10.48047/jocaaa.2026.35.03.17

# AI-Orchestrated Multi-Cloud Edge Architecture for High-Availability Financial Services Infrastructure

**Naveen Kumar**
Independent Researcher, USA

## Abstract

As legacy centralized network architectures constrain availability, latency, and security scale in financial services, the combination of cloud computing and multi-access edge computing (MEC) with AI-orchestrated network architecture is driving transition towards architectures that can meet the combined needs of real-time transaction processing with high availability and automated compliance to regulation. AI-orchestrated multi-cloud edge is the integration of machine learning-based traffic orchestration, zero-trust security automation, and distributed edge computing across a global financial services enterprise. This architecture reduces risks from centralized traffic backhauling and systemic risks from individual clouds and enables faster and more consistent security controls than customary manual operational practices. The platform provides a programmable, policy-as-code compliance framework that implements PCI-DSS, SOC 2, and ISO 27001 controls as versioned, deployable software artifacts. It dramatically reduces the distance between point-in-time audit posture and continuous regulatory compliance. The phased and distributed implementation shows that transformation at scale is possible incrementally, without impacting the processing of any financial transactions in production. By measuring these performance, availability, latency, security response time, and operational optimization attributes, this acts as a production-validated reference architecture for financial services companies that are being hamstrung by legacy infrastructures. .

**Keywords:** Multi-Cloud Edge Architecture, AI-Driven Orchestration, Zero-Trust Security Framework, Automated Regulatory Compliance, Distributed Financial Infrastructure

## 1. Introduction

Financial services infrastructure is challenged by real-time processing requirements, heavy regulation, and adversarial information threat landscapes with thousands of attack instances a day. Customary centralized hub-and-spoke network topologies have dominated enterprise financial services environments for more than a decade and have been designed using models prior to the cloud and mobile era. Due to structural dependencies, such as traffic backhauling to regional data centers, single-cloud dependencies, and manual failover, previously mentioned issues can lead to latency penalties, system fragilities, operational bottlenecks, revenue loss, and regulatory exposure. .

In a high-throughput financial processing environment, such outages incur substantial costs in lost transaction throughput, potential service level agreement (SLA) penalties, and potential damage to reputation. The volume of digital banking and real-time payment throughput is increasing as customer tolerance for performance degradation is decreasing. Cloud computing, edge intelligence, and AI-guided automation have evolved as a way to address these challenges. The characteristics of cloud computing described by Gill et al. [1], namely dynamic, metered, on-demand access to shared pools of computing resources being a primary factor, endow the elasticity required for handling the new generation of financial workloads. The same study shows that electricity consumption in cloud data centers was still

10.48047/jocaaa.2026.35.03.17

growing year-on-year by 20-25%, hinting at the importance of architectural optimization in combination with scalability. Technology trends like IoT, blockchain, and artificial intelligence within the cloud infrastructure are referred to as the "triumvirate" by Gill et al. [1] that will shape the next-generation financial system, where AI deep learning is used for resource forecasting across the global infrastructure. Multi-Access Edge Computing (MEC) addresses the problem of network latency and bandwidth limitations brought by cloud computing. According to Filali et al. [2], MEC is defined as a model to move computing and storage resources from remote centralized cloud servers to the edge of the network to support latency-sensitive real-time applications. MEC can host URLCC, eMBB, and mMTC in 5G edge computing infrastructure. Both URLCC and eMBB applications can be applied in relevant use cases such as processing of financial transactions, early detection of fraud at the network edge, and collection of data at speed at the network edge from distributed endpoints. The ETSI MEC architectural framework [2] defines the pillars of network function virtualization, software-defined networking, service function chaining, and network slicing that underlie a programmable low-latency edge computing infrastructure. .

In contrast, previous work was not able to provide a production-proven architecture to meet extreme availability, low-latency global service delivery, and automatic regulatory compliance in a multi-cloud financial setting at the same time. Existing works provide this set of goals either via a design space optimized (mainly) for latency or (mainly) for availability or security as an afterthought and not as a programmable infrastructure primitive. In addition, Service Level Agreement (SLA) violation rates, as defined in the QoS literature [1], are another metric of interest, as they define the deviation of actual from required performance. In financial settings, SLAs are contractual obligations to not only meet but also be punishable in contracts and be associated with reputational costs. .

The contributions of this article are the design of a multi-cloud edge architecture for financial services compliance; an artificial intelligence (AI) orchestrator for autonomous traffic routing, predictive scaling, and self-healing of cloud edge services; and a programmable policy-as-code security framework at the edge of the network as part of cloud edge security. .

## 2. Related Work and Methodology

Existing cloud infrastructure and network security research present the theoretical and technical basis of the proposed architecture. Gill et al. [1] discuss the merging of IoT, blockchain, and artificial intelligence into the cloud. Accordingly, they consider dynamic resource elasticity and predictive scaling essential to next-generation distributed infrastructure. Filali et al. [2] consider Multi-Access Edge Computing a structural instantiation of a solution to reduce the latency and bandwidth bottleneck that scaled cloud solutions cannot address, particularly in 5G-compliant deployments using ultra-reliable low-latency communications. Jahantigh et al. [3] characterize policy compliance and data classification concerns for cloud provisioning of IoT data streams. They also define a regulatory area to address automated policy approaches. Shaukat et al. [4] validate the benefit of machine learning over signature-based approaches for threat detection over intrusion detection, spam detection, and malware detection through empirical experiments. Zhou et al. [5] show the need for an edge computing architecture in latency-sensitive distributed sensing applications. Finally, Li et al. [6] analyze the gains from federated reinforcement learning for edge offloading in dynamic environments. He et al. [7] formalize the concept of zero-trust architecture as continuous verification decoupled from network location, while Kratzke and Quint [9] define the target running state of a proposed cloud-native deployment as scalability, elasticity, and resilience. Both sources are merged to define AI orchestration, edge intelligence, and programmable security as the trajectory of production-grade financial infrastructure design.

10.48047/jocaaa.2026.35.03.17

## 3. Legacy Architecture Constraints and Problem Formulation

Availability, performance, and security are prominent limitations of centralized financial network architectures. All inter-regional traffic must pass through a central aggregation point, resulting in round-trip latency that is architecturally irremovable regardless of the capacity of the interconnecting links in the hub and spoke model. Static capacity planning, on the other hand, wastes infrastructure costs through over-provisioning, and service quality suffers through under-provisioning with workload spikes. Furthermore, while pervasive integration of IoT and cloud computing has reached a tipping point, the architecture of these systems is still based upon static network environments that were originally designed to manage a controlled environment with limited devices (9 billion of these devices in 2017). This was expected to increase to 24 billion by 2020 [3]. These types of financial infrastructure that rely on centralized routing cannot scale to accommodate the volume of distributed device traffic without diminishing throughput or increasing latency. .

Security controls in centralized architectures are also limited. Perimeter controls such as firewalls, intrusion detection systems and VPN gateways assume the existence of a perimeter, which cannot be maintained with mobile users and workloads that have moved into the cloud. The weaknesses in perimeter detection and signature detection have led to more than 60% of attacks being detected after-the-fact, after the attack has already impacted the environment [4]. This is particularly true in the financial services sector, which has a large volume of real-time transactions and is often targeted by large-scale distributed denial-of-service attacks and real-time fraud injection attacks. Signature-based intrusion detection systems are still a common feature of legacy financial architectures but cannot discover novel or polymorphic variants of attacks in the absence of machine learning-based rules [4]. In Cisco's 2018 annual report, more than 50% of successful attacks had more than $500 million in losses for the impacted organization, exposing financial risk in rules-based, reactive security mechanisms [4]. .

Regulatory compliance requirements such as PCI-DSS and SOC 2, and often also ISO 27001, can require continuous generation of audit evidence, controlling data residency, and encrypting data in transit and at rest. This can be especially burdensome in heterogeneous and distributed systems. When IoT data streams are aggregated with cloud services, the data from the endpoints, which can be heterogeneous and decentralized, may appear in arbitrary formats that are not classified the same way, creating gaps in the audit trail that cannot be closed systematically [3]. One common measure of deviation of actual QoS performance from required performance is SLA violation, which, for non-automated SLA compliance checking, increases rapidly with the number of concurrent transactions/load [3]. .

The problem can be simplified structurally to building a globally distributed financial services system that is 1) low-latency for transactions; 2) high availability; 3) automatically regulatory compliant; and 4) federated, immune to hostile traffic, and with no proportionate increase in headcount. Machine learning (ML) can be a structural solution to the security aspect of this requirement because ML-based methodologies have a lower error rate, better prediction accuracy, and a lower false positive detection rate than customary threat detection techniques [4]. This combined set of needs demands a shift from a centralized, perimeter-based infrastructure to a distributed, intelligence-driven infrastructure.

| Constraint Dimension | Architecture Characteristic | Operational Impact |
|---|---|---|
| Performance | Hub-and-spoke traffic backhauling to central | Architecturally irremovable round-trip latency regardless of link capacity |

10.48047/jocaaa.2026.35.03.17

| | | |
|---|---|---|
| | aggregation point | |
| Capacity Planning | Static provisioning model | Over-provisioning increases expenditure; under-provisioning degrades service quality during demand peaks |
| Security Detection | Signature-based intrusion detection systems | More than 60% of attacks identified only after damage has been caused to the targeted environment |
| Threat Complexity | Perimeter-based controls with no ML augmentation | Cannot detect zero-day or polymorphic attack variants; produces elevated false positive and error rates |
| Regulatory Compliance | Manual audit documentation and point-in-time control evidence | Compliance drift between audit cycles; inconsistency across multi-cloud environments |
| Data Classification | Heterogeneous IoT transactional data streams | No standardized classification at ingestion creates audit gaps in centralized compliance tooling |

Table 1: Comparison of Legacy Architecture Constraints and Operational Impact in Financial Services Environments [3, 4]

## 3. AI-Orchestrated Multi-Cloud Edge Architecture

### 3.1 Architectural Overview

The proposed architecture eliminates centralized routing in favor of an intelligence-driven edge fabric spanning multiple cloud providers and a distributed network of edge nodes close to populations of customers. Three design principles have been applied to the architecture: multi-cloud by design, edge-first service delivery, and API-first operations. The workloads are distributed between the major public cloud providers; traffic is terminated and inspected with respect to security, and latency-sensitive processing is performed at the edge. Zhou et al. [5] perform a survey of edge computing and clever transportation systems. They argue that applications with low latency, short response time, and high efficiency (particularly at a large scale) cannot be implemented without an edge computing architecture. The distributed sensing and processing demands of smart infrastructure are directly comparable to the globally distributed financial transaction processing network, for which a central cloud routing increases architecturally irreducible latency. The number of papers in the field shows an increase from 21 in 2011 to 199 in 2019, showing the increasing adoption of distributed edge computing in transportation and edge computing applications [5]. .

### 3.2 AI-powered orchestration layer

The orchestration layer is where the action happens. It uses machine learning to collect, analyze, and optimize traffic movement in real time, evaluating latency metrics, cost, and node health to determine the best edge vs. cloud solution. Due to resource and computation constraints in edge environments, achieving efficiency is an important goal in the design of edge orchestration algorithms. Zhou et al. [5] observe a tradeoff between the sensor's real-time sensing (i.e., latency) and intelligence. The actual tradeoff between edge and cloud computation is considered when designing the system architecture itself in terms of computational workload. .

10.48047/jocaaa.2026.35.03.17

Deep reinforcement learning (RL) models can be useful for predictive scaling and task offloading. For example, Li et al. [6] show that a deep double Q-learning (DDQN)-based adaptive offloading algorithm is better than random offloading and DQN-based algorithms in minimizing computation overhead. Time is critical for image recognition: each 50-100 ms performance increase in server-side processing improves recognition accuracy by 10-20% [6]. Thus, the ability to minimize edge-side latency is closely related to downstream processing performance. The federated learning-enabled offloading model leads to an energy reduction in the range of 15%-35% compared to local processing without federated learning and provides an overall delay improvement between 15% and 20% compared to a standard federated learning model [6]. These results will be helpful for the financial edge orchestration, with both channel conditions dynamically changing and multitasking devices resembling the heterogeneous workload for different DTP nodes.

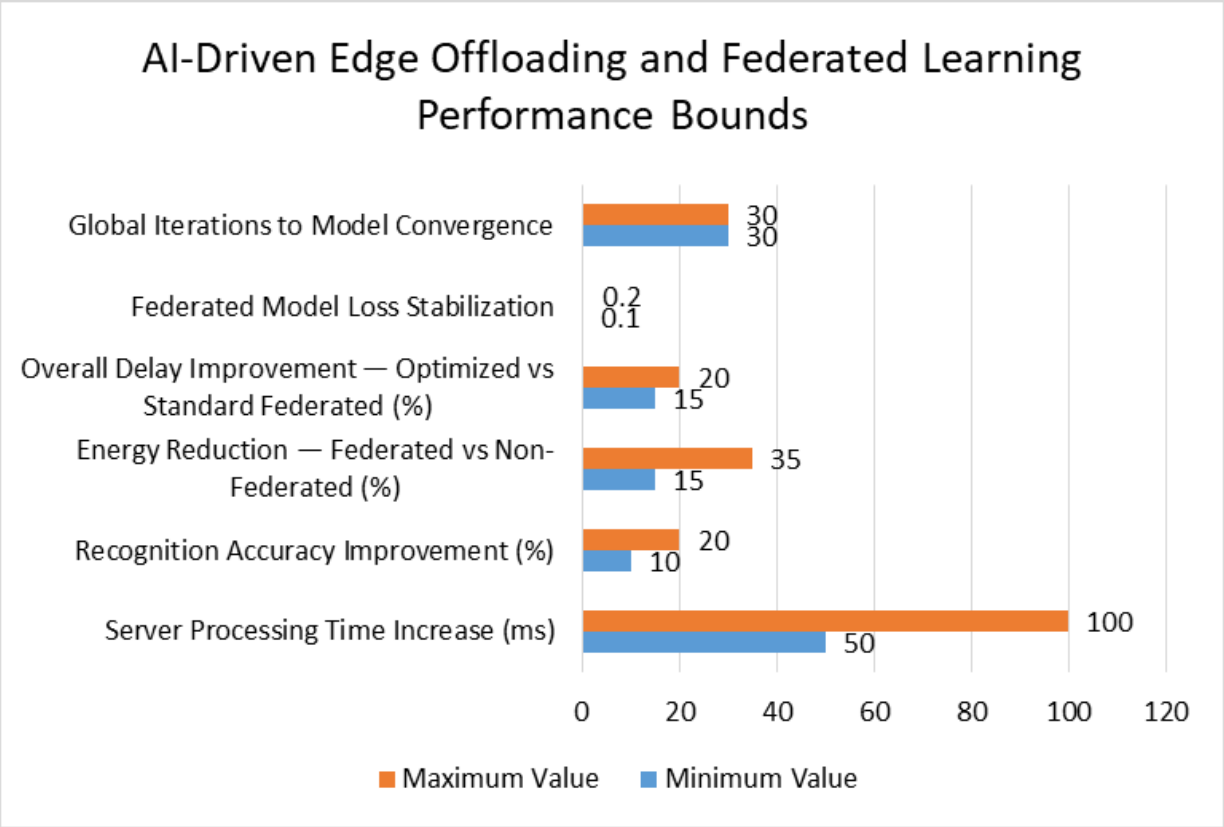


**Figure 1:** AI-Driven Edge Offloading and Federated Learning Performance Bounds [6]

### 3.3 Zero-trust security integration

Security has moved from centralized to edge-native and programmable, with a zero-trust approach that continuously verifies identity, device posture, and service authorization at all transaction boundaries. On the edge, AI models can analyze patterns in networking and transactions in real time to detect fraudulent and anomalous behavior before traffic enters core banking systems. The privacy dimensions of edge-native security are timely given that Li et al. [6] measure how federated learning can be used to train models locally on different devices without sending local data to a central server. Results show that federated learning is effective at maintaining the quality of global models while providing local privacy, where global model loss converges to 0.1-0.2 within 30 global iterations for different numbers of users.

10.48047/jocaaa.2026.35.03.17

Such a distributed privacy-preserving framework aligns with the requirements of financial services compliance environments related to data residency and confidentiality. .

# 4. Programmable Security and Regulatory Compliance Automation

## 4.1 Security Policy as Code

A conventional compliance management approach to financial infrastructure typically uses point-in-time audits and static human-maintained documentation to assert control status, without considering drift over time or the variance among multi-cloud platforms' native security tooling types. The proposed framework treats security policy as software artifacts and versions, tests, and deploys them into CI/CD and infrastructure-as-code pipelines alongside application software. .

The weaknesses of perimeter-based security models are well documented, and He et al. [7] identify the models and architectures that define the customary data center perimeter as being outdated and fundamentally inadequate for distributed cloud and mobile environments, as they define trust based on physical network location. Once authenticated under a perimeter-based model, an object is trusted until its permissions expire. Consequently, a malicious actor who passes authentication will have access to all resources. Zero Trust Architecture (ZTA) is a security model proposed by He et al. [7] to overcome the fundamental structure problem with perimeter-based models. It decouples trust from location: all users, devices, and traffic are assumed to be untrusted and must be continuously authenticated to access resources. Consequently, the principle of minimum privilege authorization assignment only grants the privileges required for execution at any given time and is verified at the boundary of a transaction [7]. .

Policies for PCI-DSS cardholder data environment requirements, SOC 2 availability and confidentiality requirements, and ISO 27001 information security management requirements are also built in as declarative policy definitions and tested against compliance rule sets automatically before being applied. Continuous monitoring is used to provide real-time audit evidence across all environments. He et al. [7] provide three technical approaches for implementing ZTA that align with their proposed policy-as-code architecture: identity authentication, access control, and trust assessment. Attribute-based access control (ABAC) is a foundational access control model for ZTA implementation, as it enables more fine-grained access control policies based on the attributes of subjects and objects and facilitates more dynamic policy enforcement across heterogeneous multi-cloud environments [7].

## 4.2 Automated Threat Response

DDoS mitigation is often performed on the network edge before attack traffic can reach internal subnets. ML-based detection models can classify attack signatures in real time and trigger mitigation actions such as rate limiting, traffic scrubbing, and anycast routing, typically within seconds of an attack. This avoids the human-in-the-loop latency characteristic of customary security operations center workflows, where, for example, the mean time to respond is measured in tens of minutes for manual environments. .

The advantage of using machine learning-based models for threat detection is based on experiments. Shaukat et al. [8] provide a detailed performance evaluation of Decision Tree (DT), Support Vector Machine (SVM), and Deep Belief Network (DBN) classifiers for intrusion detection, spam detection, and malware detection on benchmark datasets. The DBN classifier achieved a precision of 97.90% and an accuracy of 96.70% on the NSL-KDD dataset for intrusion detection, which is compared to 89.70% for the SVM classifier [8]. The Decision Tree classifier achieved the best performance for misuse detection on the KDD dataset, with an accuracy of 99.96%; while using this method of detection, it achieved better results than other classifiers [8]. When tested on the datasets used for spam filtering, the DBN classifier achieved a precision of 98.39%, an accuracy of 97.50%, and a recall of 98.02%, outperforming both SVM

10.48047/jocaaa.2026.35.03.17

and DT classifiers. The malware detection algorithms have also achieved good results with the custom datasets, with the DT classifier achieving 99.90% classification accuracy in static analysis, while the SVM classifier has achieved a recall of 100% on the Enron dataset. .

Fraud detection models running in edge nodes examine transaction characteristics and behavior for patterns against profiles of normal behavior to pre-authorize suspicious transactions before they enter processing through core systems. Shaukat et al. [8] confirmed that customary signature-based approaches suffer high false positives for polymorphic and zero-day attacks and showed that machine learning classifiers (DBN for anomaly detection and DT for misuse detection) outperform rule-based solutions in precision, recall, and accuracy. Since ML inference in edge deployments is usually real-time, it resembles the ZTA model's approach of dynamic trust assessment defined in the paper by He et al. [7], where security policies need to change depending on the behavior signals.

| Cyber Threat Category | Classifier | Dataset | Performance Outcome |
|---|---|---|---|
| Intrusion Detection | Deep Belief Network (DBN) | NSL-KDD | Precision: 97.90% — Accuracy: 96.70% |
| Intrusion Detection | Support Vector Machine (SVM) | NSL-KDD | Accuracy: 89.70% |
| Intrusion Detection | Decision Tree (DT) | KDD | Accuracy: 99.96% |
| Spam Detection | Deep Belief Network (DBN) | Enron | Precision: 98.39% — Accuracy: 97.50% — Recall: 98.02% |
| Malware Detection (Static) | Decision Tree (DT) | Custom Dataset | Accuracy: 99.90% |
| Malware Detection (Static) | Support Vector Machine (SVM) | Enron | Recall: 100% |

**Table 2:** Machine Learning Classifier Performance for Automated Threat Detection at the Financial Network Edge [7, 8]

## 5. Phased Deployment Methodology and Operational Transformation

### 5.1 Implementation Phases

This was delivered in three phases to minimize operational risk to a live financial processing environment. The pilot phase validated the AI's orchestration logic, failover, and observability instrumentation in a small, regional deployment. Latency baselines, availability baselines, and security response time baselines are measured and established to compare to the after-state.

As it entered the rollout stage, the rollout was carried out in new regions and using blue-green and canary deployments to reduce risks. The migration also included integrating the legacy core banking system, with API gateway abstractions separating legacy interfaces from new edge services of the core banking system. Similar to the movement from monolithic software architecture to cloud-native service architecture as shown by Kratzke and Quint [9], cloud-native applications are distributed, elastic, and horizontally scalable systems built with microservices and a minimum of stateful services running in self-service elastic platforms and designed with cloud-native patterns. Their systematic mapping study of the cloud-native computing ecosystem found that automation platforms, softwareization of infrastructure, and

10.48047/jocaaa.2026.35.03.17

microservice-based architectures are the dominant engineering trends in cloud-native deployment, which are the basis for the edge rollout methodology described in this paper.

Global expansion completed the transformation, extending edge coverage to over 150 countries while shutting down legacy centralized routing infrastructure. The governance moved to a policy-based model with regional and edge-tier control. The CNA property combination, as documented by Kratzke and Quint [9], of horizontal scalability, elasticity, and resilience, describes the target operational characteristics for the operation of the finalized global deployment. Now, the vendor lock-in risk, a structural limitation of multi-cloud CNA deployments [9], is also lowered through adopting standard container deployment units and orchestration tools, which enable workload portability across cloud provider boundaries without architectural rework. .

## 5.2 Organizational Enablement

Organization design was done in parallel to technical transformation, creating new roles, such as Platform Site Reliability Engineers (SREs) and Cloud Security Architects, that own automation-first operational functions. Networking, security, and application engineering teams were cross-functional, rather than organized in separate departments.

Organizational aspects of microservices adoption have been widely studied in the empirical literature. Taibi et al. [10] conducted structured interviews with 21 experienced practitioners who had been working with microservices-based architectures for at least two years. The three perspectives used here show that maintainability and scalability are ranked the highest across all groups. Implementation level—Delegation of team ownership to self-contained units was the third most important. Eleven of the 21 participants indicated a major impact [10]. Cost—An overhead cost of 20% to 30% was apparent at the beginning of the microservices journey, according to 76% of the participants. The other 24% of participants had an effort overhead lower than 10% [10]: while more expensive to implement, the majority of respondents recouped their initial investment within two to three years (33% within two years, 66% within three years) [10]. Furthermore, they report that long-term maintenance effort, which is less for these features, compensates for their higher initial cost. This directly applies to the long-term transformation of financial services, where regulation as well as business models give credence to the sustainable operational efficiency of financial services.

Investing in internal expertise in addition to vendor partnerships also helped to ensure that automation systems are understood, trusted, and improved by internal teams, rather than treated as black boxes for which teams are dependent on vendors. Taibi et al. [10] also found that effort estimation accuracy improved measurably after the introduction of microservices as a result of lower granularity in service-level estimation and an absence of cross-team synchronization dependencies, a benefit that scales with global financial infrastructure.

10.48047/jocaaa.2026.35.03.17

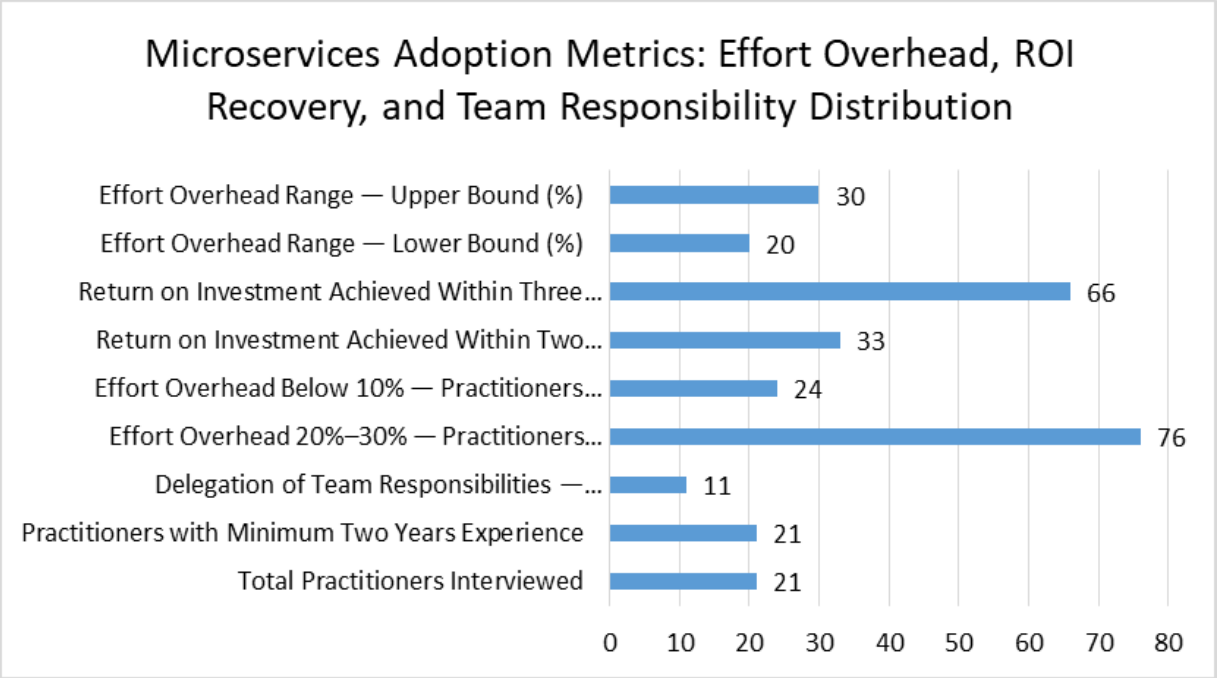


**Figure 2:** Microservices Adoption Metrics: Effort Overhead, ROI Recovery, and Team Responsibility Distribution [10]

## Conclusion

AI-orchestrated multi-cloud edge architecture fundamentally addresses the limitations of incremental optimization by disaggregating, distributing, and pushing intelligence, security enforcement, and processing power to the network edge closest to the end user, eliminating the latency and single-cloud systemic risk of traffic backhauling, while enabling security automation at a scale and speed that is infeasible with human-driven operations. A programmatic, policy-as-code compliance framework bridges the point-in-time audit posture and continuous regulatory compliance gap for PCI DSS, SOC 2, and ISO 27001 by codifying policy as versioned, testable, deployable software artifacts in a CI/CD pipeline. Zero-trust security is applied to every transaction boundary. Continuous identity verification and attribute-based access control supersede perimeter security models supported by a well-defined network edge, no longer a valid assumption in distributed cloud and mobile-first systems. The gradual rollout of cloud-native transformation across multiple geographies shows that it is feasible to perform cloud transformation without an operational halt to live transactions. The properties of horizontal scalability, elasticity, and resilience that cloud enables through workload decomposition using microservices, workload decoupling using containers, and elastic orchestration platforms produce measurable operational gains in availability, latency, and security response. For financial services organizations at the intersection of real-time payment volumes, adversarial threat landscapes, and mounting regulatory requirements, a distributed, intelligence-driven infrastructure is a structurally sound and production-validated future state. .

10.48047/jocaaa.2026.35.03.17

## References

[1] Sukhpal Singh Gill et al. "Transformative Effects of IoT, Blockchain and Artificial Intelligence on Cloud Computing: Evolution, Vision, Trends and Open Challenges." Internet of Things, 2019. [Online]. Available: https://arxiv.org/pdf/1911.01941

[2] ABDERRAHIME FILAL et al. "Multi-Access Edge Computing: A Survey," IEEE Access, vol. 8, 2020, [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9240934

[3] Motahareh Nazari Jahantigh et al. "Integration of Internet of Things and cloud computing: a systematic survey," IET Journals, 2019. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/pdfdirect/10.1049/iet-com.2019.0537

[4] KAMRAN SHAUKAT et al. "A Survey on Machine Learning Techniques for Cyber Security in the Last Decade." IEEE Access, 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9277523

[5] Xuan Zhou et al. "When Intelligent Transportation Systems Sensing Meets Edge Computing: Vision and Challenges." MDPI, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/20/9680

[6] Jie Li et al. "Task offloading mechanism based on federated reinforcement learning in mobile edge computing." Digital Communications and Networks, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864822000554

[7] Yuanhang He et al. "A Survey on Zero Trust Architecture: Challenges and Future Trends," Wireless Communications and Mobile Computing, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/6476274

[8] Kamran Shaukat et al. "Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective." [Online]. Available: https://www.researchgate.net/profile/Kamran-Shaukat/publication/344704519

[9] Nane Kratzke and Peter-Christian Quint. "Understanding cloud-native applications after 10 years of cloud computing—A systematic mapping study." Journal of Systems and Software, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0164121217300018

[10] Davide Taibi et al. "Processes, motivations, and issues for migrating to microservices architectures: An empirical investigation." IEEE Cloud Computing, 2017. [Online]. Available: https://m3s-cloud.github.io/assets/paper1.pdf