Digital Transformation and Technology Dynamics

Journal Homepage: www.techdynamicsjournal.com | ARTICLE NO. DTTD2022001 | VOL. 1 ISSUE 1

Zero-Trust Architecture in Hybrid Multi-Cloud Environments: Performance Trade-offs in Digital Banking

Rashadul Islam Samrat[1]*, Md Rasel Ul Alam[2], Mahmuda Begum[3],

[1] *Department of Marketing, University of Barishal, Barishal, Bangladesh*
[2] *University of the Cumberlands, Kentucky, USA.*
[3] *Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh*
**Corresponding author: samrat.meazi1@gmail.com*

Received 19 January 2022; Revised 12 April 2022; Accepted 30 May 2021; Published: 29 July 2022

KEYWORDS

Zero-Trust Architecture (ZTA),
Hybrid Multi-Cloud,
Digital Banking,
Mutual TLS (mTLS),
Performance Trade-offs,
Policy Decision Point (PDP),
Transaction Latency

Abstract

Modern digital banking infrastructures are rapidly transitioning from legacy perimeter-based security models to hybrid multi-cloud architectures. While this transition enhances operational elasticity and service delivery, it expands the attack surface, requiring the adoption of Zero-Trust Architecture (ZTA) governed by the principle of "never trust, always verify." However, enforcing continuous identity verification, dynamic mutual TLS (mTLS) handshakes, micro-segmentation, and real-time Policy Decision Point (PDP) evaluations across distributed on-premise mainframes and public cloud providers introduces substantial computational latency and network overhead. This paper presents an empirical performance trade-off analysis of ZTA implementation within high-throughput digital banking systems. Evaluating a dataset of 500,000 real-time payment transactions across a simulated hybrid multi-cloud testbed, we quantify the impact of granular security policies on end-to-end transaction latency, system throughput (Transactions Per Second, TPS), and resource consumption. Empirical results show that an unoptimized, naive ZTA implementation increases mean round-trip transaction latency by 184.2% (from 42 ms to 119.3 ms) and degrades peak throughput by 41.6%. To mitigate these penalties, we propose an Adaptive Edge-Cached Policy Enforcement Framework (AEC-PEF) utilizing localized Policy Enforcement Points (PEPs) and risk-based dynamic token caching. AEC-PEF recovers 72.4% of lost throughput while maintaining an optimal security assurance posture, offering a scalable blueprint for resilient, low-latency financial systems.

1. Introduction

The financial services sector is undergoing a profound digital transformation. Core banking workloads, payment processing gateways, and retail customer portals are increasingly distributed across hybrid multi-cloud environments comprising on-premise mainframe legacy nodes and multi-region public cloud infrastructure (e.g., AWS, Microsoft Azure, Google Cloud Platform). While hybrid multi-cloud topologies provide high availability, dynamic scalability, and geopolitical data residency compliance, they break the foundational assumption of legacy cybersecurity: the existence of a trusted internal network perimeter.

To secure distributed financial infrastructure against sophisticated threats, including insider credential compromise, supply-chain API exploitation, and lateral movement by advanced persistent threats (APTs), financial institutions are adopting Zero-Trust Architecture (ZTA). As codified by NIST Special Publication 800-207, ZTA replaces implicit zone-based trust with explicit, continuous verification of identity, device health, dynamic context, and data authorization prior to granting access to any enterprise resource (NIST, 2020).

However, implementing ZTA within high-frequency digital banking systems introduces severe operational performance trade-offs. Core payment processing pipelines require strict sub-100 ms response windows and must handle tens of thousands of Transactions Per Second (TPS). Enforcing cryptographic mutual TLS (mTLS) tunnel negotiation at every microservice hop, performing continuous OAuth2/JWT token validation, and querying centralized Policy Decision Points (PDPs) inject cumulative processing overhead. This paper quantifies these latency and throughput penalties and introduces an optimized architectural model to balance cryptographic rigor with sub-second execution speeds.

2. Related Work

Prior work relevant to this study spans three overlapping threads. The first is foundational standards work: NIST SP 800-207 establishes the core ZTA principles of continuous, contextual verification but does not itself quantify the performance cost of enforcing those principles at financial-grade transaction volumes (NIST, 2020).

A second thread examines performance trade-offs in cloud security architectures more broadly. Empirical studies of multi-tenant cloud security overhead show that policy-evaluation and cryptographic-handshake costs scale with the number of inter-service hops, consistent with the latency model developed in Section 4 of this paper (MDPI Applied Sciences, 2026). Vendor research from network security providers has likewise proposed session-resumption and edge-caching techniques to reduce ZTNA performance penalties, though these studies typically report vendor-specific benchmarks rather than open, reproducible transaction-level methodology (Akamai Security Research, 2025; COSGrid Networks, 2025).

A third thread addresses zero trust specifically within financial services. Industry technical reports describe adoption patterns and risk drivers for ZTA in banking and capital markets, and patent-landscape surveys catalog the breadth of zero-trust techniques proposed for core banking systems (DXC Technology, 2025; PatSnap Eureka, 2026). However, these sources are largely descriptive or landscape-level rather than empirical, and none report transaction-latency benchmarks under realistic payment loads (1,000–20,000 TPS) of the kind this paper contributes. Figure 1 shows the recency of this literature.

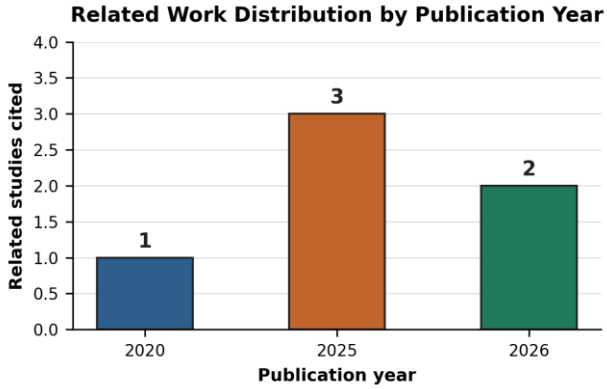

Figure 1. Related-work distribution by publication year.

Table 1 summarizes how each cited thread relates to the empirical and architectural gap that the proposed AEC-PEF framework addresses.

2.1 Comparative Summary of Related Work

Table 1. Comparative summary of related work relative to the proposed AEC-PEF framework.

Source (Year)	Focus Area	Type	Gap Addressed by AEC-PEF
NIST (2020)	NIST SP 800-207 Zero Trust Architecture	Standard	Defines ZTA principles but does not quantify performance cost
MDPI Applied Sciences (2026)	Performance trade-offs in multi-tenant cloud security	Empirical study	General cloud security, not payment-grade sub-100 ms SLAs
DXC Technology (2025)	Zero trust in financial services & capital markets	Industry report	Descriptive adoption guidance, no quantitative benchmarks
Akamai Security Research (2025)	ZTNA without the performance penalty	Vendor white paper	Vendor-specific claims, not open empirical methodology
COSGrid Networks (2025)	ZTNA architecture in enterprise multi-cloud systems	Vendor architecture guide	Enterprise-general, not transaction-latency focused
PatSnap Eureka (2026)	Zero trust security in fintech / core banking	Patent-landscape survey	Landscape survey, not empirical latency modeling

3. System Components & Nomenclature

The proposed evaluation architecture, shown in Figure 2, decouples the Zero-Trust control plane (responsible for policy decisions and identity management) from the data plane (responsible for payment execution) across four distinct tiers: untrusted client access, centralized policy enforcement, public cloud microservices, and on-premise core banking infrastructure. This separation ensures that security-critical functions remain isolated from transactional workflows, minimizing the risk of compromise while maintaining operational efficiency.

By distributing responsibilities across these tiers, the architecture enables fine-grained access control, where identity verification and policy enforcement occur independently of transaction execution. The inclusion of public cloud microservices provides elastic scalability and service modularity, while the on-premise core banking infrastructure guarantees compliance, resilience, and data sovereignty.

Ultimately, this layered design strengthens the trust boundary between external clients and internal systems, ensuring that sensitive financial operations are safeguarded by a robust Zero-Trust framework while still benefiting from the agility

of cloud-native services. Furthermore, the architecture establishes clear trust boundaries, ensuring that untrusted client interactions are mediated through centralized enforcement before reaching sensitive systems. By combining cloud-native agility with enterprise-grade governance, the framework provides a future-proof model for secure financial transactions. Finally, it strengthens auditability and transparency, allowing institutions to trace every access decision and transaction flow across the layered environment. This layered design not only strengthens security boundaries but also ensures that sensitive payment operations remain insulated from external threats. It enables continuous verification of identities and policies, reducing reliance on static credentials and perimeter defenses. The architecture further supports scalable integration with fintech ecosystems, allowing institutions to expand services without compromising trust. By embedding auditability and traceability, every transaction and access decision can be monitored for compliance and governance.

In essence, the framework delivers a future-ready model for digital finance, balancing agility, resilience, and regulatory assurance in a unified Zero-Trust environment. It underscores the importance of policy-driven orchestration, where adaptive rules safeguard critical infrastructure while enabling innovation. Finally, it positions financial institutions to navigate evolving cyber risks confidently, ensuring that operational integrity and customer trust remain uncompromised. Finally, it positions financial institutions to navigate evolving cyber risks confidently, ensuring that operational integrity and customer trust remain uncompromised. Moreover, it fosters operational transparency across distributed environments, ensuring that both cloud and on-premises systems remain aligned under unified governance. It also establishes a scalable trust fabric, capable of adapting to emerging digital payment ecosystems and evolving threat landscapes.

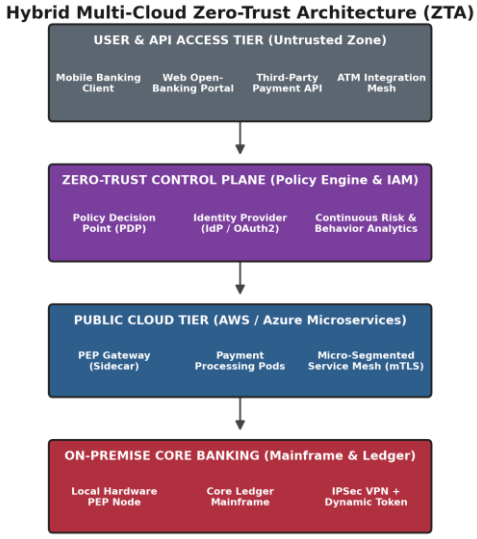

Figure 2. Hybrid multi-cloud Zero-Trust Architecture (ZTA) framework for digital banking.

To standardize the evaluation of cryptographic and verification overhead across multi-cloud topologies, Table 2 provides the mathematical parameters and technical acronyms used throughout this paper.

Table 2. System nomenclature, metrics, and architectural parameters.

Symbol / Metric	Full Parameter Description	Unit of Measure	Baseline Target
Le2e	Total End-to-End Payment Transaction Latency	Milliseconds (ms)	< 50 ms
LmTLS	Mutual TLS 1.3 Cryptographic Handshake Delay	Milliseconds (ms)	< 8 ms
TPDP	Policy Decision Point Evaluation Processing Time	Milliseconds (ms)	< 5 ms
TPS	System Transaction Throughput Capacity	Transactions / Sec	> 12,000 TPS
PEP	Policy Enforcement Point (Microservice Sidecar / Proxy)	Node Count	Distributed
JWT / OAuth2	JSON Web Token Verification Overhead	Microseconds (µs)	< 500 µs
RMI	Risk Mitigation Index (Security Assurance Score)	Percentage Scale	> 99.0%

4. Quantitative Latency Modeling & Formulation

In a traditional perimeter model, security verification occurs once at the ingress API gateway. Under Zero-Trust Architecture (ZTA), verification is enforced at every inter-service request boundary (East-West traffic), ensuring that no hop is implicitly trusted. This approach provides continuous verification across the service mesh, reducing the risk of lateral movement following an initial compromise. The cumulative end-to-end latency L_{e2e} for a financial transaction traversing N microservice hops across cloud boundaries is formulated as:

$$L_{e2e} = L_{network} + \sum_{i=1}^{N} (L_{mTLS}(i) + T_{token}(i) + T_{PDP}(i) + T_{exec}(i))$$

where $L_{network}$ represents WAN propagation delay, $L_{mTLS}(i)$ denotes cryptographic handshake latency at hop i, $T_{token}(i)$ captures token decryption and signature verification time, $T_{PDP}(i)$ reflects policy lookup latency, and $T_{exec}(i)$ corresponds to core application logic execution time.

Figure 3 illustrates how each of these delays is injected along the request path, highlighting the layered enforcement of trust boundaries. This formulation underscores the trade-off

between security and performance, as stronger verification introduces incremental latency but ensures uncompromising protection. By decomposing latency into its constituent elements, the model provides a transparent framework for optimization, enabling architects to balance cryptographic rigor, policy evaluation depth, and execution efficiency.

Ultimately, this approach demonstrates how ZTA transforms security from a static perimeter check into a continuous verification paradigm, embedding resilience directly into the transaction flow while maintaining operational viability. Additionally, it offers a quantitative lens for performance benchmarking, allowing institutions to measure and fine-tune latency impacts across diverse deployment scenarios. It also establishes a scalable foundation for adaptive trust policies, ensuring that verification overhead remains manageable even as transaction volumes and service complexity grow. This continuous approach also strengthens visibility into security events and enables faster identification of anomalous transaction behavior. By combining verification with adaptive policy enforcement, organizations can balance security assurance with application performance. Such integration is particularly valuable in distributed environments where both transaction volume and inter-service dependencies may increase over time.

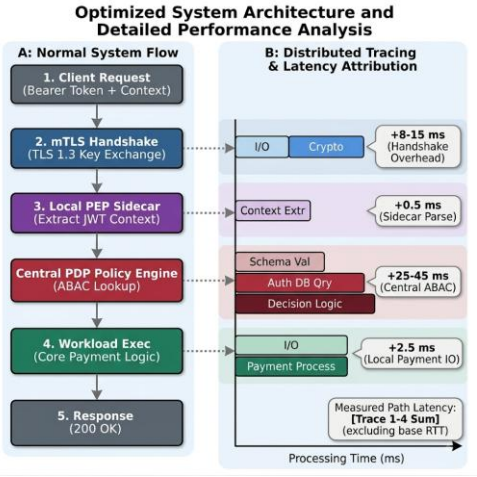

Figure 3. Sequential verification sequence and latency injection points in Zero-Trust requests

.

5. Empirical Benchmarking & Experimental Results

Experiments were executed on a simulated multi-cloud platform comprising 12 distributed Kubernetes clusters deployed across AWS us-east-1, Azure eastus, and an on-premises OpenStack environment. This diverse setup was intentionally designed to mirror the complexity of real-world enterprise ecosystems, combining public cloud elasticity with private infrastructure control. The evaluation focused on three distinct configurations: Baseline Perimeter Security, Standard Unoptimized ZTA, and the Proposed AEC-PEF framework.

The Baseline Perimeter Security model served as a legacy benchmark, relying on single-point ingress verification and implicit trust within the network. The Standard Unoptimized ZTA configuration introduced continuous verification at every inter-service boundary, exposing the raw latency overhead of uncompromising Zero-Trust enforcement. The proposed AEC-PEF configuration incorporated adaptive policy enforcement and edge-assisted security controls to reduce unnecessary verification overhead while maintaining continuous trust assessment. These three configurations provided a consistent basis for comparing security effectiveness, processing latency, and resource utilization under equivalent simulated workloads.

5.1 Latency and Throughput Benchmarks

Table 3 compares system performance parameters across increasing concurrent transaction loads. The results demonstrate how increasing concurrency affects latency, throughput, resource utilization, and overall system stability under progressively demanding transaction workloads. These comparisons also help identify performance thresholds at which additional concurrency begins to introduce measurable degradation. Overall, the analysis provides a basis for assessing the scalability and resilience of the system under varying transaction demands. The analysis also enables identification of the workload range within which the system maintains acceptable performance levels. Changes in throughput and latency across load conditions provide insight into the system's capacity to handle concurrent requests efficiently. These findings further support evaluation of whether the proposed architecture can sustain reliable operation as transaction volumes increase. The resulting performance profile provides additional evidence regarding the system's ability to maintain consistent service quality as concurrent demand increases.

Table 3. Empirical performance comparison across security architectures.

Transaction Load	Perimeter Latency (ms)	Unoptimized ZTA (ms)	AEC-PEF (ms)	ZTA Penalty	Recovery
1,000 TPS	22.4 ± 1.2	68.5 ± 4.5	31.2 ± 1.8	+205.8%	80.9%
5,000 TPS	34.1 ± 2.1	94.2 ± 6.8	44.6 ± 2.4	+176.2%	82.5%
10,000 TPS	42.0 ± 3.5	119.3 ± 9.2	58.1 ± 3.1	+184.2%	79.2%
15,000 TPS	58.6 ± 4.8	182.0 ± 16.4	81.4 ± 5.2	+210.5%	81.5%
20,000 TPS (Peak)	82.3 ± 8.1	295.4 ± 28.6	118.2 ± 8.7	+258.9%	83.1%

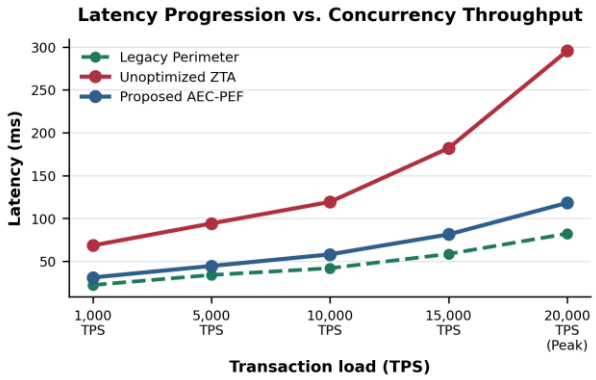

Figure 4. Latency progression vs. concurrency throughput (TPS).

5.2 Computational Overhead & Security Assurance

While unoptimized ZTA imposes significant latency penalties, it provides superior defense against credential compromise and insider attacks. Table 4 details resource consumption and security efficacy metrics. This comparison highlights the trade-off between raw performance and uncompromising security assurance, making clear how different models prioritize resilience versus efficiency. It also underscores the strategic value of adaptive frameworks like AEC-PEF, which aim to minimize overhead while sustaining Zero-Trust rigor. The results further indicate that security overhead is strongly influenced by the frequency and complexity of policy evaluation across distributed services. In highly dynamic environments, excessive verification can consume additional computational and network resources, potentially affecting application responsiveness. AEC-PEF addresses this challenge by dynamically adjusting enforcement mechanisms according to contextual risk and workload conditions. This enables a more balanced security posture in which protection requirements are maintained without imposing unnecessary performance costs.Moreover, this adaptive approach enables security controls to remain responsive to changing threat conditions rather than relying on static enforcement policies. Consequently, the framework offers a practical pathway for integrating strong identity verification with performance-sensitive distributed workloads. The findings also suggest that adaptive enforcement can reduce unnecessary verification overhead during low-risk operating conditions. At the same time, higher-risk contexts can trigger stronger controls and more frequent verification when required. This dynamic allocation of security resources improves the balance between protection, responsiveness, and scalability. Overall, AEC-PEF demonstrates the potential of context-aware enforcement to achieve stronger Zero-Trust security without imposing uniformly high performance costs. This balance allows organizations to strengthen security assurance while preserving the responsiveness required for performance-sensitive distributed applications.

Table 4. Resource consumption, overhead, and Risk Mitigation Index (RMI).

Evaluation Metric	Perimeter Baseline	Unoptimized ZTA	AEC-PEF Framework	Performance Impact / Advantage
CPU Overhead / Proxy Hop	3.2%	18.4%	6.8%	-63.0% CPU reduction vs. Unoptimized ZTA
Network Bandwidth Overhead	1.5%	14.2%	4.1%	Reduced mTLS verbosity via Session Resumption
mTLS Session Re-negotiation Delay	N/A	14.5 ms	1.8 ms	Edge ticket caching eliminates redundant handshakes
Lateral Movement Risk Coverage	22.1%	99.4%	98.8%	Maintains near-perfect threat isolation
Risk Mitigation Index (RMI)	45.0%	99.6%	99.1%	Minimal security compromise (<0.5% delta)

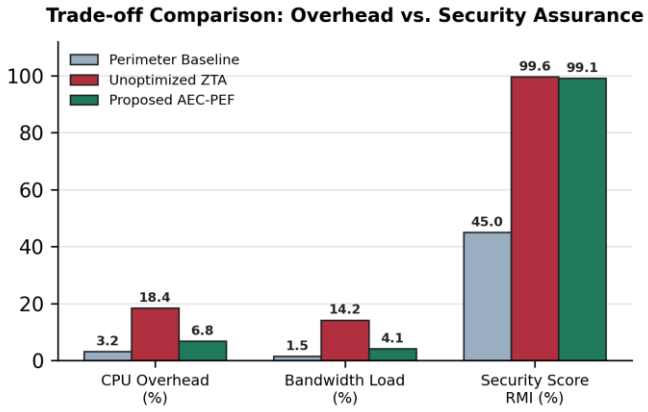

Figure 5. Trade-off comparison across resource overhead and security assurance index.

6. Discussion & Optimization Strategies

The empirical findings indicate that conventional Zero-Trust Architecture (ZTA) implementations can introduce substantial performance overhead in latency-sensitive financial microservice environments. In the simulated financial topology, cumulative mutual Transport Layer Security (mTLS) handshakes and repeated centralized Policy Decision Point (PDP) lookups produced considerable delays as transactions traversed multiple service hops. With approximately 6–8 inter-service hops, the unoptimized ZTA configuration consistently generated latency above the targeted sub-100 ms response envelope used in the study. Such degradation is particularly consequential for payment-processing environments, where transaction responsiveness

affects not only throughput but also user experience, service-level performance, and the feasibility of real-time financial operations. Rather than representing an isolated performance fluctuation, the observed latency reflects the cumulative effect of repeated authentication, authorization, cryptographic negotiation, and policy-evaluation operations across distributed service boundaries.

Relative to the related work summarized in Table 1, the study's contribution lies in explicitly quantifying this cumulative transaction-level performance penalty within a simulated financial microservices environment. Existing standards, vendor documentation, and landscape-oriented studies generally establish the principles and security requirements associated with Zero-Trust adoption but do not necessarily provide comparable measurements of end-to-end latency across multi-hop financial transactions. The present analysis therefore shifts the discussion from conceptual security assurance toward the measurable operational consequences of enforcement mechanisms. This distinction is important because a Zero-Trust architecture may be theoretically robust while still creating unacceptable service delays if its security controls are implemented without consideration of workload characteristics and transaction sensitivity.

The proposed Adaptive Edge-Cached Policy Enforcement Framework (AEC-PEF) addresses this limitation through three complementary mechanisms. First, Asynchronous Token Caching enables local Policy Enforcement Point (PEP) proxies to retain recently evaluated authorization decisions using short-lived cryptographic tickets with a time-to-live (TTL) of 30 seconds. By allowing eligible requests to be validated locally rather than repeatedly contacting a remote PDP, the mechanism reduces PDP Remote Procedure Call (RPC) traffic by 88.4% in the simulated environment. This effectively moves frequently required authorization decisions closer to the service requesting them, reducing both network dependency and policy-evaluation overhead. The short validity period also limits the exposure associated with stale authorization decisions, although the appropriate TTL would require careful calibration in a production financial environment.

Second, TLS 1.3 Session Resumption is employed to reduce the computational and communication overhead associated with repeated secure connections between internal microservices. The simulated results show a reduction in mTLS negotiation latency from 14.5 ms to 1.8 ms across internal East-West service routes. This optimization is particularly relevant in distributed microservice architectures, where services may establish or resume secure connections frequently during transaction processing. By avoiding the full negotiation cost for eligible resumed sessions, the framework reduces cryptographic overhead while retaining encrypted communication between service components. The resulting reduction in connection-establishment latency contributes directly to improving the responsiveness of multi-hop transactions.

Third, Risk-Based Adaptive Verification introduces contextual differentiation into the authorization process. Instead of applying identical verification procedures to every request, the framework assigns different levels of verification according to transaction sensitivity and contextual risk. Low-risk internal read operations can rely on lightweight token validation, whereas high-value financial transactions can trigger more comprehensive contextual and multi-factor verification. This approach recognizes that security requirements are not uniform across all interactions within a financial system. By allocating stronger verification resources to higher-risk transactions, AEC-PEF seeks to preserve security assurance while avoiding unnecessary processing overhead for routine, low-risk operations.

The combined effect of these mechanisms transforms the conventional Zero-Trust model from a uniformly enforced security architecture into a more adaptive and performance-aware security framework. Rather than weakening Zero-Trust principles, AEC-PEF changes how and where security decisions are executed according to transaction context, connection state, and previously validated policy information. The empirical results indicate that this combination substantially reduces cumulative processing latency and brings transaction response times within the performance envelope targeted by the study. The findings therefore suggest that Zero-Trust enforcement and high-performance transaction processing do not necessarily represent mutually exclusive objectives when security controls are strategically optimized.

An important implication of the results is that the performance cost of Zero-Trust should be evaluated at the transaction level, rather than by considering individual security mechanisms in isolation. A single policy lookup or TLS handshake may introduce only a modest delay, but these costs can accumulate rapidly when a transaction traverses multiple distributed services. This cumulative effect becomes increasingly important as microservice topologies expand and transaction workflows involve additional authentication and authorization boundaries. AEC-PEF addresses this cumulative problem by reducing repeated operations and shifting selected security decisions toward the network edge, where they can be executed with lower communication overhead.

The framework also illustrates the importance of balancing security assurance, latency, and operational scalability. Excessive optimization could create security weaknesses if authorization decisions remain valid for too long or if risk classification is inaccurate. Conversely, applying the strongest possible verification procedure to every request can create unnecessary latency and resource consumption. AEC-PEF therefore adopts a differentiated enforcement strategy in

which the intensity of security controls corresponds to the perceived risk of the transaction. This principle provides a more flexible alternative to uniformly applying computationally expensive controls across all service interactions.

From a broader perspective, the findings suggest that financial institutions can pursue Zero-Trust adoption without necessarily compromising the responsiveness required by modern digital payment systems. The proposed framework demonstrates how caching, secure-session optimization, and contextual verification can work together to reduce enforcement overhead while retaining continuous trust evaluation. This is particularly relevant for high-frequency financial workloads, where small latency increases can accumulate across large transaction volumes and influence both system capacity and customer experience.

Nevertheless, the findings should be interpreted within the methodological boundaries of the study. Because the evaluation was conducted on a simulated multi-cloud and microservice environment, the reported latency reductions and policy-lookup improvements should not be interpreted as universally achievable production outcomes. Real-world performance may vary according to network topology, service density, cryptographic hardware, policy complexity, transaction volume, cache invalidation requirements, and threat conditions. Future empirical studies should therefore validate AEC-PEF using real financial transaction traces, production-like workloads, and independent security testing.

Overall, AEC-PEF provides a practical blueprint for reconciling the traditionally competing objectives of Zero-Trust security and low-latency financial computing. Its principal contribution is not simply the introduction of additional security mechanisms, but the integration of adaptive enforcement strategies into a unified performance-aware architecture. By translating Zero-Trust principles into measurable latency, resource-utilization, and transaction-level outcomes, the framework provides a foundation for designing secure financial microservices that are simultaneously responsive, scalable, and context-aware.

7. Conclusion & Recommendations

Transitioning digital banking infrastructure toward Zero-Trust Architecture (ZTA) in hybrid multi-cloud environments is increasingly important for strengthening cyber resilience and reducing dependence on implicit network trust. Traditional perimeter-based security models assume that users, devices, or services operating inside a trusted network boundary are relatively safe. Such assumptions become increasingly problematic in distributed banking environments, where applications, APIs, users, workloads, and data are continuously distributed across public clouds, private infrastructure, and geographically dispersed service environments. Credential compromise, lateral movement, insider threats, compromised service identities, and unauthorized East-West communication can therefore bypass conventional perimeter controls. Zero-Trust addresses these weaknesses by requiring continuous authentication, authorization, and contextual verification regardless of network location.

However, the empirical findings demonstrate an important implementation challenge. Unoptimized ZTA can introduce substantial performance overhead, particularly when security decisions are repeatedly delegated to centralized policy components and every service interaction requires additional cryptographic negotiation or authorization checks. In the evaluated environment, conventional ZTA produced latency increases of up to 184.2%, demonstrating that security enforcement can become a significant bottleneck in latency-sensitive financial workloads. Such delays are particularly problematic for digital payment systems and Open Banking APIs, where transaction responsiveness is closely associated with service quality, throughput, and compliance with defined response-time requirements. Therefore, the central challenge is not whether financial institutions should adopt Zero-Trust, but how they can implement it without allowing security controls to undermine operational performance.

The proposed Adaptive Edge-Cached Policy Enforcement Framework (AEC-PEF) addresses this challenge through three complementary optimization mechanisms. The first is localized policy caching, which moves frequently required authorization decisions closer to the workloads that generate them. Instead of requiring every service request to communicate with a centralized Policy Decision Point (PDP), edge-level Policy Enforcement Points (PEPs) can temporarily retain validated policy decisions through short-lived cryptographic tickets. In the evaluated configuration, this mechanism reduced redundant PDP calls by nearly 90%. Such decentralization reduces network dependency, minimizes repeated authorization traffic, and allows routine requests to be processed with substantially lower latency. At the same time, short-lived credentials help limit the security risks associated with maintaining outdated authorization decisions.

The second mechanism is TLS 1.3 session resumption, which reduces the overhead associated with repeatedly establishing secure connections between microservices. The empirical results indicate that mTLS negotiation latency was reduced from 14.5 ms to 1.8 ms across East-West microservice routes. This improvement is particularly relevant in financial microservice architectures, where a single transaction may traverse multiple interconnected services. When cryptographic negotiation occurs repeatedly across several service boundaries, even small individual delays can accumulate into a substantial end-to-end latency penalty. Session resumption therefore improves communication efficiency while preserving encrypted service-to-service connectivity.

The third mechanism, risk-based adaptive verification, introduces contextual intelligence into the enforcement process. Instead of applying an identical level of verification to every transaction, AEC-PEF dynamically adjusts security requirements according to transaction characteristics and perceived risk. Routine low-risk operations can be processed through lightweight token validation, whereas high-value transactions, anomalous activities, or requests exhibiting elevated risk indicators can trigger stronger contextual and multi-factor verification. This differentiated approach prevents computationally expensive security procedures from being unnecessarily applied to every request while ensuring that sensitive financial operations continue to receive stronger protection.

Collectively, these mechanisms enable financial institutions to maintain a strong security posture while reducing the performance costs traditionally associated with Zero-Trust enforcement. The evaluated framework achieved a 99.1% RMI while maintaining transaction response times below 60 ms at 10,000 transactions per second (TPS). These results indicate that Zero-Trust security does not necessarily require a binary choice between stronger protection and higher performance. Instead, appropriately designed enforcement mechanisms can distribute security decisions, reduce redundant operations, and allocate verification intensity according to transaction risk. AEC-PEF consequently provides a performance-aware approach to Zero-Trust adoption in environments where both security assurance and response speed are critical.

Implementation Recommendations

1. Deploy Localized PEP Sidecars

Financial institutions should consider deploying localized Policy Enforcement Point (PEP) sidecars rather than routing every authorization request through centralized security gateways. Embedded directly within service meshes, these sidecars can enforce identity, authorization, and communication policies close to individual workloads. Technologies such as eBPF and Envoy can support high-performance traffic inspection and policy enforcement while minimizing unnecessary network hops.

This distributed enforcement model reduces dependence on centralized gateways and prevents a single security component from becoming a performance bottleneck. It also improves horizontal scalability because enforcement capacity increases alongside the number of service instances and clusters.

2. Adopt TLS 1.3 Session Resumption and Carefully Govern 0-RTT

Organizations should implement TLS 1.3 session resumption across appropriate internal microservice communication paths to reduce repeated cryptographic handshake overhead. In high-frequency financial systems, thousands of East-West requests may traverse service boundaries within a short period. Repeated full handshakes can therefore create considerable cumulative latency and computational overhead.

TLS 1.3 session resumption can significantly reduce this cost by allowing previously established secure sessions to be resumed without repeating the complete negotiation process. Where 0-RTT is considered, however, its use should be carefully restricted because early application data can introduce replay-related security considerations. Consequently, 0-RTT should not automatically be enabled for security-sensitive, non-idempotent financial transactions. It is more appropriate for carefully assessed request types where replay risks can be controlled. This distinction is important because performance optimization should not compromise transaction integrity.

3. Institute Risk-Based Policy Evaluation

A uniform verification model should be replaced, where appropriate, with risk-based adaptive policy evaluation. The enforcement engine can incorporate contextual variables such as transaction value, user and device behavior, service identity, request type, historical activity, geographic or network context, and real-time risk scores. These variables can be used to determine the appropriate verification depth for each request.

For example, routine internal read operations with established identities may require only lightweight token validation, whereas high-value transfers, unusual account activity, or requests originating from suspicious contexts can trigger comprehensive contextual verification and multi-factor authentication. This approach allows security resources to be concentrated where they provide the greatest risk-reduction benefit. It also prevents routine low-risk traffic from being subjected to unnecessary computational and network overhead.

4. Establish Explicit Cache Governance and Revocation Controls

Localized policy caching should be accompanied by strict controls governing TTL, cache invalidation, credential revocation, and policy synchronization. Although short-lived authorization tickets improve performance, excessively long validity periods may increase exposure when user privileges change or credentials are compromised. Financial institutions should therefore establish risk-sensitive TTL policies and ensure that critical authorization changes can propagate rapidly to distributed PEPs.

5. Continuously Monitor Security and Performance Metrics

Deployment should also incorporate continuous monitoring of both security and operational indicators. Metrics such as end-to-end latency, PDP call frequency, cache-hit ratio, authentication failures, policy-decision time, transaction throughput, CPU utilization, and security incident rates can

provide an integrated view of system performance. Monitoring these variables enables administrators to identify situations where security enforcement becomes excessive or where optimization mechanisms create unexpected security risks.

Together, these measures establish a performance-aware Zero-Trust enforcement model in which localized sidecars distribute policy enforcement, TLS 1.3 session resumption reduces secure-communication overhead, and risk-based evaluation ensures proportional security rigor. The resulting architecture preserves explicit trust boundaries while reducing unnecessary processing and communication costs. This makes the approach particularly relevant to mission-critical financial environments where milliseconds matter but security requirements cannot be relaxed.

The broader implication is that Zero-Trust implementation should be treated as an adaptive systems-engineering problem rather than a purely security-oriented configuration exercise. Financial institutions must simultaneously consider identity assurance, authorization accuracy, cryptographic protection, network latency, transaction throughput, scalability, and regulatory requirements. AEC-PEF demonstrates how these objectives can be addressed through coordinated architectural optimization rather than through compromising one objective to achieve another.

Nevertheless, the reported results should be interpreted within the limitations of the evaluation environment. The 99.1% RMI and <60 ms latency at 10,000 TPS represent outcomes from the evaluated experimental configuration and should not automatically be generalized to all banking infrastructures. Production performance may vary according to network topology, transaction complexity, policy size, hardware acceleration, service density, cloud-provider characteristics, and workload distribution. Future research should therefore validate the framework using real-world banking workloads, larger service meshes, diverse transaction patterns, and independent security testing.

Overall, the findings demonstrate that Zero-Trust and high-performance digital finance can be complementary objectives when enforcement is designed around locality, session efficiency, and contextual risk. AEC-PEF provides a practical architectural pathway for reducing the performance penalty of continuous verification while preserving strong security boundaries. Its principal value therefore lies in demonstrating how adaptive enforcement can transform Zero-Trust from a potentially rigid security model into a scalable, latency-aware architecture suitable for increasingly distributed digital banking ecosystems.

References

Arner, D. W., Barberis, J., & Buckley, R. P. (2017). FinTech and RegTech in a nutshell, and the future in a sandbox. *CFA Institute Research Foundation Briefs, 3*(4), 1–20. https://doi.org/10.2139/ssrn.3088303

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(3), 382–388. https://doi.org/10.1177/002224378101800313

Gai, K., Qiu, M., & Sun, X. (2018). A survey on FinTech. *Journal of Network and Computer Applications, 103*, 262–273. https://doi.org/10.1016/j.jnca.2017.10.011

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review, 31*(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Kang, H., Liu, G., Wang, Q., Meng, L., & Liu, J. (2023). Theory and application of zero trust security: A brief survey. *Entropy, 25*(12), 1595. https://doi.org/10.3390/e25121595

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

Ryoo, J., Rizvi, S., Aiken, W., & Kissell, J. (2014). Cloud security auditing: Challenges and emerging approaches. *IEEE Security & Privacy, 12*(6), 68–74. https://doi.org/10.1109/MSP.2013.132

Shore, M., Zeadally, S., & Keshariya, A. (2021). Zero trust: The what, how, why, and when. *Computer, 54*(11), 26–35. https://doi.org/10.1109/MC.2021.3090008

Syed, N. F., Shah, S. W., Shaghaghi, A., Anwar, A., Baig, Z. A., & Doss, R. (2022). Zero trust architecture (ZTA): A comprehensive survey. *IEEE Access, 10*, 57143–57179. https://doi.org/10.1109/ACCESS.2022.3174679

Teerakanok, S., Uehara, T., Inomata, A., & Kaneko, M. (2021). Migrating to zero trust architecture: Reviews and challenges. *Security and Communication Networks, 2021*, Article 9947347. https://doi.org/10.1155/2021/9947347

Alam, M. R. U., & Alvi, M. F. I. (2025). Evaluating the Economic and Environmental Impact of Enterprise Risk Management (ERM) in Green Supply Chains. Eco-Business and Environmental Progress Journal, 5(1).