A Survey on Edge Computing Paradigms and Technologies

SurveyForge

Abstract— Edge computing paradigms have emerged as pivotal solutions to overcome the high latency and bandwidth limitations of centralized cloud infrastructures. This survey comprehensively examines the core architectures, technologies, and application domains of edge computing, highlighting their transformative potential in enabling real-time, data-intensive applications. With a focus on frameworks like fog and mobile edge computing, the study explores the interplay between decentralized processing resources and advanced wireless networks, such as 5G and 6G, facilitating enhanced proximity to data sources. It assesses key technological enablers, including containerization and distributed intelligence, which optimize resource utilization. The challenges of securing heterogeneous edge environments are addressed through promising yet evolving methods like blockchain-based trust management and AI-driven anomaly detection. Furthermore, the integration of AI, federated learning, and energy-efficient strategies underscores the innovative directions driving edge systems. Despite these advances, ongoing challenges remain in standardization, scalability, and resource sustainability. This survey highlights the critical need for interdisciplinary research to optimize edge infrastructures and accelerate their deployment across varied sectors, from autonomous systems to smart cities, thus shaping future digital ecosystems.

Index Terms—edge computing paradigms, resource optimization strategies, real-time data processing

1 Introduction

EDGE computing represents a significant shift in the paradigm of distributed computing, addressing the increasing limitations of traditional centralized cloud architectures. This subsection delves into the foundational principles of edge computing, traces its historical evolution, and articulates its growing significance in overcoming challenges posed by modern, latency-sensitive applications. By analyzing the motivations behind its emergence and its divergence from related computing paradigms, a comprehensive understanding of edge computing's transformative potential in shaping contemporary technologies is developed.

At its core, edge computing seeks to decentralize data processing and computation, bringing these functions closer to the data source, typically at the "edge" of the network [1]. Unlike the cloud-first model that relies on remote data centers for computation, edge computing prioritizes proximity to data sources, achieving reductions in latency and network congestion. Such characteristics cater to applications where immediate responsiveness is critical—examples include autonomous vehicles, real-time health monitoring, and augmented reality systems, where processing delays could be detrimental [2]. Proximity-oriented paradigms are not only designed to meet stringent performance requirements but also mitigate issues like bandwidth bottlenecks and excessive energy consumption inherent in centralized systems [1].

The trajectory of edge computing can be traced back to its emergence as a response to the growing demands of IoT and big data-intensive applications. Initial cloud-centric architectures, while scalable and elastic, struggled to provide satisfactory quality of service (QoS) for latency-sensitive and geographically distributed use cases. Architectures such as fog computing and mobile-edge computing (MEC) emerged as intermediaries to bridge the gap between IoT endpoints and centralized data centers. Fog computing, for instance, extended cloud capabilities closer to network peripheries, utilizing localized computing nodes to reduce latency and enhance service proximity [3]. Similarly, MEC integrated processing capabilities directly within the radio access network (RAN) to serve mobile users with real-time, context-aware services [4]. These advancements signify pivotal milestones in the gradual decentralization of computational workloads, a hallmark of edge paradigms.

A distinguishing feature of edge computing is its ability to adapt to varying computational demands and resource constraints across heterogeneous environments. Unlike cloud-centric solutions that centralize resource management, edge frameworks operate within a distributed hierarchy, leveraging localized devices such as edge gateways, micro-data centers, and IoT-enabled microcontrollers. These physical and architectural considerations have enabled edge paradigms to deliver superior performance for time-critical applications while balancing energy efficiency and operational scalability [5]. Furthermore, compared to related paradigms like fog computing, edge computing often restricts the scope of processing significantly closer to end-user devices, drawing a nuanced distinction in terms of deployment layers and application focus [6]. While fog and edge computing share overlapping features, such as reduced latency and decentralization, they exhibit differences in application breadth and latency thresholds, emphasizing the need for precise architectural choices based on workload requirements [7].

The rising adoption of edge computing is driven not only by its performance benefits but also by its role in addressing emerging societal concerns, such as data privacy and regulatory compliance. By enabling localized data processing, edge frameworks inherently reduce the volume of sensitive data transmitted over long distances, alleviating privacy risks and aiding adherence to data sovereignty regulations (e.g., GDPR). This aspect is especially critical in applications like healthcare, where patient data privacy coincides with real-time processing needs [8]. However, the decentralization of edge resources also entails trade-offs, such as increased management complexity and the potential vulnerability of numerous dispersed nodes, requiring robust orchestration frameworks and security mechanisms to maintain system integrity [8].

Furthermore, edge computing's trajectory is intertwined with the progression of enabling technologies. Advances in next-generation networks, such as 5G and 6G, provide the ultra-low latency and high-bandwidth capabilities that bolster edge computing's potential in domains like autonomous mobility, AR/VR, and smart cities [9]. Similarly, developments in lightweight virtualization, containerized applications, and distributed AI models have enhanced the agility and intelligence of edge deployments, enabling sophisticated decision-making at the edge [10].

In synthesizing edge computing's foundational principles, historical milestones, and technological underpinnings, it becomes evident that this paradigm holds transformative potential in revolutionizing distributed systems. Moving forward, challenges such as standardization, resource sustainability, and seamless cloud-edge integration must be addressed. Researchers must further develop frameworks for hybrid resource orchestration, secure multi-domain edge management, and context-sensitive optimization to meet the demands of increasingly complex applications. Edge computing, while still evolving, is poised to play a foundational role in enabling the digital ecosystems of the future.

2 Architectural Foundations of Edge Computing

2.1 Layered Frameworks in Edge Computing

Layered frameworks play a pivotal role in shaping the architectural foundations of edge computing, offering a systematic approach to facilitate data processing, communication, and decision-making closer to the source. This subsection delves into the essential characteristics of hierarchical deployment models within edge computing, exploring their role in optimizing computational efficiency, reducing latency, and ensuring seamless resource distribution.

The three-layered structure commonly adopted in edge computing—comprising the device-level or sensor-edge integration, local-edge architectures, and regional-edge deployments—exemplifies the distributed yet cooperative nature of these systems. At the device or sensor level, processing resources are embedded within IoT devices or edge sensors, enabling a first layer of computation that minimizes latency and power consumption for critical tasks. This approach is particularly relevant for latency-sensitive applications such as autonomous navigation or medical monitoring, where real-time processing is necessary [5]. However, while device-layer frameworks achieve low-latency results, they are constrained by limited computation and storage capacities. This underscores the need for higher-layer collaborative frameworks to handle tasks concerning larger datasets or computationally intensive functions.

Local-edge computing nodes, often referred to as intermediary nodes, typically aggregate data from a set of nearby devices while providing enhanced processing and storage capabilities [1]. Local-edge nodes can perform tasks such as filtering, preprocessing, and lightweight analytics before transmitting relevant outputs to either regional-edge servers or cloud systems. Studies have shown that frameworks integrating micro-edge data aggregation reduce network congestion dramatically while maintaining the scalability of the edge infrastructure [9]. Nonetheless, the centralized role of these nodes raises concerns about potential single points of failure and increased pressure from multitasking coordination demands.

On a broader scale, the regional edge layer leverages edge data centers or micro-cloud sites distributed in proximity to specific geographical clusters. Regional-edge nodes act as intermediate platforms bridging numerous local nodes with centralized cloud systems. Their responsibilities include high-performance computing, collaborative storage, and multilayered workload balancing. Research has demonstrated the potential of such deployments for lowering costs associated with long-haul network communication, particularly for compute-heavy services like machine learning inference or video transcoding in urban environments [7]. However, regional-edge nodes are subject to challenges such as energy consumption and synchronization instabilities, particularly when integrating multi-tenant resources within dynamic workloads [11].

A critical strength of layered frameworks lies in their capacity to enable task offloading and distribution across multiple levels. Task offloading frameworks typically examine variables such as proximity, resource constraints, bandwidth availability, and application QoS requirements to determine the optimal layer for execution. For instance, delay-sensitive tasks may remain at device-level nodes, while batch processing or non-critical tasks can be deferred to regional or cloud setups for computation efficiency [12]. Nevertheless, achieving seamless interoperability across layers requires robust communication protocols and orchestration mechanisms. Protocols such as MQTT and CoAP have been widely adopted for lightweight communication across edge nodes, while adaptive routing frameworks further ensure effective data flow [13].

Despite the evident advantages of layered frameworks, certain trade-offs remain persistent. Resource allocation inefficiencies can emerge due to the heterogeneity of devices and system layers, necessitating advanced automation and orchestration strategies [14]. Emerging trends in edge-native AI for resource management show promise in mitigating such challenges, allowing real-time decisions within constrained environments [10]. Furthermore, developments in containerization technologies, as exemplified by lightweight Docker-based solutions, are enabling the modular deployment of services across the layered edge architecture, simplifying system scalability and adaptability [15].

Looking ahead, the integration of hierarchical edge frameworks with advanced paradigms like 6G communication and cooperative fog-to-cloud systems opens promising frontiers for edge computing. Such integrations will enhance dynamic scaling and facilitate collaborative processing across diverse global regions, ultimately supporting use cases such as metaverse applications, autonomous mobility, and resilience-oriented urban systems [16]. However, addressing gaps in interoperability, scheduling efficiency, and energy sustainability remains critical to fully realizing the potential of these architectures within decentralized computing ecosystems.

2.2 Hardware Infrastructure for Edge Systems

Edge computing represents a paradigm shift in computational design, prioritizing proximity to data sources by leveraging a distributed hardware infrastructure. Fundamental to this transformation are specialized physical components, including micro-data centers, edge gateways, and hardware accelerators, which collectively define the hardware ecosystem of edge systems. This subsection delves into these core elements, their roles, associated trade-offs, and the emerging innovations reshaping edge computing architectures, bridging insights from layered frameworks and collaborative edge-cloud systems.

Micro-data centers (MDCs) are compact server clusters strategically deployed near data generation points to reduce latency and increase computational capacity. Unlike traditional centralized cloud data centers, MDCs decentralize workloads to support latency-sensitive applications, such as autonomous driving and urban surveillance [5]. By processing high data volumes locally, they alleviate bandwidth bottlenecks while enabling rapid, localized responses. However, MDC deployments in constrained environments necessitate careful optimization of space, energy consumption, and cooling systems [1]. To handle diverse workloads like video analytics and environmental monitoring, MDCs commonly integrate general-purpose CPUs with task-specific accelerators, achieving adaptability. Nevertheless, scaling MDCs to accommodate growing computational demands while maintaining fault tolerance and compactness remains an ongoing research challenge [6].

Edge gateways play a pivotal role as intermediaries between IoT devices and higher-tier computational nodes, serving as hubs for data aggregation, preprocessing, and secure communication. Their support for wireless standards such as 5G, LoRa, and ZigBee enables uninterrupted connectivity across fragmented network topologies [4]. By incorporating lightweight computational capabilities, gateways preprocess raw data before transmission, optimizing bandwidth usage and enabling real-time responses to localized events [17]. Beyond their data-handling functionalities, edge gateways manage resource-constrained devices through workload distribution and energy-efficient operations, critical for IoT-heavy ecosystems. However, ensuring their robustness against cybersecurity threats and physical tampering remains a pressing concern, particularly given their deployment in publicly accessible settings [8].

Complementing MDCs and gateways are specialized hardware accelerators such as Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Field-Programmable Gate Arrays (FPGAs). These accelerators are indispensable for executing resource-intensive tasks like deep learning inference and real-time analytics directly at the edge. By leveraging parallel processing, they deliver high throughput and reduced energy consumption for latency-critical applications, including industrial robotics and augmented reality [18]. However, the integration of such accelerators into heterogeneous edge environments at scale poses technical challenges, including inter-hardware compatibility and resource optimization, which must be addressed to fully harness their potential [9].

Emerging hardware trends increasingly emphasize sustainability and adaptability to meet evolving edge computing demands. Low-power processors based on ARM architectures optimize energy efficiency while delivering consistent computational performance, aligning with the broader shift toward green computing [19]. Reconfigurable architectures like RISC-V microcontrollers enable task-specific adaptability, providing long-term flexibility across diverse applications [20]. Additionally, neuromorphic hardware, designed for bio-inspired computing, offers ultra-low-power neural processing, vital for sensor-driven use cases such as healthcare monitoring and precision agriculture [21].

Despite significant advancements, the hardware ecosystem of edge computing continues to face challenges. Fragmentation and a lack of standardization across hardware frameworks hinder seamless integration, complicating broader ecosystem deployment [22]. Furthermore, balancing the need for increased computational capabilities with energy efficiency remains a persistent dilemma. Future progress must focus on energy-conscious architectures, hybrid integration with cloud backends, and fault-tolerant designs to extend edge computing's capabilities to mission-critical and emerging applications. Addressing these challenges will enable next-generation hardware infrastructures to realize the full potential of edge computing in areas such as precision medicine, real-time urban management, and autonomous mobility, ensuring harmony with layered and hybrid frameworks explored in adjacent sections.

2.3 Edge-Cloud Collaborative Models

The integration of edge and cloud infrastructures has emerged as a pivotal paradigm in distributed computing, enabling systems to leverage the strengths of both computational proximity at the edge and the scalability of centralized cloud services. Edge-cloud collaborative models optimize resource utilization by strategically sharing workloads between edge nodes and cloud data centers, addressing variability in computational demand, resource constraints, and the heterogeneous nature of modern distributed applications.

At the heart of these hybrid frameworks is workload distribution, wherein computational tasks are partitioned between edge devices and cloud platforms based on various metrics, including latency sensitivity, computational complexity, and resource availability. Task partitioning approaches dynamically allocate workloads to the most suitable layer, with computationally intensive and latency-tolerant tasks being offloaded to the cloud for centralized processing, whereas latency-critical tasks are executed locally at the edge [9], [20]. This stratified partitioning improves quality of service (QoS) by reducing end-to-end latency, lowering bandwidth requirements, and ensuring efficient resource utilization.

One prominent example of such collaborative models is federated edge-cloud ecosystems. These frameworks aggregate edge nodes and cloud resources into cohesive, integrated systems that rely on shared infrastructures to optimize resource utilization and enhance reliability. Federated models enable decentralized coordination, allowing distributed edge networks to act in unison to provide support for geographically dispersed applications. This approach also facilitates the reuse of resources across edge nodes, improving fault tolerance and minimizing computational redundancy [3], [23]. However, implementing robust federated ecosystems necessitates advancements in inter-edge communication protocols and cross-layer orchestration algorithms capable of handling heterogeneity and complexity [24].

Hybrid models also explore dynamic workload balancing as a critical optimization technique. By continuously assessing resource availability and computational demand, these models employ heuristic and machine-learning-driven strategies to allocate tasks dynamically. For instance, reinforcement learning has been leveraged to predict resource constraints and balance tasks efficiently, ensuring minimal queuing delays at edge nodes while mitigating overheads on the cloud [25], [26]. Nevertheless, achieving optimal dynamic balancing remains challenging, particularly in environments characterized by fluctuating network conditions and stringent latency constraints.

Multi-cloud and multi-edge integrations represent another key advancement in edge-cloud collaboration. These frameworks incorporate diverse cloud providers and edge infrastructures to provide seamless support for global applications with heterogeneous requirements. The inclusion of multi-cloud environments fosters resource locality and redundancy by enabling applications to access the geographically nearest cloud facility, while interoperable multi-edge nodes reduce dependency on any single layer and enhance resilience. Techniques such as decentralized service caching and large-scale task offloading play a significant role in maintaining service continuity while reducing costs [27], [28].

Despite their advantages, edge-cloud collaborative models face several challenges. One of the key barriers is achieving interoperable orchestration across heterogeneous systems. These models must balance the performance trade-offs between edge and cloud layers, accounting for limitations such as limited edge storage and computational power against the cloud's inherent network latency. Additionally, such systems demand extensive coordination protocols to ensure data synchronization and consistency, especially in mission-critical or real-time applications such as autonomous driving or industrial IoT [29], [30].

Emerging trends highlight a focus on integrating advanced machine learning methodologies into edge-cloud collaborative models to achieve intelligent resource allocation. Federated learning represents a major step in this direction, utilizing decentralized model training across edge nodes without exposing private datasets to centralized cloud servers, thus maintaining data privacy while improving inference accuracy [31], [32]. Furthermore, energy efficiency is becoming a prominent consideration, with carbon-aware frameworks such as GreenScale leveraging edge-cloud infrastructures to schedule workloads in line with renewable energy availability and carbon minimization goals [33].

In conclusion, edge-cloud collaborative models embody the promise of harmonizing the unique strengths of edge computing and cloud resources, enabling efficient, scalable, and latency-aware systems. While ongoing advancements in dynamic orchestration, federated integration, and energy optimization present promising developments, challenges such as interoperability, resource heterogeneity, and security concerns require further investigation. Future research must focus on adaptive, AI-driven frameworks and standardized architectures to fully unleash the potential of hybrid edge-cloud systems in accommodating the evolving demands of modern applications.

2.4 Communication Protocols and Interconnectivity

The interconnectivity of edge computing systems relies on robust and efficient communication protocols tailored to support low-latency data transmission, adaptive routing, and reliable networking, forming the backbone of overarching edge-cloud collaborations and orchestration strategies. As edge computing decentralizes processing resources closer to data sources, these protocols ensure seamless data exchange across heterogeneous nodes and layers within the ecosystem, enabling the dynamic interactions described in hybrid edge-cloud systems and orchestrated edge infrastructures. This subsection delves into the fundamental mechanisms, challenges, and future directions inherent to communication protocols and interconnectivity in edge computing systems.

Edge computing's emphasis on ultra-low latency is paramount, particularly for latency-critical applications such as autonomous systems, immersive technologies, and industrial IoT. Lightweight, high-speed communication protocols like MQTT (Message Queuing Telemetry Transport) and CoAP (Constrained Application Protocol) have been pivotal due to their resource efficiency and minimal overhead. MQTT, designed for lightweight messaging, excels in real-time reliable data exchanges but struggles with scalability in dense deployment scenarios. Conversely, CoAP extends the benefits of HTTP to resource-constrained environments, offering multicast capabilities and proxying features ideal for device-to-device communication [1]. While efficient in isolated or well-defined setups, these protocols face limitations in highly dynamic, heterogeneous environments, underscoring the growing need for adaptive routing mechanisms to complement them.

Dynamic routing serves as a critical enabler for edge environments, ensuring efficient communication in distributed, multi-hop node architectures. Edge-specific routing approaches, such as hierarchical and adaptive protocols, dynamically allocate paths based on changing network conditions or application-specific quality of service (QoS) requirements. For example, Federated Learning (FL) frameworks rely on routing efficiency to maintain speed and data privacy amid distributed device collaboration. Topology-aware designs, which adapt to fluctuating network hierarchies, provide robust communication reliability in volatile edge ecosystems. Simulations in [34] highlight that incorporating heterogeneity-aware routing can significantly enhance communication stability amidst shifting edge-node dynamics.

Reliability, an inherently complex challenge in edge environments, further compounds the design of communication protocols. Edge nodes face susceptibility to issues like packet loss, fluctuating connectivity, and physical device failures, necessitating fault-tolerant measures such as redundant routing and data replication. While techniques like these improve system reliability, they often introduce resource trade-offs, balancing communication redundancy with resource efficiency [35]. Advanced coding techniques, such as erasure coding, also mitigate data loss risks with minimal overhead, offering balance between fault tolerance and system efficiency in edge-environment communication [36].

The advent of next-generation wireless networks, particularly 5G, marks a transformative shift in interconnectivity paradigms for edge systems. Attributes such as network slicing and ultra-reliable low-latency communication (URLLC) dynamically allocate resources across heterogeneous edge layers, enhancing service responsiveness and continuity. Beyond 5G, the scalability demands of 6G—characterized by dense device connectivity, decentralized intelligence, and AI-driven networking—highlight the necessity for cutting-edge advancements. For instance, dynamic overlay zoning approaches, as discussed in [37], leverage predictive insights to preemptively manage network load and optimize communication pathways, ensuring fluid inter-node coordination at unprecedented scales.

Security within edge communication protocols presents another critical challenge due to the distributed nature and proximity of nodes to end users. Lightweight encryption techniques and trust-based decentralized authentication systems bolster secure interactions while preserving low-latency operations. Blockchain frameworks, for instance, ensure transparent, tamper-proof trust mechanisms within IoT-edge systems without relying on centralized intermediaries [27]. However, effectively integrating such solutions into resource-constrained edge environments, while maintaining throughput and latency requirements, is an area ripe for continued innovation.

In conclusion, communication protocols and interconnectivity strategies are foundational to the seamless operation of edge computing systems, bridging the gap between decentralized edge processes and higher-layer orchestration frameworks. Achieving the trifecta of low latency, reliability, and security, while navigating the constraints of edge devices and heterogeneous architectures, requires continuous innovation. Emerging trends, including 6G networking architectures, AI-enabled optimizations, and energy-conscious designs, hold promise for enhancing the scalability and robustness of edge ecosystems. As interconnected systems advance, future research must converge on the integration of context-awareness, adaptive routing, and decentralized governance into communication designs to meet the evolving demands of edge-enabled applications.

2.5 Orchestration and Management Frameworks

Orchestration and management frameworks are fundamental to the scalability, reliability, and operational efficiency of edge computing systems, especially given the inherently decentralized and heterogeneous nature of these infrastructures. This subsection delves into the strategies, technologies, and trends shaping the orchestration and maintenance of distributed edge environments, with an emphasis on containerization, resource provisioning, fault tolerance, and sustainability.

One of the most important advancements in the orchestration of edge systems is the adoption of container-based frameworks, such as Kubernetes and EdgeX Foundry, which enable lightweight virtualization and microservice deployments tailored to distributed environments. Kubernetes, while originally designed for centralized cloud systems, has been adapted to accommodate the resource-constrained edge by introducing edge-aware extensions that allow for low-latency scheduling, fault-tolerant deployments, and dynamic scaling [38]. Furthermore, frameworks like EdgeX Foundry emphasize modularity, enabling seamless integration of application components across heterogeneous edge nodes. These approaches significantly streamline the deployment and operation of applications, maximizing efficiency while reducing deployment complexity. However, challenges remain, including ensuring consistent orchestration across highly diverse hardware environments and managing context-aware resource allocation under resource-constrained conditions [13].

Resource provisioning in edge systems is an intricate task that requires balancing computational workloads across geographically dispersed nodes. Dynamic provisioning strategies leverage real-time data on resource availability and demand, enabling the allocation of computing, storage, and bandwidth resources to achieve minimal latency and avoid task congestion [39]. Advanced algorithms, such as Lyapunov frameworks and reinforcement learning techniques, allow predictive orchestration to adjust to workload variability [25]. However, the interplay of resource constraints and application-level requirements poses a non-trivial optimization challenge. For example, while heuristic methods improve computational efficiency, they may fail to account for multi-dimensional constraints, such as energy use and task deadlines, that are essential in time-critical applications [40].

Fault tolerance plays a pivotal role in managing the distributed nature of edge systems, where the likelihood of node failure significantly increases due to environmental conditions, node mobility, or limited computational capacity. Techniques like microservice replication and stateless architecture designs enhance system resiliency by maintaining service continuity even in cases of node downtime [41]. Blockchain-based frameworks have also emerged as a promising solution for establishing trust among edge nodes, ensuring secure task migrations, and managing failure recovery through decentralized consensus mechanisms [42]. Nevertheless, the associated computational and temporal overheads of such approaches necessitate further refinement, particularly in lightweight blockchain models for edge nodes with limited resources.

Energy-efficient orchestration frameworks have garnered increasing attention due to the growing need for sustainability in edge computing. Energy-aware schedulers that leverage techniques like Dynamic Voltage and Frequency Scaling (DVFS) and localization of computational tasks based on proximity to renewable energy sources have shown promise in reducing the carbon footprint of edge systems [2]. Notably, the adoption of carbon-aware scheduling algorithms that align computational loads with renewable energy availability enables environmentally responsible orchestration, motivating innovative designs for future edge ecosystems.

Emerging trends emphasize the importance of federated orchestration frameworks to address the collaborative nature of multi-infrastructure edge environments. Federated orchestration tools extend container-based automation and policy enforcement across distributed edge nodes while accounting for latency, resource heterogeneity, and fault domains [43]. These frameworks are integral for multi-tenant environments, such as smart cities or industrial IoT settings, where diverse stakeholders share edge resources and require dynamic, policy-driven allocation mechanisms [44]. Despite these advances, achieving a standard interoperable orchestration architecture remains an open problem, compounded by the need for robust security and privacy measures.

In summary, orchestration and management frameworks enable the seamless deployment, monitoring, and optimization of distributed edge systems, serving as the backbone for efficient and reliable edge operations. While significant strides have been made in containerization, fault tolerance, and energy efficiency, pressing challenges—such as standardization, cross-layer integration, and the incorporation of novel paradigms like semantic communication—point to fertile grounds for future research. Further innovations should aim for a holistic orchestration approach that balances technical feasibilities and application-specific demands while advancing the sustainability and security of the edge computing paradigm.

3 Core Technologies and Computational Models

3.1 Virtualization and Containerization for Edge Computing

Virtualization and containerization have emerged as pivotal technologies in advancing edge computing, offering flexible solutions for efficient resource management and application deployment in constrained environments. These methodologies enable the abstraction and isolation of computing resources, facilitating optimized performance amidst the heterogeneity of edge systems. This subsection explores the foundational principles, technological implementations, and future trajectories of virtualization and containerization within edge infrastructures.

At its core, virtualization allows for the creation of virtual instances of operating systems or applications, enabling multiple workloads to coexist on a single physical host. Traditionally, hypervisor-based virtualization, such as VMware and Xen, has been prevalent; however, its resource overhead, manifested in both memory and processor utilization, makes it less suitable for the resource-constrained nature of edge deployments [45]. To address these inefficiencies, lightweight alternatives like containerization have gained traction. Containers, notably through platforms like Docker, provide application isolation at the process level rather than emulating entire operating systems, resulting in reduced computational overhead and faster initialization times [38], [46].

Containerized environments benefit edge computing in multiple ways, primarily through their ability to ensure portability and scalability across heterogeneous devices. By packaging applications and dependencies in isolated environments, containers minimize conflicts during deployment and ensure consistent performance across varying edge nodes and platforms, critical in diverse and dynamic settings [15]. From a scalability perspective, container orchestration tools such as Kubernetes and its lightweight counterpart K3s have further revolutionized deployment strategies, allowing for automated management of clusters across geographically distributed edge nodes [46]. This capability is particularly valuable in use cases with varying computational intensity, such as IoT-driven healthcare systems and autonomous vehicles, which require dynamic resource scaling to maintain low-latency service delivery [13], [47].

The modularity afforded by containerized architectures aligns seamlessly with the principles of microservices, further enhancing the adaptability of edge computing frameworks. Microservices architecture allows applications to be disaggregated into smaller, independently deployable units, enabling failure isolation and accelerated update cycles. By leveraging this granularity, critical applications such as industrial automation or vehicular systems gain robustness and precision in functionality [48], [49]. Nonetheless, despite these advancements, certain limitations persist. Containers still rely on underlying host operating systems, which may introduce security vulnerabilities if dependencies are inadequately managed. Moreover, the orchestration breadth of lightweight tools like K3s is limited when compared to traditional Kubernetes, potentially constraining performance in large-scale deployments [15].

Energy efficiency poses another critical consideration in edge use cases, particularly for battery-operated or environmentally remote nodes. The computational cost of maintaining virtualized environments can be significant, necessitating novel optimization techniques. Strategies such as clustered container scheduling and resource-efficient container designs, which focus on demand-aware deployment, have shown promise in addressing these energy challenges [50], [51]. Additionally, container-based serverless computing paradigms are increasingly being explored to handle ephemeral, event-driven workloads, leveraging lightweight environments to execute functions on-demand without persistent infrastructure requirements [16].

Emerging trends in containerized edge infrastructures suggest a convergence with cutting-edge technologies such as AI and blockchain. AI-driven orchestration frameworks, for instance, enable predictive scaling through machine learning models, while blockchain enhances trust and reliability by providing decentralized access control mechanisms [10], [52]. However, the joint implementation of these technologies introduces complexity in terms of computational overhead and inter-container communication efficiency, which warrants further research.

In conclusion, virtualization and containerization technologies are the cornerstones of modern edge computing, offering scalable and resource-efficient solutions for diverse application demands. As container platforms continue to evolve, integrating with orchestration tools and AI-based management frameworks, their ability to adapt to the intricate requirements of edge systems will only improve. Future research should focus on addressing the challenges of energy efficiency, security hardening, and orchestration scalability, all while capitalizing on the emerging synergies between containers, AI, and decentralized trust systems to shape resilient and sustainable edge architectures.

3.2 Computation Task Offloading in Edge Environments

Computation task offloading stands as a foundational capability in edge computing, enabling the efficient delegation of resource-intensive tasks from constrained devices to nearby edge nodes or, when necessary, to the cloud. By redistributing workloads, task offloading enhances system performance, reducing latency, conserving energy, and optimizing resource utilization. The inherently dynamic and distributed nature of edge environments demands advanced offloading strategies that carefully balance computational loads across heterogeneous networks while addressing latency, energy, and application-specific constraints.

At the heart of computation offloading lies task partitioning—the process of determining which components of an application should be executed locally versus offloaded to edge or cloud resources. Task partitioning frameworks, such as those employing directed acyclic graphs (DAG) to model task dependencies, analyze computational and data transmission requirements to identify optimal partition points [20]. These frameworks are particularly impactful in mobile edge computing (MEC) systems, where partitioned tasks not only reduce latency but also efficiently allocate computational overhead. However, finding the right level of granularity for partitioning remains a critical challenge, as over-partitioning can lead to excessive communication overhead, whereas under-partitioning might fail to adequately alleviate resource bottlenecks [53].

Offloading decisions often require real-time adaptability, adjusting dynamically to shifting network conditions, load distributions, and resource availabilities. Optimization for computation offloading frequently leverages heuristic algorithms or machine learning models. For instance, context-aware approaches employing reinforcement learning have demonstrated success in optimizing task allocation under multi-modal constraints, including bandwidth, energy efficiency, and computational delays [26]. Although centralized solutions offer significant optimization potential, they can introduce single points of failure and struggle to scale within larger systems. In contrast, distributed methods, including those grounded in game theory or federated optimization, promise greater scalability and resilience through decentralized decision-making [54].

Energy efficiency is a cornerstone consideration in offloading strategies, particularly given the constrained power resources of many edge devices. Balancing energy expenditure between client devices and edge infrastructure requires innovative techniques, such as dynamic voltage and frequency scaling (DVFS) or energy-aware partitioning mechanisms. Advanced methods, including Lyapunov optimization frameworks, dynamically reallocate resources to jointly minimize energy consumption and meet latency constraints in response to real-time demand variations [55]. These energy-focused offloading approaches are especially vital in mission-critical domains such as autonomous driving or telerobotics, where reliable operations depend on striking a balance between low latency and sustainable energy usage [56].

Latency occupies a central role in shaping offloading policies. Delay-sensitive tasks are often better suited for execution on nearby edge nodes, while less urgent workloads can be deferred to distant cloud resources. Delay-constraint management within MEC frameworks encompasses both transmission and computational delays, often relying on hybrid scheduling algorithms to align task deadlines with available resources [57]. Multi-tier hierarchical models, such as the edge-fog-cloud paradigm, offer adaptive architectures that enable tasks to traverse layers of computing resources in accordance with evolving latency and performance requirements [58].

Emerging trends reveal a growing integration of advanced technologies with computational offloading. Artificial intelligence (AI) is increasingly embedded into edge nodes, enabling proactive resource provisioning and predictive task scheduling based on execution patterns [18]. Concurrently, blockchain technologies are being explored to enhance trust and transparency in distributed offloading environments, providing verifiable and tamper-resistant task allocation mechanisms [16].

Despite substantial advancements, computation offloading poses ongoing challenges. A primary concern is maintaining consistency across distributed environments when offloading interdependent tasks, as disruptions in networks or resource failures can result in incomplete task execution. Moreover, the increasing demand for ultra-low latency applications in 5G and beyond intensifies the need for real-time, distributed control mechanisms. This necessitates innovative approaches to synchronize edge and cloud operations effectively [8], [59].

Future directions for research must prioritize the development of real-time task scheduling algorithms grounded in predictive analytics and energy-efficient practices, effectively bridging the gap between theoretical models and real-world implementations. As the edge computing paradigm continues to evolve, standardized benchmarking frameworks and integrated orchestration platforms will be vital to harmonizing diverse offloading strategies with application-specific performance requirements. Computation offloading remains a pivotal enabler for efficient, scalable edge computing deployments, with its continued innovation essential to fulfilling the latency and resource demands of modern applications [38], [60].

3.3 Distributed Storage and Caching Mechanisms

Distributed storage and caching mechanisms are pivotal for achieving the low-latency, high-throughput performance required in edge computing environments. By bringing data management closer to end-users and processes, these mechanisms optimize resource utilization, minimize latency, and enhance fault tolerance. This subsection explores the architectural underpinnings, strategies, and challenges of distributed storage and caching within edge systems, providing an academic analysis of their efficacy and limitations.

At the edge, distributed storage integrates diverse storage mediums such as dynamic random-access memory (DRAM), flash storage, and persistent memory to balance throughput, latency, and cost considerations [3]. Hybrid storage systems often employ tiered architectures in which high-speed, near-edge memory is used for frequently accessed data, while large-scale persistent storage accommodates infrequent data and backups. Such architectural choices significantly reduce data transmission overheads between edge and cloud layers, enabling latency-sensitive applications like autonomous driving and augmented reality [3], [25].

Edge caching solutions further exploit spatial and temporal locality by storing frequently accessed content and computational artifacts closer to users. Techniques such as cooperative caching, where multiple edge nodes synchronize and share cached data, enhance resource efficiency and reduce redundancy, especially in resource-constrained environments [35]. Adaptive caching schemes dynamically modify caching policies based on fluctuating workloads and user demand. For example, least recently used (LRU) strategies combined with machine learning-based predictions have been shown to efficiently handle bursty traffic in edge-assisted IoT settings [61]. Despite their efficacy, these approaches face scalability challenges when managing caches across geographically distributed edge nodes.

Erasure-coded storage systems play a crucial role in ensuring data reliability and fault tolerance in edge environments. By encoding data into smaller fragments and distributing them across multiple edge nodes, erasure coding achieves robust redundancy while minimizing storage overhead compared to traditional replication strategies [26]. However, the computational complexity of encoding and decoding operations impacts system performance, particularly during high request loads. Recent advancements, such as hierarchical encoding schemes, aim to mitigate this bottleneck by leveraging heterogeneous resources at multitier edge frameworks [6].

Consistency remains a core challenge in distributed systems, especially under high churn rates or failure scenarios. Weak consistency models like eventual consistency are often employed in edge systems to reduce synchronization overheads, particularly in real-time applications such as video analytics. However, eventual consistency raises concerns for applications requiring stricter guarantees, such as financial transactions or healthcare monitoring [62]. Hybrid approaches, incorporating quorum-based voting alongside adaptive consistency protocols, are emerging to tackle these trade-offs, promising a balance between performance and reliability.

Emerging trends in distributed storage and caching for edge systems include the incorporation of machine learning to predict data access patterns and adapt caching strategies accordingly. For instance, reinforcement learning algorithms are increasingly utilized to optimize cache allocation and eviction decisions in dynamic edge environments [63]. Additionally, integration with blockchain technologies is being explored to enhance trust, enabling secure data provenance and access control for edge nodes [64].

Looking ahead, system designers must address the challenges of increasing heterogeneity in storage mediums and network topologies, developing universal frameworks that abstract the complexities of hardware diversity. Further research is also required into carbon-aware storage and caching strategies, where deployment decisions are informed by sustainability goals, minimizing energy consumption alongside traditional performance metrics [33]. Building upon these insights will ensure that distributed storage and caching mechanisms remain a cornerstone of resilient and scalable edge computing systems.

3.4 Emerging Hardware for Edge Performance Optimization

The evolution of edge computing relies heavily on specialized hardware innovations that bridge the divide between escalating computational demands and the resource constraints inherent to edge environments. This subsection examines key advancements in hardware that enhance edge system performance, focusing on energy efficiency, computational capacity, and adaptability. By integrating low-power processors, hardware accelerators, neuromorphic computing architectures, and reconfigurable platforms, these technologies enable the execution of latency-critical and resource-intensive tasks in constrained environments.

Low-power processors are foundational to edge optimization, with architectures like ARM-based designs prioritizing energy efficiency and computational density to meet the needs of constrained edge environments. Techniques such as dynamic voltage and frequency scaling (DVFS) further optimize power consumption by adapting to workload variations, balancing energy efficiency and performance [19]. ARM Cortex-M processors, widely deployed in IoT edge devices, exemplify low-power performance optimized for simple data processing tasks. However, their computational limitations necessitate the integration of complementary accelerators for more demanding workloads, including machine learning tasks.

Hardware accelerators such as GPUs (Graphical Processing Units), FPGAs (Field-Programmable Gate Arrays), and ASICs (Application-Specific Integrated Circuits) play a crucial role in extending edge capabilities. GPUs excel in parallel processing and are widely used for tasks like video analytics and natural language processing [65]. However, their power demands often limit their adoption in energy-sensitive edge environments [66]. FPGAs provide a more energy-efficient alternative, offering customizable logic optimized for specific applications like deep learning inference. ASICs, typified by designs like Google's Edge TPU, deliver ultra-low-latency processing with minimal power consumption, making them well-suited for latency-sensitive use cases such as autonomous drones and augmented reality tasks [28], [67].

Emerging neuromorphic computing architectures are pushing the boundaries of edge optimization. Inspired by the human brain, these processors utilize spiking neural networks to enable event-driven computation, achieving significant energy efficiency. Neuromorphic hardware such as IBM's TrueNorth and Intel's Loihi exemplify this approach by focusing on sparse communication and localized processing, making them ideal for power-constrained devices and real-time sensor data processing [65], [68]. Despite their promise, these architectures face barriers to mainstream adoption, as they diverge significantly from conventional neural network methods, requiring software ecosystem redesigns and new training paradigms.

Reconfigurable architectures, such as those leveraging the RISC-V instruction set, offer unparalleled adaptability to evolving edge workloads. Modular extensibility allows developers to tailor systems for specific applications, from cryptographic processes to AI inference optimization [40]. Hybrid platforms combining general-purpose processing with domain-specific accelerators are gaining prominence in industrial IoT and distributed edge networks. However, the integration of reconfigurable architectures with legacy systems presents challenges in scalability and interoperability, highlighting the need for standardization [38].

A notable trend is the convergence of diverse hardware systems to create hybrid platforms that combine CPUs, GPUs, FPGAs, and ASICs, enabling optimal computational heterogeneity in distributed edge deployments. Such configurations strike an effective balance between performance efficiency and energy sustainability. Simultaneously, advancements in edge-native, AI-specific hardware platforms are tailored to frameworks such as federated learning. These systems empower distributed training and inference, enhancing privacy preservation and reducing communication overhead [68], [69].

Despite these innovations, critical challenges persist in the development of robust edge hardware ecosystems. Balancing performance, power consumption, and cooling requirements in dense deployments remains an ongoing trade-off. The need for application-specific accelerators tailored to verticals like medical imaging and real-time robotics escalates complexity in scalability. Sustainability initiatives, such as integrating edge hardware systems with renewable energy sources like solar, are emerging as critical research areas to address power limitations and environmental concerns [19].

In summary, specialized hardware technologies form the backbone of edge computing's evolution, enabling its application across diverse domains. By synergizing low-power processors, custom accelerators, reconfigurable platforms, and emerging paradigms like neuromorphic computing, these technologies meet the demands of complex and latency-sensitive tasks while addressing sustainability. The continued integration of innovative hardware with software frameworks and the prioritization of sustainable practices will play a decisive role in overcoming persistent challenges, enabling a resilient and adaptive edge ecosystem capable of supporting next-generation applications.

3.5 Collaborative and Federated Edge Learning

The integration of machine learning (ML) with edge computing has given rise to Collaborative and Federated Edge Learning (CFEL), an emerging domain that addresses the dual challenges of resource limitations and data privacy in highly distributed and heterogeneous environments. This paradigm innovatively decentralizes the training process, enabling edge devices to collaboratively train models without sharing raw data. By leveraging the proximity of computation to data sources, CFEL optimizes latency, enhances privacy, and reduces communication overhead compared to traditional cloud-based training methods [18], [31].

At the core of CFEL are Federated Learning (FL) frameworks that aggregate locally trained models across distributed edge nodes to build a global model. Unlike conventional centralized learning, FL ensures data sovereignty by transmitting only model updates (e.g., gradients) to a central server, circumventing the need to exchange sensitive raw data. Google's seminal FL framework has paved the way for broad adoption but relies heavily on the presence of a central orchestrator, which may introduce latency bottlenecks and single points of failure [70]. To address these issues, Collaborative Federated Learning (CFL) models have been proposed, enabling decentralized peer-to-peer model aggregation without strict reliance on central coordination, improving robustness and adaptability in dynamic environments [31].

Distributed learning in edge environments imposes unique computational and communication requirements. Resource-aware training represents a critical advancement, as learning algorithms must operate on devices characterized by limited computational power and memory. Lightweight neural networks (e.g., MobileNet and TinyML) and optimization techniques such as quantization and pruning have been employed to achieve efficient local training while maintaining sufficient model accuracy [31]. Nonetheless, edge devices often struggle to balance task execution and training within energy-constrained settings, necessitating innovative methods for dynamic resource allocation [25], [71].

Communication efficiency is another challenge in CFEL due to the overhead associated with iterative model updates across devices or nodes. Techniques such as gradient sparseness, model compression, and dynamic aggregation frequency have been pivotal in mitigating bandwidth constraints while ensuring convergence. Adaptive communication scheduling and the use of over-the-air computation also reduce round-trip delays in CFEL systems [72], [73]. However, these methods often encounter trade-offs between efficiency and accuracy, especially in scenarios with non-independent and identically distributed (non-IID) data across nodes, which can degrade the global model's generalization performance [70].

One promising avenue to enhance CFEL is the integration of edge-native accelerators, including GPUs, TPUs, and AI-specific chipsets, to expedite model training and inference processes. These accelerators facilitate real-time learning and inference directly on edge devices, marking a shift toward edge intelligence [31]. Furthermore, the advent of split and hierarchical learning, where complex sub-tasks of a model are divided among edge devices and proximal servers, shows potential to parallelize computation, reduce single-node strain, and improve scalability [18].

Emerging trends in CFEL include its application to edge services with stringent latency and performance objectives, such as autonomous driving, augmented reality, and healthcare diagnostics. These applications underscore the importance of real-time decision-making and adaptive learning capabilities at the edge. Nonetheless, designing systems that remain robust under mobility and intermittent connectivity, while adhering to privacy constraints, presents significant open research challenges. Blockchain frameworks for decentralized trust management are increasingly being explored to enhance the security and accountability of model updates in CFEL systems, ensuring data integrity across untrusted participants [42].

In summary, Collaborative and Federated Edge Learning exemplifies a transformative shift in distributed intelligence, providing a scalable, privacy-preserving alternative to centralized learning paradigms. Future research directions include standardizing CFEL protocols, addressing non-IID data heterogeneity, and further optimizing resource distribution to support complex edge applications. By strengthening integration with next-generation wireless networks and hardware accelerators, CFEL can further democratize AI capabilities, enabling advanced and real-time analytics directly at the edge. The continued evolution of CFEL will be indispensable in meeting the demands of highly distributed, intelligent systems in 5G and beyond [27], [74].

4 Applications and Use Cases of Edge Computing

4.1 Industrial and Manufacturing Applications

Edge computing has profoundly reshaped industrial and manufacturing sectors by addressing the intricate demands of operational efficiency, real-time decision-making, and process automation. By bringing computational resources closer to data sources on industrial shop floors, edge systems enable industries to achieve reliable, low-latency insights for maintaining productivity and innovation. This subsection investigates how edge computing supports predictive maintenance, real-time operations, supply chain management, and deployment of digital twins, while analyzing the associated technical challenges, trade-offs, and emerging trends.

Enhanced operational efficiency through edge-enabled predictive maintenance has become one of the most impactful applications in industrial environments. Traditional maintenance models rely on reactive or periodic schedules, often leading to equipment underperformance or unanticipated downtimes. Edge-based systems leverage machine sensor data—such as vibration, temperature, and sound—processed locally in near real-time for anomaly detection and prognostics. For instance, employing machine learning models at edge nodes allows for actionable predictions, reducing downtime and extending equipment life by addressing failures before they occur [31], [75]. An essential trade-off in this approach revolves around computational intensity—edge devices must balance energy consumption for AI inference and the volume of sensor data to process, requiring optimal allocation strategies [51], [55].

Edge computing's role in enabling intelligent, real-time factory operations integrates low-latency decision support with localized automation. Edge-powered robotics and Industrial Internet of Things (IIoT) devices deployed on factory floors enhance operational capabilities by enabling rapid analysis of production data. For instance, cameras and sensors on a conveyor belt can execute visual inspections using edge-based AI for defect detection. This localized processing minimizes delays compared to cloud-based approaches, which can suffer from communication bottlenecks [76]. By linking multiple edge nodes with cloud services in hybrid architectures, manufacturers achieve scalable resource pooling while maintaining low latency for high-priority tasks [30]. However, operational heterogeneity remains a challenge, given the diversity of edge devices and the need for standardized communication protocols [13].

Supply chain and logistics optimization is further augmented through the application of edge computing, where dynamic decision-making in material handling, inventory management, and routing unfolds in real-time. Logistics hubs equipped with edge devices facilitate rapid data aggregation from smart sensors and RFID tags for better tracking and management of goods. By pre-processing these data streams locally, companies can generate instant insights, such as predicting demand changes or rerouting shipments to mitigate delays. A crucial challenge here involves ensuring interoperability across geographically dispersed edge nodes, where data consistency and fault tolerance will define system reliability [7]. Solutions using blockchain technologies deployed at the edge, for example, allow for decentralized, secure data-sharing between multiple logistics entities [52].

The deployment of digital twins, which are virtual replicas of physical systems, represents an advanced edge computing application central to Industry 4.0. Digital twins empower manufacturers to simulate, monitor, and optimize industrial processes in real-time while ensuring a closed feedback loop of continuous improvement. Edge computing systems deliver low-latency synchronization of operational conditions from physical machines to their digital counterparts, enabling near-instant predictive analytics and real-time troubleshooting. A significant limitation here is the computational burden of synchronizing high-frequency sensor streams with digital models, which may overwhelm constrained edge nodes. However, innovations such as hierarchical processing paradigms—where subsets of computations are delegated to edge nodes while others execute on regional clouds—offer promising solutions [77]. Furthermore, emerging edge-AI hardware accelerators (e.g., neuromorphic processors) represent a pivotal technology for scaling digital twin deployments [76].

Looking to the future, industrial applications of edge computing must address critical challenges, including interoperability, sustainability, and scalability. Advancements in federated edge learning frameworks can optimize operational insights by securely aggregating distributed data without breaching privacy constraints [78]. Simultaneously, energy-aware orchestration strategies are essential for maintaining sustainability amid rising workload complexities [51]. With these innovations driving Industry 4.0 transformations, edge computing is poised to remain at the forefront of smart manufacturing, balancing technological sophistication with practical operational demands.

4.2 Healthcare and Biomedicine

Edge computing's integration into healthcare and biomedicine is transforming medical service delivery, addressing the critical needs for low-latency responses, resource efficiency, and stringent privacy safeguards. By leveraging edge technologies, healthcare systems are now capable of facilitating real-time patient monitoring, diagnostics, and personalized treatment, particularly benefiting underserved or remote regions. This subsection explores the profound impact of edge computing in healthcare, dissecting its application domains, inherent trade-offs, and emerging trends while positioning it within the broader context of digital healthcare transformation.

One of the most significant applications of edge computing in healthcare is real-time remote patient monitoring. Wearable devices and sensors continuously capture patient vitals such as heart rate, oxygen saturation, and glucose levels, enabling timely assessments and interventions, particularly for chronic conditions. Unlike traditional cloud-reliant approaches, edge-enabled systems process this data locally, thereby minimizing latency and delivering instantaneous alerts during emergencies such as arrhythmias or hypoglycemia [4], [17]. Despite the advances in providing real-time care, optimizing edge resource utilization while managing the energy limitations of wearables persists as a notable challenge.

Medical imaging and diagnostics represent another domain where edge computing fosters transformative benefits. Tasks demanding high computational power, such as analyzing X-rays or CT scans, are increasingly processed on edge nodes located near point-of-care facilities. This localized data processing offers faster diagnostic results compared to traditional cloud-based systems, making it an invaluable resource for underserved settings such as rural clinics or mobile units [5], [6]. However, the requirement for advanced hardware accelerators like GPUs and TPUs to handle such workloads presents challenges in terms of cost-viability and deployment scalability, underlining the need for balanced infrastructure investments.

The Internet of Medical Things (IoMT) ecosystem, comprising wearable devices and interconnected systems, is at the forefront of personalized medicine powered by edge computing. Through localized data processing, edge nodes deliver actionable insights for chronic disease management, fitness tracking, and preventive care while safeguarding privacy-sensitive data—a pivotal consideration under frameworks like GDPR and HIPAA [5], [8]. Despite these merits, achieving robust encryption and maintaining computational efficiency for IoMT devices within resource-constrained environments remain demanding tasks.

Telemedicine, especially in the wake of the challenges posed by the COVID-19 pandemic, exemplifies edge computing's capacity to enhance healthcare accessibility. Edge-enabled nodes facilitate high-quality teleconsultations by supporting real-time video streaming and locally processing diagnostic tools, ensuring service viability even in bandwidth-constrained regions [2], [45]. This mitigates latency and network bottleneck issues typical of centralized solutions. Nevertheless, the heterogeneity of edge devices and varying network conditions necessitate advanced orchestration frameworks to ensure seamless service continuity [46].

While edge computing's contributions to healthcare are groundbreaking, they also surface limitations that emphasize the need for sustained innovation. Resource heterogeneity across edge nodes often constrains scalability, affecting the uniformity of service quality for applications [22], [59]. Furthermore, improving task offloading strategies, optimizing computational energy consumption, and implementing fault-tolerant mechanisms remain critical areas for advancement [13], [50]. Federated learning and edge-based distributed AI offer promising avenues for privacy-compliant, collaborative diagnostic frameworks [18], [31].

Looking forward, the continued convergence of edge computing with technologies such as 6G networks and blockchain promises to revolutionize healthcare systems further. For example, the concept of "edge federations," where diverse resources collaboratively support real-time services, holds immense potential for scaling edge solutions to regions with limited infrastructure or during disaster recovery scenarios [16], [22]. At the same time, addressing objectives of energy sustainability and computational efficiency will be essential, which can be achieved through advancements in energy-aware algorithms, low-power edge systems, and renewable energy integration [19]. Ultimately, edge computing's transformative role in healthcare rests on its ability to balance the imperatives of real-time responsiveness, scalability, resource constraints, and ethical considerations, establishing the foundation for a more connected and patient-centric future.

4.3 Smart Cities and Urban Infrastructure

Edge computing has emerged as a cornerstone technology for smart cities, playing a vital role in managing the increasing complexity of urban infrastructure while addressing the pressing demands for efficiency, resilience, and real-time adaptability. By enabling computational tasks to be processed closer to data-generating sources, edge computing reduces latency, enhances data-driven decision-making, and fosters scalability in the deployment of intelligent urban systems. This subsection delves into the application domains of edge computing across traffic management, public safety, utilities optimization, and environmental monitoring, highlighting key technical advancements and challenges.

Traffic management represents one of the primary applications of edge computing in smart cities, aiming to alleviate congestion and optimize transportation systems. Traditional cloud-based approaches often struggle to meet the latency demands of real-time traffic flow monitoring and signal optimization. In contrast, edge-enabled systems empower intersections and road networks with localized sensors and edge nodes to execute computations at the source. Research has shown that edge systems deployed at traffic intersections can analyze data from cameras, LiDAR, and vehicular sensors, enabling real-time adaptive traffic signal control and emergency vehicle routing [79]. By processing data locally, these systems reduce communication costs and energy consumption associated with centralized cloud frameworks, while also providing rapid contextual responses. However, challenges such as the need for robust fault-tolerance mechanisms and coordination between geographically distributed edge nodes persist.

Public safety and surveillance in urban areas greatly benefit from edge computing, which facilitates real-time video analytics and situational awareness. Crimes, incidents, and disaster responses often depend on split-second decisions, which are constrained by the high latency of cloud-reliant systems. Edge-based surveillance infrastructures process live feeds from high-resolution cameras to detect anomalies, such as unauthorized access or crowding in public spaces [62]. These systems adopt AI-enhanced models hosted at edge locations to perform real-time recognition tasks, including object detection or event classification, while preventing bandwidth saturation by allowing only critical metadata to be sent to the cloud. While such systems dramatically enhance urban security, issues of privacy and ethical considerations around data ownership and algorithmic biases remain major areas of concern [64].

The optimization of urban utilities, including energy distribution and water management, is another critical domain where edge computing demonstrates significant value. Edge nodes integrated with IoT-enabled utility grids facilitate predictive analytics for resource allocation while minimizing operational inefficiencies. For instance, edge-assisted operational intelligence in smart grids enables dynamic load balancing and energy trading, aligning energy consumption with renewable energy availability [5]. Similarly, water networks equipped with edge sensors allow real-time monitoring of leaks and contamination events. The tradeoff between computational overhead and achieving ultralow latency in these systems poses a key challenge, often requiring intelligent resource scheduling [25].

Environmental monitoring powered by edge devices contributes extensively to mitigating the adverse effects of urbanization. Edge-connected air quality sensors, noise detectors, and weather stations collaborate to provide localized insights with minimal latency. Research demonstrates that distributed edge networks can efficiently process data streams to identify pollution hotspots or predict hazardous conditions, guiding policy decisions and urban planning [31]. However, ensuring reliable communication in heterogeneous environments with fluctuating connectivity remains a critical obstacle.

Despite these advancements, emerging trends reveal both opportunities and challenges. The integration of artificial intelligence with edge systems, such as federated learning, represents a promising direction for urban infrastructure, as it enables continuous training of models across distributed nodes while preserving privacy [31]. However, as the scale and density of edge deployments in cities increase, so do concerns over resource constraints, including energy efficiency and the carbon footprint of edge infrastructures [33]. Another pressing issue lies in achieving regulatory compliance for data privacy and security in decentralized architectures, particularly as these systems interact with sensitive civic data under varying jurisdictional requirements [64].

In summary, edge computing offers transformative capabilities for smart cities by enabling efficient and scalable management of critical urban infrastructure. Traffic optimization, public safety, utilities, and environmental monitoring benefit immensely from edge-driven innovation through reduced latency, real-time analytics, and resource-efficient processing. However, challenges related to ethical considerations, privacy safeguarding, and energy sustainability demand sustained research efforts. As future systems adopt 6G networks and advanced AI-augmented edge frameworks, smart city infrastructure will continue to evolve toward higher intelligence, adaptability, and inclusivity.

4.4 Autonomous Systems and Mobility

Advancements in autonomous systems and mobility (ASM) illustrate the pivotal role of real-time data processing and decision-making. Edge computing, by colocating computational resources closer to sensors and actuators, addresses the stringent latency, bandwidth, and reliability requirements inherent to ASM applications. This paradigm represents a transformative shift, enabling seamless operations for autonomous vehicles (AVs), drones, and intelligent transportation systems (ITS), thereby advancing beyond the limitations of traditional cloud-centric models.

Autonomous vehicles, a cornerstone of ASM, rely on continuous and high-speed interactions among perception, decision-making, and control modules to navigate safely through complex and dynamic urban environments. These processes demand ultra-low-latency computation for time-sensitive tasks such as obstacle detection, object classification, and real-time trajectory planning. Traditional cloud-based architectures face significant bottlenecks due to communication delays and the constraints of centralized processing. Edge computing alleviates these limitations by offloading critical tasks to localized edge nodes, positioned closer to vehicles and the surrounding infrastructure. For example, edge-enabled real-time video analytics optimize hazard detection, such as identifying pedestrians or nearby vehicles, enabling quicker reaction times and safer operations [36].

The integration of vehicle-to-everything (V2X) communication further enhances the utility of edge computing in ASM. V2X encompasses vehicle-to-infrastructure (V2I), vehicle-to-vehicle (V2V), and vehicle-to-cloud (V2C) communication paradigms, fostering collaborative and predictive vehicular operations. Edge nodes deployed at road intersections, traffic lights, or cellular base stations aggregate and process data from multiple vehicles to facilitate real-time collision avoidance and adaptive traffic flow management. Furthermore, federated systems enable vehicles to share locally trained AI models, promoting collective intelligence while maintaining stringent data privacy standards [68], [70]. Emerging frameworks, such as the Application-Centric Edge-Cloud Collaboration (ACE) architecture, improve orchestration by leveraging multi-tier computational layers for seamless scalability and mobility-aware data routing [80].

Additionally, drones and unmanned aerial vehicles (UAVs) leverage edge computing to manage complex functionalities, such as real-time path planning, object recognition, and environment mapping. Collaborative, edge-driven approaches distribute workloads across UAV teams, enhancing energy efficiency, extending operational lifespans, and supporting critical missions like search-and-rescue or industrial inspections. These systems achieve heightened reliability through decentralized load-balancing and fault-tolerant mechanisms that mitigate node failures in dense and dynamic operational environments [81]. This localized processing framework ensures high responsiveness for UAV tasks while minimizing dependence on distant cloud servers [82].

Fleet management and logistics further exemplify the benefits of edge-based architectures. These systems utilize edge-enabled analytics to optimize routes, reduce fuel consumption, and plan predictive maintenance schedules. Edge-integrated telemetry systems grant fleet operators real-time oversight of vehicle performance metrics, allowing rapid responses to operational anomalies [38]. Moreover, edge nodes dynamically reallocate resources and prioritize tasks to reduce delays and meet service-level agreements (SLAs), ensuring minimal disruption to logistics operations [83].

Despite its advantages, edge computing for ASM still faces pressing challenges. The inherent heterogeneity of edge devices and vehicular systems complicates orchestration efforts and the establishment of standardized protocols. Achieving interoperability across diverse automotive technologies remains a critical issue. Furthermore, mobility-centric scenarios, such as the dynamic handovers required during V2V and V2I communications, introduce resource allocation conflicts and stability concerns. Addressing these challenges necessitates the development of advanced provisioning algorithms capable of predicting traffic conditions and adaptive computational demands in real time [81].

Emerging trends in ASM highlight key areas of innovation, including the adoption of 6G connectivity, AI-native hardware accelerators, and blockchain-based trust systems. 6G networks are expected to fortify edge systems with ultra-reliable low-latency communication (URLLC) and unprecedented bandwidth, enabling smoother interactions across vast networks of AVs and UAVs [9], [37]. Concurrently, the development of lightweight AI models and neuromorphic processors is anticipated to enhance the energy efficiency of edge-assisted inference tasks [84]. Blockchain technology further supports trustworthy and tamper-resistant data exchange, bolstering the reliability and resilience of cooperative vehicular systems [70].

In conclusion, edge computing serves as a foundational enabler for autonomous systems and mobility by addressing the latency and reliability challenges inherent to these technologies. As ASM ecosystems continue to mature, advancing dynamic edge orchestration, integrating specialized hardware, and establishing highly scalable communication frameworks will be critical to meeting the demands for functionality, safety, and resilience. These innovations promise to unlock unprecedented capabilities in autonomous transportation, logistics, and intelligent mobility solutions.

4.5 Media, Gaming, and Immersive Experiences

Edge computing has emerged as a transformative enabler for media delivery systems, gaming platforms, and immersive technologies such as Augmented Reality (AR), Virtual Reality (VR), and the Metaverse. By addressing the stringent latency, bandwidth, and computational requirements inherent in these domains, edge computing has made possible new levels of interactivity, realism, and personalization that were once unattainable under conventional cloud-centric architectures.

Traditional cloud models have struggled to meet the low-latency demands of real-time gaming and interactive media due to the geographical separation between centralized data centers and end users. In contrast, edge computing brings storage and computation closer to end devices, significantly minimizing network delays and supporting greater Quality of Experience (QoE). For ultra-low latency gaming, edge servers process game state updates, physics simulations, and AI behaviors within milliseconds, providing seamless and responsive user interactions. Studies [4], [12] have also shown that localized rendering and content synchronization at edge nodes reduce bandwidth-intensive data exchanges, opening avenues for large-scale multiplayer gaming experiences.

Immersive technologies like AR and VR further highlight the indispensability of edge computing. These applications require real-time processing and rendering of 3D environments, high-resolution visuals, and continuous tracking of user motion to deliver a lag-free experience. For instance, edge-assisted VR systems exploit proximity to preprocess and stream immersive environments, ensuring frame rates above 90 frames per second and latency tolerance below 20 ms, which are necessary to avoid motion sickness or degraded interactivity [2], [9]. Moreover, AR-enabled services like real-time navigation and industrial training rely on edge support for rapid contextual data fusion and object recognition. These applications leverage edge nodes to reduce offloading latency while integrating sensory data streams, advancing situational responsiveness and efficiency [30], [72].

Furthermore, media delivery systems, exemplified by content streaming platforms, have reaped considerable advancements via edge caching and content delivery optimizations. Through distributed edge storage, content providers can serve geographically localized users with higher throughput and reduced load times. Video streaming services profit from trans-coding capabilities at the edge, adapting content to heterogeneous device capabilities and network conditions in real-time [39]. Emerging trends like personalized content recommendations, inseparable from user-specific preferences and contextual data analytics, further emphasize the role of edge-based platforms in enhancing audience engagement [7].

The Metaverse, envisioned as persistent, collective, and interactive virtual spaces, places unprecedented demands on computational and network infrastructures. Supporting functionalities including 3D avatar rendering, synchronous collaboration, and real-time spatial audio depends on the distributed processing and low-latency synchronization benefits provided by edge nodes. Recent investigations into edge-cloud collaborative models highlight the potential for dynamic workload partitioning and localized computation to balance resource utilization and scalability [71]. However, edge-enablement in the Metaverse still presents open challenges in areas such as consistency maintenance across distributed realms and energy efficiency for sustained operation.

Despite these successes, several limitations persist in deploying edge solutions across these domains. Trade-offs between computation resource allocation and energy efficiency are particularly critical for portable end-user devices, such as VR headsets or handheld gaming consoles, which operate within stringent power constraints [8]. Additionally, heterogeneity in user hardware and network environments complicates the implementation of scalable edge-native designs [15]. Moreover, resource contention among concurrent immersive experiences remains a bottleneck, necessitating sophisticated task orchestration and resource scheduling frameworks [43]. Addressing these multifaceted challenges demands collaborative innovation across hardware and algorithmic domains.

Looking ahead, the convergence of edge computing with Artificial Intelligence (AI) and federated learning techniques is anticipated to catalyze breakthroughs in this sector. AI-optimized edge platforms can predict user interactions with minimal data exchange, fostering efficient resource use and personalized media services [70], [72]. Additionally, advances in multi-access edge computing (MEC) and 6G networks are likely to bolster the technical feasibility of seamless Metaverse applications [85]. As these technologies mature, edge computing will undoubtedly play a foundational role in redefining how users engage with virtual worlds, media content, and gaming ecosystems.

4.6 Agriculture and Environmental Sustainability

Edge computing is revolutionizing agriculture and environmental sustainability by facilitating on-site, real-time data processing, thereby enhancing the efficiency, precision, and responsiveness of various practices. Historically, these sectors have relied on centralized cloud infrastructures, which introduced latency issues and proved less effective in rural or remote areas with inconsistent connectivity. By decentralizing computational resources through edge computing, stakeholders can leverage localized data insights for real-time decision-making, thereby reducing resource wastage, enhancing productivity, and addressing the challenges posed by connectivity limitations.

A prominent application of edge computing in agriculture is precision farming, which harnesses networks of edge-enabled IoT devices to monitor soil conditions, crop health, and environmental variables such as temperature and humidity in real time. For instance, edge-assisted systems process multispectral imaging data from drones to detect indications of crop stress caused by nutrient deficiencies or pest infestations. Unlike cloud-based approaches that rely on large-scale data backhauling, these systems analyze imagery data locally on edge nodes, providing instantaneous, actionable insights to farmers [2], [17]. Additionally, precision farming integrates real-time soil moisture sensors with irrigation control algorithms deployed at the edge to dynamically regulate water distribution, minimizing water waste and lowering energy consumption [48].

Edge computing also brings transformative improvements to livestock management. IoT-powered collars embedded with edge intelligence facilitate continuous monitoring of livestock health and behavior by analyzing data such as movement patterns and body temperature directly at the edge. This enables the immediate detection of potential anomalies, including early signs of diseases or suboptimal husbandry conditions, empowering farmers to implement timely interventions that improve animal welfare and prevent losses [13].

Beyond agriculture, edge computing is indispensable to environmental monitoring and disaster response. Distributed edge networks equipped with sensors are utilized to track critical environmental parameters—such as air and water quality, precipitation, and carbon dioxide levels—while processing data locally to avoid latency-related delays. This capability proves especially valuable in time-sensitive scenarios, such as the early detection of floods, forest fires, or other environmental hazards [86], [87]. Furthermore, edge computing contributes to renewable energy management within agricultural ecosystems. Farms using solar or wind energy benefit from edge-localized energy models that optimize consumption, coordinate local storage, and manage grid-level feed-ins, thus improving energy use efficiency [33].

Despite these advantages, the adoption of edge computing in agriculture and environmental sustainability encounters notable challenges. Edge nodes often face resource constraints, such as limited computational power and energy availability, particularly in rural deployments [8]. Additionally, the heterogeneity of IoT devices and data sources across these sectors necessitates advanced edge orchestration frameworks to integrate diverse streams and manage resource contention effectively. Research efforts leveraging frameworks like FogBus2 and architectures such as EdgePier have showcased scalable methodologies for optimizing real-time data processing at the edge [88], [89].

Emerging developments in artificial intelligence (AI) and machine learning (ML) at the edge present substantial opportunities for advancing these domains further. For instance, edge AI platforms such as OpenEI enable localized inference and predictions for latency-sensitive applications like pest detection or weather anomaly forecasting without relying heavily on cloud connectivity [90]. However, these advancements remain computationally demanding, prompting the need for lightweight and resource-efficient AI models tailored to constrained edge environments [10].

To unlock the full potential of edge computing in agriculture and environmental sustainability, it is essential to address key challenges, including resource sustainability, fault tolerance, and the standardization of edge architectures to ensure interoperability across diverse applications. The development of energy-efficient edge systems—incorporating renewable energy sources and adaptive power management—will be crucial to minimizing the environmental footprint of large-scale edge deployments [19]. In summary, edge computing stands as a transformative enabler for sustainable agriculture and environmental management, offering innovative solutions to drive food security, ecological balance, and long-term resilience.

5 Security, Privacy, and Ethical Considerations

5.1 Data Sovereignty and Privacy Regulations

The rise of edge computing has transformed the way data is processed, stored, and transmitted, particularly by enabling localized data management closer to the source. However, this paradigm shift introduces complex legal and ethical challenges related to data sovereignty and privacy enforcement. Data sovereignty refers to the principle that data is subject to the laws and governance structures of the jurisdiction where it is collected or processed. Given that edge systems operate across geographic boundaries and involve numerous stakeholders, compliance with local and international regulations such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and other region-specific mandates is critical. This subsection delves into how these regulations shape the design of edge computing systems and evaluates privacy-preserving mechanisms to address these challenges effectively.

Localized data governance is a core pillar of data sovereignty, especially in edge computing systems that rely on distributed computational nodes across multiple regions. Edge systems are increasingly required to process and store data within specific jurisdictional boundaries to comply with laws like GDPR, which mandates strict localization of personal data and imposes penalties for non-compliance. For instance, mechanisms such as geo-fencing ensure that data does not leave a designated geographic boundary by dynamically routing data through edge nodes only within permitted regions [1]. Similarly, the CCPA's emphasis on consumer rights over personal data further complicates matters by necessitating systems that allow users to access, delete, and restrict the sharing of collected data. A significant trade-off here is balancing enforcement of these localized governance rules with seamless data flow across nodes, particularly in applications requiring cross-border functionality, such as global IoT systems [27].

To comply with these regulations, privacy-preserving technologies have emerged as vital enablers in edge systems. Differential privacy, for example, introduces statistical noise into aggregated datasets to protect individual data points while still enabling analytical utility. This technique is particularly relevant in sectors like healthcare and smart cities where sensitive data must be anonymized before analysis [8], [11]. Homomorphic encryption, which allows computations to be performed on encrypted data without requiring decryption, is another promising method for securing sensitive information processed at the edge. Despite its significant computational overhead, hardware advancements such as specialized encryption accelerators have made homomorphic encryption more feasible for edge devices [10]. Similarly, technologies like federated learning combine localized data processing with global model aggregation, ensuring sensitive data never leaves the edge node. This decentralized approach has been pivotal in addressing privacy constraints while adhering to regulatory requirements across jurisdictions [31].

However, global edge deployments present substantial regulatory challenges. Jurisdictional conflicts often arise from inconsistencies in privacy standards and data localization laws between regions. For instance, while GDPR emphasizes data protection through explicit user consent and opt-in requirements, other jurisdictions may operate on opt-out frameworks, creating potential compliance conflicts for globally distributed edge networks. Proposed solutions include creating regulatory sandboxes for experimenting with multi-jurisdictional deployments or adopting blockchain-based systems to ensure transparency and traceability for data transactions across borders [52].

Looking ahead, harmonizing regulatory compliance with system efficiency in edge computing demands deeper integration of technical and policy perspectives. Emerging trends, such as carbon-aware computing, highlight the need to design systems not only compliant with privacy regulations but also sustainable and socially responsible [75]. Developing modular privacy-preserving frameworks that adapt to jurisdiction-specific rules would enhance scalability while meeting legal requirements. Furthermore, cross-border collaborations between policymakers, technical architects, and industries could standardize privacy and sovereignty protocols, reducing ambiguity in global edge deployments. As edge computing continues to expand into critical domains like healthcare, autonomous systems, and defense, these innovations will remain central to ensuring user trust and regulatory compliance on a global scale.

5.2 Security Threats and Attack Surface in Edge Computing

The paradigm shift toward edge computing has introduced new opportunities for operational efficiency and latency mitigation but has also significantly expanded the security threat landscape. By decentralizing computational resources closer to end users, edge environments inherently increase the attack surface, exposing distributed nodes to a variety of vulnerabilities. This subsection explores these security threats, highlighting the most prominent attack vectors that impact edge infrastructures. Furthermore, it evaluates existing mitigation mechanisms, emphasizing the inherent trade-offs and identifying opportunities for future advancements to secure these ecosystems, especially in the context of trust and authentication challenges discussed in the subsequent section.

Edge systems are particularly vulnerable to a wide range of external and internal attack vectors due to their distributed architecture and frequent deployment in untrusted or semi-trusted environments. Unlike centralized cloud systems, edge nodes often operate under significant resource constraints, limiting their ability to implement robust security protocols. This makes them attractive targets for attackers. Among the primary attack vectors, Distributed Denial of Service (DDoS) attacks remain one of the most significant threats, capable of overwhelming edge nodes and services with malicious traffic. By leveraging botnets, attackers can compromise multiple devices to amplify the impact of DDoS attacks, as discussed in [8]. Additionally, edge computing infrastructures are increasingly vulnerable to malware propagation and ransomware attacks, where weak authentication or outdated software is exploited by attackers to inject harmful payloads into the system.

Physical and device-level vulnerabilities present unique challenges in edge environments. With their proximity to end users, edge devices are particularly susceptible to physical tampering, unauthorized access, and hardware exploitation. For example, attackers using side-channel techniques that analyze physical-layer signals can extract sensitive data or compromise cryptographic implementations. These risks are further magnified in deployments where physical security measures are inadequate, especially in remote or undersupervised locations [8]. Even basic physical tampering can compromise an edge device's integrity, potentially cascading the effects throughout the interconnected network.

Communication protocols within edge systems are also prone to significant threats, including Man-in-the-Middle (MITM) attacks and data interception. Given that edge computing heavily relies on decentralized decision-making and collaborative data sharing, inter-node communication becomes a critical vulnerability. Lightweight protocols commonly employed in edge-IoT networks, such as MQTT, are often designed with a focus on performance rather than security, leaving systems open to eavesdropping and message manipulation [4]. Attackers can exploit this vulnerability by injecting themselves between communicating nodes, rerouting, corrupting, or manipulating data. This risk is especially pronounced in mobile edge computing systems, where dynamic mobility and frequent handovers add layers of complexity to securing communication channels [4].

To counter these threats, several mitigation strategies have been proposed and analyzed. Lightweight cryptographic techniques specifically designed for resource-constrained edge nodes show promise in protecting communication channels against eavesdropping and message alteration. However, striking a balance between stringent security requirements and the operational limitations of edge nodes remains a major challenge. Trust management mechanisms, including blockchain-based frameworks, have also been investigated to ensure tamper-proof validation of interactions between edge nodes and connected devices [8].

Proactive threat modeling and monitoring are critical to mitigating risks specific to edge computing environments. Real-time monitoring systems, powered by AI-enabled intrusion detection frameworks, offer effective tools for identifying anomalies and predicting potential attack patterns [31]. These systems leverage advanced pattern recognition and statistical models, making them well-suited for securing multi-node interactions. However, their vulnerability to adversarial attacks, wherein malicious actors deliberately manipulate input data to evade detection, underscores the importance of advancing research on resilient AI frameworks.

Looking forward, securing edge computing necessitates a focus on resilience against sophisticated attacks and sustaining service availability. Dynamic resource reallocation mechanisms, such as those explored for resilient fog networks [41], serve as an effective defense against cascading attacks caused by single-node failures. Additionally, federated learning-based collaborative security updates have shown the potential to enhance distributed threat awareness while protecting the privacy of data at the edge [18].

In conclusion, while edge computing provides a solution to latency and efficiency challenges by bringing computational resources closer to data sources, its distributed and heterogeneous nature introduces new attack surfaces requiring innovative and adaptive security solutions. The interplay of resource constraints, mobility challenges, and the increased exposure inherent to edge deployments demands an equilibrium between rigorous security measures and operational performance. Future advances in edge security will depend on scalable, lightweight, and collaborative approaches that address both physical vulnerabilities and evolving communication risks. These innovations will ultimately work in tandem with the trust and authentication mechanisms explored in the following section, contributing to a more secure and resilient edge ecosystem.

5.3 Trust Management and Authentication in Distributed Architectures

Establishing trust and ensuring robust authentication in distributed edge computing architectures is an intricate challenge, compounded by the decentralized, heterogeneous, and resource-constrained nature of edge systems. These systems must navigate dynamically shifting contexts, untrusted operating environments, and diverse participants, from IoT devices to regional edge servers, that may not share pre-established trust relationships. This subsection explores frameworks and mechanisms devised to address these challenges, focusing on Zero-Trust paradigms, distributed trust through blockchain, and lightweight authentication schemes.

Zero-Trust architectures have emerged as a foundational strategy for securing access in distributed environments. Unlike traditional network security paradigms that assume implicit trust within the network boundary, Zero-Trust models require verification of every entity, whether user, device, or application, regardless of its location within or outside the architecture. In edge networks, Zero-Trust frameworks operate at multiple layers, employing strategies such as adaptive authentication protocols and role-based access controls to mitigate risks arising from compromised edge nodes or adversarial actors [8]. By enforcing granular access policies and using telemetry to monitor system integrity, these frameworks improve resilience across nodes, though their implementation is computationally demanding, particularly for resource-constrained edge IoT devices [64].

Blockchain and other distributed ledger technologies (DLTs) have emerged as a transformative approach to establishing decentralized trust across heterogeneous edge nodes. Unlike conventional Public Key Infrastructure (PKI) reliant on centralized certificate authorities, blockchain frameworks enable tamper-proof, transparent trust mechanisms by cryptographically recording interactions and transactions on immutable ledgers. For instance, smart contracts within a blockchain ecosystem can automate authentication processes and establish trust guarantees without relying on intermediaries [64]. For edge environments, the integration of lightweight consensus algorithms such as Delegated Proof-of-Stake (DPoS) or Practical Byzantine Fault Tolerance (PBFT) has reduced latency and resource consumption, making such mechanisms feasible for some latency-sensitive edge use cases [8]. However, the trade-offs between energy usage and latency inherent in DLTs remain an open challenge, particularly in IoT-heavy systems where devices often have limited battery life and computational power [23]. Emerging approaches such as hierarchical blockchains coupled with partitioned ledgers that delegate resource-efficient operations to subchains demonstrate promise in addressing these limitations.

Lightweight and context-aware authentication protocols are particularly essential in edge systems, where resource constraints impose significant limitations on traditional cryptographic methods. Solutions such as elliptic curve cryptography and one-pass mutual authentication techniques have demonstrated reduced computational overhead without sacrificing security, enabling their deployment in environments with severely constrained resources, such as wearable IoT devices [91]. Additionally, context-driven schemes leverage environmental factors like device behavior, location, and activity patterns to authenticate participants dynamically while reducing reliance on data-intensive methods like frequent credential exchanges [2]. However, challenges persist in ensuring reliable context measurements in noisy or adversarial settings, where environmental spoofing attacks may compromise authentication reliability.

Despite these advances, trust management in distributed edge systems remains an ongoing research frontier with significant challenges. Guaranteeing scalability across millions or billions of nodes, accommodating node heterogeneity, mitigating insider threats, and achieving reliable operation in adversarial environments demand further innovation. Emerging techniques leveraging federated learning for collaborative anomaly detection and trustworthiness evaluations between nodes hold potential for improving both security and reliability [31]. Additionally, integrating multidimensional trust metrics, encompassing device integrity, behavioral patterns, and cryptographic credentials, could further enhance trust granularity while reducing false positives in dynamic application domains such as autonomous vehicles [8].

In conclusion, trust management and authentication in edge computing require a multilateral approach that integrates Zero-Trust principles, distributed trust architectures like blockchain, and lightweight yet adaptive authentication protocols. Addressing the scalability and resource optimization challenges inherent in these models will be critical as the edge paradigm continues to expand. Furthermore, early experimental findings suggest promising avenues in leveraging collective intelligence through decentralized and collaborative frameworks to bolster the trustworthiness of edge ecosystems. Future research must focus on balancing security rigor with system efficiency to achieve practical and sustainable solutions for this evolving frontier.

5.4 Privacy-Ethical Dilemmas in Edge Computing

The widespread adoption of edge computing presents complex privacy and ethical dilemmas due to its decentralized nature. By enabling data processing closer to end-users, edge systems achieve localized decision-making, reduced latency, and improved bandwidth efficiency. However, these advantages also magnify ethical challenges concerning privacy preservation, algorithmic transparency, fairness, and data autonomy. Addressing these challenges demands a comprehensive understanding of how edge computing redefines the socio-technical landscape of decentralized decision-making.

Privacy emerges as a foremost ethical concern because data in edge environments is processed and stored across heterogeneous, geographically dispersed devices, often with varying levels of security and oversight. Unlike centralized cloud systems, where a unified entity enforces robust security protocols, edge computing operates with fragmented responsibility models, creating increased vulnerability to data breaches and unauthorized access. Edge data frequently resides in resource-constrained devices, such as IoT sensors, which are susceptible to physical tampering and cyberattacks. Federated learning approaches have been proposed as a privacy-preserving mechanism by eliminating the need to transfer raw data to central servers. However, federated learning has its own limitations, such as communication overhead and risks of model parameter leakage, potentially exposing sensitive information [68], [70]. Similarly, advanced techniques like differential privacy and homomorphic encryption can address specific privacy issues but introduce significant computational overhead, which resource-starved edge devices may struggle to handle effectively [19].

Alongside privacy concerns, algorithmic bias in decentralized AI systems poses a critical and underexplored ethical challenge. Many edge-enabled AI algorithms, especially those used in sensitive domains such as surveillance or personalized healthcare, depend on locally collected device data. This localized training approach often lacks the diversity of larger, centralized datasets, increasing the risk of reinforcing systemic biases [65], [84]. For instance, models trained on data from specific devices may exhibit discriminatory tendencies when applied to broader populations. Although distributed learning frameworks like collaborative federated learning aim to protect privacy through data decentralization, they cannot inherently guarantee representativeness or fairness across diverse edge nodes [68]. These biases raise ethical questions about accountability—who bears the responsibility for the inequitable decisions of an edge-based system? Such concerns emphasize the necessity for transparent deployment standards that can mitigate and address potential biases.

Transparency and explainability in edge-based decision-making are indispensable for fostering trust and ethical compliance. The decentralized and distributed nature of edge systems introduces substantial challenges in tracing decision-making processes back to specific models or datasets. Solutions like blockchain-enabled provenance tracking have emerged as tools to provide a decentralized means of ensuring transparency and traceability [37]. However, these solutions often introduce trade-offs, such as higher computational and communication burdens, which can conflict with the latency-sensitive demands of applications like autonomous vehicles [1]. Similarly, the integration of explainable AI (XAI) frameworks into edge systems remains an obstacle due to the resource limitations of edge devices. Generating interpretable outputs in real-time scenarios often strains local computational resources, hampering seamless deployment.

The ethical dilemmas in edge computing are deeply intertwined with questions of data ownership and control. Edge nodes process vast amounts of user-generated data, ranging from biometric details in wearable devices to environmental measurements in smart cities. These systems leverage such data to enable localized decision-making, yet ethical tensions arise regarding whether users retain autonomy over their own data. Concepts like data marketplaces, where users can manage and potentially monetize their data, offer an intriguing pathway to resolving these tensions, but their implementation is still in its infancy and may invite exploitation without strong regulatory safeguards [36]. Additionally, the distributed nature of edge environments complicates privacy compliance across jurisdictions, as nodes operating across diverse regions may simultaneously have to adhere to legally disparate frameworks, such as GDPR and CCPA [27].

As edge computing continues to evolve, research must focus on harmonizing the technological benefits of edge architectures with robust ethical frameworks. Addressing privacy-ethical challenges will require interdisciplinary collaboration between engineers, lawyers, and ethicists to design systems that optimize fairness, preserve privacy, and engender trust. Privacy-centric AI techniques tailored to the resource constraints of edge architectures, along with adaptive models for resource allocation, represent promising directions for mitigating these ethical concerns. Emerging trends such as explainable edge AI and blockchain-augmented transparency frameworks are poised to redefine the balance between usability and ethical accountability. Ultimately, to realize the transformative potential of edge computing, stakeholders must holistically address privacy-ethical dilemmas while ensuring equitable access, transparency, and social responsibility at the core of this rapidly expanding paradigm.

5.5 Intrusion Detection and Real-Time Threat Responses

Intrusion detection and real-time threat response have become imperative in edge computing, where the distributed and resource-constrained nature of nodes introduces complex attack vectors and increases the risk of exploitation. Unlike centralized systems, edge environments are inherently decentralized, exposing a wide, heterogeneous attack surface. This subsection explores contemporary methodologies for intrusion detection and mitigation, emphasizing anomaly detection models, collaborative threat intelligence, and adaptive incident response mechanisms.

Intrusion detection in edge systems operates under stringent latency and computational constraints, necessitating lightweight anomaly detection models. Machine learning (ML)-based approaches, particularly those leveraging both supervised and unsupervised learning, have emerged as primary enablers of real-time detection. Supervised learning models such as support vector machines (SVMs), often trained on historical attack patterns, are highly effective at classifying known threats but struggle with zero-day attacks due to reliance on labeled datasets. Conversely, unsupervised techniques, including clustering algorithms and deep autoencoders, excel in recognizing abnormal behaviors indicative of unknown intrusions, making them indispensable for edge-native intrusion detection systems (IDS) [92], [93].

One promising approach is the integration of federated learning (FL) into anomaly detection frameworks. FL enables decentralized training of detection models across diverse edge nodes without transferring raw data to a central location. This paradigm enhances privacy while ensuring that the intrusion detection models remain relevant across heterogeneous environments. For instance, collaborative FL frameworks allow edge devices to consolidate local threat intelligence, improving detection accuracy while preserving privacy [31], [70]. However, challenges stemming from model drift across nodes, computational heterogeneity, and communication overhead remain critical obstacles.

Collaborative threat intelligence within edge ecosystems plays a vital role in real-time intrusion management. Techniques leveraging blockchain-based logging and consensus protocols have emerged to provide a tamper-proof mechanism for sharing threat insights across nodes, thereby improving collective resilience. Blockchain systems offer decentralized trust management, reducing reliance on centralized network controllers vulnerable to single points of failure [42], [93]. However, these systems often face scalability concerns, as the computational cost of blockchain validation may conflict with the low-latency requirements of edge systems. Lightweight cryptographic techniques and selective consensus mechanisms are under research to address these trade-offs.

Real-time threat responses require the deployment of incident management frameworks capable of isolating compromised nodes, updating system configurations, and restoring integrity with minimal impact on system performance. Dynamic threat response solutions often utilize graph-based models to localize the spread of attacks within edge networks, allowing for rapid containment [92]. For example, zero-trust architectures adapted for edge systems ensure that every node must continually authenticate its communications, reducing lateral movement opportunities for attackers [93]. Automated response mechanisms, such as patch management systems that leverage predictive algorithms to proactively detect and resolve software vulnerabilities, also play an essential role in maintaining system resilience.

Despite significant advances, open challenges persist. The high variability of attack types across application domains, ranging from autonomous vehicles to industrial IoT, demands domain-specific intrusion detection models [8]. Additionally, limited computational resources at edge nodes constrain the deployment of resource-intensive detection algorithms, necessitating the development of energy-efficient methods such as neuromorphic computing and lightweight AI accelerators [18], [31]. Future research should prioritize hybrid approaches that integrate anomaly detection with predictive analytics, enabling preemptive responses to emerging threats.

The growing complexity of edge environments, coupled with their dynamic and distributed nature, underscores the need for advanced and adaptive frameworks for intrusion detection and threat response. By synthesizing innovations in ML, federated intelligence, and blockchain, along with energy-efficient implementations, edge systems can achieve scalable, robust, and privacy-preserving security solutions. As edge computing continues to evolve, establishing standardized practices and cross-domain models for intrusion handling will play an essential role in ensuring secure and reliable systems across diverse applications.

5.6 Sustainability and Energy-Efficient Security

As edge computing systems scale to meet the growing demands of latency-sensitive and resource-constrained applications, the sustainability of security mechanisms emerges as a critical consideration. The energy limitations of edge devices, combined with the computational overhead of traditional security mechanisms, necessitate innovative approaches that balance robust security with resource efficiency. Building on advancements in adaptive security design, this subsection explores state-of-the-art approaches to energy-efficient security in edge computing, evaluates associated trade-offs, and outlines emerging trends and research challenges shaping this dynamic field.

Energy-efficient encryption techniques represent a cornerstone of sustainable security strategies for edge systems. Lightweight cryptographic algorithms, such as elliptic curve cryptography (ECC) and optimized variations of symmetric encryption algorithms, are increasingly employed due to their lower computational requirements compared to conventional methods like RSA. For example, ECC delivers equivalent security levels with significantly smaller key sizes, reducing energy consumption during encryption and decryption processes [8]. Similarly, energy-aware adaptations of block ciphers, such as Advanced Encryption Standard (AES), focus on tailoring operational complexity to the capabilities of individual edge devices, mitigating energy overuse [55]. However, these techniques must navigate the trade-offs between computational simplicity and the level of security provided, especially when confronting sophisticated adversaries capable of exploiting potential vulnerabilities.

Beyond encryption, sustainable threat mitigation strategies tailored for energy-constrained environments have seen significant advancements. Intelligent resource allocation algorithms, often leveraging reinforcement learning frameworks or heuristic-based optimization strategies, are being deployed to efficiently distribute security tasks across edge nodes. Such approaches aim to simultaneously balance computational effort, energy consumption, and security robustness. Notably, task scheduling and security provisioning techniques guided by Lyapunov optimization frameworks dynamically allocate resources to ensure energy-efficient security operations [55]. Nonetheless, these methods often require extensive computational profiling, and their scalability can become a challenge within larger, heterogeneous edge networks, particularly when dynamic workloads cause frequent reconfigurations.

Integrating renewable energy sources into edge systems has also gained traction as a promising avenue for aligning sustainability goals with secure operations. Frameworks such as GreenScale propose carbon-aware scheduling policies, wherein computationally intensive security tasks are executed during periods of abundant renewable energy availability to reduce environmental impact [33]. However, the intermittent and unpredictable nature of renewable energy sources introduces complexity, necessitating predictive models to optimize energy availability and workload planning while maintaining robust security guarantees.

Achieving energy-efficient security further necessitates a focus on reconciling performance with reliable security measures. Mechanisms such as intrusion detection systems (IDS) and real-time threat response are often power-intensive, presenting challenges for their deployment on edge nodes with constrained energy reserves. Recent advances in anomaly-based IDS leverage machine learning algorithms explicitly optimized for edge environments, achieving real-time threat identification with minimal energy overhead [10]. Additionally, collaborative frameworks for sharing threat intelligence across distributed edge nodes reduce the computational burden on individual devices by avoiding redundant processing. Yet, these systems require robust governance mechanisms to prevent the propagation of outdated or erroneous threat data.

Emerging technological innovations promise to enhance energy efficiency across the edge security landscape. Neuromorphic processors and edge AI accelerators have demonstrated capabilities in reducing power consumption for critical tasks such as behavioral monitoring and anomaly detection [10]. Furthermore, blockchain frameworks are increasingly explored for not only enabling decentralized trust across resource-constrained environments but also for incorporating energy-efficient consensus algorithms that reduce computational waste [42].

Despite these progressions, several open challenges remain. Lightweight security models must continuously adapt to evolving attack techniques to sustain their resilience over time. The dynamic and heterogeneous nature of edge environments—characterized by mobility, variable workloads, and diverse device capabilities—complicates predictive modeling for energy-efficient threat mitigation. Efforts to establish standardized metrics to assess and balance energy consumption against security robustness are still nascent and require interdisciplinary attention. Moreover, the integration of renewable energy-aware frameworks with existing edge security solutions must be closely examined to ensure their feasibility and reliability in real-world applications.

Looking forward, the convergence of AI-based resource management strategies with energy-efficient security innovations offers a compelling direction for the future of sustainable edge computing. By leveraging adaptive, context-aware models to dynamically optimize security overheads, next-generation systems can strike an effective balance between operational sustainability and data protection. These ongoing advancements underscore the importance of multidisciplinary research in addressing the complex interplay between security, energy efficiency, and the infrastructural demands of edge computing.

6 Optimization Strategies and Resource Management

6.1 Resource Allocation and Scheduling Mechanisms

Efficient resource allocation and scheduling in edge computing environments serve as critical enablers of system performance, ensuring optimal utilization of constrained computational, storage, and networking resources while maintaining the quality of service (QoS) and addressing latency-sensitive requirements. This subsection examines the diverse methodologies for resource allocation and scheduling, addressing the trade-offs, challenges, and emerging trends in this domain.

Edge environments, characterized by their distributed, heterogeneous, and resource-constrained nature, demand real-time and adaptive mechanisms to effectively allocate tasks across nodes. Traditional centralized approaches, effective in cloud environments, are often inadequate in edge contexts due to their latency and scalability limitations. Consequently, distributed and decentralized resource management strategies have emerged as a focal area. Stochastic optimization techniques, such as Markov Decision Process (MDP)-based frameworks, play a pivotal role in dynamic settings, formulating scheduling policies that respond to fluctuating workloads and network conditions. For instance, delay-optimal scheduling policies leveraging MDPs have shown significant potential in minimizing task execution delays and power consumption, balancing transmission and computation costs dynamically [94].

Task offloading, a cornerstone of resource allocation in edge computing, has seen extensive research investigating decision mechanisms for distributing computational tasks between resource-constrained devices and nearby edge nodes. Should tasks be entirely processed locally, partially offloaded, or executed in the cloud? Various heuristic, evolutionary, and machine learning-based frameworks have been proposed to answer this question. Edge-centric implementations that integrate computation offloading with resource-aware scheduling leverage algorithms like stochastic gradient descent or deep reinforcement learning for task partitioning and resource allocation. These methods achieve near-optimal trade-offs between energy consumption and task latency [47], [95].

In addition to offloading, multi-resource optimization frameworks simultaneously address the constraints of processing power, memory, bandwidth, and energy availability. Joint computation-communication models, frequently modeled as mixed-integer nonlinear programming (MINLP) problems, optimize edge resources while meeting real-time application demands. Techniques leveraging convex relaxations and gradient-based optimizations, such as BSUM or penalty convex-concave procedures, have demonstrated improvements in system efficiency compared to traditional allocation methods [51], [96]. These methods highlight trade-offs between computational complexity and real-world applicability, raising questions about scalability in geographically dispersed edge ecosystems.

Furthermore, multi-tenant resource sharing has gained momentum as edge platforms increasingly support diverse applications with competing demands. Mechanism design theories, including auction-based approaches, have been proposed to mediate fair and efficient allocation among tenants. For example, combinatorial auctions ensure Pareto-efficient resource allocation while adhering to bounded latency constraints, albeit at the cost of increased computational overhead [97].

Emerging trends point to the integration of AI-driven frameworks for real-time resource orchestration. Reinforcement learning-based approaches, which adaptively learn task scheduling policies, exhibit promise in handling unpredictable workloads while optimizing resource usage. The adoption of federated learning models further enhances collaborative allocation decisions across geographically distributed nodes, preserving data locality and privacy [10]. Additionally, blockchain technologies are emerging as trust-enabling mechanisms to facilitate decentralized cooperation and secure resource allocation in multi-stakeholder environments [52].

Despite these advancements, challenges persist. The dynamic mobility of edge devices introduces unpredictability, complicating task scheduling decisions. Mobility-aware adaptations, for example, mobility-centric resource load balancing, aim to reduce service disruptions at the expense of computational overhead [98]. Moreover, the inherent heterogeneity of edge hardware introduces disparities in performance, further complicating resource pooling. Hybrid scheduling frameworks that account for device capabilities, energy constraints, and SLA requirements will therefore be critical in addressing these challenges.

As edge computing adoption intensifies across IoT, smart cities, and industrial domains, the need for scalable, low-latency, and energy-efficient allocation mechanisms remains paramount. Future investigations should emphasize decentralized approaches that leverage predictive modeling and quantum-inspired optimization algorithms to tackle the NP-hard nature of complex scheduling problems. Ultimately, the confluence of AI, federated resource sharing, and trust computation frameworks represents a promising frontier for advancing resource allocation methodologies in edge computing ecosystems.

6.2 Energy-Efficient Architectures and Designs

Energy efficiency plays a pivotal role in the architecture and operational strategies of edge computing systems, particularly given the increasing demand for computational resources at the network edge and the constraints posed by limited energy availability. As edge environments aim to balance high performance with resource conservation, this subsection explores energy-efficient designs and resource management techniques, assesses comparative approaches, and highlights challenges and emerging trends in achieving sustainable edge computing.

A fundamental aspect of energy efficiency lies in hardware optimization tailored for low power consumption. Devices equipped with ARM-based low-power processors and specialized accelerators, such as GPUs and TPUs, effectively reduce energy requirements while maintaining robust computational capabilities. Additionally, emerging architectures like neuromorphic computing and domain-specific accelerators have demonstrated promise in managing energy-intensive tasks such as AI inference with minimal power usage [5], [19]. However, while these innovations excel in task-specific scenarios, they often face challenges related to reduced generalizability and high upfront costs, limiting their widespread adoption.

Dynamic power scaling techniques, such as Dynamic Voltage and Frequency Scaling (DVFS), further augment energy efficiency by adjusting processor voltage and frequency to match varying workload demands. Widely adopted across edge computing environments, DVFS helps balance computational effectiveness with power savings, particularly for latency-tolerant applications [17]. However, these approaches must contend with trade-offs in processing speed and Quality of Service (QoS). For time-critical systems like autonomous vehicles and medical monitoring, predictive algorithms are needed to maintain real-time responsiveness while minimizing energy consumption.

Complementing these methods are carbon-aware scheduling strategies, which align computational workloads with periods of high renewable energy availability or low carbon intensity. These systems effectively integrate distributed scheduling with renewable energy forecasting to reduce the environmental footprint of edge operations, while optimizing execution latency in heterogeneous environments [19], [58]. Despite their potential, practical challenges in achieving real-time renewable energy predictions and coordinating geographically distributed edge nodes persist, underscoring the need for further development.

Energy harvesting innovations further enhance sustainable edge designs by capturing ambient energy sources—such as solar, kinetic, or thermal energy—and optimizing their use via power-aware scheduling and energy storage techniques. These solutions have gained particular relevance in remote or off-grid applications, such as precision agriculture and environmental monitoring, where extended operational lifespans are essential [19], [99]. Nevertheless, current energy harvesting technologies encounter limitations in efficiency and scalability, particularly when applied to high-load computational scenarios, highlighting the ongoing need for more advanced breakthroughs.

Distributed and collaborative energy-efficient practices are also gaining traction, with resource-sharing models playing a pivotal role in minimizing redundancy and extending device lifespans. Federated edge architectures, for example, enable dynamic resource pooling across nodes through energy-optimized orchestration algorithms, reducing the energy costs associated with centralized data transmission [22], [25]. However, the security and reliability of these distributed systems, particularly in adversarial environments, remain challenging areas requiring robust solutions.

Looking to the future, machine learning's predictive capabilities offer promising avenues for precision resource forecasting in energy-efficient edge systems. By training models on historical workload and energy consumption data, computational tasks can be dynamically optimized to conserve energy without sacrificing performance [20], [60]. Additionally, blockchain-based frameworks provide opportunities for decentralized energy trading among edge nodes, fostering cooperative resource sharing while maintaining trust and accountability [16], [31].

Energy-efficient architectures and strategies form a cornerstone of sustainable edge computing, balancing power conservation with optimal system performance. However, addressing the intricate challenges of scalability, cost, and computational efficiency remains critical. Future research should prioritize interdisciplinary approaches—combining quantum-inspired algorithms, renewable energy integration, and decentralized decision-making frameworks—to meet the evolving demands of energy-efficient edge environments, while aligning with broader sustainability goals.

6.3 Communication Optimization in Distributed Systems

Communication optimization in distributed edge computing environments plays a pivotal role in ensuring the efficiency, scalability, and reliability of edge systems. The inherent geographic distribution of edge nodes presents numerous challenges, including high latency, limited bandwidth, and vulnerability to network disruptions. This subsection delves into optimization methodologies tailored to address these challenges by enhancing data exchange efficiency in resource-constrained and dynamic settings.

Efficient data aggregation techniques are instrumental in mitigating the communication overhead in distributed systems. By consolidating redundant or irrelevant information, these techniques reduce the volume of data transmitted between edge devices and neighboring nodes or the cloud. For instance, methods such as compression, deduplication, and hierarchical aggregation have demonstrated significant potential in minimizing bandwidth utilization while preserving the integrity of information [3], [100]. Cooperative aggregation schemes, wherein local edge nodes share preprocessing responsibilities, further enhance the system's resilience against network congestion by leveraging localized computation [17]. However, such approaches must strike a balance between computational overhead at the edge and the resultant communication savings.

Low-latency communication protocols are central to achieving time-sensitive data transfer in geographically dispersed edge networks. Protocols such as MQTT and CoAP, designed for lightweight communication, are widely adopted in edge systems to reduce protocol overhead and expedite data transmission [38]. Furthermore, advancements in adaptive routing strategies, such as Software-Defined Networking (SDN) and shortest-path algorithms, dynamically optimize data flow based on current network conditions. These strategies are particularly effective in enhancing responsiveness in latency-critical applications like autonomous driving and augmented reality [78]. However, the limited scalability of certain protocols and the complexities introduced by multi-hop communication across edge nodes remain key limitations, necessitating further research.

Task offloading with network constraints is another critical area of optimization, where edge devices offload computation-intensive tasks to remote nodes or the cloud while accounting for communication delays and bandwidth availability. Multi-objective optimization frameworks leveraging Lyapunov-based scheduling [25] or heuristic algorithms [20] have demonstrated remarkable efficacy. These solutions optimize task placement by dynamically assessing network bandwidth, computational capacity, and latency requirements. Nonetheless, the trade-offs inherent in task partitioning—balancing energy consumption, bandwidth usage, and processing latency—continue to challenge the deployment of universally optimal frameworks [101].

Hierarchical edge-cloud interactions further augment communication optimization by introducing multi-layered architectures that adaptively offload tasks between edge nodes and central cloud data centers. By localizing the processing of less resource-intensive tasks, such architectures significantly alleviate network congestion while reserving cloud resources for large-scale data analytics and storage [27]. Federated edge-cloud ecosystems enhance performance further by enabling collaborative processing between multiple edge nodes and cloud servers, ensuring robustness and scalability in high-demand environments [30]. However, integrating hierarchical models with dynamic workload conditions across heterogeneous systems remains non-trivial due to the complexities in synchronization and coordination.

Emerging trends emphasize the integration of advanced machine learning models for predictive communication optimization. AI-driven techniques can predict network states and adapt data flows to preempt congestion, reducing latency and enhancing fault-tolerance [31], [72]. Moreover, blockchain-based communication frameworks are being explored to establish decentralized, trust-worthy coordination across distributed systems while maintaining communication reliability [64]. Despite the promise of these methods, challenges related to computational overhead and scalability must be addressed to realize their full potential.

Future research directions in communication optimization should center on hybrid approaches that integrate advanced AI, dynamic routing strategies, and resource-aware task offloading. These solutions must also address the heterogeneity of edge environments and support seamless interoperability. As latency-critical applications proliferate, ensuring robust and low-latency communication under varying network conditions will solidify communication optimization as a cornerstone of next-generation edge computing systems.

6.4 QoS and SLA Management in Edge Systems

Efficient Quality of Service (QoS) and Service Level Agreement (SLA) management are crucial for ensuring the reliability, scalability, and performance of edge computing systems, particularly given their heterogeneous, geographically distributed, and dynamic nature. Unlike centralized cloud environments, edge systems must contend with fluctuating resource availability, diverse workload demands, and stringent latency constraints. This subsection explores key methodologies, frameworks, and challenges for maintaining QoS and SLA compliance in edge ecosystems, emphasizing the trade-offs and emerging research directions that align with the overarching goals of communication optimization and adaptive scalability.

QoS management focuses on ensuring metrics such as latency, throughput, availability, and fault tolerance meet user-defined performance expectations. SLA-aware resource allocation frameworks have been widely adopted, utilizing mathematical optimization, AI-driven prediction, and rule-based heuristics. Recent research [102] demonstrates how multi-objective scheduling algorithms can balance computation costs, energy consumption, and delay violations, providing promising solutions for joint optimization in heterogeneous systems. However, the dynamic and decentralized nature of edge environments complicates resource stability and reduces prediction accuracy, posing challenges for integrating such frameworks into large-scale deployments.

Dynamic resource provisioning policies tailored to ensure SLA adherence play a pivotal role in bridging variability in workload demands. For instance, SLA-aware frameworks leveraging Lyapunov optimization techniques [36] provide probabilistic guarantees to manage computational queue lengths, ensuring low-latency performance for critical applications. Reinforcement learning (RL) has also emerged as a valuable approach for adaptive resource reallocation, enabling these systems to handle fluctuations in workload with minimal manual intervention. RL frameworks [83] enrich resource orchestration by logically clustering heterogeneous resources, thus simplifying SLA-aware decisions. Nonetheless, scalability challenges persist, especially for computationally expensive algorithms in expansive edge networks.

Real-time monitoring and telemetry are essential to ensuring continuous QoS compliance and mitigating potential SLA violations. State-of-the-art orchestration engines, such as Kubernetes extensions [103], incorporate telemetry data, including resource utilization and network conditions, into their decision-making processes, effectively reducing SLA violations. Complementary research [24] has introduced open-source benchmarking tools for runtime analysis, further enabling real-time anomaly detection. However, achieving low-latency failure detection and rapid reaction in real-world implementations remains a substantial challenge.

During resource contention or overload events, SLA violations can be mitigated using graceful service degradation mechanisms, which prioritize minimal acceptable QoS levels while avoiding complete service failure. Strategies like hierarchical resource pooling and fallback offloading [35] help redirect overflow workloads from edge nodes to regional or centralized cloud layers. These mechanisms enhance system resilience during extreme conditions but may compromise latency guarantees for time-sensitive services such as autonomous systems or industrial IoT applications.

AI-driven predictive models offer another crucial dimension to QoS management. Emerging frameworks use historical telemetry data for forecasting possible SLA violations and dynamically adjusting allocation strategies. For instance, federated learning models applied to SLA adaptation [68] enable decentralized decision-making while preserving user data privacy. Empirical evidence suggests that such predictive frameworks significantly reduce SLA violations compared to traditional centralized approaches. However, their deployment in energy-constrained edge devices presents ongoing challenges, especially in achieving an optimal trade-off between prediction computations and resource efficiency.

Despite significant advances, QoS and SLA management in edge computing continue to face numerous challenges, including standardization, fault recovery, and balancing energy efficiency with SLA compliance. As the scope of edge systems expands with new technologies such as 6G and Internet-of-Everything (IoE), dynamic orchestration frameworks [37] and predictive SLA contracts that can anticipate evolving user needs will become indispensable. Additionally, decentralized trust mechanisms for SLA validation [70] and quantum-inspired models for nonlinear optimization [84] represent promising frontiers for addressing complex QoS challenges.

In conclusion, effective QoS and SLA management serves as a cornerstone for enabling reliable and scalable edge computing systems, complementing efforts in communication optimization and dynamic resource adaptation. Integrating theoretical advancements, such as predictive analytics and distributed orchestration, with robust and scalable real-world implementations will be essential in shaping future edge ecosystems capable of meeting the demands of next-generation applications.

6.5 Scalability and Dynamic Resource Adaptation

Scalability and dynamic resource adaptation are fundamental elements for ensuring the resilience and efficiency of edge computing systems, particularly under the heterogeneous and dynamic nature of user demands and edge environments. As edge systems strive to handle fluctuating workloads and enable seamless resource mobility across diverse devices, a multifaceted approach to scaling and adapting resources is paramount.

Horizontal and vertical scaling mechanisms form the bedrock of scalability in edge computing. Horizontal scaling refers to the expansion of system capacity through the addition of new edge nodes, enabling a distributed handling of increasing workloads. This approach is particularly common in geographically dispersed nodes, ensuring localized data processing and reduced latency [4], [45]. Vertical scaling, on the other hand, focuses on enhancing the computational capacity of existing edge nodes, including adding processing units or adjusting task prioritization policies dynamically. While both strategies are effective, horizontal scaling excels in maintaining low-latency communication for geographically diverse tasks, whereas vertical scaling is more suitable for resource-intensive but clustered workloads. However, the trade-off lies in the higher coordination overhead of horizontal scaling compared to the hardware-dependent constraints of vertical scaling [12].

Dynamic orchestration frameworks are essential for enabling scalability across heterogeneous devices. Context-aware orchestration, for instance, leverages real-time environmental and workload context to prioritize resource allocation and ensure application-specific performance. Techniques like container-based deployment, using lightweight virtualization frameworks such as Docker, have demonstrated their potential in scaling edge systems quickly and efficiently while maintaining resource isolation [15], [104]. Moreover, microservices architectures allow the modular adaptation of individual services, particularly under variable user loads, enabling decentralized and fault-tolerant scaling. However, challenges arise in maintaining inter-service communication efficiency and fault resilience, especially during rapid scalability events [13].

Mobile device integration into edge ecosystems is an increasingly significant trend for resource adaptation. Leveraging mobile devices as ephemeral edge nodes introduces unique opportunities to scale edge systems in resource-constrained environments. Mobile nodes can offload computation and participate in distributed processing scenarios, but their dynamic availability necessitates robust mobility management strategies to ensure uninterrupted service provisioning [105]. Furthermore, mechanisms to mitigate variability introduced by mobile node participation, such as predictive resource allocation and distributed caching, are critical to optimize performance amidst flux [106].

The role of predictive and AI-driven decision-making is pivotal in tackling the complexity of adaptive resource allocation. Federated learning frameworks deployed at the edge enable resource prediction models tailored to heterogeneous workload patterns while maintaining data privacy [70]. In addition, reinforcement learning methods have gained attention as tools for adapting scaling policies dynamically based on real-time workload observations and network states. These intelligent approaches significantly enhance resource optimization while reducing manual intervention and operational lag [31].

Emerging trends also include leveraging decentralized trust mechanisms for scalability. Blockchain frameworks provide secure collaboration among heterogeneous edge nodes, enabling trust-enhanced resource sharing across multi-stakeholder environments [42]. Furthermore, blockchain smart contracts can automate resource trading and allocation agreements, enhancing the dynamic adaptability of edge ecosystems under varying loads while maintaining transactional transparency and accountability.

Despite these advancements, significant challenges persist. A critical limitation is the inherent energy inefficiency associated with frequent scaling, particularly in resource-constrained edge nodes. Techniques like energy-aware scheduling and dynamic voltage scaling are essential to mitigate the environmental impact of intense scaling operations while ensuring performance sustainability [25]. Additionally, the coordination complexity inherent in multi-tier edge environments, exacerbated by geographical dispersion and device heterogeneity, underscores the need for further research into unified and scalable orchestration frameworks [44].

Looking ahead, the convergence of advanced AI techniques, robust decentralized mechanisms, and novel energy-efficient designs holds promise for tackling the multifaceted challenges of scalability and resource adaptation in edge computing. Future research should explore hybrid models integrating horizontal and vertical scaling strategies with predictive orchestration frameworks. Furthermore, interdisciplinary innovations, such as quantum-inspired resource optimization and precision scaling for latency-critical systems, will be pivotal in transforming theoretical advancements into scalable, real-world deployments.

6.6 Emerging Trends in Optimization and Resource Management

Emerging trends in optimization and resource management within edge computing systems are reshaping how computational, storage, and network resources are utilized to meet the increasing demands of latency-sensitive, energy-efficient, and scalable applications. While traditional optimization models have predominantly focused on static allocation strategies, the inherently dynamic and heterogeneous edge environments demand novel paradigms that can adeptly address fluctuating workloads, resource constraints, and decentralized architectures. This subsection delves into these advancements, highlighting prominent techniques such as AI-driven optimization, decentralized frameworks, precision prediction models, and quantum-inspired approaches.

A transformative shift in resource allocation is evident with the adoption of federated learning (FL). FL facilitates distributed resource optimization while upholding data privacy, leveraging decentralized machine learning algorithms to balance workloads across heterogeneous edge nodes without sharing raw data. This approach mitigates bandwidth usage and privacy concerns, making it particularly advantageous in regulatory-intensive contexts such as healthcare and finance [90]. However, federated systems grapple with challenges like communication overhead stemming from model synchronization and disparities in computational capacities among nodes, which can impact both model convergence speed and resource consumption.

In parallel, blockchain-based trust mechanisms are redefining collaboration in multi-stakeholder edge environments. Blockchain frameworks enable secure, transparent, and auditable resource-sharing agreements through immutable smart contracts, fostering seamless task offloading and inter-node cooperation [42]. While these capabilities enhance reliability and trust, the computational and latency overhead associated with blockchain technologies currently limit their feasibility for resource-constrained edge settings, necessitating further optimization.

Another critical advancement is precision resource prediction, propelled by machine learning (ML) and predictive analytics. These techniques empower edge systems to forecast workload fluctuations and resource needs with high accuracy, enabling just-in-time provisioning and minimizing inefficiencies. Deep reinforcement learning, in particular, has shown promise in autonomous task scheduling, dynamically adapting systems to evolving demands [10]. Nonetheless, reconciling the computational cost of these advanced models with the real-time operational constraints inherent in edge environments remains a key research challenge.

Emerging as a disruptive methodology, quantum-inspired optimization introduces innovative opportunities for addressing complex, non-linear resource management problems in edge computing. Drawing inspiration from quantum computing principles like superposition and entanglement, these techniques, such as quantum annealing, show potential for solving multi-variable scheduling problems with enhanced speed and accuracy compared to traditional approaches [37]. Although still in its infancy, quantum-inspired optimization is gaining traction as a promising area for advancing resource efficiency in next-generation edge systems.

Equally significant is the growing emphasis on energy-aware optimization frameworks. As sustainability becomes a priority, strategies like carbon-aware scheduling—which aligns computational tasks with periods of low-carbon energy availability—are being explored to lower the environmental impact of edge systems [33]. By integrating renewable energy sources and employing techniques such as dynamic voltage scaling for energy-efficient processors, these frameworks balance environmental and operational goals [19]. However, such measures often require balancing trade-offs with user-centric Quality of Service (QoS) metrics, including latency and throughput, complicating their practical deployment.

Lastly, multi-layered orchestration frameworks are emerging as a response to the growing complexity of edge-cloud continuum interactions. These hierarchical approaches dynamically distribute workloads across cloud, fog, and edge nodes, ensuring scalable and reliable performance across diverse application domains [22], [43]. Leveraging containerized microservices, modern orchestration systems enhance resource sharing and application deployment agility. Nonetheless, they face enduring challenges related to inter-layer communication overheads and achieving fault-tolerant service continuity across disparate layers.

In conclusion, the integration of AI, blockchain, predictive modeling, quantum-inspired techniques, and energy-aware frameworks represents the vanguard of advancements in resource optimization and management within edge computing. While these innovations promise transformative potential, formidable challenges—including computational overheads, scalability limitations, and environmental considerations—underline the need for cohesive, energy-efficient, and resilient optimization ecosystems. Future research must aim to synthesize these cutting-edge approaches, paving the way for holistic and sustainable edge computing paradigms while seamlessly aligning with scalability and adaptation innovations discussed earlier.

7 Integration and Emerging Trends in Edge Systems

7.1 Artificial Intelligence and Edge-Native Learning

Artificial intelligence (AI) has fundamentally redefined the potential of edge computing, heralding a paradigm shift toward intelligent, context-aware, and autonomous edge-native systems. By embedding AI capabilities at the network edge, these systems leverage advanced learning frameworks to perform computationally intensive operations such as inference, training, and decision-making in proximity to data sources. This subsection explores the integration of AI in edge-native learning environments, focusing on advancements in federated learning, lightweight neural architectures, and distributed intelligence. Furthermore, it critically examines the trade-offs, challenges, and emerging trends that shape the implementation of AI in resource-constrained, privacy-sensitive edge environments.

One of the cornerstone innovations in edge-native AI is federated learning (FL), a decentralized machine learning paradigm where edge devices collaboratively train global models without sharing raw data. This approach preserves user privacy, a critical concern in applications like healthcare and autonomous vehicles, while reducing network bandwidth consumption. Federated learning frameworks, such as those applied in EdgeFL systems, adapt classical FL algorithms to the unique challenges of edge computing, including hardware heterogeneity and intermittent connectivity [2], [78]. Extensions to FL, such as semi-federated or hierarchical techniques, further bridge the gap between cloud and edge by partitioning tasks across layers based on computational and communication constraints [1], [11].

Despite its promises, federated learning faces fundamental trade-offs. For instance, frequent exchange of model updates between devices introduces new challenges related to energy efficiency and resource allocation. Communication-efficient strategies, such as gradient sparsification, quantization, and periodic local updates, have been applied to mitigate these challenges [107]. However, reducing communication often leads to slower convergence rates, emphasizing the need for adaptive, context-aware protocols that balance efficiency and performance. Moreover, ensuring fairness in FL—where devices with diverse computational capacities contribute equitably—is an ongoing area of exploration [10].

Another critical area of advancement is the development of lightweight AI models explicitly tailored for resource-limited edge devices. Techniques such as model pruning, knowledge distillation, and neural architecture search (NAS) have led to the emergence of compact yet high-performing models capable of running efficiently on constrained hardware [10], [31]. For example, pruning algorithms eliminate redundant parameters in pre-trained networks, while NAS designs optimal architectures by exploring trade-offs between latency, accuracy, and energy consumption. These methods are increasingly augmented with reinforcement learning to autonomously explore configuration landscapes [53]. Integrating such models with hardware accelerators like GPUs and TPUs further enhances performance, supporting applications ranging from real-time video analytics to intelligent transportation systems [76].

Distributed intelligence enables real-time contextual decision-making by coupling edge-native learning processes with IoT ecosystems. This architecture shifts the AI capability closer to devices, enabling immediate responses in latency-sensitive applications such as augmented reality or collaborative vehicular networks. However, embedding distributed AI systems requires addressing data heterogeneity and scalability challenges. Innovative approaches, such as collaborative inference frameworks that partition neural network layers across multiple edge devices, effectively mitigate these issues while improving inference latency [30].

Emerging trends in edge-native learning include sustainable AI, which optimizes energy usage without compromising computational effectiveness, and explainable AI (XAI), which enhances transparency and accountability in distributed decision-making [40]. Furthermore, the convergence of edge computing with blockchain infrastructures introduces decentralized trust mechanisms crucial for verifying AI model provenance and training integrity [52]. Lastly, the advent of 6G networks, offering ultra-reliable and low-latency connectivity, will accelerate the deployment of federated and distributed edge AI frameworks across diverse application domains [73], [94].

In conclusion, AI-driven edge-native learning represents a transformative shift in distributed system design, enabling applications that demand low latency, enhanced privacy, and localized intelligence. While significant strides have been made, challenges related to resource optimization, fairness, and explainability persist. Future research must focus on developing adaptive architectures that unify edge and cloud capabilities, ensuring scalability and robust performance amidst the dynamic demands of next-generation intelligent systems.

7.2 Next-Generation Networks and Edge Optimization

The integration of next-generation wireless networks, specifically 5G and the emerging vision for 6G, represents a transformational leap for the edge computing paradigm, offering ultra-low latency, enhanced bandwidth, and seamless connectivity. These advancements form the backbone for real-time, data-intensive, and latency-sensitive applications such as autonomous vehicles, immersive augmented reality (AR), metaverse platforms, and mission-critical industrial automation. This subsection explores the evolving interplay between next-generation networks and edge computing, emphasizing their co-dependencies, technological synergies, and the optimization opportunities they introduce.

5G networks have redefined the operational capabilities of edge computing by addressing critical performance bottlenecks in legacy network architectures. By implementing enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC), and massive Machine Type Communication (mMTC), 5G facilitates unprecedented advancements in data transfer speed, latency reduction, and connectivity density [9], [78]. URLLC, in particular, addresses stringent latency requirements while ensuring reliability, making it indispensable for use cases like telemedicine and automated manufacturing [9]. Furthermore, advanced features such as network slicing allow for the creation of virtualized logical networks optimized for specific edge applications. For example, tailored slices for AR-based gaming or autonomous vehicle operations showcase a critical leap in resource utilization and service personalization [60].

Building on the capabilities of 5G, 6G networks promise to further evolve edge computing into a hyper-converged ecosystem with peak data rates exceeding 1 Tbps, sub-millisecond latencies, and AI-driven resource optimization. These enhancements will redefine edge computing infrastructures, enabling dynamically adaptive service provisioning and data processing in response to real-time conditions [108]. A pivotal innovation in 6G, Distributed AI (DAI) frameworks, seamlessly integrates with edge intelligence paradigms to support intelligent decision-making and workload management across the network [18]. Additionally, emerging enablers such as over-the-air (OTA) computing and semantic communication frameworks enhance data efficiency by prioritizing meaningful information exchange instead of raw data throughput, thereby reducing redundancy—a vital development for constrained edge deployments [109].

These network-layer advancements complement the adoption of multi-tiered architectures that integrate edge, fog, and cloud layers. Such hierarchical configurations facilitate the delegation of computation-intensive tasks, minimizing latency while distributing resource loads effectively [5], [6]. Multi-tier architectures are further enhanced by dynamic zoning enabled through 5G and 6G capabilities, where deployment zones are adaptively reconfigured according to fluctuating user demand and environmental constraints. For instance, these mechanisms have demonstrated significant potential in disaster-response scenarios, ensuring latency-efficient coordination [59]. Technologies such as software-defined networking (SDN) and network function virtualization (NFV) further enable scalable, flexible management of resources across edge ecosystems [45].

The intersection of edge AI and network slicing highlights key optimization opportunities. Decentralized deep reinforcement learning techniques facilitate intelligent orchestration of computational and network resources, maintaining Quality of Service (QoS) under varied and constrained conditions [38], [60]. However, addressing the inherent challenges posed by workload heterogeneity and resource constraints at the edge remains a crucial area of exploration. Future innovations must refine energy-efficient scheduling methods and enhance collaboration across multi-tier systems [19].

Additionally, federated orchestration strategies unlock collaborative potential in edge-cloud ecosystems. 6G's capacity to support massive device interconnectivity is expected to streamline federated learning systems, promoting privacy-aware applications [31]. However, operationalizing such capabilities necessitates solutions to challenges like network interoperability, trust management, and hybrid platform security risks [8].

In conclusion, next-generation wireless networks will not only amplify the technical performance of edge computing but also drive new paradigms of deployment and interaction at the edge. The introduction of 6G signals a shift from hierarchical structures to dynamic, intelligent, and self-organizing ecosystems, integrating edge optimization as a core network function. Future research must explore cutting-edge domains such as quantum-inspired resource optimization and align high-performance edge networks with sustainability objectives [19]. These advancements will solidify next-generation networks as the essential enabler of a transformative era in edge computing.

7.3 Blockchain and Decentralized Trust Management

The integration of blockchain and decentralized trust management frameworks in edge computing represents a transformative step towards enhancing security, privacy, and autonomous governance in distributed systems. By leveraging blockchain's immutable and decentralized ledger capabilities, edge ecosystems can achieve robust trust mechanisms across highly heterogeneous, resource-constrained environments. This subsection explores the synergies between blockchain and edge computing, analyzing current implementations, inherent trade-offs, and future potential.

Blockchain's foundational attributes, such as distributed consensus, cryptographic verification, and tamper-proof ledgers, align well with the needs of edge ecosystems, where centralized authority is impractical and trust must be established across vast distributed nodes. In this context, blockchain addresses trust challenges by ensuring authenticated and verifiable operations among edge nodes without requiring direct oversight from a centralized entity [8]. For example, decentralized peer-to-peer mechanisms managed through blockchain reduce the threats of single points of failure, which are prevalent in traditional centralized models. This capability is crucial for secure task offloading, resource sharing, and collaborative decision-making in edge environments.

One prominent application area is the use of blockchain for securing sensitive data exchanges in IoT-enabled edge systems, where millions of nodes such as sensors, devices, and gateways interact constantly. Blockchain platforms, such as Ethereum and Hyperledger, provide distributed trust frameworks that facilitate immutable data provenance and secure communication between these IoT-edge devices [71]. Smart contracts—self-executing code stored on blockchains—further enhance automation by enabling conditional mechanisms for resource usage, task offloading agreements, and real-time energy trading among edge nodes. These benefits are illustrated by "smart energy grids," where blockchain-backed contracts dynamically allocate renewable energy resources while minimizing energy wastage [3].

However, deploying blockchain in edge ecosystems reveals significant technical trade-offs, particularly concerning resource constraints. Consensus mechanisms like proof-of-work (PoW) incur high computational and energy costs, which are unsuitable for resource-constrained edge nodes. Alternatives, like proof-of-stake (PoS) and lightweight consensus algorithms, are emerging as scalable options to mitigate the energy consumption and latency inherent in traditional blockchain systems [64]. Furthermore, factors such as storage limitations necessitate off-chain solutions, including state channels or sharding techniques, to optimize blockchain integration in edge scenarios.

An essential advantage of integrating blockchain in edge computing is enhanced trust management in collaborative and multi-stakeholder environments. Systems such as WedgeChain allow for real-time chain auditing, access authorization, and fault-tolerant consensus, creating transparent and tamper-resistant records [6]. Notably, decentralized trust models facilitated by blockchain are critical in managing heterogeneous IoT deployments, where mutual distrust between devices or regions poses a challenge to collaborative governance [23]. Blockchain establishes audit trails that reinforce the integrity and ownership of data while ensuring compliance with localized data sovereignty regulations.

Future trends point toward the closer integration of blockchain with advanced computing paradigms such as edge AI. Blockchain-enabled federated learning frameworks provide decentralized nodes with mechanisms to collaboratively train machine learning models without exposing sensitive user data. This integration enhances privacy preservation while fostering cooperative intelligence across edge systems [110]. Scalability in such models can be achieved through hybrid edge-cloud infrastructures where blockchain networks operate at the edge and cloud layers synergistically [30].

Despite these opportunities, challenges remain in the form of latency overhead and network scalability under high transaction volumes. Emerging concepts like quantum-safe blockchains and zero-knowledge proofs hold promise in overcoming some of these limitations. Furthermore, interdisciplinary advancements in communication protocols, including 5G and software-defined networking (SDN), can provide the necessary backbone for seamless blockchain deployment in edge ecosystems [4].

In conclusion, blockchain and decentralized trust management present groundbreaking opportunities to address foundational trust and security challenges in edge computing systems. By combining secure, transparent mechanisms with edge-specific customizations such as lightweight consensus algorithms, blockchain technologies can significantly enhance the resilience and autonomy of distributed edge environments. However, the field demands continuing research to address scalability, energy efficiency, and operational complexity in real-world applications. Addressing these issues will ensure blockchain's enduring role in shaping next-generation edge systems [2], [3].

7.4 Expanding Application Domains and Interdisciplinary Integration

The continuous evolution of edge computing is marked by its expanding adoption across diverse application domains and the integration of interdisciplinary approaches, which together enhance its versatility and impact. By capitalizing on inherent characteristics such as low-latency processing and proximity to data sources, edge paradigms have emerged as pivotal solutions for tackling domain-specific challenges while unlocking novel functionalities across fields like quantum research, environmental sustainability, immersive multimedia, and autonomous systems.

A transformative area of exploration in edge computing is its synergy with quantum technologies, which promises unparalleled computational efficiencies. The convergence of edge and quantum computing seeks to blend proximity processing with advanced quantum algorithms, optimizing their delivery in domains such as industrial IoT, healthcare, and smart cities. Specifically, this integration distributes intermediate computation from centralized quantum data centers to resource-constrained edge nodes, reducing latency while balancing resource demands. For instance, leveraging the probabilistic modeling capabilities of quantum devices alongside low-latency edge architectures could significantly enhance fault-tolerant applications such as energy grid monitoring or molecular simulations. However, foundational challenges, such as secure data encoding, heterogeneity abstraction, and quantum-optimized task scheduling, must be addressed to establish seamless and reliable interactions between quantum devices and distributed edge infrastructures.

Edge computing's role in resilience and sustainability has gained prominence, particularly in addressing pressing environmental and disaster response challenges. Emerging applications include real-time wildfire detection, flood prediction, precision agriculture, and renewable energy optimization, where distributed edge systems analyze rich spatio-temporal datasets for hyper-localized decision-making. For example, AI-enabled sensors deployed across edge nodes can monitor environmental changes, enabling timely interventions to mitigate natural disasters [19], [111]. Hybrid frameworks that integrate edge platforms with renewable energy systems not only enhance operational efficiencies but also contribute to significant reductions in carbon emissions [19]. However, balancing energy efficiency with system responsiveness, particularly under variable computational requirements, remains a challenge requiring advancements in energy-aware orchestration and elastic resource scaling protocols.

In the domain of immersive media technologies, including augmented reality (AR), virtual reality (VR), and interactive gaming, edge computing continues to play a vital role. The integration of distributed architectures with low-latency computational techniques underpins seamless and engaging user experiences. For instance, edge-driven innovations, such as real-time feedback pipelines in AR/VR applications, improve rendering efficiency and minimize jitter through localized image processing [26], [38]. Hierarchical caching mechanisms tailored to user preferences further enhance streaming delivery. However, as immersive technologies evolve toward complex paradigms like the Metaverse and collaborative virtual environments, ultra-low latency demands and the need for adaptive resource allocation intensify, highlighting the necessity of further empirical research and innovations in edge ecosystems.

Edge computing also underpins the development of autonomous systems operating across transportation, robotics, and aerospace, where real-time decision-making is crucial. Edge infrastructures support capabilities such as localization, object recognition, and path planning for autonomous vehicles, drones, and maritime systems. Federated learning frameworks adapted for resource-constrained environments facilitate the collaborative training of predictive models across autonomous nodes while preserving data privacy [68]. Despite these advances, challenges such as sensor fusion, coordination latencies, and dynamic system adaptation within urban and industrial landscapes remain critical barriers to large-scale application.

Collaborative edge-cloud ecosystems further shape the integration of edge computing into multidisciplinary workflows, optimizing resource utilization across disparate task requirements. Strategies such as computational offloading and service partitioning between edge, fog, and cloud layers ensure scalability for diverse applications, from biomedical data analysis to industrial automation [26], [112]. These approaches mitigate communication overhead and enhance resource flexibility across hybrid systems. Nevertheless, efficiently coordinating multi-tier resource hierarchies remains a complex undertaking, underscoring the need for advanced orchestration frameworks capable of dynamically adapting to workload variability.

Future directions for edge computing emphasize extending its interdisciplinary connections. Neuromorphic computing, federated intelligence, and blockchain-augmented trust mechanisms represent critical areas for innovation. Neuromorphic hardware holds potential for enhancing edge efficiency in AI-heavy applications by mimicking biological neural processes, while federated intelligence frameworks enable collaborative, privacy-preserving computations across distributed edge setups [68]. Blockchain, as explored earlier, offers robust provenance and trust capabilities, crucial for high-stakes domains such as healthcare, finance, and public safety [37]. Advancing these paradigms will require multidisciplinary efforts, combining insights from distributed systems design, sustainability science, and human-computer interaction to build edge architectures that are as socially impactful as they are technically robust.

In conclusion, edge computing's growing incorporation into diverse application domains, coupled with its alignment with interdisciplinary technologies, underscores its transformative potential. As key sectors adopt edge technologies to achieve real-time, optimized processing capabilities, carefully balancing scalability, adaptability, and ethical considerations will shape the trajectory of edge ecosystems. This multidisciplinary focus ensures edge computing's role as a cornerstone of next-generation distributed systems.

7.5 Sustainability and Resource Management in Modern Edge Systems

The global rise of interconnected edge computing systems has undoubtedly revolutionized operational efficiency across industries, but its increasing energy demands raise critical concerns regarding sustainability. Balancing progressive technological advancements with environmental stewardship has now become a cornerstone of edge system design. This subsection explores energy-efficient resource management strategies, assesses the carbon footprint of edge operations, and highlights emerging trends and frameworks for achieving sustainable edge deployments.

Edge computing's proximity to data sources inherently reduces long-distance transmissions to centralized servers, offering baseline energy savings compared to traditional cloud computing. However, the distributed nature of edge architectures leads to an exponential proliferation of edge nodes, creating significant challenges in terms of energy consumption. Hence, energy-sustainable edge architectures have emerged as a primary focus. Techniques such as dynamic power scaling, which adjusts the operational voltage and frequency of edge nodes based on workload intensity, have shown promise. For example, frameworks utilizing Dynamic Voltage and Frequency Scaling (DVFS) methods demonstrate energy reductions while maintaining performance [113]. Similarly, low-power hardware solutions incorporating ARM-based processors and resource-specific accelerators strike a balance between computational efficiency and reduced energy usage in resource-constrained environments [114].

The carbon footprint of edge operations also demands critical attention. Unlike cloud data centers offering centralized control over sustainability strategies, the distributed and heterogeneous nature of edge systems complicates energy optimization. Carbon-aware scheduling, which coordinates task loads to align with periods of renewable energy availability or low-carbon-intensity power, emerges as a sustainable edge computing paradigm. Recent advancements apply predictive energy profiling and renewable energy integration in microgrids to adapt workloads effectively. For example, edge-enabled smart cities deploying renewable-centric micro-data centers significantly reduce their emissions footprint while improving application scalability [7]. However, the adoption of carbon-aware solutions is often constrained by challenges in real-time energy source monitoring and multi-region synchronization.

Federated resource management for scalability represents a pivotal direction in sustainable edge system design. Frameworks that facilitate collaborative resource sharing among edge nodes dynamically redistribute workloads, improving the overall energy efficiency of multi-layer edge networks. Techniques such as collaborative computation offloading—where less-loaded nodes assist their overloaded neighbors—reduce resource wastage caused by isolated task processing [30]. When integrated with blockchain for decentralized trust management, such strategies enhance coordination reliability while enabling micro-level energy accountability [42].

Despite promising developments, sustainability in edge computing confronts several persistent challenges. High inter-node communication overheads and inefficient routing protocols can erode energy savings, particularly in geographically dispersed deployments. Techniques like over-the-air computation and lightweight edge-native protocols are emerging to mitigate these inefficiencies [72]. Further, edge nodes must overcome dual constraints of computational viability and energy efficiency, necessitating advancements in scalable optimization algorithms for task scheduling. For instance, incorporating reinforcement learning-driven adaptive scheduling models into edge ecosystems has offered improvements in prediction accuracy and energy utilization, but further practical validation is needed to address deployment barriers [25].

At the intersection of sustainability and resource management lies the imperative for multidisciplinary innovations. Efforts to converge artificial intelligence with energy-efficient hardware, as seen in lightweight AI accelerators for edge environments, exemplify how technological synergy can enhance both computational and environmental performance [18]. Future edge systems must also consider circular hardware design, where lifecycle management focuses on reusability and recycling of edge-specific components to reduce environmental waste. Furthermore, establishing global frameworks for energy-efficient edge orchestration—leveraging carbon reporting standards and cross-border power-sharing models—represents an untapped area for widespread adoption.

In summary, the sustainability of modern edge systems hinges on harmonizing dynamic energy management strategies with eco-conscious design principles. While advancements such as carbon-aware task scheduling, energy harvesting, and federated resource coordination are transformative, addressing the inherent challenges in scalability, heterogeneity, and real-time energy profiling remains critical. With interdisciplinary innovations and robust policy frameworks, edge systems can transcend their current limitations, laying a sustainable foundation for next-generation computational paradigms.

8 Conclusion

This survey has explored the multifaceted domain of edge computing paradigms by systematically examining their foundational principles, architectural designs, core technologies, applications, security implications, and optimization strategies. Edge computing has emerged as a transformative paradigm that addresses the limitations of centralized cloud architectures, such as high latency, bandwidth inefficiencies, and reduced support for privacy-critical tasks. By bringing computational resources closer to data sources, edge computing paradigms enable low-latency data processing, reduced network congestion, and enhanced contextual awareness, presenting an indispensable framework for modern data-intensive and real-time applications [5], [100].

A critical comparative analysis reveals the spectrum of approaches within edge computing, such as fog computing and mobile edge computing (MEC). While fog computing offers hierarchical and geographically distributed resources that facilitate scalable deployments for latency-sensitive IoT applications [6], MEC emphasizes resource provisioning near mobile network users, particularly in the context of 5G and beyond, to support bandwidth-intensive services like video streaming, augmented reality, and vehicular communications [49], [78]. Despite their complementary strengths, these approaches must be better aligned to effectively harmonize network, computation, and storage resources, as emphasized in works highlighting multi-access architectures [3], [9]. While their adoption has enabled revolutionary applications in healthcare, smart cities, and autonomous systems, a fragmented landscape with limited standardization poses significant challenges to broader deployment.

Moreover, this work has delved into the technological enablers underpinning edge computing advancements, such as containerization, distributed caching, and task offloading. Innovative mechanisms like microservices and lightweight containers have proven critical in optimizing resource-constrained environments [46], [48]. However, achieving optimal trade-offs between energy efficiency and performance remains an unresolved challenge. Computational models integrating dynamic voltage and frequency scaling, collaborative edge-cloud task allocation, and energy-aware orchestration are promising directions for optimization [47], [51].

The security and privacy challenges of edge systems have also been analyzed extensively, as the decentralized and heterogeneous nature of edge infrastructures inherently increases their attack surfaces. Notable advances such as blockchain-enhanced trust mechanisms and AI-driven anomaly detectors have showcased their potential to secure edge environments [10], [52]. However, these approaches face scalability and implementation challenges when integrated into large, resource-constrained edge networks. Additionally, ethical considerations surrounding algorithmic fairness, data ownership, and transparency warrant further attention as edge systems increasingly influence socially significant domains [1], [115].

Emerging trends underscore the continued relevance and evolution of edge computing paradigms. Integration with cutting-edge technologies, such as AI for edge-native learning, 6G-enabled ultra-reliable low-latency networks, and quantum-inspired resource optimization, paves the way for ground-breaking possibilities [16], [31], [113]. Furthermore, the convergence of edge computing with innovative domains like the metaverse, federated intelligence, and energy sustainability represents exceptional opportunities for interdisciplinary exploration [10], [16].

In conclusion, while edge computing paradigms show immense promise in addressing the limitations of existing computational models, several challenges remain. These include achieving global standardization, maintaining energy-efficient deployments, mitigating security vulnerabilities, and resolving affordability concerns for scalable adoption. Future research must focus on developing universally interoperable frameworks, advancing AI-augmented edge orchestration mechanisms, and devising sustainable, lightweight edge architectures to meet diverse application demands. The trajectory of edge computing, as featured throughout this survey, lies at the nexus of innovation and integration, and its role as a critical enabler across industries will only deepen as new challenges emerge. [1], [116].

REFERENCES

[1] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, and D. S. Nikolopoulos, "Challenges and opportunities in edge computing," *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 20–26, 2016. 1, 2, 3, 4, 15, 17, 24, 29

[2] C. Yao, Y. Zhang, H. Luo, T.-Y. Chen, G.-H. Chen, H.-T. Chen, Y.-J. Wang, H.-Y. Wei, and C.-T. Chou, "Management and orchestration of edge computing for iot: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 10, pp. 14 307–14 331, 2023. 1, 6, 11, 13, 14, 17, 24, 27

[3] R. Naha, S. Garg, D. Georgakopoulos, P. Jayaraman, L. Gao, Y. Xiang, and R. Ranjan, "Fog computing: Survey of trends, architectures, requirements, and research directions," *IEEE Access*, vol. 6, pp. 47 980–48 009, 2018. 1, 4, 8, 21, 26, 27, 29

[4] Y. Mao, C. You, J. Zhang, K. Huang, and K. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, pp. 2322–2358, 2017. 1, 3, 11, 13, 16, 23, 27

[5] A. V. Dastjerdi, H. Gupta, R. Calheiros, S. Ghosh, and R. Buyya, "Fog computing: Principles, architectures, and applications," *ArXiv*, vol. abs/1601.02752, 2016. 1, 2, 3, 11, 12, 20, 25, 29

[6] M. R. Mahmud and R. Buyya, "Fog computing: A taxonomy, survey and future directions," *ArXiv*, vol. abs/1611.05539, 2016. 1, 3, 8, 11, 25, 26, 29

[7] L. U. Khan, I. Yaqoob, N. H. Tran, S. M. A. Kazmi, N. D. Tri, and C. Hong, "Edge-computing-enabled smart cities: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 7, pp. 10 200–10 232, 2019. 1, 2, 10, 13, 28

[8] R. Román, J. López, and M. Mambo, "Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges," *Future Gener. Comput. Syst.*, vol. 78, pp. 680–698, 2016. 2, 3, 7, 11, 14, 15, 16, 17, 18, 19, 26

[9] V. Q. Pham, F. Fang, H. Vu, M. Le, Z. Ding, L. Le, and W. Hwang, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2019. 2, 3, 4, 13, 25, 29

[10] S. Iftikhar, S. Gill, C. Song, M. Xu, M. Aslanpour, A. Toosi, J. Du, H. Wu, S. Ghosh, D. Chowdhury, M. Golec, M. Kumar, A. Abdelmoniem, F. Cuadrado, B. Varghese, O. Rana, S. Dustdar, and S. Uhlig, "Ai-based fog and edge computing: A systematic review, taxonomy and future directions," *ArXiv*, vol. abs/2212.04645, 2022. 2, 6, 14, 15, 19, 20, 24, 25, 29

[11] C. Mouradian, D. Naboulsi, S. Yangui, R. Glitho, M. Morrow, and P. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, pp. 416–464, 2017. 2, 15, 24

[12] Y. Deng, X. Chen, G. Zhu, Y. Fang, Z. Chen, and X. Deng, "Actions at the edge: Jointly optimizing the resources in multi-access edge computing," *IEEE Wireless Communications*, vol. 29, pp. 192–198, 2022. 2, 13, 23

[13] M. Goudarzi, M. Palaniswami, and R. Buyya, "Scheduling iot applications in edge and fog computing environments: A taxonomy and future directions," *ACM Computing Surveys*, vol. 55, pp. 1 – 41, 2022. 2, 5, 6, 10, 11, 14, 23

[14] B. G. S. Costa, J. Bachiega, L. R. Carvalho, M. Rosa, and A. P. F. Araujo, "Monitoring fog computing: a review, taxonomy and open challenges," *Comput. Networks*, vol. 215, p. 109189, 2022. 2

[15] R. Mahmud and A. Toosi, "Con-pi: A distributed container-based edge and fog computing framework," *IEEE Internet of Things Journal*, vol. 9, pp. 4125–4138, 2021. 2, 6, 14, 23

[16] Y. Wang and J. Zhao, "Mobile edge computing, metaverse, 6g wireless communications, artificial intelligence, and blockchain: Survey and their convergence," *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, pp. 1–8, 2022. 3, 6, 7, 11, 21, 29

[17] H. Gupta, A. V. Dastjerdi, S. Ghosh, and R. Buyya, "ifogsim: A toolkit for modeling and simulation of resource management techniques in the internet of things, edge and fog computing environments," *Software: Practice and Experience*, vol. 47, pp. 1275 – 1296, 2016. 3, 11, 14, 21

[18] Y. Han, X. Wang, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, pp. 869–904, 2019. 3, 7, 9, 11, 16, 18, 25, 28

[19] N. Shalavi, G. Perin, A. Zanella, and M. Rossi, "Energy efficient deployment and orchestration of computing resources at the network edge: a survey on algorithms, trends and open challenges," *ArXiv*, vol. abs/2209.14141, 2022. 3, 8, 9, 11, 14, 17, 20, 21, 24, 26, 27

[20] X. Li, M. Abdallah, Y.-Y. Lou, M. Chiang, K. T. Kim, and S. Bagchi, "Dynamic dag-application scheduling for multi-tier edge computing in heterogeneous networks," *ArXiv*, vol. abs/2409.10839, 2024. 3, 4, 7, 21

[21] R. Cárdenas, P. Arroba, and J. L. Risco-Martín, "Bringing ai to the edge: a formal m&s specification to deploy effective iot architectures," *Journal of Simulation*, vol. 16, pp. 494 – 511, 2021. 3

[22] X. Cao, G. Tang, D. Guo, Y. Li, and W. Zhang, "Edge federation: Towards an integrated service provisioning model," *IEEE/ACM Transactions on Networking*, vol. 28, pp. 1116–1129, 2019. 3, 11, 21, 24

[23] J. Zhang and K. Letaief, "Mobile edge intelligence and computing for the internet of vehicles," *Proceedings of the IEEE*, vol. 108, pp. 246–261, 2019. 4, 17, 26

[24] M. Jansen, A. Al-Dulaimy, A. Papadopoulos, A. Trivedi, and A. Iosup, "The spec-rg reference architecture for the compute continuum," *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pp. 469–484, 2022. 4, 22

[25] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 23, pp. 2131–2165, 2021. 4, 5, 8, 9, 12, 21, 23, 28

[26] S. Bi, L. Huang, and Y. Zhang, "Joint optimization of service caching placement and computation offloading in mobile edge computing systems," *IEEE Transactions on Wireless Communications*, vol. 19, pp. 4947–4963, 2019. 4, 7, 8, 27

[27] L. Bittencourt, R. Immich, R. Sakellariou, N. Fonseca, E. Madeira, M. Curado, L. Villas, L. Silva, C. A. Lee, and O. Rana, "The internet of things, fog and cloud continuum: Integration and challenges," *ArXiv*, vol. abs/1809.09972, 2018. 4, 5, 10, 15, 18, 22

[28] J. Xu, L. Chen, and P. Zhou, "Joint service caching and task offloading for mobile edge computing in dense networks," *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pp. 207–215, 2018. 4, 8

[29] P. Fondo-Ferreiro, F. Gil-Castiñeira, F. J. Gonzélez-Castaño, D. Candal-Ventureira, J. Rodriguez, A. J. Morgado, and S. Mumtaz, "Efficient anchor point deployment for low latency connectivity in mec-assisted c-v2x scenarios," *IEEE Transactions on Vehicular Technology*, vol. 72, pp. 16 637–16 649, 2023. 4

[30] A. Ndikumana, N. H. Tran, T. Ho, Z. Han, W. Saad, D. Niyato, and C. Hong, "Joint communication, computation, caching, and control in big data multi-access edge computing," *IEEE Transactions on Mobile Computing*, vol. 19, pp. 1359–1374, 2018. 4, 10, 13, 22, 25, 26, 28

[31] D. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, T. Jiang, J. Crowcroft, and P. Hui, "Edge intelligence: Architectures, challenges, and applications," *arXiv: Networking and Internet Architecture*, 2020. 4, 9, 10, 11, 12, 15, 16, 17, 18, 21, 22, 23, 25, 26, 29

[32] D. Liu, H. Kong, X. Luo, W. Liu, and R. Subramaniam, "Bringing ai to edge: From deep learning's perspective," *ArXiv*, vol. abs/2011.14808, 2020. 4

[33] Y. G. Kim, U. Gupta, A. McCrabb, Y.-B. Son, V. Bertacco, D. Brooks, and C.-J. Wu, "Greenscale: Carbon-aware systems for edge computing," *ArXiv*, vol. abs/2304.00404, 2023. 4, 8, 12, 14, 19, 24

[34] J. Wu, F. Dong, H. Leung, Z. Zhu, J. Zhou, and S. Drew, "Topology-aware federated learning in edge computing: A comprehensive survey," *ACM Computing Surveys*, vol. 56, pp. 1 – 41, 2023. 5

[35] K. Poularakis, J. Llorca, A. Tulino, I. Taylor, and L. Tassiulas, "Joint service placement and request routing in multi-cell mobile edge computing networks," *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 10–18, 2019. 5, 8, 22

[36] C.-F. Liu, M. Bennis, M. Debbah, and H. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Transactions on Communications*, vol. 67, pp. 4132–4150, 2018. 5, 12, 18, 22

[37] M. Farhoudi, M. Shokrnezhad, T. Taleb, R. Li, and J. Song, "Discovery of 6g services and resources in edge-cloud-continuum," *ArXiv*, vol. abs/2407.21751, 2024. 5, 13, 17, 22, 24, 27

[38] F. Liu, G. Tang, Y. Li, Z. Cai, X. Zhang, and T. Zhou, "A survey on edge computing systems and tools," *Proceedings of the IEEE*, vol. 107, pp. 1537–1562, 2019. 5, 6, 7, 9, 13, 21, 26, 27

[39] Y. Cai, J. Llorca, A. Tulino, and A. Molisch, "Dynamic control of data-intensive services over edge computing networks," *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pp. 5123–5128, 2022. 5, 13

[40] K. Toczé and S. Nadjm-Tehrani, "A taxonomy for management and optimization of multiple resources in edge computing," *Wirel. Commun. Mob. Comput.*, vol. 2018, pp. 7 476 201:1–7 476 201:23, 2018. 5, 9, 25

[41] M. Ebrahim and A. Hafid, "Resilience and load balancing in fog networks: A multi-criteria decision analysis approach," *Microprocess. Microsystems*, vol. 101, p. 104893, 2022. 5, 16

[42] A. V. Rivera, A. Refaey, and E. Hossain, "A blockchain framework for secure task sharing in multi-access edge computing," *IEEE Network*, vol. 35, pp. 176–183, 2020. 5, 10, 18, 19, 23, 24, 28

[43] A. Ullah, T. Kiss, J. Kovács, F. Tusa, J. Deslauriers, H. Dagdeviren, R. Arjun, and H. Hamzeh, "Orchestration in the cloud-to-things compute continuum: taxonomy, survey and future directions," *Journal of Cloud Computing*, vol. 12, pp. 1–29, 2023. 6, 14, 24

[44] M. A. U. Rehman, M. S. U. Din, S. Mastorakis, and B. Kim, "Foggyedge: An information-centric computation offloading and management framework for edge-based vehicular fog computing," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, pp. 78–90, 2023. 6, 23

[45] Y. Mao, C. You, J. Zhang, K. Huang, and K. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, pp. 2322–2358, 2017. 6, 11, 23, 26

[46] Z. Wang, M. Goudarzi, J. Aryal, and R. Buyya, "Container orchestration in edge and fog computing environments for real-time iot applications," *ArXiv*, vol. abs/2203.05161, 2022. 6, 11, 29

[47] N. Sathyavageeswaran, R. D. Yates, A. D. Sarwate, and N. Mandayam, "Timely offloading in mobile edge cloud systems," *ArXiv*, vol. abs/2405.07274, 2024. 6, 20, 29

[48] S. Pallewatta, V. Kostakos, and R. Buyya, "Placement of microservices-based iot applications in fog computing: A taxonomy and future directions," *ACM Computing Surveys*, vol. 55, pp. 1 – 43, 2022. 6, 14, 29

[49] M. A. Khan, E. Baccour, Z. Chkirbene, A. Erbad, R. Hamila, M. Hamdi, and M. Gabbouj, "A survey on mobile edge computing for video streaming: Opportunities and challenges," *IEEE Access*, vol. 10, pp. 120 514–120 550, 2022. 6, 29

[50] A. Abouaomar, S. Cherkaoui, A. Kobbane, and O. A. Dambri, "A resources representation for resource allocation in fog computing networks," *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2019. 6, 11

[51] B. Kopras, B. Bossy, F. Idzikowski, P. Kryszkiewicz, and H. Bogucka, "Task allocation for energy optimization in fog computing networks with latency constraints," *IEEE Transactions on Communications*, vol. 70, pp. 8229–8243, 2022. 6, 10, 20, 29

[52] H. Xue, D. Chen, N. Zhang, H. Dai, K. Y. S. F. R. I. of Ministry of Management, U. of Electronic Science, T. China, U. Windsor, L. University, and W. University, "Integration of blockchain and edge computing in internet of things: A survey," *ArXiv*, vol. abs/2205.13160, 2022. 6, 10, 15, 20, 25, 29

[53] J. Wang, J. Pan, F. Esposito, P. Calyam, Z. Yang, and P. Mohapatra, "Edge cloud offloading algorithms," *ACM Computing Surveys (CSUR)*, vol. 52, pp. 1 – 23, 2018. 7, 25

[54] X. Gong, "Delay-optimal distributed edge computing in wireless edge networks," *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pp. 2629–2638, 2020. 7

[55] A. Abouaomar, S. Cherkaoui, Z. Mlika, and A. Kobbane, "Resource provisioning in edge computing for latency-sensitive applications," *IEEE Internet of Things Journal*, vol. 8, pp. 11 088–11 099, 2021. 7, 10, 19

[56] S. Suman, Stefanović, S. Došen, and P. Popovski, "Analysis and optimization of the latency budget in wireless systems with mobile edge computing," *ICC 2022 - IEEE International Conference on Communications*, pp. 5029–5034, 2022. 7

[57] C. Cicconetti, M. Conti, and A. Passarella, "Architecture and performance evaluation of distributed computation offloading in edge computing," *ArXiv*, vol. abs/2109.09415, 2020. 7

[58] D. Kimovski, R. Math'a, J. Hammer, N. Mehran, H. Hellwagner, and R. Prodan, "Cloud, fog, or edge: Where to compute?" *IEEE Internet Computing*, vol. 25, pp. 30–36, 2021. 7, 21

[59] Y. Lin, W. Feng, Y. Chen, N. Ge, Z. Feng, and Y. Gao, "Edge information hub-empowered 6g ntn: Latency-oriented resource orchestration and configuration," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4241–4259, 2024. 7, 11, 26

[60] Q. Liu, T. Han, and E. Moges, "Edgeslice: Slicing wireless edge computing network with decentralized deep reinforcement learning," *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pp. 234–244, 2020. 7, 21, 25, 26

[61] Q. Li, Y. Zhang, Y. Li, Y. Xiao, and X. Ge, "Capacity-aware edge caching in fog computing networks," *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 9244–9248, 2020. 8

[62] M. Hu, Z. Luo, A. Pasdar, Y. C. Lee, Y. Zhou, and D. Wu, "Edge-based video analytics: A survey," *ArXiv*, vol. abs/2303.14329, 2023. 8, 12

[63] J. Shao and J. Zhang, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Communications Magazine*, vol. 58, pp. 20–26, 2020. 8

[64] A. Alwarafy, K. A. Althelaya, M. Abdallah, J. Schneider, and M. Hamdi, "A survey on security and privacy issues in edge-computing-assisted internet of things," *IEEE Internet of Things Journal*, vol. 8, pp. 4004–4022, 2020. 8, 12, 16, 22, 26

[65] A. R. Khouas, M. R. Bouadjenek, H. Hacid, and S. Aryal, "Training machine learning models at the edge: A survey," *ArXiv*, vol. abs/2403.02619, 2024. 8, 9, 17

[66] L. Zeng, S. Ye, X. Chen, and Y. Yang, "Implementation of big ai models for wireless networks with collaborative edge computing," *IEEE Wireless Communications*, vol. 31, pp. 50–58, 2024. 8

[67] M. Zhang, J. Cao, X. Shen, and Z. Cui, "Edgeshard: Efficient llm inference via collaborative edge computing," *ArXiv*, vol. abs/2405.14371, 2024. 8

[68] B. T. Hasan and A. K. Idrees, "Federated learning for iot/edge/fog computing systems," *ArXiv*, vol. abs/2402.13029, 2024. 9, 12, 17, 22, 27

[69] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proceedings of the IEEE*, vol. 107, pp. 2204–2239, 2018. 9

[70] M. Chen, H. Poor, W. Saad, and S. Cui, "Wireless communications for collaborative federated learning," *IEEE Communications Magazine*, vol. 58, pp. 48–54, 2020. 9, 12, 13, 14, 17, 18, 22, 23

[71] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on multi-access edge computing for internet of things realization," *IEEE Communications Surveys & Tutorials*, vol. 20, pp. 2961–2991, 2018. 9, 13, 26

[72] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys & Tutorials*, vol. 22, pp. 2167–2191, 2020. 9, 13, 14, 22, 28

[73] M. S. Elbamby, C. Perfecto, C.-F. Liu, J. Park, S. Samarakoon, X. Chen, and M. Bennis, "Wireless edge computing with latency and reliability guarantees," *Proceedings of the IEEE*, vol. 107, pp. 1717–1737, 2019. 9, 25

[74] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Towards an intelligent edge: Wireless communication meets machine learning," *ArXiv*, vol. abs/1809.00343, 2018. 10

[75] S. Srirama, "A decade of research in fog computing: Relevance, challenges, and future directions," *Software: Practice and Experience*, vol. 54, pp. 23 – 3, 2023. 10, 15

[76] M. Zhang, F. Zhang, N. Lane, Y. Shu, X. Zeng, B. Fang, S. Yan, and H. Xu, "Deep learning in the era of edge computing: Challenges and opportunities," *ArXiv*, vol. abs/2010.08861, 2020. 10, 25

[77] M. R. Mahmud and R. Buyya, "Modelling and simulation of fog and edge computing environments using ifogsim toolkit," *ArXiv*, vol. abs/1812.00994, 2018. 10

[78] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5g networks: New paradigms, scenarios, and challenges," *IEEE Communications Magazine*, vol. 55, pp. 54–61, 2016. 10, 21, 24, 25, 29

[79] Z. Kostić, A. Angus, Z. Yang, Z. Duan, I. Seskar, G. Zussman, and D. Raychaudhuri, "Smart city intersections: Intelligence nodes for future metropolises," *Computer*, vol. 55, pp. 74–85, 2022. 11

[80] W. Liu, J. Geng, Z. Zhu, Y. Zhao, C. Ji, C. Li, Z. Lian, and X. Zhou, "Ace-sniper: Cloud–edge collaborative scheduling framework with dnn inference latency modeling on heterogeneous devices," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, pp. 534–547, 2024. 12

[81] H. Xing, L. Liu, J. Xu, and A. Nallanathan, "Joint task assignment and resource allocation for d2d-enabled mobile-edge computing," *IEEE Transactions on Communications*, vol. 67, pp. 4193–4207, 2019. 13

[82] Z. Nezami, K. Zamanifar, K. Djemame, and E. Pournaras, "Decentralized edge-to-cloud load balancing: Service placement for the internet of things," *IEEE Access*, vol. 9, pp. 64 983–65 000, 2020. 13

[83] S. Shen, Y. Ren, Y. Ju, X. Wang, W. Wang, and V. C. M. Leung, "Edgematrix: A resource-redefined scheduling framework for sla-guaranteed multi-tier edge-cloud computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 41, pp. 820–834, 2023. 13, 22

[84] W. Li, H. Hacid, E. Almazrouei, and M. Debbah, "A comprehensive review and a taxonomy of edge machine learning: Requirements, paradigms, and techniques," *AI*, 2023. 13, 17, 22

[85] M. M. Adam, L. Zhao, K. Wang, and Z. Han, "Beyond 5g networks: Integration of communication, computing, caching, and control," *China Communications*, vol. 20, pp. 137–174, 2022. 14

[86] C. Chang, S. Srirama, and R. Buyya, "Internet of things (iot) and new computing paradigms," in *Fog and Edge Computing*, 2018, pp. 1–23. 14

[87] S. Gill, "A manifesto for modern fog and edge computing: Vision, new paradigms, opportunities, and future directions," *ArXiv*, vol. abs/2109.12195, 2021. 14

[88] M. Goudarzi, Q. Deng, and R. Buyya, "Resource management in edge and fog computing using fogbus2 framework," *ArXiv*, vol. abs/2108.00591, 2021. 14

[89] S. Becker, F. Schmidt, and O. Kao, "Edgepier: P2p-based container image distribution in edge computing environments," *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pp. 1–8, 2021. 14

[90] X. Zhang, Y. Wang, S. Lu, L. Liu, L. Xu, and W. Shi, "Openei: An open framework for edge intelligence," *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 1840–1851, 2019. 14, 24

[91] E. Covi, E. Donati, H. Heidari, D. Kappel, X. Liang, M. Payvand, and W. Wang, "Adaptive extreme edge computing for wearable devices," *Frontiers in Neuroscience*, vol. 15, 2020. 17

[92] B. Huang, Z. Li, P. Tang, S. Wang, J. Zhao, H. Hu, W. Li, and V. Chang, "Security modeling and efficient computation offloading for service workflow in mobile edge computing," *ArXiv*, vol. abs/1907.02506, 2019. 18

[93] X. Jin, C. Katsis, F. Sang, J. Sun, A. Kundu, and R. Kompella, "Edge security: Challenges and issues," *ArXiv*, vol. abs/2206.07164, 2022. 18

[94] J. Liu, Y. Mao, J. Zhang, and K. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1451–1455, 2016. 20, 25

[95] Y. Mao, J. Zhang, S. Song, and K. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 5994–6009, 2017. 20

[96] W. Wen, Y. Cui, T. Q. S. Quek, F. Zheng, and S. Jin, "Joint optimal software caching, computation offloading and communications resource allocation for mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 7879–7894, 2020. 20

[97] H. Qiu, K. Zhu, N. C. Luong, C. Yi, D. Niyato, and D. I. Kim, "Applications of auction and mechanism design in edge computing: A survey," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, pp. 1034–1058, 2021. 20

[98] Z. Nezami, E. Chaniotakis, and E. Pournaras, "When computing follows vehicles: Decentralized mobility-aware resource allocation for edge-to-cloud continuum," *ArXiv*, vol. abs/2404.13179, 2024. 20

[99] V. Q. Pham, R. Ruby, F. Fang, D. C. Nguyen, Z. Yang, M. Le, Z. Ding, and W. Hwang, "Aerial computing: A new computing

paradigm, applications, and challenges," *IEEE Internet of Things Journal*, vol. 9, pp. 8339–8363, 2022. 21

[100] A. Zaslavsky, C. Perera, and D. Georgakopoulos, "Sensing as a service and big data," *ArXiv*, vol. abs/1301.0159, 2013. 21, 29

[101] A. R. Nandhakumar, A. Baranwal, P. Choudhary, M. Golec, and S. Gill, "Edgeaisim: A toolkit for simulation and modelling of ai models in edge computing environments," *ArXiv*, vol. abs/2310.05605, 2023. 21

[102] F. Hoseiny, S. Azizi, M. Shojafar, and R. Tafazolli, "Joint qos-aware and cost-efficient task scheduling for fog-cloud resources in a volunteer computing system," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, pp. 1 – 21, 2021. 22

[103] M. Zhang, J. Cao, L. Yang, L. Zhang, Y. Sahni, and S. Jiang, "Ents: An edge-native task scheduling system for collaborative edge computing," *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*, pp. 149–161, 2022. 22

[104] Q. Qu, R. Xu, S. Nikouei, and Y. Chen, "An experimental study on microservices based edge computing platforms," *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 836–841, 2020. 23

[105] Y. Seraj, S. Fadaei, B. Safaei, A. Javadi, A. M. H. Monazzah, and A. Hemmatyar, "Limo: Load-balanced offloading with mape and particle swarm optimization in mobile fog networks," *ArXiv*, vol. abs/2408.14218, 2024. 23

[106] K. Ma and J. Xie, "A multi-layered distributed computing framework for enhanced edge computing," 2024. 23

[107] M. Asim, Y. Wang, K. Wang, and P. qiu Huang, "A review on computational intelligence techniques in cloud and edge computing," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, pp. 742–763, 2020. 25

[108] M. Ishtiaq, N. Saeed, and M. A. Khan, "Edge computing in iot: A 6g perspective," *ArXiv*, vol. abs/2111.08943, 2021. 25

[109] N. G. Evgenidis, N. A. Mitsiou, V. I. Koutsioumpa, S. A. Tegos, P. Diamantoulakis, and G. Karagiannidis, "Multiple access in the era of distributed computing and edge intelligence," *ArXiv*, vol. abs/2403.07903, 2024. 25

[110] K. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6g: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications*, vol. 40, pp. 5–36, 2021. 26

[111] J. Moura and D. Hutchison, "Fog computing systems: State of the art, research issues and future trends, with a focus on resilience," *J. Netw. Comput. Appl.*, vol. 169, p. 102784, 2019. 27

[112] C. Shan, R. Gao, Q. Han, Z. Yang, J. Zhang, and Y. Xia, "Kces: A workflow containerization scheduling scheme under cloud-edge collaboration framework," *ArXiv*, vol. abs/2401.01217, 2024. 27

[113] J. Cao, W. Feng, N. Ge, and J. Lu, "Delay characterization of mobile-edge computing for 6g time-sensitive services," *IEEE Internet of Things Journal*, vol. 8, pp. 3758–3773, 2020. 28, 29

[114] J. Gamazo-Real, R. T. Fernández, and A. M. Armas, "Comparison of edge computing methods in internet of things architectures for efficient estimation of indoor environmental parameters with machine learning," *ArXiv*, vol. abs/2403.08810, 2023. 28

[115] C. Wang, Z. Yuan, P. Zhou, Z. Xu, R. Li, and D. O. Wu, "The security and privacy of mobile-edge computing: An artificial intelligence perspective," *IEEE Internet of Things Journal*, vol. 10, pp. 22 008–22 032, 2023. 29

[116] B. Varghese, E. de Lara, A. Ding, C.-H. Hong, F. Bonomi, S. Dustdar, P. Harvey, P. Hewkin, W. Shi, M. Thiele, and P. Willis, "Revisiting the arguments for edge computing research," *IEEE Internet Computing*, vol. 25, pp. 36–42, 2021. 29