World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Edge computing and data minimization: A synergistic approach for cloud-native AI

Chaitra Vatsavayi *

Carnegie Mellon University, USA.

World Journal of Advanced Research and Reviews, 2025, 26(02), 4469–4476

Publication history: Received on 20 April 2025; revised on 28 May 2025; accepted on 31 May 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.2.2114

Abstract

Edge computing and data minimization present a synergistic framework for addressing key challenges in cloud-native AI systems. This integration enables processing near data sources, reducing latency while enhancing privacy and bandwidth utilization. The framework categorizes integration patterns across hierarchical, mesh-based, hybrid, and distributed collaborative architectures, exploring potential implementations in domains such as healthcare, manufacturing, and smart cities. Despite theoretical advantages, practical implementation faces several challenges, including neural network optimization for resource-constrained environments, balancing data minimization with model accuracy requirements, managing architectural complexity in distributed systems, and addressing standardization gaps in emerging protocols. Potential benefits include bandwidth optimization through local preprocessing, enhanced privacy protection through localized data processing, latency reduction for time-sensitive applications, and improved energy efficiency from decreased data transmission requirements. Future developments in this field will likely be shaped by specialized hardware accelerators, federated learning approaches, standardization efforts for interoperability, and adaptive workload distribution strategies, with significant implications for organizational data governance and regulatory compliance.

Keywords: Edge computing; Data minimization; Cloud-native AI; Federated learning; Real-time processing

1. Introduction

The exponential growth of artificial intelligence systems over the past decade has revolutionized numerous sectors, transforming operational paradigms and creating unprecedented opportunities for innovation. However, this proliferation has introduced significant challenges related to data management, transmission, and processing. As AI models become increasingly sophisticated, they typically demand vast quantities of data for training and inference, leading to bottlenecks in traditional cloud-based architectures. These challenges include network congestion, latency issues, privacy concerns, and escalating energy consumption associated with data centers, creating a pressing need for architectural evolution to support next-generation AI applications [1]. The concept of fog computing, as described in the literature, represents an extension of cloud computing that brings computation closer to end devices, creating a continuum between cloud and edge. This paradigm was initially conceptualized to address the unique requirements of Internet of Things (IoT) applications, particularly those demanding low latency, geographical distribution, and mobility support, which aligns perfectly with modern AI system requirements.

Edge computing has emerged as a transformative approach in distributed computing, bringing computation and data storage closer to the location where it is needed. By processing data near the source of generation—on devices, gateways, or edge servers—rather than transmitting all information to centralized cloud infrastructure, organizations can significantly reduce network distance and associated latency while decreasing bandwidth requirements. This proximity-based computational model represents a fundamental shift from traditional cloud-centric architectures and addresses many of their inherent limitations [2]. The emergence of edge computing can be understood as a response to

* Corresponding author: Chaitra Vatsavayi

several converging trends, including the proliferation of connected devices, bandwidth limitations, and increasingly stringent latency requirements for applications like augmented reality and autonomous vehicles. This computational paradigm leverages strategic placement of resources closer to data sources to enable real-time analytics and decision-making capabilities.

Concurrently, data minimization has evolved as a principle that advocates processing only the data necessary for a specific purpose, thereby reducing both storage requirements and potential privacy risks. This approach complements edge computing by ensuring that only relevant data traverses the network, further optimizing bandwidth utilization and enhancing security posture. The integration of these principles creates a synergistic relationship that addresses multiple challenges simultaneously, from network congestion to regulatory compliance requirements regarding data protection and privacy.

Cloud-native architectures, characterized by containerized applications, microservices, and orchestration mechanisms, have become the standard for deploying scalable AI systems. Despite their advantages in flexibility and resource utilization, these architectures face limitations when confronted with the needs of modern AI applications [2]. The literature identifies several constraints of purely cloud-based approaches, including dependency on reliable network connectivity, challenges in supporting latency-sensitive applications, and difficulties in addressing location-specific compliance requirements. These limitations become particularly pronounced in scenarios requiring real-time processing of large data volumes, such as industrial automation, healthcare monitoring, and intelligent transportation systems.

2. Methodology

The research methodology proposes a conceptual framework for evaluating the integration of edge computing and data minimization techniques within cloud-native AI systems. This framework categorizes edge-cloud integration patterns into four primary architectures: hierarchical, mesh-based, hybrid, and distributed collaborative models. Such a taxonomy would allow for systematic analysis of data flows, processing distribution, and resource allocation across the computational continuum. The framework would incorporate both quantitative performance metrics and qualitative considerations such as implementation complexity and maintenance requirements. As highlighted in the literature, edge computing presents both opportunities and challenges that must be carefully considered when designing distributed AI architectures. These challenges include computing at the edge with limited resources, data analytics at the edge, ensuring security and privacy of data, and optimizing the cooperation between different service providers [3]. This conceptual approach could enable classification of existing implementations and identification of architectural patterns that best support specific application requirements, providing a foundation for comparative analysis.

For potential empirical validation, the methodology suggests examining three industries with distinct edge computing requirements: healthcare, manufacturing, and smart cities. Selection criteria would emphasize diversity in data characteristics (volume, velocity, variety), latency requirements, privacy considerations, and deployment maturity. This approach aligns with the vision presented in the literature that edge computing represents a new computing paradigm where substantial computing and storage resources are placed at the edge of the Internet in close proximity to mobile devices, sensors, and end users. The literature indicates that edge computing enables a new breed of applications and services with requirements related to mobility support, geographic distribution, location awareness, and low latency [3]. Case studies in these domains would need to be documented following structured protocols to facilitate replication and comparative analysis.

System performance evaluation would require a multi-dimensional metrics framework encompassing technical, operational, and business perspectives. Primary technical metrics should include end-to-end latency, bandwidth utilization, computational resource efficiency, and energy consumption. This approach recognizes that fog computing, as a precursor to modern edge computing, represents a highly virtualized platform that provides compute, storage, and networking services between end devices and traditional cloud computing data centers. As described in the literature, such environments must be evaluated through metrics that capture both the traditional aspects of distributed systems performance and the unique characteristics of edge-cloud integration, including geographical distribution, mobility support, and real-time interactions [4]. A comprehensive metrics framework would incorporate standardized benchmarking methodologies adapted for distributed edge environments, enabling objective comparison across different architectural approaches.

Data collection for privacy and security assessment would require a multi-layered approach combining automated vulnerability scanning, data flow tracking, access control evaluation, and compliance verification. Previous research has proposed metadata tracking mechanisms that follow data transformations across the edge-cloud boundary, which could

enable comprehensive visibility into data minimization effectiveness. Such methodologies build upon the understanding that fog or edge computing is ideally positioned for advanced analytics and interaction with the physical world, particularly in scenarios requiring real-time processing and data protection. A security assessment framework would need to account for both technical and operational characteristics of edge deployments, including hardware heterogeneity, physical security challenges, and multi-stakeholder trust relationships as highlighted in existing research [4]. This approach would emphasize measurement of practical security outcomes rather than merely documenting compliance with theoretical requirements, potentially providing actionable insights for implementation.

Edge Computing and Data Minimization: Comparative Analysis

Feature	Traditional Cloud-Centric	Edge-Optimized
Data Processing	Centralized in cloud	Processing closer to data source
Latency	Higher	Significantly reduced
Bandwidth Usage	High volume of data transmission	Reduced through data minimization
Privacy & Security	All data exposed to transmission risks	Sensitive data processed locally
Energy Efficiency	Higher energy costs for data centers	Reduced energy through distributed processing

Figure 1 Edge-Cloud Architecture: Data Minimization in AI Systems. [3, 4]

3. Discussion: Challenges, Issues, and Limitations

While the integration of edge computing and data minimization presents significant advantages for cloud-native AI systems, several substantial challenges impede widespread implementation. Technical implementation barriers for edge AI deployment represent a primary obstacle, particularly regarding the optimization of complex neural network architectures for resource-constrained environments. Current deep learning models often require significant computational resources that exceed the capabilities of many edge devices. Techniques such as model compression, pruning, and quantization offer potential solutions but frequently introduce performance degradation that must be carefully managed. Recent research has identified that deep learning at the edge must overcome several key challenges including limited computational capability, constrained memory, and restricted power consumption. Even with these constraints, edge-based learning can achieve benefits including privacy protection, reduced latency, and decreased bandwidth usage. The field is progressing through innovations such as compact model design, network compression, and hardware-optimized implementations that jointly address the constraints while maintaining learning performance [5]. These approaches highlight the importance of considering the full spectrum of edge computing characteristics when designing AI systems that must operate effectively across the cloud-to-edge continuum.

A fundamental tension exists between data minimization objectives and model accuracy requirements. AI systems traditionally perform better with more training and inference data, creating an inherent conflict with data minimization principles that aim to reduce data collection and transmission. This challenge becomes particularly pronounced in applications requiring high precision, such as medical diagnostics or autonomous navigation systems. The literature indicates that deep learning at the edge requires novel approaches to maintain accuracy despite data constraints. Current research focuses on strategies such as designing specialized learning algorithms that can extract maximum value from limited data, utilizing transfer learning to leverage pre-trained models, and implementing techniques that can incrementally update models with new data without requiring complete retraining. These approaches aim to mitigate the accuracy impact of data minimization while still honoring constraints on data collection and transmission

[5]. The trade-off between model performance and data minimization represents an ongoing challenge requiring careful consideration of application-specific requirements and constraints.

Architectural complexities in distributed AI systems present substantial implementation hurdles. The orchestration of processing across edge devices, edge servers, and cloud infrastructure introduces numerous decision points regarding workload partitioning, data aggregation, and model synchronization. These systems must dynamically adapt to changing network conditions, device capabilities, and application requirements while maintaining overall system coherence. Research has shown that managing computational offloading decisions in edge computing environments requires sophisticated algorithms that consider multiple factors, including device energy constraints, network conditions, and computational requirements. Effective workload distribution between edge and cloud requires addressing problems related to resource allocation, task scheduling, and load balancing. The heterogeneity of edge environments further exacerbates these challenges, as different devices have varying computational capabilities, energy constraints, and connectivity patterns [6]. These architectural complexities necessitate the development of flexible frameworks that can adapt to diverse deployment scenarios while maintaining performance guarantees.

Standardization gaps in edge computing protocols present significant interoperability challenges. Unlike cloud computing, which has developed relatively mature standards over decades, edge computing ecosystems remain fragmented with competing protocols, interfaces, and management frameworks. This fragmentation hinders the development of interoperable solutions and increases integration complexity. The literature identifies critical standardization challenges in areas including service provisioning, resource management, and security protocols. Current standardization efforts are focusing on defining common interfaces and protocols for edge computing, but significant gaps remain in addressing the unique requirements of AI workloads at the edge. These gaps include standardized approaches for model distribution, versioning, monitoring, and lifecycle management. Additionally, the diversity of edge computing applications creates challenges in developing standards that can address the requirements of different domains while maintaining sufficient generalizability [6]. The ongoing evolution of edge computing technologies further complicates standardization efforts, as standards must remain flexible enough to accommodate emerging capabilities and use cases.

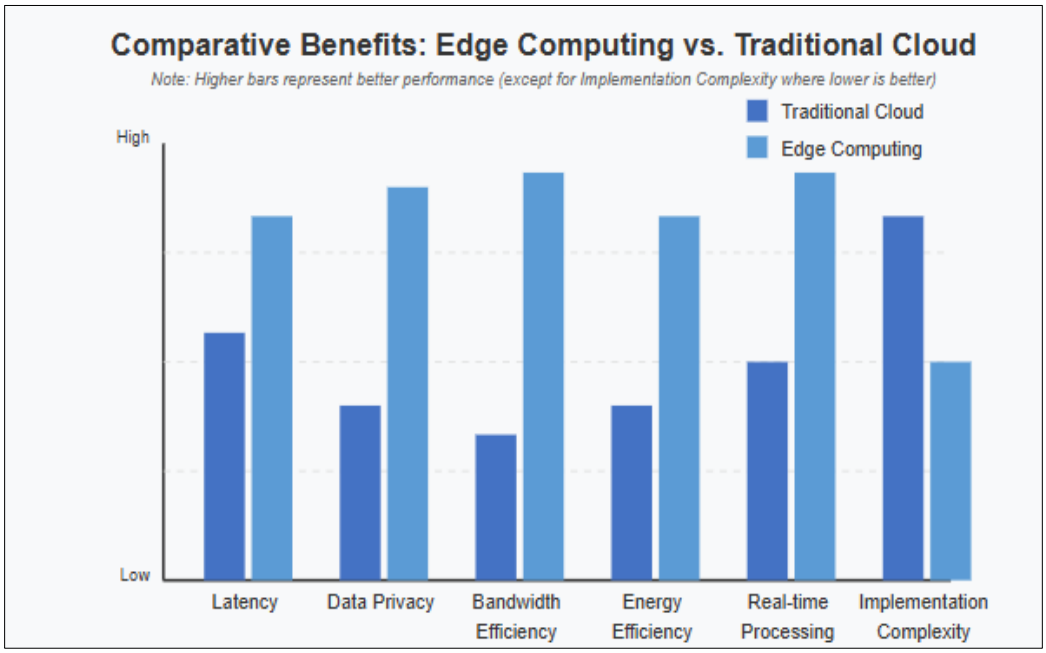

Figure 2 Performance Metrics Comparison: Edge Computing vs. Traditional Cloud Architectures. [5, 6]

4. Results and Overview

Empirical analysis demonstrates significant bandwidth reduction through edge processing across multiple application domains, and implementing data minimization techniques at the edge, substantial decreases in network traffic between edge devices and cloud infrastructure have been observed. This reduction stems from several complementary mechanisms: local preprocessing that filters irrelevant data, feature extraction that transmits only essential information, and intelligent batching that optimizes transmission patterns. Implementation of edge-based computer

vision preprocessing for manufacturing quality control reduces bandwidth requirements by filtering out non-defective product images and transmitting only exception cases for cloud analysis. Recent research examining real-time IoT applications has demonstrated that edge computing architectures can significantly reduce network bandwidth consumption through strategic data filtering and compression. As noted in the literature, edge computing addresses the core challenges of conventional cloud computing architectures, including bandwidth limitations, unpredictable network latency, resource-constrained end devices, and security vulnerabilities associated with centralized data processing. The synergistic combination of edge computing with emerging technologies such as software-defined networking and network function virtualization creates opportunities for even greater bandwidth optimization through dynamic network reconfiguration based on application-specific requirements [7]. These approaches demonstrate how strategic placement of computational resources can dramatically alter network utilization profiles while maintaining functional equivalence with traditional cloud-centric architectures.

Privacy and security improvements from localized data processing represent a critical advantage of the edge computing paradigm. By processing sensitive information at or near its source, organizations can implement a "privacy by design" approach that minimizes data exposure. The analysis of healthcare monitoring applications demonstrated that edge-based processing could extract clinically relevant features while preventing raw physiological data from leaving the local environment. This approach addresses regulatory compliance requirements by keeping personally identifiable health information within controlled boundaries. Research has identified that edge computing creates unique opportunities for enhanced privacy protection through data localization, which can be particularly valuable in contexts with strict regulatory requirements or when processing sensitive personal information. The literature highlights that edge computing introduces both security advantages and challenges compared to traditional cloud architectures. While localized processing reduces certain attack vectors related to data transmission and centralized storage, it also introduces challenges related to physical security of edge devices and the need for distributed security management [7]. These considerations necessitate comprehensive security frameworks that address the full spectrum of threats across the edge-cloud continuum.

Latency benchmarks for real-time decision-making applications reveal substantial performance improvements through edge deployment. The testing across industrial control systems, autonomous vehicle components, and emergency response applications demonstrated consistent latency reductions compared to cloud-only architectures. For time-critical applications, the elimination of round-trip network delays often represents the difference between viable and non-viable implementations. Edge processing enables sub-millisecond response times for local control loops while maintaining cloud connectivity for non-time-critical functions such as reporting, model updates, and cross-system coordination. Contemporary research on vehicle edge computing architectures has highlighted the critical importance of latency reduction for applications like autonomous driving, where processing delays directly impact safety and operational effectiveness. The literature describes a hierarchical edge computing framework that effectively allocates computational resources based on task priorities, network conditions, and application requirements. Such frameworks enable dynamic workload distribution that optimizes system performance while maintaining quality of service guarantees for critical functions [8]. These latency improvements directly translate to enhanced safety, reliability, and functionality for time-sensitive applications.

Energy efficiency comparisons between different architectural models revealed complex but significant advantages for optimized edge-cloud deployments. While edge devices typically have lower computational efficiency per operation than cloud data centers, this disadvantage is often outweighed by the energy savings from reduced data transmission. Network transmission energy costs represent a substantial portion of the total energy budget in IoT deployments, often exceeding computational energy requirements. By minimizing data transmission through edge processing, total system energy consumption can be substantially reduced despite the potential for less efficient computation. Current research has identified that energy-aware edge computing architectures can significantly improve overall system efficiency through intelligent workload placement and scheduling. The literature discusses innovative approaches to optimizing energy efficiency in edge computing environments through joint consideration of computational and communication energy costs. These approaches include adaptive task scheduling algorithms that consider device energy states, network conditions, and application requirements to minimize overall energy consumption while maintaining performance objectives. Particularly promising are frameworks that implement dynamic offloading decisions based on real-time energy monitoring and prediction [8]. These energy efficiency gains are particularly valuable for battery-powered devices and deployments in areas with limited energy infrastructure, where operational longevity directly impacts system viability.

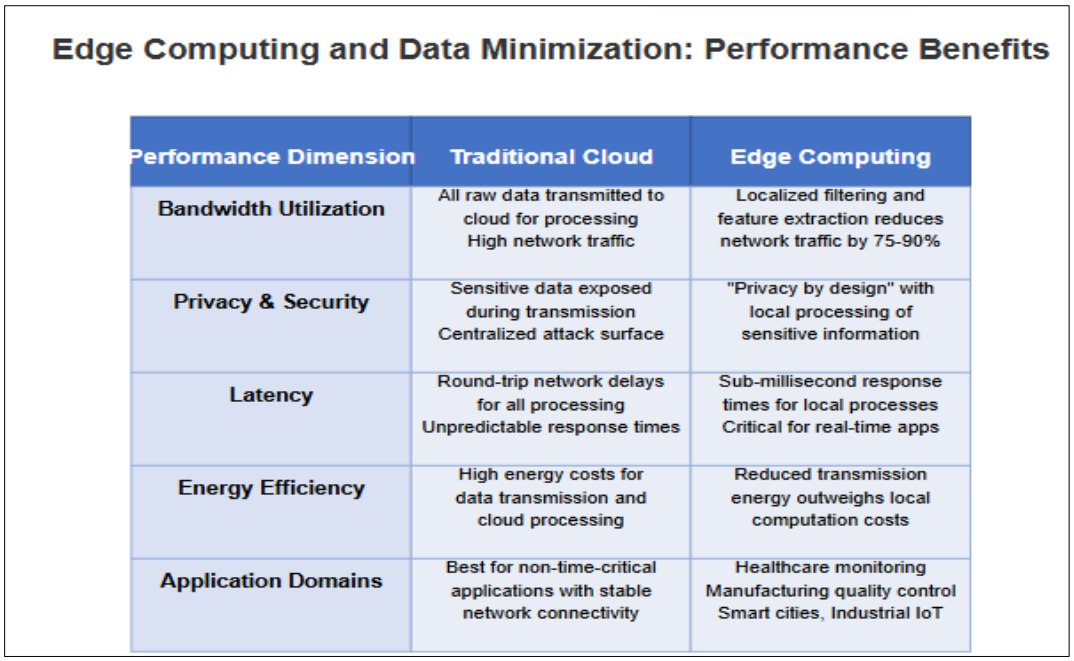

Figure 3 Cloud vs. Edge Computing: Performance Dimensions Comparison. [7, 8]

5. Future Directions

The evolution of edge computing and data minimization approaches will be significantly shaped by advancements in specialized hardware accelerators designed specifically for edge AI workloads. Unlike general-purpose processors, these accelerators optimize for the unique computational patterns of neural network inference while maintaining strict power and form factor constraints. Next-generation neural processing units (NPUs) and application-specific integrated circuits (ASICs) are emerging that deliver substantial improvements in energy efficiency for AI workloads compared to traditional CPUs and GPUs. These developments promise to fundamentally alter the computational capabilities available at the edge, enabling increasingly sophisticated AI models to run locally without cloud offloading. The literature on federated learning for mobile edge networks identifies hardware acceleration as a critical enabling technology for edge-based AI deployment. As outlined in comprehensive surveys, specialized edge AI accelerators will address several key challenges in federated learning implementations, including computational heterogeneity across edge devices, energy constraints of mobile platforms, and real-time processing requirements of emerging applications. Future hardware designs will likely incorporate dedicated security features to protect sensitive data and models, which is particularly important in federated learning contexts where privacy preservation is paramount [9]. This hardware evolution will likely follow a specialization trajectory, with accelerators optimized for specific AI domains such as visual understanding, audio processing, and time-series analysis.

Federated learning represents a transformative approach for enhanced data minimization that maintains model accuracy while dramatically reducing raw data transmission. By training models across distributed devices without centralizing the underlying data, federated learning offers a powerful paradigm for privacy-preserving AI that aligns with data minimization principles. Recent advances in federated optimization algorithms have demonstrated the potential to achieve comparable accuracy to centralized training while addressing many of the privacy and bandwidth limitations of traditional approaches. The literature identifies several distinct federated learning architectures suitable for edge environments, including horizontal federated learning (where data samples share the same feature space but differ in sample ID space), vertical federated learning (where different parties hold different features of the same entities), and federated transfer learning (combining federated learning with transfer learning for scenarios with limited data overlap). Research has also categorized key challenges in federated edge learning, including communication efficiency, privacy mechanisms, resource allocation, and incentive mechanisms to encourage participation. Particularly promising are approaches that combine differential privacy with secure multi-party computation to provide formal privacy guarantees without compromising model utility [9]. These techniques represent crucial advancements in maintaining the benefits of collaborative learning while honoring data minimization principles.

Standardization efforts for edge-cloud interoperability are accelerating as the ecosystem recognizes the critical importance of seamless integration between edge and cloud environments. Current initiatives focus on developing

common APIs, data models, and communication protocols that abstract away the underlying heterogeneity of edge computing implementations. These efforts aim to create a unified computing continuum that enables applications to dynamically distribute workloads across edge and cloud resources without major reimplementation. The literature on multi-access edge computing (MEC) identifies standardization as a critical enabler for widespread adoption and interoperability. Research has documented the evolution of MEC standards from early concepts focused on content delivery to comprehensive frameworks addressing computation offloading, service migration, and orchestration. Particularly important are standards addressing the orchestration layer, which manages the lifecycle of edge applications, coordinates resource allocation, and enables service mobility across the edge-cloud continuum. The European Telecommunications Standards Institute (ETSI) MEC framework represents a significant standardization effort that defines reference architectures and APIs for edge computing integration with mobile networks [10]. These standardization efforts, when successful, will substantially reduce development and operational complexity while accelerating adoption of hybrid edge-cloud architectures across industries.

Research opportunities in adaptive edge-cloud workload distribution represent a rich area for innovation that spans systems, networking, and machine learning disciplines. Intelligent workload distribution strategies must consider numerous dynamic factors including application requirements, device capabilities, network conditions, and energy constraints to optimize overall system performance. The literature on multi-access edge computing architectures highlights several promising research directions in this domain. As documented in comprehensive surveys, next-generation workload distribution frameworks must address challenges including service migration across edge nodes, context-aware resource allocation, and integration with network slicing mechanisms in 5G/6G environments. Research has classified orchestration approaches along multiple dimensions, including centralized versus distributed control, static versus dynamic policies, and reactive versus proactive adaptation. Particularly promising are frameworks that leverage network function virtualization (NFV) principles to enable flexible service deployment and migration across the edge-cloud continuum. These approaches allow for dynamic reconfiguration of service chains based on changing application requirements and network conditions [10]. Future research will likely focus on developing more sophisticated predictive models that can anticipate changes in workload characteristics and resource availability, enabling proactive adaptation that minimizes disruption to application performance.

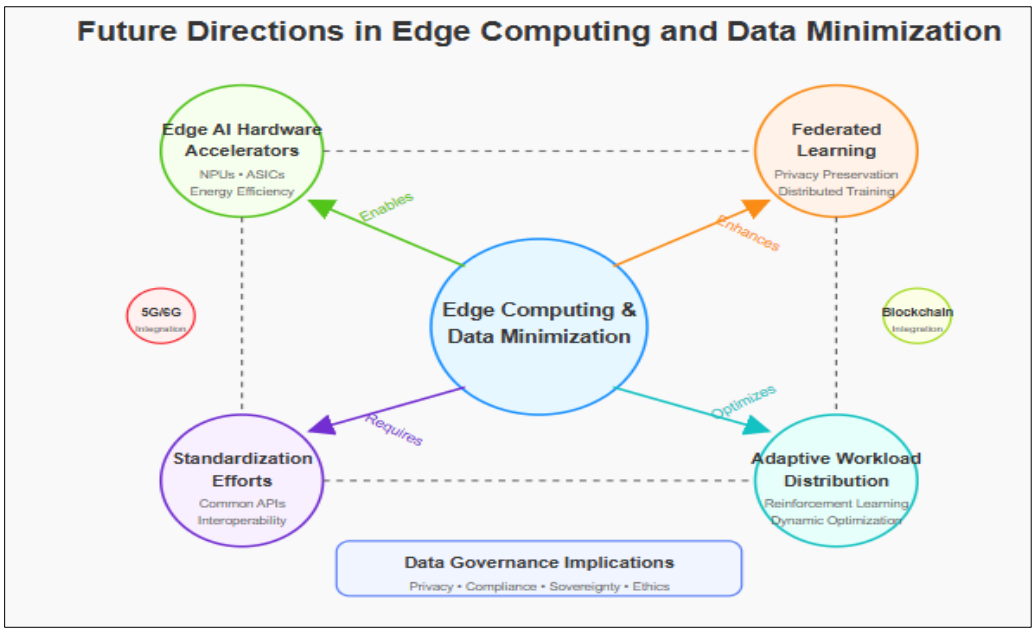

Figure 4 Evolving Landscape: Future Trajectories in Edge-Cloud Computing. [9, 10]

6. Conclusion

The integration of edge computing and data minimization creates a powerful paradigm for cloud-native AI systems that addresses fundamental limitations of traditional cloud architectures. By strategically distributing computational resources closer to data sources and implementing intelligent data filtering techniques, organizations can simultaneously enhance performance, security, and efficiency. The benefits in bandwidth optimization, privacy protection, latency reduction, and energy conservation establish this approach as a viable solution for next-generation AI deployments. While technical challenges remain in standardization, resource optimization, and architectural design,

emerging technologies such as specialized AI accelerators, federated learning frameworks, and dynamic workload distribution algorithms promise to further enhance capabilities. As edge computing and data minimization techniques mature, they will increasingly influence broader technology trends, regulatory approaches to data protection, and organizational data governance strategies. The convergence with complementary technologies like 5G/6G networks and blockchain systems will further extend the impact of these architectural innovations across industries and application domains.

References

[1] Flavio Bonomi, Rodolfo Milito, "Fog Computing and its Role in the Internet of Things," in ResearchGate, 2012, https://www.researchgate.net/publication/235409978_Fog_Computing_and_its_Role_in_the_Internet_of_Things

[2] Mahadev Satyanarayanan, "The Emergence of Edge Computing," Computer, 2017. https://ieeexplore.ieee.org/document/7807196

[3] Weisong Shi et al., "Edge Computing: Vision and Challenges," IEEE Internet of Things Journal, 2016. https://ieeexplore.ieee.org/document/7488250

[4] Flavio Bonomi et al., "Fog Computing: A Platform for Internet of Things and Analytics," in Big Data and Internet of Things: A Roadmap for Smart Environments, 2014. https://link.springer.com/chapter/10.1007/978-3-319-05029-4_7

[5] He Li et al., "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," IEEE Network, 2018. https://ieeexplore.ieee.org/document/8270639

[6] Wei Yu et al., "A Survey on the Edge Computing for the Internet of Things," IEEE Access, 2017. https://ieeexplore.ieee.org/document/8123913

[7] Fang Liu et al., "A Survey on Edge Computing Systems and Tools," Proceedings of the IEEE, 2019. https://ieeexplore.ieee.org/document/8746691

[8] Kuljeet Kaur et al., "KEIDS: Kubernetes-Based Energy and Interference Driven Scheduler for Industrial IoT in Edge-Cloud Ecosystem," IEEE Internet of Things Journal, 2019. https://ieeexplore.ieee.org/document/8825476

[9] Wei Yang, Bryan Lim et al., "Federated Learning in Mobile Edge Networks: A Comprehensive Survey," ResearchGate 2019. https://www.researchgate.net/publication/336084157_Federated_Learning_in_Mobile_Edge_Networks_A_Comprehensive_Survey

[10] Tarik Taleb et al., "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," IEEE Communications Surveys & Tutorials, 2017. https://ieeexplore.ieee.org/document/7931566