Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 42–50

Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/qmvvay81

A Lightweight Zero-Trust Framework for Small-to-Medium Enterprises on AWSAzure Hybrid Clouds

Bibek Kumar Katwal[1,*], Rajesh Chhetry [1],

[1]Islington College, Kathmandu, Nepal
[*]*Correspondence: bibek21k@gmail.com*

ARTICLE HISTORY

Received: 27 March 2026 Revised: 30 April 2026
Accepted: 17 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Katwal, B. K. and Chhetry, R. (2026). 'A lightweight zero-trust framework for small-to-medium enterprises on AWSAzure hybrid clouds' *Islington Journal of Multidisciplinary Research*, 1(1), pp. 42–50. Available at: https://doi.org/10.67556/qmvvay81

ABSTRACT

Small-to-medium enterprises (SMEs) in developing economies are digitizing rapidly while their security capacity lags, leaving credential-based attacks and cross-cloud misconfiguration largely uncontested. Zero Trust Architecture (ZTA) is the recognized strategic response, yet documented implementations overwhelmingly assume enterprise budgets and specialist teams that resource-constrained firms cannot sustain. This study aimed to design and empirically evaluate a lightweight, open-source ZTA framework for SMEs operating on a hybrid AWSAzure infrastructure, demonstrating both technical viability and economic feasibility without recourse to specialist security personnel. A pragmatist, design-science approach combined four-sprint Agile development, load testing at 5, 25, and 50 concurrent users using Locust 2.x, STRIDE threat modelling, and a structured SME feasibility self-assessment anchored in ENISA (2021) guidance. The four-component framework comprised an OAuth 2.0 identity service, an RBAC policy engine, an access gateway, and an administrative dashboard with append-only audit logging, built entirely from open-source software and free-tier cloud services. The prototype achieved 100% access-control accuracy across 500 test requests and a mean end-to-end authentication latency of 187 ms at peak load, within the 200 ms imperceptibility threshold. Five of six STRIDE threat categories were fully mitigated; Denial of Service carried medium-residual risk addressable via free-tier DDoS protection. Total monthly deployment cost was approximately USD 43, one to two orders of magnitude below comparable commercial ZTA platforms. A credible, NIST SP 800-207-aligned Zero Trust posture is technically and economically feasible for SMEs without specialist security personnel, contributing a deployable open-source artefact and a validated costperformance model replicable across emerging economies.

Keywords: Zero trust architecture, hybrid cloud security, open-source cybersecurity, identity-centric access control, SME security economics

JEL Classification: O33, Technological Change: Choices and Consequences; L86, Information and Internet Services; M15, IT Management

Introduction

The convergence of affordable cloud computing and mobile connectivity has enabled small-to-medium enterprises (SMEs) in developing economies to digitize business operations at unprecedented speed. Nepal is no exception. The Government of Nepals Digital Nepal Framework (Government of Nepal, Ministry of Communication and Information Technology 2019) sets out an explicit national agenda for digital transformation across sectors, and SMEs, which the Asian Development Bank (2020) identifies as a cornerstone of the Nepali economy, are increasingly adopting cloud-hosted productivity tools, e-commerce platforms, and remote-access architectures. Rapid digitization without commensurate investment in cybersecurity, however, creates systemic exposure.

CERT-NP (2021) documents a rising volume of reported incidents, and the Kathmandu Post (2024) reports a marked spike in cybercrime cases nationally, with credential theft and unau-

thorized access featuring prominently. Thapa (2025) further finds that Nepali SMEs face acute cybersecurity challenges arising from limited budgets and scarce specialist skills.

Traditional perimeter-based security models, which assume that everything inside a corporate network can be trusted, are poorly suited to hybrid-cloud environments where resources span multiple providers and users connect from diverse, uncontrolled endpoints. The United States National Institute of Standards and Technology formalized Zero Trust Architecture (ZTA) as the strategic response to this structural problem (Rose et al. 2020). ZTA replaces implicit network trust with continuous, identity-centric access decisions enforced at the resource level. Zero Trust has moved firmly into the mainstream of enterprise security practice: Gartner (2023) reports that a majority of organizations worldwide have now implemented or begun implementing a zero-trust strategy. Commercial ZTA products from established vendors offer mature implementations but typically require recurring subscription fees together with specialist staff for configuration and ongoing management, conditions that most Nepali SMEs simply cannot meet.

A growing body of research confirms this gap between ZTAs theoretical benefits and its practical accessibility for resource-constrained organizations. Buck et al. (2021), in a multivocal literature review, identify complexity and cost as the primary recurring barriers to zero-trust adoption and highlight a troubling shortage of empirical, implementation-level evidence. Mehraj and Banday (2020) propose lightweight conceptual frameworks but stop short of empirical evaluation. Syed et al. (2022), in a comprehensive survey, document a range of cloud-native ZTA components but observe that most target enterprise environments. For a typical Nepali SME, operating with a single IT generalist, no dedicated security function, and tight capital constraints, the practical choice is frequently between an unaffordable enterprise ZTA product and no structured access control at all. The latter leaves the organization reliant on perimeter assumptions that hybrid-cloud working has already invalidated, exposing it to precisely the credential-based attacks that dominate the regional threat statistics.

No peer-reviewed study, to the authors knowledge, has built and empirically validated a complete, deployable, open-source ZTA framework explicitly designed for SMEs operating on hybrid AWSAzure infrastructure in a developing-economy context. To close this gap, the study develops and evaluates a deployable zero-trust framework tailored to SME constraints. Its three research objectives are: (RO1) to design a lightweight ZTA framework that meets NIST SP 800-207 core tenets while remaining deployable by a single engineer without specialist security training; (RO2) to measure the performance and cost characteristics of such a framework under realistic SME workloads; and (RO3) to evaluate the extent to which the framework mitigates threats identified through STRIDE analysis. The corresponding research questions are: (RQ1) Can a lightweight ZTA framework meet NIST SP 800-207 core tenets while remaining deployable by a single engineer without specialist security training? (RQ2) What are the measurable performance and cost characteristics of such a framework under realistic SME workloads? (RQ3) To what extent does the framework mitigate threats identified through STRIDE analysis?

Literature Review

Zero Trust Architecture: Conceptual Origins and Standards

The Zero Trust model was first articulated by Kindervag (2010) at Forrester Research, who coined the phrase and proposed a conceptual framework centered on micro-segmentation and least-privilege access. The model gained institutional legitimacy when NIST published SP 800-207 (Rose et al. 2020), which defines ZTA through seven core tenets: treating all resources as external; enforcing least-privilege access; inspecting and logging all traffic; performing dynamic authentication; collecting telemetry; and maintaining an inventory of authorized devices and identities. SP 800-207 further defines the three logical components of a ZTA, namely the Policy Decision Point (PDP), the Policy Enforcement Point (PEP), and the Policy Administrator (PA), which together form the architectural template for the present study.

Complementary guidance from other authoritative bodies reinforces the NIST framework. The UK National Cyber Security Centre (NCSC 2021) published zero-trust architecture design principles for practitioners, emphasizing an identity-centric, and least-privilege approach for multi-provider environments, directly applicable to hybrid AWSAzure deployments. The European Union Agency for Cybersecurity (ENISA 2021) produced guidance specifically addressing cybersecurity challenges and recommendations for SMEs, acknowledging that comprehensive security maturity is an aspirational target for most small organizations and advocating pragmatic, prioritized implementation paths. Taken together, these standards constitute the internationally recognized benchmark against which the present framework is designed and evaluated.

Existing Lightweight ZTA Proposals and Adoption Barriers

Research on ZTA adaptation for resource-constrained environments reveals a persistent pattern: conceptual proposals are common, but empirically validated and deployable artefacts are rare. Mehraj and Banday (2020) present a conceptual zero-trust strategy for cloud computing but offer no prototype or empirical evaluation, limiting practical utility. Syed et al. (2022) conduct a comprehensive survey of ZTA architectural components and deployment models, yet find that virtually all documented implementations assume enterprise-scale resources and expertise, leaving a clear and unaddressed gap for SME-oriented solutions. Buck et al. (2021) provide a multivocal literature review that identifies cost, complexity, and a shortage of implementation-level case studies as the dominant barriers to broader zero-trust adoption, particularly among smaller organizations.

More recent practitioner-oriented work has begun to address specific ZTA components in constrained contexts. Gilman

and Barth (2017) provide foundational guidance on building zero-trust networks and note that mature ZTA policy engines ultimately incorporate dynamic, context-aware decisions, including device posture signals, user behavior analytics, and threat intelligence feeds, alongside static role assignments. Their finding that mean authentication overhead below 200 ms is imperceptible to end users in interactive applications directly informs the performance target adopted in this study. The convergence of these findings suggests that a static, RBAC-configured Phase 1 ZTA is both a rational first target for SMEs and a meaningful contribution to a literature dominated by either enterprise-scale implementations or purely conceptual proposals.

Cybersecurity in Developing-Economy SME Contexts

The cybersecurity challenges of SMEs in developing economies constitute a distinct and underserved research domain. Kshetri (2020) examines these economies distinctive constraints, including limited budgets, skills shortages, and weak institutional capacity, and identifies them as primary impediments to effective cybersecurity posture. Within Nepal specifically, Thapa (2025) documents the cybersecurity challenges confronting SMEs, and the Asian Development Bank (2020) characterizes the resource constraints typical of the sector. CERT-NP (2021) data on rising incident volumes and the Kathmandu Post (2024) reporting of a spike in cybercrime cases provide the quantitative regional threat context. Together, these sources converge on a clear and actionable requirement: any security solution intended for Nepali SMEs must be low-cost, low-complexity, and operable without specialist staff, criteria that commercial ZTA platforms consistently fail to satisfy.

Identity and Access Management Building Blocks

The technical building blocks underpinning the proposed framework are individually well established and proven at scale. OAuth 2.0 (Hardt 2012) is the dominant standard for delegated authorization in modern API architectures, providing a widely supported mechanism for identity assertion. RFC 7519 (Jones, Bradley and Sakimura 2015) specifies JSON Web Tokens (JWTs) as a compact, self-contained credential format suited to stateless, distributed systems, a key property for a ZTA access gateway that must make high-frequency policy decisions with minimal latency. Role-Based Access Control (RBAC), introduced by Sandhu et al. (1996) and consolidated by Ferraiolo, Kuhn and Chandramouli (2007), remains the most widely deployed access control model in enterprise environments and is well-suited as a policy engine substrate within a Zero Trust design. The deliberate use of static RBAC as a Phase 1 implementation is a scope decision appropriate to the SME target environment; dynamic, context-aware policy extensions are identified as future work.

Performance Benchmarking in Security Architectures

Latency introduced by authentication and authorization layers is a persistent engineering concern in ZTA implementations. Gilman and Barth (2017) note that authentication overhead remains low relative to overall request time and is generally imperceptible to end users in interactive applications when mean latency stays below 200 ms, the threshold adopted in this study. Syed et al. (2022) acknowledge throughput overhead as an inherent characteristic of ZTA enforcement layers, acceptable when user-perceptible latency remains within target. For load generation and measurement, Locust (2023) provides an open-source, Python-based HTTP load-testing tool widely used in prior security architecture benchmarking studies and adopted here for all performance measurements.

Synthesis and Research Gap

Existing scholarship highlights three recurring patterns. First, the conceptual and standards foundations of ZTA are mature, with NIST SP 800-207 providing an authoritative architectural template and ENISA (2021) offering pragmatic, SME-oriented implementation guidance. Second, the open-source building blocks required to construct a ZTA control plane, including OAuth 2.0, JWT, RBAC, and load-tested API gateways, are individually proven and freely available. Third, and most critically, no existing study integrates these elements into a single, complete, empirically evaluated framework that is simultaneously: (a) deployable on hybrid AWSAzure infrastructure; (b) operable by a non-specialist within free-tier and entry-level cloud budgets; and (c) validated in a developing-economy SME context. Prior work is either conceptual (Mehraj and Banday 2020), enterprise-scoped in its surveyed implementations (Syed et al. 2022), or barrier-focused rather than solution-oriented (Buck et al. 2021). This study closes that integration gap by designing, building, and measuring a working artefact against explicit accuracy, performance, security, and cost criteria.

Materials and Methods

Research Philosophy and Design

The research adopts a pragmatist epistemology (J. W. Creswell and J. D. Creswell 2018; Saunders, Lewis and Thornhill 2019), holding that the value of a security framework is determined by its demonstrated performance in the target context rather than by theoretical alignment with an idealized model. The research follows an applied design-science strategy (Bryman 2016), in which the primary contribution is a functioning system evaluated against pre-specified utility criteria. This orientation is appropriate because the deliverable is not a survey instrument or statistical model but a functioning system whose scientific contribution lies in demonstrating what is measurably achievable and reproducible.

The study evaluates four operationally defined variables. Access control accuracy measures the proportion of test requests correctly classified as ALLOW or DENY, including both legitimate and adversarial traffic; any value below 100% indicates either false-positive grants (a security failure) or false-negative denials (an availability failure). Authentication latency records the end-to-end elapsed time in milliseconds from client request to gateway response, measured at mean

and 95th-percentile levels, with a target threshold of 200 ms drawn from Gilman and Barth (2017). Threat mitigation coverage assesses the proportion of STRIDE threat categories rated Fully Mitigated through structured analysis of the implementation against each threat vector. Monthly deployment cost captures the sum of AWS and Azure line-item charges at January 2025 public pricing, denominated in USD. Ethical considerations are fully addressed by the exclusive use of synthetic test data and fabricated credentials throughout all evaluation phases; no real user data, organizational systems, or production environments were accessed at any point. The complete artefact is reproducible by any practitioner following the published 47-step deployment guide.

System Architecture and NIST SP 800-207 Mapping

The proposed architecture consists of four integrated components deployed across a hybrid AWSAzure environment, mapping directly onto the NIST SP 800-207 logical model (Rose et al. 2020):

- **Identity Service (AWS EC2 t2.micro), Policy Decision Point [Identity]:** A FastAPI application implementing the OAuth 2.0 authorization code flow with PKCE (RFC 7636). Issues RS256-signed JWTs with a 15-minute access-token lifetime. Password storage uses Bcrypt with a work factor of 12. Supports multifactor authentication via TOTP (RFC 6238). This component is the sole authority for identity assertions in the framework.
- **RBAC Policy Engine (AWS Lambda), Policy Decision Point [Policy]:** A Python function evaluating access requests against a role-to-resource permission matrix stored in Amazon DynamoDB. Evaluates five contextual dimensions on every request: resource, action, role, time window, and IP range. Returns an ALLOW or DENY decision in under 10 ms at the 95th percentile under all tested loads.
- **ZTA Access Gateway (Azure App Service B1), Policy Enforcement Point:** A FastAPI reverse proxy intercepting all API calls, validating JWT signatures against the public JWKS endpoint, invoking the Policy Engine, and forwarding or rejecting requests accordingly. Implements per-token rate limiting (60 requests per minute) and appends a structured, immutable audit record to a PostgreSQL database on each decision.
- **Administrative Dashboard (Azure Static Web Apps), Policy Administrator:** A Next.js application providing real-time visibility into access decisions, role assignments, and security alerts via a read-only PostgreSQL connection. This component enables an IT generalist to monitor the ZTA posture without dedicated security operations tooling.

A typical request traverses the framework as follows: the client authenticates against the Identity Service and receives a short-lived JWT; each subsequent API call carries this token to the Access Gateway, which validates the RS256 signature against the public JWKS endpoint, extracts subject and role claims, and invokes the Policy Engine; the Policy Engine evaluates the request against the five-dimension permission matrix and returns ALLOW or DENY; the Gateway forwards or rejects the call and writes an immutable record to the audit log. Because the token is self-contained and policy evaluation is entirely stateless, every request is authorized independently, satisfying the core Zero Trust principle that no implicit trust is granted by network location or prior session state. The deliberate distribution of components across two cloud providers also demonstrates that the framework is genuinely hybrid, mitigating provider lock-in.

Agile Development Lifecycle

Development followed a four-sprint Agile cycle. This incremental approach was chosen deliberately: each security control could be tested in isolation before integration, surfacing defects early and keeping each sprints scope small enough to be understood and maintained by a single engineer, the same operating constraint the target organizations face. Table A documents the sprint sequence, deliverables, and the functional acceptance validation gate applied before proceeding to each subsequent sprint.

Table A: Agile Sprint Sequence, Deliverables, and Inter-Sprint Validation Gates

Sprint	Focus Area	Key Deliverable	Validation Gate
1	Identity Service and JWT issuance	FastAPI OAuth 2.0 + PKCE endpoint; RS256 token issuance; TOTP MFA	Token issued with correct claims; expiry enforced at 15 minutes
2	RBAC Policy Engine and DynamoDB schema	Lambda function; five-dimension permission matrix in DynamoDB	Correct ALLOW/DENY for all representative role–resource pairs
3	Access Gateway integration and audit logging	FastAPI reverse proxy; append-only PostgreSQL audit log; rate limiter	Audit record completeness; rate-limit enforcement confirmed
4	Dashboard, deployment guide, and full system testing	Next.js dashboard; 47-step deployment guide validated end-to-end	Dashboard data consistency; end-to-end load test passed

Evaluation Protocol

Performance testing

Load was generated using Locust 2.x across three scenarios: 5 concurrent users (baseline SME load), 25 users (moderate growth), and 50 users (peak stress). Each scenario ran for 10 minutes with a

30-second ramp-up. Metrics collected: mean and 95th-percentile end-to-end authentication latency (ms), requests per second (RPS), and error rate. A parallel baseline measurement (Identity Service only, without gateway or policy engine) established the overhead attributable to ZTA enforcement components.

Security evaluation

STRIDE threat modelling (Shostack 2014) was applied to the complete request flow. Each of the six categories, namely Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege, was rated Fully Mitigated, Partially Mitigated, or Not Addressed, with supporting evidence drawn directly from the implementation.

SME feasibility assessment

A structured readiness self-assessment was derived from the ENISA (2021) SME cybersecurity recommendations, adapted for the Nepali context through mapping to the national digital-transformation priorities set out in the Digital Nepal Framework (Government of Nepal, Ministry of Communication and Information Technology 2019). Criteria covered operational simplicity, documentation completeness, and the feasibility of day-to-day management without dedicated security personnel.

Cost calculation

Monthly costs were calculated using AWS and Azure public pricing as of January 2025: EC2 t2.micro (USD 8.47), AWS Lambda invocations at SME volume (USD 0.00, free tier), DynamoDB on-demand reads/writes (USD 1.25), Azure App Service B1 (USD 13.14), Azure Static Web Apps (USD 0.00, free tier), Azure PostgreSQL Flexible Server Burstable B1ms (USD 14.17), and estimated data transfer (USD 6.00), yielding a total of USD 43.03 per month.

Results

Access Control Accuracy (RQ1, RQ3)

The RBAC policy engine correctly evaluated all 500 test requests (250 ALLOW, 250 DENY) with zero false positives and zero false negatives, yielding 100% access control accuracy. ALLOW decisions covered valid role-resource-action combinations; DENY decisions covered expired tokens (n = 50), insufficient role (n = 100), out-of-window access (n = 50), and IP-range violations (n = 50). No policy bypass was observed during adversarial testing, which encompassed token replay, JWT algorithm confusion, and privilege escalation attempts.

The adversarial results warrant specific elaboration. Token replay attempts were defeated by the combination of short token lifetimes and server-side expiry validation: a captured token became operationally worthless within fifteen minutes. JWT algorithm-confusion attempts, in which an attacker substitutes a weaker or null algorithm in the token header, were rejected because the gateway pins verification to RS256 and the published public key, refusing any non-conforming token. Privilege-escalation attempts, in which an authenticated low-privilege user requested resources mapped exclusively to higher roles, were uniformly denied by the policy engines explicit least-privilege matrix. These results compare favorably with Gilman and Barth (2017), who report that well-configured RBAC-based policy engines routinely achieve perfect accuracy on static permission matrices; the present study confirms this finding in the specific context of a hybrid cloud deployment at SME scale and under adversarial conditions, directly addressing RQ3.

Performance Under Load (RQ2)

Table 1 summarizes end-to-end authentication latency and throughput across the three load scenarios. The 200 ms mean-latency target (Gilman and Barth 2017) was met at all load levels. At peak stress (50 concurrent users), mean latency reached 187 ms, just 6.5% below the threshold, confirming that the framework remains responsive under realistic peak SME load. The 95th-percentile latency at peak load, however, reached 341 ms, indicating that a minority of requests exceed the mean target under stress. Reducing this tail latency through policy-result caching and connection pooling is identified as the primary production-readiness priority (see Section 5.2).

Table 1: Load Test Results, Showing End-to-End Authentication Latency and Throughput Across All Load Scenarios

Users	Mean Latency (ms)	95th pct (ms)	RPS	Overhead vs Baseline	Target Met?
5	47	89	4.8	-18%	Yes
25	112	198	22.1	-41%	Yes
50	187	341	43.6	-64%	Yes (mean)

The 64% throughput reduction at 50 concurrent users reflects the computational cost of three sequential operations, namely JWT validation, policy engine invocation, and gateway forwarding, relative to a direct API call without ZTA enforcement. This overhead is an acknowledged and expected characteristic of ZTA enforcement layers (Syed et al. 2022) and is acceptable given that the user-perceptible latency target was met. Compared with deployments cited by Gilman and Barth (2017), where authentication overhead can exceed 200 ms even at low concurrency, the present frameworks stateless, Lambda-based policy evaluation provides a meaningful efficiency advantage, particularly at the low-to-moderate concurrency levels typical of SME workloads.

Security Evaluation: STRIDE Threat Analysis (RQ3)

Table 2 presents the STRIDE analysis results. Five of six threat categories were fully mitigated; Denial of Service alone carries medium-residual risk.

Table 2: STRIDE Threat Analysis, Showing Mitigation Status and Control Mechanisms for Each Threat Category

STRIDE Category	Threat Vector	Mitigation Status	Control Mechanism
Spoofing	Credential stuffing / identity impersonation	**Fully Mitigated**	OAuth 2.0 + optional TOTP-based MFA
Tampering	API payload or token injection	**Fully Mitigated**	RS256 JWT signature verification at gateway
Repudiation	Log deletion or alteration	**Fully Mitigated**	Append-only PostgreSQL audit log; immutable timestamps
Information Disclosure	Token interception in transit	**Fully Mitigated**	TLS 1.3 enforcement + 15-minute access-token lifetime
Denial of Service	Volumetric or distributed flooding	**Partially Mitigated**	Per-token rate limiting (60 req/min); CloudFlare free-tier DDoS protection recommended
Elevation of Privilege	RBAC bypass or role escalation	**Fully Mitigated**	Policy engine with explicit five-dimension least-privilege matrix

Each mitigation merits brief elaboration. Spoofing is defeated by combining OAuth 2.0 credential validation with optional TOTP-based MFA: an attacker requires both valid credentials and the second factor to impersonate a legitimate user. Tampering is addressed by RS256 JWT signature verification; any modification to a tokens payload invalidates the signature and triggers immediate rejection at the gateway. Repudiation is countered by the append-only PostgreSQL audit log, which records every access decision with an immutable timestamp, making it infeasible for a user to credibly deny an action they performed. Information Disclosure is limited by enforcing TLS 1.3 in transit and by issuing only short-lived (15-minute) access tokens, narrowing the exploitation window for any intercepted credential. Elevation of Privilege is prevented by the policy engines explicit least-privilege matrix, which evaluates five dimensions, namely resource, action, role, time window, and IP range, on every request without exception. Denial of Service is the sole category carrying medium-residual risk: the per-token rate limiter blunts low-volume abuse but cannot absorb a distributed volumetric flood. CloudFlares free-tier DDoS protection is the recommended mitigation, providing meaningful volumetric defense without incurring additional cost.

Cost Feasibility (RQ2)

Table 3: Monthly Deployment Cost Breakdown by Component (AWS and Azure Public Pricing, January 2025)

Framework Component	Cloud Service	Monthly Cost (USD)
Identity Service	AWS EC2 t2.micro	8.47
RBAC Policy Engine	AWS Lambda (SME volume)	0.00 (free tier)
Policy Data Store	AWS DynamoDB on-demand	1.25
Access Gateway	Azure App Service B1	13.14
Admin Dashboard	Azure Static Web Apps	0.00 (free tier)
Audit Database	Azure PostgreSQL Flexible Server Burstable B1ms	14.17
Data Transfer	Estimated cross-cloud egress	6.00
Total Monthly Cost	*AWS + Azure (January 2025 pricing)*	**USD 43.03**

At USD 43.03 per month, the frameworks deployment cost is one to two orders of magnitude below the recurring cost of comparable commercial ZTA platforms: enterprise ZTA subscriptions from vendors such as Zscaler Private Access, Cloudflare for Teams, or Palo Alto Prisma Access typically range from USD 500 to over USD 5,000 per month at SME-relevant seat counts, before factoring in professional services for configuration and onboarding. The cost structure of the present framework is dominated by compute (Azure App Service B1: USD 13.14) and managed database (Azure PostgreSQL: USD 14.17), both of which can be reduced further by migrating to spot instances or shared-tier services as organizational maturity increases.

An important structural advantage of the framework concerns cost scaling. Commercial ZTA platforms commonly apply per-user or minimum-seat pricing that grows proportionally with headcount, whereas the frameworks cost is driven by compute and database tiers that remain effectively flat across the entire SME usage range tested. Consequently, the relative cost advantage tends to widen rather than narrow as an organization adds users up to the capacity ceiling established in the load tests, reinforcing the economic case for the target segment.

SME Feasibility Assessment

Assessed against the readiness criteria drawn from ENISA (2021), the framework addresses the majority of applicable controls fully, a small number partially (specifically those requiring endpoint detection and response capabilities beyond the frameworks intended scope), and a few that are not applicable to a cloud-hosted SME (such as physical security controls, which are addressed through cloud provider service-level agreements). The complete 47-step deployment guide was validated end-to-end in 4 hours and 35 minutes by a single engineer with intermediate cloud experience, directly confirming RQ1: a non-specialist can establish a credible, NIST SP 800-207-aligned Zero Trust posture without bespoke configuration consulting. The administrative dashboard provides sufficient real-time visibility for an IT generalist to monitor and respond to access anomalies without dedicated security operations personnel.

Discussion

Findings in Relation to Research Objectives

RO1 / RQ1, Deployability without specialist staff. The framework was deployed in under five hours by a single engineer with intermediate cloud experience, confirming that a non-specialist can establish a credible Zero Trust posture by following the 47-step documented guide. This finding addresses the complexity barrier identified by Buck et al. (2021) as the primary impediment to SME-level zero-trust adoption. By pre-selecting open-source components, targeting free-tier and entry-level cloud services, and providing complete step-by-step documentation, the framework removes the consultation dependency that makes commercial ZTA onboarding prohibitive for the target segment. The deliberate choice of static RBAC over dynamic policy is the key design decision: it eliminates the need for threat intelligence feeds, behavioral analytics platforms, or dedicated security operations tooling while preserving the core Zero Trust property of per-request, identity-centric authorization.

RO2 / RQ2, Performance and cost under realistic SME loads. Mean latency of 187 ms at 50 concurrent users confirms the framework is responsive under peak SME load, consistent with Gilman and Barths (2017) sub-200 ms imperceptibility threshold. The 95th-percentile latency of 341 ms at peak load is the primary performance limitation and indicates that tail-latency optimization, specifically policy-result caching at the Lambda layer and connection pooling at the gateway, is a necessary step before production deployment in high-frequency API environments. The 64% throughput reduction at peak load, while substantial, is consistent with Syed et al.s (2022) characterization of ZTA enforcement overhead and is acceptable given that user-perceptible latency met its target. At USD 43.03 per month, the cost represents savings of 90 to 99 per cent relative to commercial alternatives at equivalent SME seat counts, with a flat scaling profile that becomes increasingly advantageous as headcount grows.

RO3 / RQ3, Threat coverage under STRIDE analysis. Five of six STRIDE categories are fully mitigated; Denial of Service carries medium-residual risk but is addressable via CloudFlares free-tier DDoS protection at no additional cost. This profile compares favorably with Mehraj and Bandays (2020) conceptual framework, which addresses spoofing and information disclosure but does not discuss repudiation controls or systematic privilege-escalation prevention. The five fully mitigated categories, namely Spoofing, Tampering, Repudiation, Information Disclosure, and Elevation of Privilege, correspond precisely to the threat vectors most frequently cited in Nepals regional cybercrime statistics (CERT-NP 2021; Kathmandu Post 2024), reinforcing the frameworks operational relevance to its target context.

Scalability and Broader Applicability

This frameworks scalability is bounded at the upper end by the Azure App Service B1 tier, which the load tests demonstrate is sufficient for up to 50 concurrent users at sub-200 ms mean latency. Organizations with high-frequency API workloads, such as fintech or e-commerce platforms with sustained traffic above 40 RPS, would need to scale the gateway horizontally (via multiple App Service instances behind Azure Front Door) and implement policy-result caching at the Lambda layer. These extensions would increase monthly cost modestly but keep total expenditure well within the SME budget envelope: Azure Front Door Standard adds approximately USD 35 per month, keeping total monthly cost below USD 100 while extending capacity to several hundred concurrent users.

Geographic applicability beyond Nepal is strong. The framework is constructed entirely from globally available AWS and Azure free-tier and entry-level services, and the qualitative cost advantage over commercial ZTA is likely to hold across South Asia, Sub-Saharan Africa, and South-East Asia, regions that share Nepals profile of high digital-adoption growth, limited specialist security talent, and constrained capital. Direct application to other jurisdictions requires re-mapping the ENISA (2021)-derived feasibility checklist against local data-protection and cybersecurity regulations, a task supported by the structured, modular design of the checklist itself.

Limitations

These findings should be interpreted in light of several important limitations. First, all performance testing was conducted on a development workstation rather than live cloud infrastructure; production latency figures may differ, though cloud-native deployment on co-located AWS and Azure resources would be expected to reduce cross-cloud latency and improve throughput relative to the values reported. Second, the static RBAC policy engine is a deliberate and consequential simplification: full NIST SP 800-207 compliance ultimately requires dynamic, context-aware policy decisions incorporating real-time device posture signals, user behavior analytics, and threat intelligence feeds, capabilities that substantially increase both system complexity and monthly cost, placing them outside the SME target profile of the present study. Third, the study is geographically scoped to Nepali SMEs, and cost calculations reflect January 2025 pricing; organizations in other regions may encounter different relative costs depending on regional cloud availability and data-residency requirements.

Conclusion

This research was undertaken to address a concrete and consequential gap: the absence of a complete, empirically validated, open-source Zero Trust Architecture framework designed for the resource constraints of SMEs operating on hybrid AWSAzure infrastructure in developing economies. That gap has been closed. The framework was designed through a four-sprint Agile lifecycle, implemented using proven open-source components at free-tier and entry-level cloud cost, and evaluated against pre-specified accuracy, performance, threat-coverage, and cost criteria.

This study advances the field in three ways. First, it demonstrates that a NIST SP 800-207-aligned Zero Trust posture is technically achievable for SMEs without specialist security personnel, validated by 100% access-control accuracy under both legitimate and adversarial conditions and by sub-200 ms mean authentication latency under peak load. Second, it establishes economic feasibility at USD 43.03 per month with a flat-scaling cost structure that makes the framework increasingly competitive as organizations grow within the tested capacity range. Third, it delivers a reproducible, open-source artefact with a 47-step deployment guide, enabling Nepali SMEs and comparable organizations across South Asia to adopt a credible Zero Trust posture without external consultancy or capital expenditure.

Future research could build on this work in three principal directions: (1) lightweight dynamic policy signals, specifically device posture inference from user-agent analysis and login-time behavioral patterns, that can be incorporated without specialist tooling; (2) a longitudinal production study measuring framework performance, maintenance burden, and security incident outcomes in a live Nepali SME environment over twelve or more months; and (3) cross-jurisdictional adaptation of the feasibility checklist, beginning with South Asian regulatory contexts that share Nepals data-protection

and digital-transformation profile.

Practical Implications and Future Research

Practical implications for SME practitioners. The framework provides an immediately deployable path to Zero Trust that demands no security budget beyond USD 43 per month and no specialist staff beyond a single IT generalist with intermediate cloud experience. Organizations operating in Nepal and comparable developing-economy contexts can reduce their exposure to the credential-based attacks that dominate regional cybercrime statistics without first acquiring enterprise-level resources. The administrative dashboard and append-only audit log provide the observability needed for ongoing governance without a dedicated security operations function.

Policy implications for regulators and decision-makers. The Digital Nepal Framework (Government of Nepal, Ministry of Communication and Information Technology 2019) and equivalent national digital-transformation agendas in other developing economies could meaningfully incorporate open-source ZTA frameworks as a recommended cybersecurity baseline for SMEs. The cost-performance model validated in this study provides a quantitative foundation for such policy recommendations. At a regional level, the frameworks replicability across South Asian SMEs suggests potential for inclusion in sector-level cybersecurity readiness programmes.

Research implications and future scholarly directions. The design-science evaluation protocol, combining functional acceptance testing, tiered load testing, STRIDE threat modelling, and a structured ENISA-derived feasibility checklist, constitutes a reusable methodological template for future SME ZTA evaluations, regardless of jurisdiction. Four specific future research directions are identified: (1) lightweight device posture integration using user-agent parsing and login-time pattern detection without specialist tooling; (2) federated identity extensions to support multi-tenant SME deployments; (3) machine-learning-based anomaly detection operating on the existing audit log without additional infrastructure; and (4) cross-jurisdictional feasibility studies adapting the checklist to local regulatory and institutional contexts across South Asia, Sub-Saharan Africa, and South-East Asia.

Disclosure Statement

Conflict of Interest: The authors declare no conflicts of interest.

AI Tools Disclosure: Large language models were used solely for language improvement (grammar and clarity checking) and literature-search support. No research content, design, analysis, or findings were generated by AI tools; all such intellectual work is the authors own.

Authors Contributions: Bibek Kumar Katwal conceived the study, designed and implemented the framework, conducted all evaluations, and wrote the manuscript. Rajesh Chhetry supervised the research, advised on methodology and threat-model design, and reviewed the manuscript. Both authors read and approved the final version submitted for publication.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. **Competing Interests:** None declared.

Data Availability Statement: The source code, deployment guide, and load-test datasets supporting the findings of this study will be made openly available in a public repository upon acceptance. The repository link is omitted in this version to preserve anonymity during peer review.

Acknowledgements

The author thanks Rajesh Chhetry for supervisory guidance throughout this research, and Islington College, Kathmandu (affiliated to London Metropolitan University) for institutional support.

References

Asian Development Bank (2020). *Small and Medium-Sized Enterprises in Nepal*. ADBI Working Paper 1166. Accessed: 15 December 2025. Tokyo: Asian Development Bank Institute. https://www.adb.org/sites/default/files/publication/623281/adbi-wp1166.pdf (visited on 12/15/2025).

Bryman, Alan (2016). *Social Research Methods*. 5th. Oxford: Oxford University Press.

Buck, Christoph, Christian Olenberger, André Schweizer, Fabiane Völter and Torsten Eymann (2021). "Never Trust, Always Verify: A Multivocal Literature Review on Current Knowledge and Research Gaps of Zero-Trust." In: *Computers & Security* 110, p. 102436. 10.1016/j.cose.2021.102436.

CERT-NP (2021). *Annual Report 2020/2021*. Tech. rep. Accessed: 27 November 2025. Kathmandu: CERT-NP. https://www.cert.gov.np/uploads/files/Annual%20Report%202077-78.pdf (visited on 11/27/2025).

Creswell, John W. and J. David Creswell (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th. Thousand Oaks, CA: SAGE Publications.

ENISA (2021). *Cybersecurity for SMEs: Challenges and Recommendations*. Tech. rep. Accessed: 3 January 2026. Athens: European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications/cybersecurity-for-smes (visited on 01/03/2026).

Ferraiolo, David, D. Richard Kuhn and Ramaswamy Chandramouli (2007). *Role-Based Access Control*. 2nd. Norwood, MA: Artech House.

Gartner (Apr. 2023). *Gartner Survey Reveals 63% of Organizations Worldwide Have Implemented a Zero-Trust Strategy*. Press release. Stamford, CT.

Gilman, Evan and Doug Barth (2017). *Zero Trust Networks: Building Secure Systems in Untrusted Networks*. Sebastopol, CA: O'Reilly Media.

Government of Nepal, Ministry of Communication and Information Technology (2019). *Digital Nepal Framework: Unlocking Nepal's Growth Potential*. Tech. rep. Accessed: 14 December 2025. Kathmandu: Government of Nepal. https://mocit.gov.np (visited on 12/14/2025).

Hardt, Dick (2012). *The OAuth 2.0 Authorization Framework*. Tech. rep. RFC 6749. Accessed: 28 April 2026. Fremont, CA: IETF. https://www.rfc-editor.org/rfc/rfc6749 (visited on 04/28/2026).

Jones, Michael, John Bradley and Nat Sakimura (2015). *JSON Web Token (JWT)*. Tech. rep. RFC 7519. Accessed: 28 April 2026. Fremont, CA: IETF. https://www.rfc-editor.org/rfc/rfc7519 (visited on 04/28/2026).

Kathmandu Post (Aug. 2024). *Cybercrime Cases Spike in Nepal*. *Kathmandu Post*. Accessed: 7 May 2026. https://kathmandupost.com/national/2024/08/21/cybercrime-cases-spike-in-nepal (visited on 05/07/2026).

Kindervag, John (2010). *No More Chewy Centers: Introducing the Zero Trust Model of Information Security*. Tech. rep. Cambridge, MA: Forrester Research.

Kshetri, Nir (2020). "Cybersecurity in Emerging Economies: In Search of a Solution." In: *Computer* 53.3, pp. 74–78.

Locust (2023). *Locust: An Open-Source Load Testing Tool*. Accessed: 28 April 2026. https://locust.io/ (visited on 04/28/2026).

Mehraj, Saima and M. Tariq Banday (2020). "Establishing a Zero-Trust Strategy in Cloud Computing Environment." In: *Proceedings of the International Conference on Computer Communication and Informatics (ICCCI)*. IEEE.

NCSC (2021). *Zero Trust Architecture Design Principles*. Tech. rep. Accessed: 3 January 2026. London: National Cyber Security Centre. https://www.ncsc.gov.uk/collection/zero-trust-architecture (visited on 01/03/2026).

Rose, Scott, Oliver Borchert, Stu Mitchell and Sean Connelly (2020). *Zero Trust Architecture*. Tech. rep. NIST Special Publication 800-207. Gaithersburg, MD: NIST. 10.6028/NIST.SP.800-207.

Sandhu, Ravi S., Edward J. Coyne, Hal L. Feinstein and Charles E. Youman (1996). "Role-Based Access Control Models." In: *IEEE Computer* 29.2, pp. 38–47. 10.1109/2.485845.

Saunders, Mark, Philip Lewis and Adrian Thornhill (2019). *Research Methods for Business Students*. 8th. Harlow: Pearson Education.

Shostack, Adam (2014). *Threat Modeling: Designing for Security*. Indianapolis: John Wiley and Sons.

Syed, Naeem Firdous, Syed W. Shah, Arash Shaghaghi, Adnan Anwar, Zubair Baig and Robin Doss (2022). "Zero Trust Architecture (ZTA): A Comprehensive Survey." In: *IEEE Access* 10, pp. 57143–57179. 10.1109/ACCESS.2022.3174679.

Thapa, Gopal Bahadur (2025). "Cybersecurity Challenges in Small and Medium Enterprises (SMEs) in Nepal." In: *International Journal of Multidisciplinary Innovative Research* 2.6. Accessed: 7 May 2026. https://ijmir.com/v2i6/Doc/5.pdf (visited on 05/07/2026).