth of cloud environments and data centers, these challenges are joining the importance of providing cost-effective and efficient services.

TABLE V. Comparison of Different Resource Management Methods

Parameter	Traditional Resource Management	Dynamic Resource Management (SDN, AI)
Resource Allocation	Static, manual allocation	Dynamic, real-time allocation based on current demand
Workload Handling	Reactive, may lead to over/under-provisioning	Proactive, with AI forecasting demand to optimize usage
Energy Efficiency	Low, resource over-provisioning common	High, optimized resource use and server consolidation
Scalability	Limited for large-scale operations	Highly scalable, automated adjustments as demand fluctuates
Load Balancing	Static or round-robin algorithms	Intelligent, dynamic load balancing using AI and SDN
Management Complexity	High, requiring significant manual intervention	Low, automated processes reduce human management needs

1) Dynamic resource scaling

Cloud computing naturally employs dynamic scaling, which means resources are procured in a real-time fashion depending on the current availability, thus giving a very high utilization rate and minimal wasted resources. Historical architectures place capacity high above demand levels, thus the provision of excess capacity for a short time [53]. Dynamic scaling, on the other hand, permits cloud platforms to provide a change in resource allocation, for instance, computing capacity, memory, and storage by expanding itself during active hours and contracting during some other non-active hours. This real-time responsiveness makes it possible to ensure that applications get the right resources as needed to meet their performance needs while avoiding unnecessary use of resources at times when application demand is low. To implement dynamic scaling, you can use vertical scaling, which incorporates additional resources to the already existing virtual machines, or horizontal scaling, which incorporates other virtual machines or containers. Both are crucial for ensuring gathered service-level agreements and minimizing the expenses of cloud infrastructures' functioning [55]. Cloud platforms can allow businesses to set up more flexible systems as they can map resource usage to loads over time, a lot better than a traditional on-premises environment, and therefore the current levels of data transmission or user activity do not need to be constantly tweaked [56]. This is one of the primary benefits of the flexibility that is achieved by cloud computing; since many companies only wish to use specific resources for a finite amount of time, it is efficient and cost-effective to pay for only that amount of service.

2) Live migration

For efficient allocation of resources, live migration of VMs is an important function that transfers workloads from the congested host to other less congested hosts without interrupting the service.

Live migration is a technology that lets the VM move from one physical server to another while it is still operational, so that users can access applications on the VM and interact with them even as the VM changes physical hosts [57]. This load redistribution among the servers is functional in avoiding system congestion as well as improving the total system performance. In a cloud or data center, the workload could be random; therefore, some servers will end up being over-employed while others will be under-employed. When those imbalances occur without live migration, the result is an inefficient use

of resources, potential performance degradation, or worse, service unavailability [58]. Through the constant check of the server's usage, the management of resources can immediately recognize that a server is on the brink of its resource limit, and the transfer of VMs to more capable servers can occur. Table VI shows the comparison of different security mechanisms.

This is made possible by contemporary virtualization techniques, by which VMs can retain all such parameters as network connections, memory states, and storage contents when they are being moved. Hence, users can be very limited or experience no dos at all, something very important in today's high-availability applications such as e-commerce, fintech, and healthcare [59]. Moreover, live migration helps energy efficiency in the sense that consolidated servers mean that several under-utilized servers can be shut down, or put into sleep mode to minimize energy consumption without diminishing the benefits [60]. In general, live VM migration is a preeminent technique for revamping resource orchestration, balancing loads, and preserving services in screen and data center domains.

TABLE VI. COMPARISON OF DIFFERENT SECURITY MECHANISMS

Parameter	Traditional Security (Firewalls, VPNs)	Emerging Security (SDN, Blockchain)
Security Mechanism	Based on static configurations and fixed routes	Dynamic, with real-time updates via SDN and blockchain tech
Vulnerability to Attacks	Higher, more exposed to static attack vectors	Lower, centralized control allows for real-time threat response
Traffic Isolation	Limited traffic isolation in shared environments	High, SDN and NFV enable network function isolation
Intrusion Detection	Manual, based on predefined rules	Automated, with AI-based threat detection and response
Flexibility	Low, difficult to change or update security rules	High, SDN allows flexible and fast security policy updates
Data Integrity	Requires multiple systems for integrity checks	Enhanced through blockchain's immutability

B. Load Balancing

Traffic distribution is very important to avoid congestion and optimize the use of the resources in the cloud and data center. This utility manages workloads in a manner that distributes them over the many servers or resources making up a system; load balancing avoids a scenario in which one server or resource becomes saturated, a situation that results in decreased performance or complete failure [61]. There are several approaches one is likely to come across when wishing to attain proper load distribution. This is accomplished by perhaps the simplest method—Round Robin, where fresh incoming requests are distributed circularly around all the servers. It is quite simple to work with, but it can distribute the traffic unevenly since it is not dependent on the current server's load or capacity [62].

The simplest is First In/First Out, which means that the traffic is sent to the server with the least current load, while Least Connections is even more refined and sends traffic to the server with the smallest number of connections, as it is targeted at environments with large variances in workload. This helps to prevent overloading busier servers and, at the same time, makes full use of the less busy servers [63]. This is followed by IP Hashing, where clients are always routed to a particular server depending on the client's IP address. This is especially valuable when the applications need to have session persistence.

Weighted Round Robin is the superior version of the basic Round Robin algorithm in that it is possible to give a dynamic weight to each server, thus making more powerful servers rotate more frequently. The last one at present Ubiquitous Cloud environment, Dynamic Load balancing by using SDN. The ability to monitor traffic flow in real-time and dynamically optimize the routing of the traffic by the actual current state of the network, greatly augmenting the efficiency and minimizing the latency, is the core feature of SDN [64]. These techniques as a whole make certain that no resources are wasted, and capacity is not left unused, while averting situations where the system is overloaded and consequently offers a stable and reliable performance. These techniques are as follows:

1) Round Robin

Among the most basic load-balancing algorithms that can be applied to form cyclic sequences of incoming requests among multiple servers, the Round Robin algorithm is used. It works based on feeding the next request to the next available server, and hence, going around in circles with the list of available servers [65]. After the last server has been tried and no response has been given, the program goes back to the first server to attempt to convey the message so that all the servers are used at an equal frequency. This is easy to apply; it does not need any information like the current load or processing ability of the server, so it is suitable where all the servers have similar abilities. The Round Robin scheduling technique is quite efficient if the incoming requests are of the same size and duration [66]. However, it poses some problems for optimal solutions where servers are of different capabilities or when requests are different in terms of resource demands. In such situations, better scheduling algorithms such as the Weighted Round Robin or the Least Connection algorithms might be applied to incorporate such changes [67]. Nevertheless, Round Robin still retains its status as one of the most applied load-balancing algorithms because of its simplicity and rather good applicability in complex systems, whereas Round Robin is mostly applicable to homogeneous server systems.

2) Least connection

The least connections load-balancing algorithm distributes the traffic to the server with the minimum link connections and thereby is more dynamic. While Round Robin binds the requests with equal intervals, no matter the state of the server, Least Connections looks into how many requests the server to which it forwards the request is handling at a given time [36, 63]. Because of this, it is most beneficial in situations where the traffic might be unpredictable or where single requests may require different amounts of resources. If there is a new request,

the load balancer determines the number of new connections in each server and distributes the current request to the server with fewer connections. This prevents any particular server from being overloaded as well, since the many servers that are already processing numerous requests from parties will receive the new ones [68]. Therefore, the least connections guarantees a proper workload distribution between servers and the shortest response times. It is most useful when there are sudden variations in application traffic [69] and in overlong connection-based applications like web applications, where the connection between the client and server may be sustained for a long time without a break.

3) SDN-based load balancing

SDN controllers provide a more intelligent load balancing since they change their policies depending on current traffic patterns. With the controller, decisions have to be made regarding load balancing, traffic flow, and server conditions across the network, all in real-time, unlike with earlier methods that involved static load balancing [70]. This level of awareness enables SDN controllers to offer traffic rerouting and adapt to the high and low traffic flows, the availability of servers, and congestion within the network. SDN for centralized control continuously changes the routing and load balancing policy to adjust resource utilization [71]. For instance, if a single server is experiencing a high influx of traffic, it can be easily redirected to other servers that would be less occupied with flows of traffic, hence avoiding congestion among several servers. Furthermore, controllers in software-defined networks can program algorithms of their choice and use concrete rules to distribute the traffic evenly in terms of latency, bandwidth utilization, and current server performance [72]. SDN also boasts of flexibility in terms of scalability, and integration with cloud services is an essential requirement for current complex as well as distributed environments [73]. As a result, load balancing can be done efficiently by using deputies in real-time, which is possible through SDN, resulting in high availability, optimized allocation of resources, and improved user experience in dynamic networks.

C. Energy-Efficient Resource Management

Data center energy consumption is a crucial challenge for immense server farms; therefore, efficient energy management methods are crucial.

1) Green scheduling

Designing the work and related processes, taking into account when the server is most energy efficient or when renewable sources of energy can be used to power the data centers and cloud structures, is one of the ways through which energy consumption can be reduced [74]. This is otherwise known as energy-aware scheduling and essentially seeks to utilize the computing resources to the fullest while at the same time addressing the question of environmental management and overall costs. The power of assessment to decrease greenhouse gas emissions meant that production planning had to obtain a green approach [75]. For instance, energy-demanding functions can be set to run during the optimum clean energy, such as in the foundation, middle, and summer seasons when the renewable energy is abundant. Likewise, tasks can be shifted to a time of minimum grid energy demand, thereby cutting down the utilization of non-renewable energy forms. Also, in optimizing administrative tasks, a lower energy server or processor can be implemented to perform tasks or be allocated to workloads to decrease the energy usage of the system. This dynamic scheduling not only aids sustainable operations but also minimizes energy expenses for businesses by using time-of-use energy rating [76]. The increased utilization of this model in different companies also creates a green IT environment and serves other environmental objectives, such as the lowering of greenhouse gases. Energy-efficient scheduling may save power, but it has some drawbacks. Scheduling high-demand processes during periods of low demand to take advantage of green power may result in longer wait times, which might have an impact on applications that require speedy replies. As a result, we must find a balance between improving energy efficiency, which can save 25–35% of power, and the potential 10-20% increase in task delays when pushing tasks to greener times.

2) Dynamic Voltage and Frequency Scaling (DVFS)

DVFS is a mechanism used to vary CPU operation with fluctuating traffic so that more energy is saved during idle time. Dvfs means DVFS, which allows varying the processor's voltage and working frequency depending on the tasks executed to prevent the servers from consuming more energy than required at lighter computing loads [77]. The downside of low demand is that the system scales down the processing frequency and voltage of the CPU to minimize power consumption, but still ensures it can take on workloads adequately.

Current studies conducted in cloud data centers show that DVFS can reduce CPU energy usage by 25–35% at low-utilization periods while preserving QoS levels [77]. In addition, real-time DVFS optimization systems have been proven to cut overall server power use by 18–22%, contributing significantly to total data center energy savings.

Although DVFS saves power, it also decreases CPU frequency during low loads, which may result in performance deterioration or increased response time for services in real-time. Comparisons indicate that adopting harsh DVFS policies decreases energy consumption by up to 30%, but may result in 8–12% speed implications in latency-sensitive applications [78]. As a result, DVFS must be designed for workload characteristics, giving up maximum power savings for confined QoS standards. Table VII shows the comparison of different energy-efficient techniques.

When the load is high, the response to it is also increased, so that during periods of high load, high performance is achieved. This dynamic adjustment, therefore, means that while a high-power consumption is needed for computational purposes, the energy efficiency is kept low, hence providing great energy-saving solutions, especially for facilities such as data centers where the load could fluctuate most of the time [79]. DVFS is most useful

in saving energy for servers that are usually over-provisioned to cater to maximum loads, but most of the time, they are lightly loaded. Balancing power consumption hence makes DVFS effective [78] in reducing costs, and promoting sustainability, which makes it vital in energy efficiency computer systems.

3) Server consolidation

During low traffic periods, data centers can allocate most workloads to fewer servers; hence, the number of active servers would be greatly reduced, thus most of the energy saving would be achieved [80]. This technique, also known as server consolidation or workload consolidation, applies dynamic transfer of the workload when the workload's level is low, the computer can shut down this extra server or put it into sleep mode.

A multi-cloud operational analysis from [81] revealed that dynamic server consolidation lowered the number of active physical servers by up to 45% at off-peak hours. This method resulted in average energy savings of 30–40% and increased hardware lifespan by lowering operational heat load. Table VIII shows the comparison of different scalability parameters.

Consequently, power that would otherwise be used to cool buildings, distribute power, and other uses on the servers is saved. The traffic to data centers varies from high to specific times and is significantly low at other times [81]. As opposed to all servers operating consistently, workload consolidation takes the processing to the most efficient servers possible during low utilization hours. Here, system virtualization technologies are useful in this process due to the enablement of multiple virtual machines by a few physical servers so that the workloads can be reshuffled without discontinuing services. Not only does this strategy help to cut energy costs, but the amount of wear and tear that reduced activity places on hard-working servers can also lead to longer life spans for expensive equipment that otherwise would be running non-stop [82]. Furthermore, it provides usefulness to a wider range of sustainability endeavors by reducing the impact of data centers on the environment while expanding the growth of offered services comparatively cheaper and less carbon-intensive.

TABLE VII. COMPARISON OF DIFFERENT ENERGY-EFFICIENT APPROACHES

Parameter	Traditional Data Centers	Energy-Efficient Cloud/Data Centers (AI, Green Tech)
Energy Usage Monitoring	Manual, reactive monitoring	Real-time, AI-driven predictive monitoring
Server Utilization	Inefficient, often over-provisioned	Optimized, with server consolidation and dynamic scaling
Power Management	Limited, basic power-saving techniques	Advanced, with dynamic voltage scaling and AI optimization
Renewable Energy Integration	Low, sporadic use of renewables	High, dynamic integration with solar/wind based on availability
Cooling Optimization	Basic, non-automated cooling strategies	AI-driven, adaptive cooling systems
Overall Efficiency	Lower, energy waste due to over-provisioning	Higher, minimal waste due to intelligent resource distribution

TABLE VIII. COMPARISON OF DIFFERENT SCALABILITY PARAMETERS

Parameter	Traditional Data Centers	Emerging Cloud/Data Centers (AI, SDN)
Horizontal Scalability	Limited, manual scaling of infrastructure	High, automated scaling based on real-time needs
Vertical Scalability	Limited by physical hardware capacity	Flexible, and supports both vertical and horizontal scaling with AI
Traffic Management Scalability	Requires manual configuration updates	Automated scaling and adjustments based on traffic with SDN
Cloud Expansion	Time-consuming, often manual	Fast, flexible cloud expansion with minimal manual intervention
Cost Efficiency of Scaling	Expensive, due to hardware needs and over-provisioning	Cost-effective with dynamic scaling, reducing over-provisioning

VI. EMERGING TECHNOLOGIES IN OPTIMIZATION

New technologies are the key drivers of the changes in optimization in many areas, especially in cloud computing, data centers, and networking. As networks and infrastructures develop an increasing degree of dynamics, more and more, the old methods of resource management and performance optimization do not suffice [83]. Today, these challenges are solvable by implementing ML, AI, quantum computing, and blockchain as innovative technologies. These technologies improve the opportunity to fine-tune all or any aspects, depending on resource usage to energy management, network paths, and security. In this vein, we propose that organizational improvement is contingent on these developments and offers improved efficiency, scalability, and sustainability [84]. Of all these Acoustics technologies [85], there is an extraordinary impact in routing and resource management, because cloud infrastructures and data centers demand adaptive intelligent systems for great volumes of traffic and dynamic workload [86]. These technologies are able to implement real-time decision-making, predictive analysis, and adaptive resource optimization to ensure optimal efficiency of the programs despite high load.

A. ML and AI for Routing and Resource Management

ML and AI have become the trending technologies used for routing and resource management of cloud computing and data centers. Given the scale and increasing complexity of these infrastructures, ML and AI provide leverage in better and more efficient handling of resources [72]. Prior approaches to routing and resource allocation use predefined procedures or allocations often based on rigid criteria, and these are inadequate for organizations' dynamic modern settings. While this is impossible for a human controller to do within the time span that is required, ML and AI can analyze enormous amounts of network data in real-time and make decisions regarding traffic organization that would reduce congestion and prevent the wastage of resources. In routing, AI applications or systems are capable of learning

through algorithms such as reinforcement learning of the best channel for relaying the data based on other parameters, such as congestion and latency, among others [38]. This dynamic adaptation allows the networks to quickly reroute traffic, therefore avoiding considerable delay when communicating between the servers. Furthermore, the AI can forecast new network-related problems in advance and dynamically distribute loads to avoid them. In the field of resource management, AI will analyze past patterns of a program's usage and dynamically allocate the CPU, memory, and storage required for that application accordingly [87]. These predictive models help cloud providers to escape over-provisioning, which consumes more energy and resources, and under-provisioning, which would cause system overload and poor performance. Resource management based on AI also increases efficiency in energy usage by bundling tasks during slow hours so that all servers that are not in use can be shut down to save energy [84]. This leads to enhanced functionality and resource utilization, cost reduction in the facility's operation, and hence a practical and durable physical system.

Although ML-based models work well, they have major challenges in real-world applications. These models required high-quality training data. Model generalization across different network environments is another issue that can cause performance degradation. In addition, latency-sensitive systems limit their use due to the cost of computation, training time, and complexity.

1) Reinforcement Learning (RL) for routing

The basic RL algorithms are now being employed to perform re-route strategies in a more enhanced and dynamic environment of the network. Unlike other routing techniques, which use fixed protocols or predetermined algorithms to guide the routing process, RL algorithms can learn in an online fashion from the network to which it is applied [88]. This is a learning-by-doing process whereby the RL agent (in this case routing algorithm) takes time to select the best possible routing decisions depending on the feedback it receives from the environment, for instance, congestion, delay, and packet losses. The most basic notion of RL is that it involves an agent that takes actions in an environment to maximize a cumulative reward [89].

With regard to routing, the RL agent decides on the direction of transferred data packets depending on the current status in the network, such as traffic intensity or bandwidth capacity. They recommend that after the action is performed, the agent gets feedback in terms of a payoff, which could be either a reduction in delay time or an increase in throughput [72]. If the routing decision optimized these parameters, then the reward is positive, endorsing that action. On the other hand, equating a negative reward to poor performance on the contention cost (e.g., the routing leads to more packet loss or congestion), the agent knows how to avoid such decisions next time it will be in the same scenario [90]. They incorporate the principles of RL [91], and therefore the algorithms are capable of acquiring near-optimal routes when interacting with their network environment in the long term. When the structure of the network changes or when the flow of traffic varies, the RL agent's decision-making policy changes, making it ideal for dynamic networks [92].

RL-based routing systems can predict traffic patterns to avoid congestion, reroute traffic in the case of network failure, and generally improve the overall performance of the network [93]. This adaptive learning capacity of the RL algorithm proves most useful when the network is in constant flux, something commonly witnessed in cloud computing architecture or large data center networks. Through repeated learning steps of the routing plan and the subsequent optimization of that plan, RL algorithms can dramatically increase the effectiveness of an existing network, lower latency, and, in general, improve the quality of the service offered to the users. However, this model has drawbacks, with slow convergence and instability during the learning phase. Prior to convergence, the use of exploration policies can lead to suboptimal routing decisions. Also, RL models are highly sensitive to how the reward function is set up and to changes in the environment. The cost of computation in continuous learning remains a significant challenge. Fig. 5 shows the reinforcement learning framework.

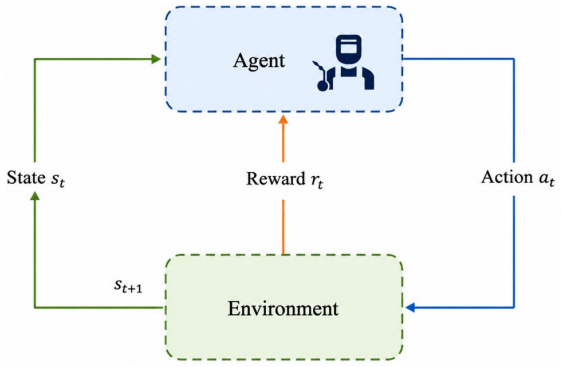

Fig. 5. Reinforcement learning framework: Agent-environment interaction.

2) Predictive resource allocation

ML models are very useful in predicting future demand due to resource allocation and thus can help in load balancing of the cloud and data center systems. Many traffic patterns, user behaviors, and usage of the resources in the past can be analyzed with the help of ML, and the forecast of the demand shortly will be highly accurate [94]. They assist in the real-time management of resources due to the fact that the system is equipped to handle busy periods and periods with fewer clients. It is a characteristic of workloads of cloud infrastructures to vary time by time within a day, and hence, static resource provisioning causes inefficiencies. When the demand is high and unexpected, more traffic flows to these systems, which causes bottlenecks [84, 95]. On the same note, if demand is below expected levels, this means that some of the resources employed to meet that demand are left idle and therefore lead to wastage of energy and resources and higher operating costs. Table IX shows the comparison of different latency management parameters.

TABLE IX. COMPARISON OF DIFFERENT LATENCY MANAGEMENT PARAMETERS

Parameter	Traditional Networks	SDN and AI-Driven Networks
Latency Sensitivity	Lower sensitivity to real-time changes in traffic	High sensitivity, with real-time traffic adjustments
Latency Control	Limited ability to control or reduce latency	Highly effective, with SDN enabling fast route adjustments
Congestion Handling	Reactive, often leads to congestion	Proactive, with AI predicting congestion and rerouting traffic
Priority Traffic Handling	Fixed priority settings, not adaptable	Dynamic prioritization of traffic based on real-time conditions
Latency for Real-Time Apps	Higher latency, less optimal for real-time apps	Low latency, ideal for applications like online gaming and AI workloads

Random forests and decision trees overcome these challenges since they take data patterns and learn the correlation between different factors controlling the use of resources. For example, supervised learning approaches can be trained to look for patterns associated with enhanced traffic, time physical events, or user actions. Once trained, the model can then make forecasts of what demand for the services will be in the future, so the system can proactively assign more processing power, storage, or network I/O as needed in advance to avoid performance degradation. Likewise, in low-demand conditions, capacity can be contracted, therefore leading to low utilization of energy and hence low cost [96]. They also improve load balancing because the load forecast outcomes of the ML models provide estimations of high traffic and ensure good dissemination of the incoming requests among the servers. Such an approach prevents any particular server from getting overloaded, hence enhancing the system utilization, cutting down on latency, and making the use of the system more convenient for the end user [97]. Overall, the implementation of more powerful and effective ML algorithms to manage our resources mainly makes the extraction of more solid and dependable cloud as well as data center solutions more economically sound.

B. *Edge Computing Integration*

The primary advantage of edge computing is that computation is shifted to the network boundary, including the end-user side, by localization in local servers or devices other than the cloud data centers. This shift minimizes latency and enhances the speed at which computations are undertaken, thus acting in favor of such applications as Internet of Thing (IoT), automotive cars, and augmented reality. Edge computing allows data to be processed nearer to its source, thereby cutting out unnecessary transmission of data at large distances, which spares bandwidth and lowers latency rates [98]. However, the combination of edge computing with traditional centralized cloud resources poses new routing and resource distribution problems. The proper task distribution between edge and cloud is a constantly changing process, and thus leadership demands intelligent and flexible solutions to implement [99]. There is the problem of data placement, or, in other words, deciding at which tier of the network a particular data processing task should be accomplished.

This decision-making process takes into consideration factors like resources available at the edge, state of the network, and task characteristics like lateness and data size. Similarly, network optimization is more challenging in hybrid systems where data is always in transit across the edge device and cloud server. They also need to be able to respond to existing conditions and status, such as congestion or traffic loads in servers, to facilitate the flow of data [100]. New strategies for achieving these aims include AI-based routing, as well as the active use of methods of predictive analytics for facilitating the system's reaction to changes in conditions and for optimizing resource consumption and data storage organization.

C. *Network Function Virtualization (NFV)*

NFV is the ability to virtualize selected network service components, for example, firewalls, load balancers, and routers, by carefully controlling the underlying physical network, sometimes termed as a substrate network [54, 101]. Concisely, these network functions were implemented in traditional hardware appliances that were expensive, inflexible, and unsuitable for expansion. NFV, if it relocates these services from hardware and operates them as VMs or containers on simple servers, revolutionizes the manner of networks and makes them more manageable and less costly [102].

Making it possible to control and manage the resources dynamically, NFV provides the necessary leverage for enabling data centers to accommodate dynamic and variable workloads and traffic loads. For example, during high traffic, it becomes easier and very flexible to add more instances of the network service, such as a firewall or load balancer [103].

On the other hand, when it is off-peak, then such incidents, in some cases, can be oversimplified, thus saving on, for instance, energy costs. This flexibility is essential to resource utilization as it enables data centers to provide exactly the amount of computation and interconnectivity as is required in real-time.

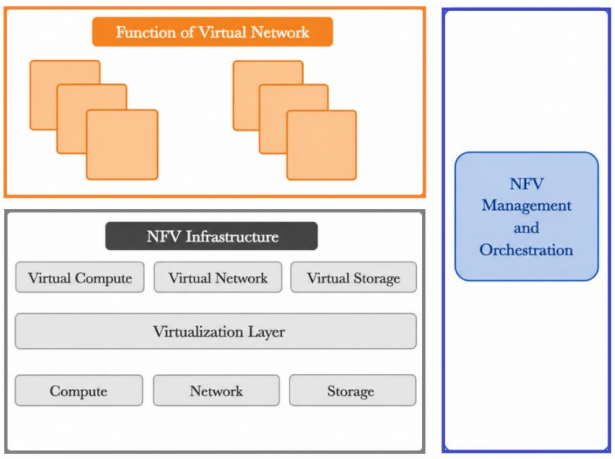

Fig. 6. NFV architecture: A layered approach to virtualization.

Correspondingly, the applications of NFV for the deployment of modification, and management of VNFs without necessarily requiring new hardware integration speed up innovation and cut down the duration necessary to commercialize a new service. This also assists network managers in making network changes more effectively, since it provides operators with dynamic changes to their network configurations, which has led to improved cost control, timely service delivery, and the ability to add or remove networks based on the demands of the market [104]. This makes NFV a mandatory solution for today's and future data centers, helping them to adapt easily to traffic fluctuation and optimize resource usage constantly. Fig. 6 depicts the NFV architecture.

Table X evaluates various resource management and scheduling techniques by examining their algorithm type, objective functions, simulation or platform environment, workload context, and reported improvements. This table shows how different methods deal with intelligent scheduling, load balancing, resource use, and energy efficiency in modern cloud and data center environments.

TABLE X. COMPARATIVE ANALYSIS OF RESOURCE MANAGEMENT AND SCHEDULING APPROACHES

Ref.	Algorithm category	Technique/ method	Objective function(s)	Simulation/ platform	Dataset/workload	Reported improvement
[55]	ML approach	VM allocation	Energy, resource utilization	Cloud simulation environment	Variable cloud workloads	Improved autoscaling efficiency
[57]	Live migration framework	Live migration management	Load balancing	Cloud computing framework	VM migration scenarios	Better service continuity
[77]	DVFS	DVFS-based scheduling	Energy efficiency	Cloud computing model	Dynamic load conditions	Reduced power usage
[81]	Consolidation	Server consolidation techniques	Energy	Cloud data center studies	Variable workloads	Reduced active servers
[97]	ML-centric scheduling	ML-based resource management	Resource efficiency	Cloud resource management environment	Historical workload patterns	Better proactive allocation
[105]	Hybrid metaheuristic	QHRMOF	Energy, load balancing	Cloud scheduling framework	Task scheduling workload	Improved energy-aware scheduling

VII. CHALLENGES, OPPORTUNITIES, AND FUTURE DIRECTIONS

Despite the significant advancements made in routing and resource allocation in the cloud as well as data center networks, there are still several barriers. Due to the increasingly complicated structure of cloud infrastructures and the dispersion of data centers, it becomes hard to mitigate such issues. However, they also reveal a possibility of huge potential for growth and development in this constantly developing area. This section briefly discusses some of the main problems, reviews the emerging areas of novelty, and identifies potential future research directions based on the analysis of the state-of-the-art of the published literature. A current problem is, for example, the complexity and scale of current cloud and data center networks. Due to the increase in pervious accessibility of appliances and data traffic, conventional methods of routing as well as management of resources have been sacrificed. It is much easier said than done to sustain high performance on distributed data centers and efficient utilization of resources at the same time; it would need some very elaborate algorithms as well as live analysis [106]. Thus, to provide secure and efficient network capabilities, new solutions should be adaptable and efficient in scale. The fourth challenge is energy efficiency because data centers are among the biggest energy consumers in the world. Despite advances in techniques such as dynamic resource allocation and workload consolidation, innovation advances are needed. Areas including energy demand forecasting and AI-integrated renewable energy resources for cloud computing must be regarded as disruptive research initiatives since they can cut the carbon footprint of computing infrastructures drastically. However, it is not a secret that all such lucrative business ideas are associated with constant security and privacy issues [107]. Since data centers manage ever more significant amounts of very sensitive information, the balancing of resource needs with adequate security measures is a challenging task. Security-aware routing as well as privacy-preserving resource management are now integrated into optimization frameworks, which are still emerging. In addition, researchers can therefore consider discovering AI self-organizing networks where, other than predicting traffic patterns and resource requirements, machine learning algorithms control networks with minimum interference of humans [108]. Furthermore, edge computing integration means that new possibilities in routing and resource management will arise, especially for designing and implementing the mixed architecture that combines the edge and cloud resources.

A. Challenges

1) Scalability

The ever-growing growth of data centers makes network management a huge task in large-scale networks. The current routing and resource allocation approaches, which rely on preconceptions about organization structures, functional silos, and a slower rate of change in environmental variables, break down when the scale involves millions of servers and other connected equipment. At the same time, routing decisions, load balancing, and handling overall traffic become tremendously difficult, especially with the increased number of interconnected nodes [109]. In such large-scale system environments, every small amendment in routing and resource management can produce dramatic

consequences, summarized as system performance erosion, traffic jams, reduced resource utilization, and over-provisioning.

More specifically, consider the novel characteristics of today's challenging landscape, such as cloud-native architectures, containerized workloads, and microservices applications. These modern architectures are leaf architectures of high distribution and constantly evolving, comprising thousands of small, often independently deployable, and highly interactive executable components that require dynamic scaling. These are not environments that can be serviced well by stale or static mechanisms, as most traditional static methods are slow to adapt to the ever-shifting workloads and traffic [110]. Resources in these architectures are managed to optimize the placement of containers and microservices in the network to reduce latency, fully distribute traffic, and prevent resource contention. To overcome these challenges, therefore, more flexible and dynamic approaches are inevitable. Real-time traffic monitoring, demand forecasting, and route and resource distribution are emerging as potential core application areas in big networks for which AI-driven automation is starting to show potential. These approaches allow the data centers to scale up with more proficiency, also helping them keep up a high standard of service about performance, regardless of the growth in size and myriad complexities [111]. Through such intelligent and adaptive solutions, data centers are able to manage their way around bottlenecks, resource usage, and fundamental scalability that are needed in supporting today's cloud-native systems.

2) Latency and QoS

Computing with low latency is a crucial requirement for applications that face strict temporal constraints, such as multimedia and video-on-demand, online gaming, and especially financial trading. Some of the persistent issues include ensuring that the data packets arrive at a destination in as short a time as possible as is possible and meet certain QoS parameters [112, 113]. Static and slow-adaptive routing protocols fail to respond to changes in real-time traffic adequately. Therefore, such protocols may cause latency peaks and traffic jams or decrease the QoS, primarily in the case of mission-critical applications where predictability is a key to success.

To overcome these problems, routing and resource management strategies should not only reduce delay by precisely allocating resources where they are needed most but should also preferentially serve traffic with certain quality of service characteristics over traffic with other qualities. For example, traffic originating from a real-time financial transaction may be first in line, while traffic originating from a large file download may be lower down the priority chain [114]. This needs routing algorithms to be adaptive and intelligent that obtain the state of the network and make decisions on the proper route for the packet in a real-time manner.

Two of the main issues related to the consistencies of QoS by using advanced resource orchestration across highly fluctuating workload applications are the dispersed data centers and emerging edge computing environments. Edge computing is intended to be a data processing paradigm that occurs closer to the end user, thus reducing latency; however, performing computations close to the end user requires managing many nodes. Managing the routing of traffic across the distributed environments and guaranteeing that latency-sensitive applications get the bandwidth and processing required implies the need to achieve a certain level of intelligence, traffic flow metrics, AI-based optimization of the network in real-time, as well as prediction metrics for the nature of that traffic for evolving network conditions [115]. Most importantly, satisfying low-latency requirements and preserving QoS means embracing more sophisticated, dynamic methods of managing resources and their utilization for real-time traffic to make sure application delivery is never impacted.

3) Energy efficiency

The problem of high energy demands experienced by mass data centers can be attributed to environmental and financial issues. As improvements in the microprocessor design of the hardware and the resource management algorithms have been made, the sheer size of the cloud infrastructure still presents data centers as large consumers of power. These facilities, with their continual use of cooling systems, energy to run the equipment, and twenty-four-hour operations, are great consumers of energy all around the world [116]. Techniques such as server virtualization and DVFS are rather efficient in energy consumption saving because of better server utilization and adaptive power management in the function of workloads.

Nevertheless, the above-mentioned methods are optimal only up to a point, where significant additional energy and performance reductions present a significant challenge. To this end, distributed intelligent systems are required for efficient workload management to avoid high utilization of power during low power demands [117]. One of the approaches includes workload allocation, which shields the energy-intensive servers or data centers that are supplied by renewable energy sources. This proactive management assists in minimizing wasted energy in the servers so that only the necessary ones are switched on; the rest are either shut down or put on energy-saving modes.

If we fast forward to the future where edge computing, fog networks, and other related architectures are becoming more of a reality than a futuristic dream, then the management of energy usage becomes a much harder proposition. Taking place closer to users and being more distributed, edge nodes need to accomplish computational tasks faster while consuming as little energy as possible. Compared to centralized data centers, these nodes cannot apply large-scale cooling systems or energy-efficient nodes; thus, one has to become more inventive. Improving the energy efficiency at both data center and edge levels is vital for future cloud and edge infrastructure, while keeping latency down and performance high [118]. Any further improvement in energy efficiency is likely to be realized with the help of machine learning-based energy conservation, prognostics, and renewables if the conflicting goals between performance, cost, and sustainability are to be met.

4) Security and privacy

Today, security forms one of the most important considerations, given the growing use of SDN and NFV in clouds and data center networks. Although these technologies offer enhanced flexibility, scalability, and automation through centralizing control and NFV, they have their inherent risks [119]. The centralized control plane in SDN, which controls and coordinates the entire network, becomes a tempting and relatively easy-to-exploit point of link. Denial of this control plane may result in the availability, integrity, and confidentiality of data affecting large sections of the network.

Furthermore, with the reliance on ML to make decisions on routing and resource allocation, networks are vulnerable to new threat models, including adversarial actions. In such an instance, the criminals can feed the ML models with wrong data or inputs and make wrong decisions. It can cause misrouting or resource wasting in the case of communication, or may even lead to network breakdowns [120]. Protection against such attacks will formally involve designing better defenses in the machine learning models; moreover, these attacks should be watched in real-time to counter them efficiently.

Another major issue is protecting data, especially in the shared cloud environment where many clients/users/Organizations share space on the physical hardware. Such surroundings require that tenants be kept apart to eliminate the possibility of data breaches by some of the other renters. However, it is often challenging to sustain this seclusion while at the same time guaranteeing the proper distribution of resources available in the organization [121]. Virtualization security is especially significant in attempting to provide superior resource utilization concurrent with data confidentiality; other measures include encryption and secure multi-party computation.

Meeting security and privacy needs while optimizing cloud and data center networks and efficiently supporting SDN, NFV, and ML in the current fast-growing environment calls for a layered solution approach. This entails using IDS, a secure SDN controller, encryption methods, and advanced machine learning models to provide security to the network without losing much on performance [122]. As these technologies progress in the future, it will be increasingly important to have solutions to address the security issues with these technologies to provide secure and reliable cloud and data center infrastructures.

5) Interoperability

With the growth of the cloud ecosystem, the need to integrate many cloud providers, network topologies, and service models is creating a crucial need. A cut above is the hybrid cloud, where private and public clouds are both used, and the multi-cloud, in which many cloud providers are active simultaneously [123]. This, linked with the uplifting trend in edge computing, which moves the processing of data to the end user level, poses substantial problems regarding consistent and economic cooperation across diverse infrastructures.

It is crucial in such environments to be able to allow workloads to connect and transition between these platforms, enforcing data uniformity and also giving a unified user interface. However, to maintain population health, several barriers have to be surmounted first. It also varies across cloud providers, and each of them has its own traditional vendor technologies and standards to make integration without another layer of complexity [124]. For instance, APIs, storage formats, and network protocols may be different across cloud services, which makes the movement of data and apps difficult.

Furthermore, routing and resource optimization are issues that are far more complex when operating in these diverse environments. Every cloud provider and infrastructure has unique resource provisioning, traffic flow rules, and distinctive optimization patterns that affect the overall optimization of the ecosystem. Scheduling workloads across a multi-cloud or hybrid cloud environment, optimally and with as little latency as possible, can be quite challenging [125]. That is why it calls for advanced orchestration instruments that can command resources in various settings and turn to real-time data to make rational choices of how to navigate traffic and manage resources.

Additionally, security and data privacy are less straightforward in these complex systems because of the migration of data and applications across multiple clouds and networks. The effective integration of the platforms will thus need to be defined with references to standards that can help define a common platform on which interoperability without compromise on security can be achieved. Overcoming these challenges requires investment in the creation of open standards, the adoption of technologies such as Kubernetes, and the use of cross-platform resource orchestration, security, and traffic [111]. That way, they can get the level of interoperability that is needed while at the same time ensuring that they are able to properly route traffic and manage resources when dealing with an ever-expanding cloud environment that is becoming more diverse.

Many optimization issues in cloud and DSN are multi-objective in nature. A generic formulation can be written as shown in Eq. (1), which is subjected to system constraints on bandwidth, compute resources, QoS, and service requirements.

$$\min \left\{ E(x), L(x), C(x) \right\} \tag{1}$$

Since these objectives often conflict, many studies seek Pareto-optimal solutions, where improving one objective worsens at least one other. A common aggregation strategy is the weighted-sum model as shown in Eq. (2).

$$\min J(x) = w_1 E(x) + w_2 L(x) + w_3 C(x) \tag{2}$$

where the weights reflect application priorities.

B. Opportunities

While challenges persist, they also open the door for significant opportunities in cloud and data center optimization. As the demand for cloud services continues to rise, the industry is ripe for innovation in several key areas.

1) AI and ML

Routing and resource management in cloud networks have benefited from AI and ML, adding new layers of capabilities for traffic flow optimization, resource pre-assignment, and the general performance of the network. Using AI and ML in cloud networks, Cloud networks can shift from being managed reactively and instead have the ability to predict and manage the network from a more proactive standpoint as a result of constant AI and ML analysis of network trends. The most revolutionary use of AI and ML in cloud networks is predictive analysis. It allows networks to predict future traffic patterns, react to demand surges, and thus minimize performance issues before they occur using Machine learning models trained on historical data [126]. I agree with this because, with an understanding of what is to come, networks can reroute traffic intelligently so that paths can be changed proactively and networks can achieve optimal traffic flow, therefore reducing latency. For instance, the reinforcement learning algorithms can learn from the network performance and change the routing policies resulting from the analysis of traffic patterns and availability of resources in real time.

Other than traffic control, predictive analytics is widely applicable throughout the area and resource allocation. AI systems-based algorithms operating in the background can estimate the total computational needs of an application or service based on trending analysis of workload and usage patterns. This predictive capacity results in networks being able to adjust resources up or down in computation as required to meet SLAs but with no over-provisioning and a drain on power. AI can also determine the load distribution for multi-cloud or hybrid-cloud settings, to run the load on the nearest, most suitable data center or edge nodes [127]. Another important advantage is the facility of performing predictive maintenance by means of AI. Thanks to the analysis of big historical and real-time data, AI can predict patterns indicating potential network problems—traffic increase or, for instance, abnormal server activity. This capability allows for anticipating conspicuous failures that lead to notifications of the administrators or could prompt automatic control measures and thus minimize time lost, hence offering consistent service delivery.

Moreover, ML models perform optimization very effectively and are capable to learn and updating the changes in the network environment. Static models have been antithetical to dynamic traffic and resource variability that hampers operational effectiveness. On the other hand, with ML algorithms, it is possible to process real-time, rather large-scale data on traffic loads, latencies, and server statuses, and make corrections to routing and load balancing dynamically. When a server is loaded or some path is busy, the ML model can send all traffic to other, less-loaded resources to avoid latency. While the inclusion of AI and ML solutions in the cloud networks increases the performance by decreasing the latency and maintaining high QoS, the overall user experience is improved as well [128]. In this way, with the help of predictive maintenance, dynamic resource allocation, and intelligent traffic management, the agility, resilience, and cost efficiency of the cloud infrastructures may be improved at the same time that they may be oriented, with an even higher level of success, to real-time needs.

2) Edge and fog computing

Edge and fog computing are innovative approaches to the cloud and data center networks that have potential increases in effectiveness, speed, and reduced costs in contrast to conventional centralized cloud systems. These technologies decentralize or distribute data processing closer to the edges of the network or intermediate nodes, hence minimizing long-haul trips to centralized public cloud servers. This proximity reduces latency, which is important in latency-sensitive applications such as self-driving cars, smart cities, smart devices as well as augmented realities. Since most data is processed at the edge or at neighboring fog nodes, edge computing entails lower latency and improved responsiveness [129]. For instance, in Self-driving cars, edge computing provides real-time computation, keeping the control of vehicle processing different information received from the sensory systems close to the system so that immediate decisions are made without recourse to distant servers. Additionally, edge computing also has a significant role in smart cities where the huge amount of data collected from the sensors used in that application, including intelligent traffic management systems or public security, can be analyzed in real-time to prevent congestion and enhance the functionality of an urban city.

Besides, edge and fog computing relieve centralized cloud data centers' loads by distributing workloads among numerous small nodes. This decentralized approach not only equalizes load but also reduces the use of network bandwidth, operating expenses, and energy consumption. Because it provides localized data processing for certain uses or areas, it can support new business models that promote advancement in industries like retail, logistics, healthcare, and manufacturing, among others. It is, therefore, necessary to implement dynamic resource management in efficient edge and fog systems with versatile architectural models that will be capable of adapting to workload and network changes, as well as the priority of different services [130]. Such frameworks control and achieve the highest levels of system performance and QoS without stressing a single node, employing edge-based load balancing and resource orchestration. They also improve QoS because they reduce tensions, deny services from centralized networks, and guarantee the delivery of services at optimum usage times.

3) Energy-saving technologies

Energy efficiency and sustainability are important factors for cloud and data center providers since the organization requires a lower impact on the environment

without compromising performance. As more organizations turn to cloud technologies, and the size of data centers continues to expand, power intake and, consequently, emissions have spiked. This has led to these people finding ways of improving the outcomes, where results include power usage, designing manufacturing hardware that uses as little power as possible, finding ways to efficiently cool these systems, and designing algorithms that can change dynamically to use fair amounts of power while meeting performance needs. One of the biggest unutilized chances for sustainability in cloud data centers is the use of green energy, such as solar and wind energy [131]. Through these renewable sources, data centers can decrease the use of fossil fuels, thereby making a tremendous improvement in cutting their carbon footprint. However, the main issue resides in the time variation of renewable energy source availability. To this, systems with improved methods for the management of dynamic energy are being set to ensure that they take into consideration the dynamism of the availability of renewable resources and the demands of the workloads. These systems can proactively schedule workload distributions towards data centers or servers that are backed up by renewable energy during distinct power production times, thereby decreasing the energy expenses and the environmental impact.

Furthermore, it illustrates that AI would be a feasible solution for enhancing the energy performance of data centers. Data centers can use historical as well as real-time data samples in workload through the use of ML algorithms to determine the amount of energy that is likely to be consumed at different points in time. For instance, AI can detect instances where the usage is low and then rearrange non-essential processes to lower-Asian resources or servers as a way of saving energy within an organization that uses significant power on its intensive structures [132]. In addition, AI can control the cooling systems since it can predict thermal loads and apply cooling efforts according to the result, hence reducing the amount of cooling used unnecessarily. When renewable energy and AI coordination are implemented, data centers can regulate their utilization rate in regard to energy and the performance necessary for the task. It also provides a better environmental design and solution for the cloud operators to meet both the environmental and operational requirements, for the development of green cloud computing systems.

4) 5G and beyond

Cloud and data center networks' high-bandwidth, low-latency applications are expected to obtain new support from the 5G networks as their deployment is expected to transform the cloud and data center networks. Assuming that the key 5G service enablers that currently include Ultra-Reliable Low-Latency Communication (URLLC) and massive Machine-Type Communication (mMTC) will create a bulging demand for innovative data handling mechanisms at the (Cloud Service Provider) CSP level. The transition to 5G offers new possibilities but it also needs new approaches to routing and resource management to make the networks ready to carry the traffic load of the future and to provide service level quality which is expected from networks. 5G networks will cover significant service areas such as autonomous transport, smart environments, industrial systems, and real-time healthcare solutions, which require truly reliable and low-latency ultra-high data rate solutions [133]. These applications demand amazing reliability, performance, and responsiveness of the underlying network. Therefore, CSPs will need to look for new ways of orchestrating traffic across 5G networks, edge devices, and the core data center. The traditional centralized cloud model could be too slow to support millisecond latency requirements that are expected to be supported by 5G systems. Hence, edge computing will be another critical enabler that aims at providing the needed computing resources nearer to the data source to respond quickly and reduce the load on centralized data centers. Challenges arising from data traffic exchange between edge nodes and massive central datacenters, and the dynamic requirements of a spectrum of 5G-connected devices and applications, require efficient and intelligent AI-based network management solutions. They will require the flexibility to change routing tables, allocate resources optimally, and predict congestion on the network [134]. With the ongoing advancement of 5G to 6G and subsequent generational networks, the need for such intelligent and autonomous network implementation solutions is set to rise. The continued progressive growth and integration of smart devices and data traffic will require next-generation traffic analysis, proactive and predictive routing, and resource provisioning for the networks to provide higher performance and scalable characteristics for supply. Table XI shows the the comparison of optimization methods.

TABLE XI. CRITICAL COMPARISON OF REVIEWED OPTIMIZATION METHODS

Approach Category	Scalability	Computational Complexity	Stability	Deployment Challenges
Traditional heuristics	Moderate	Low	Stable	Limited adaptability
SDN-based	High	Moderate	Fast operational response	Controller bottlenecks
ML-based	Potentially high	Moderate to high	Model-dependent	Data quality
DRL-based	High in dynamic settings	High	May be slow	Training cost, reproducibility
Hybrid methods	Strong optimization capability	High	Iterative	Tuning complexity,

The reviewed techniques show considerable potential for making routing and resource management better in cloud and data center networks. However, their real-world effectiveness depends on a few key factors. Heuristic and

traditional methods are easy to use and take up little space, but they often have difficulties scaling and adapting in environments that change quickly. SDN, ML, DRL, and hybrid optimization methods, on the other hand, are more adaptable and efficient; they also require extra computation, take longer to train, and may have problems with convergence and stability. In the real world, deploying them is even harder because of controller bottlenecks, data needs, interoperability issues, and integration problems in heterogeneous cloud and multi-cloud environments.

C. Future Directions

The future of routing and resource management in cloud and data center networks will be shaped by ongoing research in several critical areas. The following future directions are likely to dominate the field.

1) AI-powered autonomous networks

Autonomous networking is also predicted to be a major trend in future networking, as self-organizing networks are fully controlled and operated by artificial intelligence. These networks will be able to work with self-optimization, self-healing, as well as self-scaling features without the need for human interference. Automation will gradually replace many human activities, which, when it comes to the operation of networks, will increase the efficiency of the networks, decrease the costs of running the networks, as well as reduce the role of errors in the networks, thus promoting the creation of more sustainable networks. One of the primary elements in these autonomous networks will include a self-learning routing protocol. These algorithms will grow smarter in the future and will be capable of recognizing real-time optimal routing in terms of traffic patterns, network conditions, and usage. Through steady learning, the AI systems will anticipate traffic congestion before it happens, and the networks will necessarily change the routing plans and maintain the precise flow. Such a level of intelligence will ensure that the desired changes in the traffic patterns, such as during peak times as well as incidences of network failure, will be addressed, hence enhancing interaction and reducing latency. A further essential characteristic of fully Autonomous networks is smart resource management. It will also be possible to use predictive analytics and get the best percentage estimates as AI-applied systems work out how best to allocate resources based on changing demand. For example, the network can increase computational capacity as well as the bandwidth during a certain period of time of high utilization to maximize its usability, while at other low-demand periods can decrease computation and bandwidth utilization without wastage.

This dynamic scaling feature will not only increase the efficiency and interactivity but at the same time cut down substantial energy-using costs and at the same time avoid resource overuse expenses. Self-learning routing algorithms together with automated resource provisioning mechanisms will enable the future, autonomous networks necessary for the management of increasingly complex tasks. These networks will redefine how infrastructures are run, thus shifting the business and service providers' concentration from manual Network Operations to one that will be progressive, measurable, and affordable. Fig. 7 shows the evolution of networks.

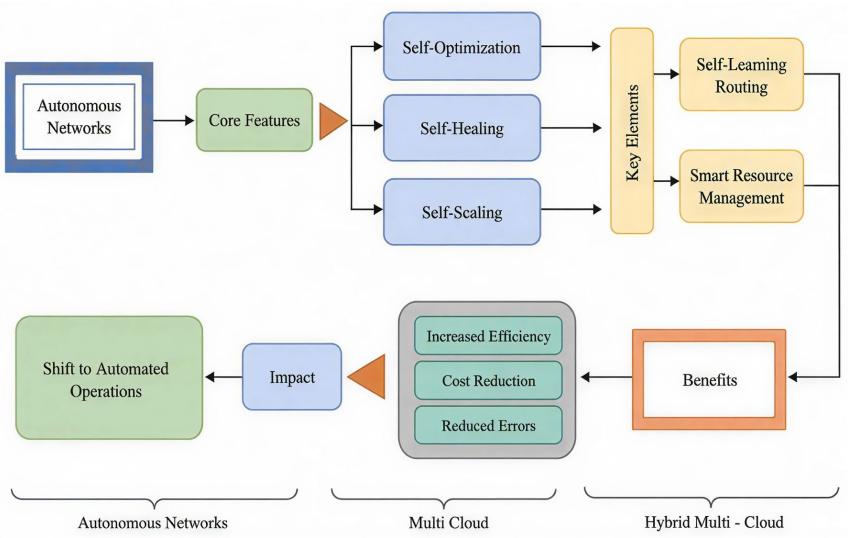

Fig. 7. The evolution of networks: From autonomous to self-managed.

2) Hybrid cloud and multi-cloud optimization

Given the fact that enterprises keep on extending their cloud adoptions to hybrid and multi-cloud topologies, future studies are likely to concentrate not only on correcting the interoperability limitations between different clouds but also on bringing operational efficiency to resource provisioning across multiple clouds. In such architectures, enterprises employ one or more public, private, and edge clouds, possibly from different vendors. On the same note, this approach has its set of pros as they include flexibility, scalability, and cost-optimization advantages, but also has unpredictability in terms of resource management in different cloud platforms, and expected communication modes. The first domain, which will require further research, is improving inter-cloud load

balancing. To optimize workload in a multi-cloud environment, the workload needs to be dynamically ported across different clouds by taking into account important criteria such as latencies, cost, and resource availability. This implies that sophisticated load-balancing needs to be put in place for intelligently directing traffic flow between the clouds. What has to be avoided is a situation in which one or more clouds are overloaded, while at the same time, all possible resources should be utilized to achieve optimality.

Another crucial factor is cross-cloud orchestration. Coordinating resources in heterogeneous environments for workload coherence in a tool that can fully and automatically handle the deployment of applications across multiple clouds. This orchestration should include multiple architectures, APIs, and service delivery models used by various cloud vendors. New work in this area will focus on enhancing orchestration systems so that different clouds can effectively coordinate, so that applications can be run as needed while being located in different clouds. Network slicing is also being envisioned as a critical technology to support optimizations in the utilization of resources across hybrid and multi-cloud networks. This technique involves generating application-specific virtual and isolated network slices on the physical shared substrate. Through network slicing, enterprises can allocate network resources for the different cloud settings and applications with precision to the workload without wastage or oversupply.

With more enterprises leaning towards hybrid and multi-cloud environments, the area of study consisting of interoperability, load balancing, orchestration, and network slicing will prove to be valuable to advance the enhancement of resource optimization on diverse cloud solutions for enterprises. Fig. 8 shows the hybrid multi-cloud.

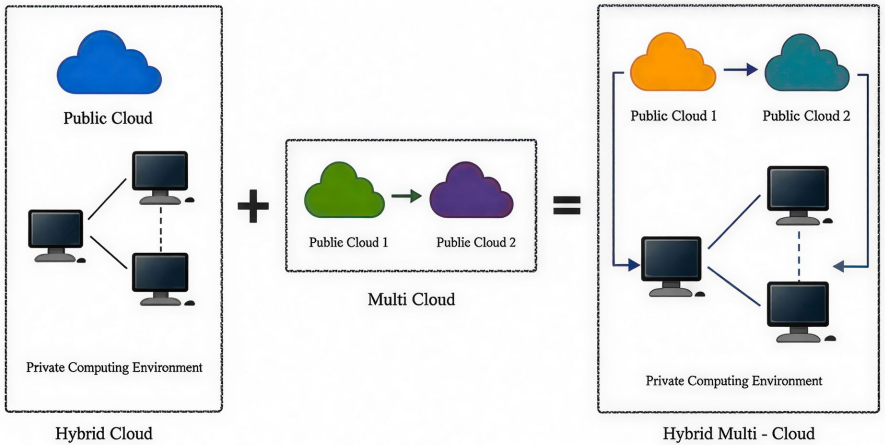

Fig. 8. Hybrid multi-cloud: Merging public and private clouds.

3) Blockchain for secure resource management

Blockchain has the potential to increase the reliability of cloud and data center environments and become increasingly important as enterprises transition to the network-blended hybrid multi-cloud environment. Given the consolidation and immutability that blockchain offers, it can support new paradigms for managing resources in a secure, reliable, and transparent manner, which will help enterprises build trust in one another and validate actions while having instant recourse to a clear record of their actions and the use of their resources across different cloud environments. Cloud networks are one of the main areas of application of blockchain technology, and one of the most important ones is the decentralized trust level. In conventional environments, trust is created for all consumers and cloud providers through centralized mechanisms, which are at greater risk of being hacked, fraudulent, or mismanaged. Blockchain solutions allow a decentralized system of sharing data in which trust is established through transparent data sharing and receipt of corresponding verified data. Blockchain could also enable safe communication between multiple cloud services and consumers in a hybrid or multiple-service environment by registering and auditing every exchange through a distributed database. Instead of an opaque structure for making decisions and keeping track of them, this ledger would enable clear, transparent documentation of action taken in each case.

Another revolutionary element of the blockchain concept is smart contracts, which have the ability to revolutionarily transform cloud networks. Smart contracts on the blockchain are self-simple-minting digital contracts whereby the terms of the agreement are encoded into the formulation of the contract. They can also be employed to develop policies, resource provisioning, and the overall bill for the management of the cloud. For instance, smart contracts could delegate work, such as storage or computing resources for storage, access, or computing based on set parameters regarding customer demand or payment confirmation, while guaranteeing fair customer billing in fair measure only for services metered. Through automation, Blockchain minimizes the possibility of disputes and inaccuracies, besides doing away with the middlemen, ensuring efficiency and economization.

However, it is noteworthy that through the use of blockchain technology, data integrity in cloud networks can be improved. Because data recorded in the blockchain is uncontrollable and unchangeable, blockchain

technology can be used to monitor and report the usage of resources across the cloud structures to guarantee protected data integrity. They can be very useful in compliance-heavy environments, particularly given this level of auditing transparency. Fig. 9 shows the blockchain powered could automation.

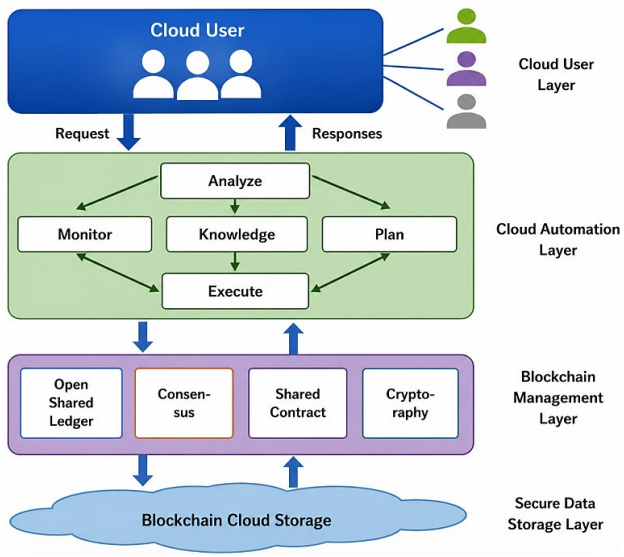

Fig. 9. Blockchain-powered cloud automation: A decentralized approach to secure resource management.

4) Quantum computing and networking

Machine learning and quantum computing are relatively new technologies in cloud computing that have the potential to revolutionize data centers through the provision of vast and distinctly superior computational speeds that are impossible with normal computing systems. This quantum advantage could let the cloud providers solve some of the intricate problems in domains including cryptography, molecular modeling, and intensive information interpretation with great velocity and efficiency. However, with the development of quantum networking technologies, cloud infrastructures are going to face more routing and resource management problems, which require new approaches and methods to manage quantum data. Traditional data is the basis of classical data, and on the other hand, quantum data in some way has a distinct nature of its information storage in qubits, which is known as quantum bits. Qubits can be in many states numerically at the same time because of quantum superposition and entanglement, which makes quantum computers solve large problems concurrently. Quantum data, therefore, is very susceptible to noise and other interferences in the environment, hence the difficulty in transmitting quantum data over networks. Future quantum networks will need new routing algorithms that reflect the qubit's susceptibility to noise sources, guaranteeing that quantum information can be transmitted as effectively and with as little loss as possible.

New approaches to routing will have to be developed, and resource management will have to be adapted to meet the demands of quantum computing workloads in future quantum data centers. The original cloud resources management methodologies, which were developed for the classical processors and the data streams, will be inadequate to control the quantum resources quantum processors, and quantum memory, for instance. There will be a need to incorporate quantum-conscious scheduling to request allocation to permit quantum workload contention across quantum Meadow as well as classical computing, alongside influencing supremacy and decoherence losses of qubits.

VIII. Conclusion

Optimizing routing and resource management in cloud and data center networks is vital for ensuring the performance, scalability, and sustainability of modern IT infrastructure. As cloud services continue to grow in demand and complexity, the limitations of traditional routing and resource allocation methods have become increasingly apparent. Technologies such as SDN, NFV, and ML have emerged as transformative solutions, enabling more dynamic, flexible, and efficient network management. These advancements allow data center operators to make real-time adjustments to network traffic, optimize resource utilization, and improve energy efficiency, addressing some of the most pressing challenges faced by cloud infrastructure today. However, despite these advances, significant challenges remain. Scaling these technologies to accommodate ever-expanding networks, ensuring low latency for time-sensitive applications, managing the enormous energy demands of large data centers, and protecting network infrastructures from security threats are critical issues that need further attention. Additionally, the growing complexity of cloud environments, particularly in hybrid and multi-cloud deployments, necessitates more sophisticated solutions for resource management and interoperability. At the same time, the rapid evolution of technologies such as artificial intelligence, edge computing, 5G, and blockchain presents exciting new opportunities for optimizing cloud and data center networks. AI-driven automation can enable self-optimizing networks capable of dynamically adapting to changing conditions, while edge computing and 5G can reduce latency by bringing computation closer to the end user. Blockchain technology offers the potential to enhance security and transparency in decentralized cloud infrastructures, providing new ways to manage and allocate resources securely. In light of these advancements and challenges, future research must focus on developing scalable, secure, and energy-efficient solutions that leverage these emerging technologies. The integration of AI and machine learning into network management systems, the widespread adoption of edge computing and 5G, and the exploration of blockchain for secure resource management all offer promising avenues for further investigation. By addressing these challenges and seizing these opportunities, we can create the next generation of cloud and data center networks that are not only more efficient and adaptable but also sustainable and secure, meeting the growing demands of the digital world.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

FUNDING

The author extends his appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2023/01/26209).

ACKNOWLEDGMENT

Thanks to Prince Sattam bin Abdulaziz University for supporting this research work.

REFERENCES

[1] M. N. Pathan *et al*., "Priority based energy and load aware routing algorithms for SDN enabled data center network," *Comput. Netw.*, vol. 240, 110166, 2024.

[2] T. O. Agboola, F. C. T. Mezue, S. B. Adebayo, J. Adegede, and O. C. Oyeniran, "Technical challenges and solutions to TCP in data center," *Int J Inf Technol Comput Eng*, pp. 36–46, 2024.

[3] Y. Feng, Y. Qi, H. Li, X. Wang, and J. Tian, "Leveraging federated learning and edge computing for recommendation systems within cloud computing networks," in *Proc. 3rd Int. Symp. Comput. Appl. Inf. Syst. (ISCAIS 2024)*, 2024, vol. 13210, pp. 279–287.

[4] P. Borra, "An overview of cloud computing and leading cloud service providers," *Int. J. Comput. Eng. Technol. (IJCET)*, vol. 15, no. 3, pp. 122–133, 2024.

[5] Y. Liu, T. Yu, Q. Meng, and Q. Liu, "Flow optimization strategies in data center networks: A survey," *J. Netw. Comput. Appl.*, vol. 226, 103883, 2024.

[6] Y. Zhang *et al*., "FLAIR: A fast and low-redundancy failure recovery framework for inter data center network," *IEEE Trans. Cloud Comput.*, vol. 12, no. 2, pp. 737–749 2024.

[7] A. Sharma, "Advancements in routing protocols: Analyzing emerging technologies and their practical applications in modern networks," *Academia*, 2024.

[8] T. Narcisse, S. Etienne, B. K. Trinité, A. Olivier, and K. Adama, "An intelligent load balancing strategy to improve performance and QoS in SD-DCN (software defined-data center network)," *Far East J. Appl. Math.*, vol. 117, no. 2, pp. 149–167, 2024.

[9] Ž. Bojović, *Application of Network Function Virtualization in Modern Computer Environments*. Boston-Delft: Now Publishers, 2024.

[10] D. Mustafa, "Machine learning-driven strategies for efficient resource management in cloud data centers," M.S. thesis, Concordia Univ., 2024.

[11] P. Zhou, L. Lin, Z. Zhang, Y. Deng, and T. He, "GHB: A cost-effective and energy-efficient data center network structure with greater incremental scalability," *Cluster Comput.*, vol. 27, no. 1, pp. 91–107, 2024.

[12] Z. Li, J. Huang, S. Wang, and J. Wang, "Achieving low latency for multipath transmission in RDMA based data center network," *IEEE Trans. Cloud Comput.*, vol. 12, no. 1, pp. 337–346, 2024.

[13] S. U. Khan, Z. U. Khan, M. Alkhowaiter, J. Khan, and S. Ullah, "Energy-efficient routing protocols for UWSNs: A comprehensive review of taxonomy, challenges, opportunities, future research directions, and machine learning perspectives," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 36, no. 7, 102128, 2024.

[14] O. Chidolue, P. E. Ohenhen, A. A. Umoh, B. Ngozichukwu, A. V. Fafure, and K. I. Ibekwe, "Green data centers: sustainable practices for energy-efficient IT infrastructure," *Eng. Sci. Technol. J.*, vol. 5, no. 1, pp. 99–114, 2024.

[15] R. Buyya, S. Ilager, and P. Arroba, "Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads," *Softw. Pract. Exp.*, vol. 54, no. 1, pp. 24–38, 2024.

[16] S. S. Nair, "Challenges and concerns related to the environmental impact of cloud computing and the carbon footprint of data transmission," *J. Comput. Sci. Technol. Stud.*, vol. 6, no. 1, pp. 195–199, 2024.

[17] A. Rajagopalan *et al*., "Empowering power distribution: Unleashing the synergy of IoT and cloud computing for sustainable and efficient energy systems," *Results Eng.*, vol. 21, 101949, 2024.

[18] M. Yenugula, "Data center power management using neural network," *Int. J. Adv. Acad. Stud.*, vol. 3, pp. 320–325, 2021.

[19] H. Nawaz, M. A. Ali, S. I. Rai, and M. Maqsood, "Comparative analysis of cloud based SDN and NFV in 5G networks," *Asian Bull. Big Data Manage.*, vol. 4, no. 1, 2024.

[20] J. Zhang, A. Ouda, and R. Abu-Rukba, "Authentication and key agreement protocol in hybrid edge–fog–cloud computing enhanced by 5G networks," *Future Internet*, vol. 16, no. 6, 209, 2024.

[21] I. Wang, K. S. Batta, and J.-l. Tsai, "Data center network architectures for large-scale distributed machine learning," *HAL*, 2024.

[22] R. Ding *et al*., "RD-Probe: Scalable monitoring with sufficient coverage in complex datacenter networks," in *Proc. ACM SIGCOMM 2024 Conf.*, 2024, pp. 258–273.

[23] N. Blach *et al.*, "A high-performance design, implementation, deployment, and evaluation of the Slim Fly network," in *Proc. 21st USENIX Symp. Netw. Syst. Design Implement. (NSDI 24)*, 2024, pp. 1025–1044.

[24] D. Das, B. Sahoo, S. Roy, and S. Mohanty, "Performance analysis of an OpenFlow-enabled network with POX, Ryu, and ODL controllers," *IETE J. Res.*, vol. 70, no. 12, pp. 8538–8555, 2024.

[25] Y. Guo, "Towards leaner data centers: Energy efficiency and carbon savings through network optimization," Ph.D. dissertation, Univ. California, San Diego, 2024.

[26] M. Zhao, Z. Han, and X. Du, "A survey of data center network topology structure," in *Proc. 2023 25th Int. Conf. Adv. Commun. Technol. (ICACT)*, 2023, pp. 303–309.

[27] M. Gupta, R. Kumar, D. Dobriyal, and P. Pandey, "Design strategies and performance enhancement techniques for spine-leaf architecture: A review," *NEU J. Artif. Intell. Internet Things*, vol. 2, no. 4, 2023.

[28] X. Han, Q. Huangpeng, Q. Gao, Y. Fu, and X. Duan, "Study of data center communication network topologies using complex network propagation model," *Front. Phys.*, vol. 11, 1174099, 2023.

[29] I. Buzhin, V. Antonova, V. Gnezdilov, Y. B. Mironov, and E. Gayfutdinov, "A way to implement network services in a data center," in *Proc. 2024 Wave Electron. Appl. Inf. Telecommun. Syst. (WECONF)*, 2024, pp. 1–5.

[30] D. Yu, W. Shao, and G. Shen, "When electronic spine-leaf meets optical torus: A hybrid optical-electronic data center network," in *Proc. 2023 Asia Communications and Photonics Conference/2023 International Photonics and Optoelectronics Meetings (ACP/POEM)*, 2023, pp. 1–4.

[31] Y. C. Tian and J. Gao, "Network routing architecture," in *Proc. Network Analysis and Architecture*, Springer, 2023, pp. 221–273.

[32] M. D. Tache, O. Păscuțoiu, and E. Borcoci, "Optimization algorithms in SDN: Routing, load balancing, and delay optimization," *Appl. Sci.*, vol. 14, no. 14, 5967, 2024.

[33] H. T. Duong, "Research and build a three-layer network system for businesses," B.S. thesis, Vietnam-Korea Univ. Inf. Commun. Technol., 2024.

[34] M. Elmadani and S. O. Sati, "Data center lab using VxLAN data plane and BGP-EVPN control plane," in *Proc. 2023 4th Int. Conf. Data Anal. Bus. Ind. (ICDABI)*, 2023, pp. 354–358.

[35] P. Boryło *et al*., "SDNRoute: Proactive routing optimization in software defined networks," *Comput. Commun.*, vol. 225, pp. 250–278, 2024.

[36] H. Iesar *et al*., "Revolutionizing data center networks: Dynamic load balancing via Floodlight in SDN environment," in *Proc. 2024 5th Int. Conf. Advancements Comput. Sci. (ICACS)*, 2024, pp. 1–8.

[37] B. Balakiruthiga, P. Deepalakshmi, S. N. Mohanty, D. Gupta, P. P. Kumar, and K. Shankar, "Segment routing based energy aware routing for software defined data center," *Cogn. Syst. Res.*, vol. 64, pp. 146–163, 2020.

[38] Y. Wang, Y. Li, T. Wang, and G. Liu, "Towards an energy-efficient data center network based on deep reinforcement learning," *Comput. Netw.*, vol. 210, 108939, 2022.

[39] A. H. Alomari, S. K. Subramaniam, N. Samian, R. Latip, and Z. A. Zukarnain, "Dual-phase resource allocation algorithm in software-defined network SDN-enabled cloud," *IEEE Access*, vol. 11, pp. 102301–102315, 2023.

[40] A. Sharma, S. Tokekar, and S. Varma, "A comprehensive survey on network resource management in SDN enabled data centre network,"

in *Proc. 6G Enabled Fog Computing in IoT: Applications and Opportunities*, 2023, pp. 333–353.
[41] F. Lin *et al*., "Fast, scalable and robust centralized routing for data center networks," *IEEE/ACM Trans. Netw.*, vol. 31, no. 6, pp. 2624–2639, 2023.
[42] Z. Jia, Q. Liu, and Y. Sun, "sRetor: A semi-centralized regular topology routing scheme for data center networking," *J. Cloud Comput.*, vol. 12, no. 1, 150, 2023.
[43] P. Arabas, T. Jóźwik, and E. Niewiadomska-Szynkiewicz, "Router activation heuristics for energy-saving ECMP and valiant routing in data center networks," *Energies*, vol. 16, no. 10, 4136, 2023.
[44] F. Wang *et al*., "Dynamic distributed multi-path aided load balancing for optical data center networks," *IEEE Trans. Netw. Service Manag.*, vol. 19, no. 2, pp. 991–1005, 2021.
[45] Y. Liang, M. Lu, Z. J. M. Shen, and R. Tang, "Data center network design for internet-related services and cloud computing," *Prod. Oper. Manag.*, vol. 30, no. 7, pp. 2077–2101, 2021.
[46] S. Gharehpasha, M. Masdari, and A. Jafarian, "Power efficient virtual machine placement in cloud data centers with a discrete and chaotic hybrid optimization algorithm," *Cluster Comput.*, vol. 24, no. 2, pp. 1293-1315, 2021.
[47] J. Han, K. Xue, W. Wang, R. Li, Q. Sun, and J. Lu, "RateMP: Optimizing bandwidth utilization with high burst tolerance in data center networks," in *Proc. IEEE INFOCOM 2024*, 2024, pp. 1361–1370.
[48] Y. Wang, X. Wang, H. Li, Y. Dong, Q. Liu, and X. Shi, "A multi-service differentiation traffic management strategy in SDN cloud data center," *Comput. Netw.*, vol. 171, 107143, 2020.
[49] Y. Guo, K. Huang, C. Hu, J. Yao, and S. Zhou, "Traffic engineering in dynamic hybrid segment routing networks," *Comput. Mater. Continua*, vol. 68, no. 1, pp. 655–670, 2021.
[50] M. R. Majma and E. Soltani Nejad, "SEMTE: Scalable and extended modular traffic engineering in software-defined data center networks," *Photon. Netw. Commun.*, vol. 42, pp. 143–166, 2021.
[51] C. Zhang, Y. Wang, H. Wu, and H. Guo, "An energy-aware host resource management framework for two-tier virtualized cloud data centers," *IEEE Access*, vol. 9, pp. 3526–3544, 2020.
[52] P. Zhang, M. Zhou, and X. Wang, "An intelligent optimization method for optimal virtual machine allocation in cloud data centers," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1725–1735, 2020.
[53] M. Aldossary, "A review of dynamic resource management in cloud computing environments," *Comput. Syst. Sci. Eng.*, vol. 36, no. 3, 2021.
[54] S. Yang, F. Li, S. Trajanovski, R. Yahyapour, and X. Fu, "Recent advances of resource allocation in network function virtualization," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 2, pp. 295–314, 2020.
[55] D. Saxena and A. K. Singh, "A proactive autoscaling and energy-efficient VM allocation framework using online multi-resource neural network for cloud data center," *Neurocomputing*, vol. 426, pp. 248–264, 2021.
[56] S. Chhabra and A. K. Singh, "Dynamic resource allocation method for load balance scheduling over cloud data center networks," *J. Web Eng.*, vol. 20, no. 8, pp. 2269–2284, 2021.
[57] T. He and R. Buyya, "A taxonomy of live migration management in cloud computing," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 1–33, 2023.
[58] A. Satpathy, M. N. Sahoo, A. Mishra, B. Majhi, J. J. Rodrigues, and S. Bakshi, "A service sustainable live migration strategy for multiple virtual machines in cloud data centers," *Big Data Res.*, vol. 25, 100213, 2021.
[59] J. O. Gutierrez-Garcia and A. Ramirez-Nafarrate, "Collaborative agents for distributed load management in cloud data centers using live migration of virtual machines," *IEEE Trans. Serv. Comput.*, vol. 8, no. 6, pp. 916–929, 2015.
[60] T. Wood, K. Ramakrishnan, P. Shenoy, and J. Van der Merwe, "CloudNet: Dynamic pooling of cloud resources by live WAN migration of virtual machines," *ACM SIGPLAN Notices*, vol. 46, no. 7, pp. 121–132, 2011.
[61] S. M. Shetty and S. Shetty, "Analysis of load balancing in cloud data centers," *J. Ambient Intell. Humanized Comput.*, vol. 15, no. 1, pp. 973–981, 2024.
[62] A. Javadpour *et al*., "An energy-optimized embedded load balancing using DVFS computing in cloud data centers," *Comput. Commun.*, vol. 197, pp. 255–266, 2023.
[63] A. N. Quttoum *et al*., "ABLA: Application-based load-balanced approach for adaptive mapping of datacenter networks," *Electronics*, vol. 12, no. 17, 3689, 2023.
[64] L. Mao, R. Chen, H. Cheng, W. Lin, B. Liu, and J. Z. Wang, "A resource scheduling method for cloud data centers based on thermal management," *J. Cloud Comput.*, vol. 12, no. 1, 84, 2023.
[65] T. H. Chopra and P. V. Lahande, "Performance evaluation of service broker policies in cloud computing environment using round robin," in *Proc. Int. Conf. Soft Comput. Eng. Appl.*, 2023, pp. 201–213.
[66] R. Kabir, R. G. Kim, and M. Nikdast, "RISA: Round-robin intra-rack friendly scheduling algorithm for disaggregated datacenters," in *Proc. SC'23 Workshops*, 2023, pp. 1512–1520.
[67] Z. Han, G. Yang, and F. Zuo, "A review of virtual resource management research in cloud data centers," *J. Phys. Conf. Ser.*, vol. 2562, no. 1, 012023, 2023.
[68] T. Wang, X. Fan, K. Cheng, X. Du, H. Cai, and Y. Wang, "Parameterized deep reinforcement learning with hybrid action space for energy efficient data center networks," *Comput. Netw.*, vol. 235, 109989, 2023.
[69] J. Zerwas, C. Györgyi, A. Blenk, S. Schmid, and C. Avin, "Duo: A high-throughput reconfigurable datacenter network using local routing and control," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 7, no. 1, pp. 1–25, 2023.
[70] A. Shirmarz and A. Ghaffari, "Performance issues and solutions in SDN-based data center: A survey," *J. Supercomput.*, vol. 76, no. 10, pp. 7545–7593, 2020.
[71] B. Girish, B. Subramanaya, S. Rohan, and V. N. Priya, "Optimizing data center load balancing in cloud computing with SDN controller," in *Proc. 2024 Int. Conf. Knowl. Eng. Commun. Syst. (ICKECS)*, 2024.
[72] K. Sathupadi, "AI-driven energy optimization in SDN-based cloud computing for balancing cost, energy efficiency, and network performance," *Int. J. Appl. Mach. Learn. Comput. Intell.*, vol. 13, no. 7, pp. 11–37, 2023.
[73] Q. Du, X. Cui, H. Tang, and X. Chen, "Review of load balancing mechanisms in SDN-based data centers," *J. Comput. Commun.*, vol. 12, no. 1, pp. 49–66, 2024.
[74] A. Alarifi *et al*., "Energy-efficient hybrid framework for green cloud computing," *IEEE Access*, vol. 8, pp. 115356–115369, 2020.
[75] I. Alvarez-Meaza, E. Zarrabeitia-Bilbao, R. Rio-Belver, and G. Garechana-Anacabe, "Green scheduling to achieve green manufacturing: Pursuing a research agenda by mapping science," *Technol. Soc.*, vol. 67, 101758, 2021.
[76] A. Montazerolghaem, M. H. Yaghmaee, and A. Leon-Garcia, "Green cloud multimedia networking: NFV/SDN based energy-efficient resource allocation," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 3, pp. 873–889, 2020.
[77] M. S. Ajmal, Z. Iqbal, F. Z. Khan, M. Bilal, and R. M. Mehmood, "Cost-based energy efficient scheduling technique for dynamic voltage and frequency scaling system in cloud computing," *Sustain. Energy Technol. Assessments*, vol. 45, 101210, 2021.
[78] J. Masoudi, B. Barzegar, and H. Motameni, "Energy-aware virtual machine allocation in DVFS-enabled cloud data centers," *IEEE Access*, vol. 10, pp. 3617–3630, 2021.
[79] S. Kumar *et al*., "Energy efficient model for balancing energy in cloud datacenters using dynamic voltage frequency scaling (DVFS) technique," in *Proc. 3rd Doctoral Symp. Comput. Intell. (DoSCI 2022)*, 2022, pp. 533–540.
[80] A. A. Khan and M. Zakarya, "Energy, performance and cost efficient cloud datacentres: A survey," *Comput. Sci. Rev.*, vol. 40, 100390, 2021.
[81] N. Chaurasia, M. Kumar, R. Chaudhry, and O. P. Verma, "Comprehensive survey on energy-aware server consolidation techniques in cloud computing," *J. Supercomput.*, vol. 77, pp. 11682–11737, 2021.
[82] S. Jangiti and S. S. VS, "EMC2: Energy-efficient and multi-resource-fairness virtual machine consolidation in cloud data centres," *Sustain. Comput. Inform. Syst.*, vol. 27, 100414, 2020.
[83] K. H. K. Reddy, A. K. Luhach, V. V. Kumar, S. Pratihar, D. Kumar, and D. S. Roy, "Towards energy efficient smart city services: A software defined resource management scheme for data centers," *Sustain. Comput. Inform. Syst.*, vol. 35, 100776, 2022.
[84] I. Hamzaoui, B. Duthil, V. Courboulay, and H. Medromi, "A survey on the current challenges of energy-efficient cloud resources management," *SN Comput. Sci.*, vol. 1, pp. 1–28, 2020.

[85] Q. Gang, A. Muhammad, Z. U. Khan, M. S. Khan, F. Ahmed, and J. Ahmad, "Machine learning-based prediction of node localization accuracy in IIoT-based MI-UWSNs and design of a TD coil for omnidirectional communication," *Sustainability*, vol. 14, no. 15, 9683, 2022.

[86] M. Aman, Q. Gang, Z. Shang, Z. U. Khan, M. S. Khan, and I. Ullah, "Realization of RSSI based, three major components (Hx, Hy, Hz) of magnetic flux created around the MI-TD coil," in *Proc. 2023 IEEE Int. Conf. Electr. Autom. Comput. Eng. (ICEACE)*, 2023, pp. 1012–1017.

[87] Y. Kumar, S. Kaul, and Y. C. Hu, "Machine learning for energy-resource allocation, workflow scheduling and live migration in cloud computing: State-of-the-art survey," *Sustain. Comput. Inform. Syst.*, vol. 36, 100780, 2022.

[88] T. Thein, M. M. Myo, S. Parvin, and A. Gawanmeh, "Reinforcement learning based methodology for energy-efficient resource allocation in cloud data centers," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 10, pp. 1127–1139, 2020.

[89] N. K. Pandey, M. Diwakar, A. Shankar, P. Singh, M. R. Khosravi, and V. Kumar, "Energy efficiency strategy for big data in cloud environment using deep reinforcement learning," *Mob. Inf. Syst.*, vol. 2022, no. 1, 8716132, 2022.

[90] H. Yang, W. D. Zhong, C. Chen, A. Alphones, and X. Xie, "Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5677–5689, 2020.

[91] W. U. Rahman, Q. Gang, Z. Feng, Z. U. Khan, M. Aman, and I. Ullah, "A Q-learning-based multi-hop energy-efficient and low collision MAC protocol for underwater acoustic wireless sensor networks," in *Proc. 2023 20th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, 2023, pp. 872–877.

[92] Z. U. Khan, M. Aman, W. U. Rahman, F. Khan, T. Jamil, and R. Hashim, "Machine learning-based multi-path reliable and energy-efficient routing protocol for underwater wireless sensor networks," in *Proc. 2023 Int. Conf. Front. Inf. Technol. (FIT)*, 2023, pp. 316–321.

[93] Q. Ding, R. Zhu, H. Liu, and M. Ma, "An overview of machine learning-based energy-efficient routing algorithms in wireless sensor networks," *Electronics*, vol. 10, no. 13, 1539, 2021.

[94] T. Kamble, S. Deokar, V. S. Wadne, D. P. Gadekar, H. B. Vanjari, and P. Mange, "Predictive resource allocation strategies for cloud computing environments using machine learning," *J. Electr. Syst.*, vol. 19, no. 2, 2023.

[95] R. S. S. Dittakavi, "Deep learning-based prediction of CPU and memory consumption for cost-efficient cloud resource allocation," *Sage Sci. Rev. Appl. Mach. Learn.*, vol. 4, no. 1, pp. 45–58, 2021.

[96] S. Jayaprakash, M. D. Nagarajan, R. P. d. Prado, S. Subramanian, and P. B. Divakarachari, "A systematic review of energy management strategies for resource allocation in the cloud: Clustering, optimization and machine learning," *Energies*, vol. 14, no. 17, 5322, 2021.

[97] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," *J. Netw. Comput. Appl.*, vol. 204, 103405, 2022.

[98] Z. Zhou, M. Shojafar, J. Abawajy, H. Yin, and H. Lu, "ECMS: An edge intelligent energy efficient model in mobile edge computing," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 1, pp. 238–247, 2021.

[99] Y. Jararweh, "Enabling efficient and secure energy cloud using edge computing and 5G," *J. Parallel Distrib. Comput.*, vol. 145, pp. 42–49, 2020.

[100] W. Duan, X. Gu, M. Wen, Y. Ji, J. Ge, and G. Zhang, "Resource management for intelligent vehicular edge computing networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9797–9808, 2021.

[101] R. Souza, K. Dias, and S. Fernandes, "NFV data centers: A systematic review," *IEEE Access*, vol. 8, pp. 51713–51735, 2020.

[102] I. Setiawan, B. Kar, and S.-H. Shen, "Energy-efficient softwarized networks: A survey," arXiv preprint arXiv:2307.11301, 2023.

[103] G. Sun, R. Zhou, J. Sun, H. Yu, and A. V. Vasilakos, "Energy-efficient provisioning for service function chains to support delay-sensitive applications in network function virtualization," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6116–6131, 2020.

[104] B. Li, B. Cheng, X. Liu, M. Wang, Y. Yue, and J. Chen, "Joint resource optimization and delay-aware virtual network function migration in data center networks," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 3, pp. 2960–2974, 2021.

[105] U. K. Lilhore *et al*., "QHRMOF: A quantum-inspired hybrid multi-objective framework for energy-efficient task scheduling and load balancing in cloud computing," *J. Cloud Comput.*, vol. 14, no. 1, 54, 2025.

[106] H. Ballani *et al*., "Sirius: A flat datacenter network with nanosecond optical switching," in *Proc. Annu. Conf. ACM SIGCOMM*, 2020, pp. 782–797.

[107] D. Jiang, Y. Wang, Z. Lv, W. Wang, and H. Wang, "An energy-efficient networking approach in cloud services for IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 928–941, 2020.

[108] W. X. Liu, J. Cai, Q. C. Chen, and Y. Wang, "DRL-R: Deep reinforcement learning approach for intelligent routing in software-defined data-center networks," *J. Netw. Comput. Appl.*, vol. 177, 102865, 2021.

[109] L. Shalev, H. Ayoub, N. Bshara, and E. Sabbag, "A cloud-optimized transport protocol for elastic and scalable HPC," *IEEE Micro*, vol. 40, no. 6, pp. 67–73, 2020.

[110] X. Xue *et al*., "ROTOS: A reconfigurable and cost-effective architecture for high-performance optical data center networks," *J. Lightw. Technol.*, vol. 38, no. 13, pp. 3485–3494, 2020.

[111] D. Abts and J. Kim, *High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities*, Springer Nature, 2022.

[112] S. Shukla, M. F. Hassan, D. C. Tran, R. Akbar, I. V. Paputungan, and M. K. Khan, "Improving latency in Internet-of-Things and cloud computing for real-time data transmission: A systematic literature review (SLR)," *Cluster Comput.*,vol. 26, no. 5, pp. 2657–2680, 2023.

[113] W. ur Rahman, Q. Gang, Z. Feng, Z. U. Khan, M. Aman, and M. Bilal, "A MACA-based energy-efficient MAC protocol using Q-learning technique for underwater acoustic sensor network," in *Proc. 2023 IEEE 11th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, 2023, pp. 352–355.

[114] B. Charyyev, E. Arslan, and M. H. Gunes, "Latency comparison of cloud datacenters and edge servers," in *Proc. GLOBECOM 2020*, 2020, pp. 1–6.

[115] A. A. Laghari, X. Zhang, Z. A. Shaikh, A. Khan, V. V. Estrela, and S. Izadi, "A review on Quality of Experience (QoE) in cloud computing," *J. Reliable Intell. Environ.*, vol. 10, no. 2, pp. 107–121, 2024.

[116] A. Katal, S. Dahiya, and T. Choudhury, "Energy efficiency in cloud computing data centers: A survey on software technologies," *Cluster Comput.*, vol. 26, no. 3, pp. 1845–1875, 2023.

[117] S. Suryadevara, "Energy-proportional computing: Innovations in data center efficiency and performance optimization," *Int. J. Adv. Eng. Technol. Innov.*, vol. 1, no. 2, pp. 44–64, 2021.

[118] S. Bharany *et al*., "A systematic survey on energy-efficient techniques in sustainable cloud computing," *Sustainability*, vol. 14, no. 10, 6256, 2022.

[119] P. Sun, "Security and privacy protection in cloud computing: Discussions and challenges," *J. Netw. Comput. Appl.*, vol. 160, 102642, 2020.

[120] M. S. Sheikh, J. Liang, and W. Wang, "Security and privacy in vehicular ad hoc network and vehicle cloud computing: A survey," *Wireless Commun. Mobile Comput.*, vol. 2020, no. 1, 5129620, 2020.

[121] A. Masood, D. S. Lakew, and S. Cho, "Security and privacy challenges in connected vehicular cloud computing," *IEEE Commun. Surv. Tut.*, vol. 22, no. 4, pp. 2725–2764, 2020.

[122] P. Yang, N. Xiong, and J. Ren, "Data security and privacy protection for cloud storage: A survey," *IEEE Access*, vol. 8, pp. 131723–131740, 2020.

[123] N. E. H. Bouzerzour, S. Ghazouani, and Y. Slimani, "A survey on the service interoperability in cloud computing: Client-centric and provider-centric perspectives," *Softw. Pract. Exp.*, vol. 50, no. 7, pp. 1025–1060, 2020.

[124] G. Yang, M. A. Jan, A. U. Rehman, M. Babar, M. M. Aimal, and S. Verma, "Interoperability and data storage in internet of multimedia things: Investigating current trends, research challenges and future directions," *IEEE Access*, vol. 8, pp. 124382–124401, 2020.

[125] L. Poutievski *et al*., "Jupiter evolving: Transforming Google's datacenter network via optical circuit switches and software-defined networking," in *Proc. ACM SIGCOMM 2022 Conf.*, 2022, pp. 66–85.

[126] W. Lu, L. Liang, B. Kong, B. Li, and Z. Zhu, "AI-assisted knowledge-defined network orchestration for energy-efficient data center networks," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 86–92, 2020.
[127] D. Soni and N. Kumar, "Machine learning techniques in emerging cloud computing integrated paradigms: A survey and taxonomy," *J. Netw. Comput. Appl.*, vol. 205, 103419, 2022.
[128] U. K. Padyana, H. P. Rai, P. Ogeti, N. S. Fadnavis, and G. B. Patil, "AI and machine learning in cloud-based Internet of Things (IoT) solutions: A comprehensive review and analysis," *Integr. J. Res. Arts Humanities*, vol. 3, no. 3, pp. 121–132, 2023.
[129] H. Kuchuk and E. Malokhvii, "Integration of IoT with cloud, fog, and edge computing: A review," *Adv. Inf. Syst.*, vol. 8, no. 2, pp. 65–78, 2024.
[130] S. Ahmad, "A review on edge to cloud: Paradigm shift from large data centers to small centers of data everywhere," in *Proc. 2020 Int. Conf. Inventive Comput. Technol. (ICICT)*, 2020, pp. 318–322.
[131] H. Cheng, B. Liu, W. Lin, Z. Ma, K. Li, and C. H. Hsu, "A survey of energy-saving technologies in cloud data centers," *J. Supercomput.*, vol. 77, no. 11, pp. 13385–13420, 2021.
[132] X. Chang, S. Yang, Y. Jiang, X. Xie, and X. Tang, "Research on key energy-saving technologies in green data centers," in *Proc. 2020 IEEE Int. Conf. Smart Cloud (SmartCloud)*, 2020, pp. 111–115.
[133] S. Malathy *et al*., "A review on energy management issues for future 5G and beyond network," *Wireless Netw.*, vol. 27, pp. 2691–2718, 2021.
[134] A. Mahmood *et al*., "Industrial IoT in 5G-and-beyond networks: Vision, architecture, and design trends," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4122–4137, 2021.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).