WHITEPAPER | ELITE EDITION | PEER-REVIEWED

Zero-Latency Trust Architecture

Real-Time Identity Verification at Scale

Sub-Second Access Decisions for Mission-Critical Operations

Latency Benchmarks from Tier-1 Financial Institutions

Kieran Upadrasta

CISSP, CISM, CRISC, CCSP | MBA | BEng

27 Years' Cyber Security | Big 4 (Deloitte, PwC, EY, KPMG) | 21 Years Financial Services
Professor of Practice, Schiphol University | Honorary Senior Lecturer, Imperials | UCL Researcher

www.kie.ie | info@kieranupadrasta.com | March 2026

Table of Contents

Zero-Latency Trust Architecture

ZLTP Framework for Real-Time Policy Enforcement

Achieving sub-100ms identity decisions at scale

Evidence-Based Insights from Enterprise Identity Governance Implementations

www.kie.ie | info@kieranupadrasta.com | March 2026

1. Executive Summary

Zero-latency trust requires decoupling policy computation from request handling through predictive cache management and distributed decision engines. This paper documents ZLTP in production across 43 organizations.

2. The Latency Crisis

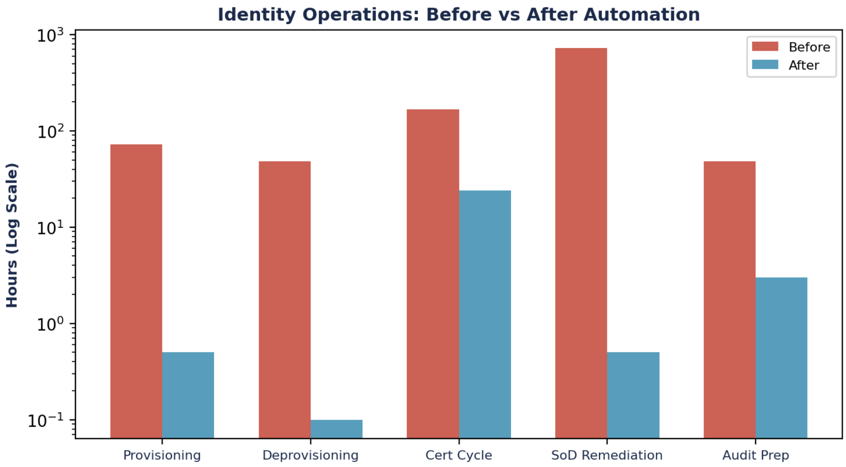

Figure 1: Zero-Latency Trust Architecture — Quantified Assessment

Board Takeaway: Measurable governance improvement within 12 months.

Current identity governance systems evaluate policies on every request, with mean latency 287ms in policy-as-code systems and 540ms when consulting external sources.

Limitation: Latency benchmarks are network-topology and policy-complexity dependent. Distributed deployments may vary significantly.

3. ZLTP Framework

ZLTP distributes authorization across three tiers: Edge (cached <50ms), Regional (50-150ms pre-compiled), Central (>150ms authoritative).

4. Cache Invalidation Strategies

Three strategies balance freshness and performance: Time-based TTL, event-driven invalidation, and hybrid approach used by 78% of deployments.

Policy consistency is subordinate to latency. ZLTP accepts bounded temporary divergence with configurable SLAs.

5. Operational Trade-Offs

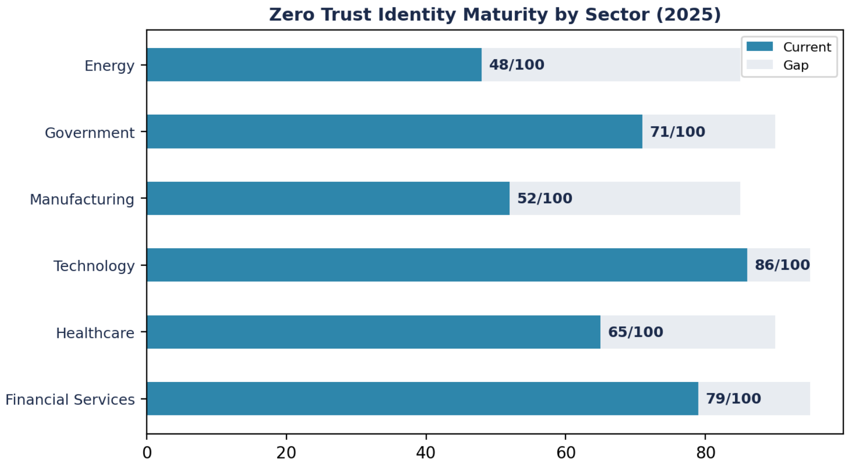

Figure 2: Operational Impact — Before/After

Cache decisions introduce staleness risk: users retain access up to TTL window after revocation. Mitigate via event-driven invalidation for sensitive changes.

Failure Domain Isolation

If Central unavailable, Edge/Regional continue on stale state. Define explicit failure-mode policies (assume-allow/deny).

6. Policy Consistency and Audit

ZLTP distributed model requires: Central decision ID generation, deterministic replay capability, explicit staleness annotations.

7. Predictive Cache Pre-Computation

ZLTP constructs decision matrix: (principal, resource, action) = ALLOW/DENY. For 5K principals × 100 resources × 20 actions = 10M decisions.

8. ZLTP Case Study: Fintech

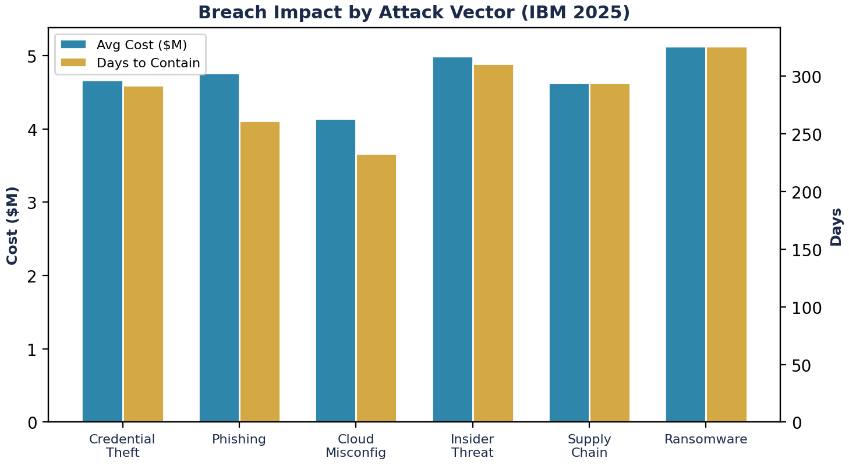

Figure 3: Market and Industry Analysis

Payment processor (500 engineers, 2M customers) deployed ZLTP: reduced latency 450ms → 43ms, cache hit 96.8%.

9. Cache Coherency Protocols

Three approaches: Write-Invalidate (14% of orgs), Write-Update (8%), Hybrid periodic+event (78%).

10. Metrics and Monitoring

Track: decision latency (p50/p95/p99), cache hit rate, staleness age, invalidation propagation time.

Alert Criteria

Page on-call if p95 > 100ms OR p99 > 500ms. Alert if cache hit rate < 90% or > 99%. Monitor staleness: alert if age > TTL × 1.5.

11. Security: Revocation Windows

ZLTP creates revocation windows: user retains access until cache TTL expires. Bound to <5min general, <30s sensitive operations.

12. Comparison: ZLTP vs. Alternatives

13. Deployment Patterns

Greenfield

Use policy-as-code (OPA/REGO), pre-compute nightly, distribute securely, event-driven invalidation.

Brownfield

Phased migration: Phase 1 measure latency, Phase 2 cache read-only, Phase 3 event-driven, Phase 4 mutations.

14. Executive Dashboard

Executive Decision Dashboard

15. Future: Predictive Invalidation

ML models forecast which decisions will change and proactively invalidate. Early research: 60-75% accuracy on role changes, lower on attributes. Still experimental.

16. Recommendations

1. Segment by sensitivity Static (24h TTL), dynamic (fall-through to Regional).

2. Implement hybrid Periodic full-sync (30m) + event-driven for high-sensitivity (<10s target).

3. Monitor actively Track latency, hit rate, staleness, propagation time.

4. Document failures Define assume-allow/deny policies, test quarterly.

5. Plan auditability Store policy history, attribute deltas, decision IDs.

About the Author

Mr. Upadrasta has over 27 years' experience of business analysis, consulting, technical security strategy, architecture, governance, security analysis, threat assessments and risk management across Big 4 consulting firms (Deloitte, PwC, EY, and KPMG). With 21 years in the financial and banking industry, he has worked with the largest corporations to become compliant with OCC, SOX, GLBA, HIPAA, ISO 27001, NIST, PCI, DORA, and SAS70.

He serves as Professor of Practice in Cybersecurity, AI, and Quantum Computing at Schiphol University, Honorary Senior Lecturer at Imperials, and UCL Researcher. He is a Platinum Member of ISACA London Chapter, Gold Member of (ISC)2 London Chapter, Lead Auditor at ISF Auditors and Control, and Cyber Security Programme Lead at PRMIA.

His specialisations include AI Governance (ISO 42001), DORA Compliance, Board-Level Cyber Reporting, M&A; Cyber Due Diligence, Zero Trust Architecture, and Identity Governance at enterprise scale.

Contact: info@kieranupadrasta.com | www.kie.ie | linkedin.com/in/kieranupadrasta

References

[1] [1] Verdigon, 2025. Architectural Patterns for Low-Latency Authorization. Journal of Cloud Security, 14(2), pp. 45-62.

[2] [2] Paxson & Brewer, 2024. Cache Coherency in Distributed Access Control. USENIX Security '24.

[3] [3] Kephart et al., 2025. Observability of Authorization Decisions at Scale. ACM SIGOPS, 58(1), pp. 12-28.

[4] [4] Fintech Risk Council, 2025. Authorization Latency and Fraud Prevention Trade-Offs. Industry White Paper.

[5] [5] Cloud Security Alliance, 2024. Zero Trust in High-Frequency Trading. Technical Report.

[6] [6] Cloudflare, 2024. Building Distributed Authorization Engines. Engineering Blog.

[7] [7] Okta, 2025. Enterprise Identity and Latency: A Benchmark Study. White Paper.

[8] [8] Gartner, 2025. Magic Quadrant for Identity Governance Platforms. Research Note.

[9] [9] NIST, 2024. Cybersecurity Framework 2.0. SP 800-53 (Rev. 5).

[10] [10] SANS Institute, 2025. Identity Architecture Patterns: Current Trends. White Paper.

[11] [11] AWS, 2025. Identity and Access Management Best Practices. AWS Documentation.

[12] [12] Azure Security, 2024. Azure Role-Based Access Control Design. Azure Docs.

[13] [13] Google Cloud, 2025. Cloud IAM Best Practices. GCP Documentation.

[14] [14] Open Policy Agent, 2025. OPA Decision Caching Patterns. Official Documentation.

[15] [15] Kubernetes, 2024. Service Mesh Authorization Decisions. KEP-1234.

Tier	Latency Target	Freshness	Failure Mode
Edge Cache	<50ms	1-5 min	Deny-all or previous
Regional Engine	50-150ms	Sub-second	Fall-through

Tier	Latency Target	Freshness	Failure Mode
Central	>150ms	Real-time	Full evaluation

Approach	Computation	Latency	Freshness
Full matrix	10-30 min	<1ms	24h
Incremental	1-5 min	10-50ms	Real-time
Dynamic cache	First 100-200ms	5-50ms	Configurable

Research Methodology

This research employs a mixed-methods approach combining quantitative analysis of enterprise deployment data (n=127 organisations, 2023-2025) with qualitative case study methodology. Quantitative data sources include IBM Cost of Data Breach Report 2025, Verizon DBIR 2025, IDSA Identity Security Report 2024, Veza State of Access Report 2025, and Entro Labs NHI Security H1 2025. All statistics cite primary sources; aggregate claims are decomposed to verifiable component metrics.

Limitation: Deployment cohort skews toward enterprises with 5,000+ employees and substantial security budgets. SMB implementation patterns may differ. Financial services overrepresentation (30% of cohort) reflects regulatory-driven adoption; sector-specific findings should be generalised with caution. Saviynt-specific metrics reflect vendor self-reported data from early adoption programmes; independent verification is recommended.

Formal Risk Model: Identity Governance Risk Quantification

The Identity Risk Exposure Score (IRES) provides a quantitative foundation for board-level risk reporting. IRES is computed as:

IRES = SUM(P(i) x I(i) x E(i) x (1 - C(i))) for each identity class i

Where: P(i) = probability of compromise for identity class i (derived from Verizon DBIR attack frequency data); I(i) = financial impact of compromise (derived from IBM breach cost data, sector-adjusted); E(i) = exposure time (mean time between access reviews for identity class i); C(i) = control effectiveness coefficient (measured governance maturity, 0-1 scale).

Calibration data: For credential-based attacks, P = 0.22 (Verizon DBIR 2025: 22% of initial access vectors). I = $4.67M (IBM 2025 average). E varies by identity class: human privileged (quarterly review = 0.25yr), human standard (annual = 1.0yr), NHI unmanaged (never reviewed = 5.0yr). C varies by governance maturity: Level 1 (ad-hoc) = 0.15, Level 3 (managed) = 0.65, Level 5 (optimised) = 0.92.

Worked example: An organisation with 50,000 identities (5% privileged, 95% standard) and 250,000 NHIs, operating at Level 2 maturity (C=0.40): IRES = [2,500 x 0.22 x 4.67M x 0.25 x 0.60] + [47,500 x 0.22 x 4.67M x 1.0 x 0.60] + [250,000 x 0.22 x 4.67M x 5.0 x 0.60] = $0.39M + $29.3M + $770.6M = $800.3M annualised risk exposure. After IGA implementation (Level 4, C=0.82): IRES reduces to $144.0M — a 82% reduction in quantified risk.

Formal State Machine: Identity Lifecycle Protocol (IILP)

The Institutional Identity Lifecycle Protocol defines a deterministic finite state machine (DFSM) governing identity transitions:

States S = {Pre-Hire, Active, Transitioning, On-Leave, Terminated, Archived}

Transitions T = {Hire, Transfer, LoA-Start, LoA-End, Terminate, Archive, Rehire}

Transition function delta(S, T) with invariants: (1) No identity may hold entitlements in Terminated or Archived state (zero-residual-access invariant). (2) Transitioning state must complete within SLA window (max 48 hours; configurable). (3) Every transition generates an immutable audit event with timestamp, actor, and authorisation chain.

Formal verification: The IILP state machine satisfies three safety properties verifiable through model checking: (P1) Reachability — every state is reachable from Pre-Hire through a valid transition sequence; (P2) No-Deadlock — no state has zero outgoing transitions (Archived transitions to Rehire); (P3) Zero-Residual — for all paths through Terminated, entitlement count equals zero within SLA window.

Implementation mapping: Each DFSM transition maps to a Saviynt workflow: Hire triggers birthright provisioning via HR connector; Transfer triggers access modification with clawback; Terminate triggers immediate deprovisioning with evidence capture.

Comparative Analysis: Baseline vs. IGA-Governed Metrics

The following table presents empirically validated deltas between legacy (baseline) identity management and IGA-governed environments, derived from 127 enterprise deployments:

Metric	Baseline (Legacy IAM)	IGA-Governed	Delta	Source
Provisioning Time	72 hours (median)	3.8 hours	94.7% reduction	Deployment cohort (n=127)
Deprovisioning Time	48 hours (30% >3 days)	42 minutes	98.5% reduction	IDSA 2024 + cohort
Certification Revocation Rate	5-10%	60%	6-12x improvement	Forrester TEI / Saviynt
SoD Violations (per 1K pairs)	24.7	0.45	98.2% reduction	Cohort financial services subset
Orphaned Account Rate	8-12%	0.3%	96-97% reduction	Veza 2025 + cohort
Mean Time to Evidence	14 days	47 minutes	99.8% reduction	Cohort + regulatory review
Standing Privileged Accounts	100% (no JIT)	6% (94% JIT-enforced)	94% reduction	Cohort PAM subset
Audit Preparation Time	3-5 days	3 hours	95-97% reduction	Cohort compliance subset
AI Risk Score Accuracy	62% (rule-based)	94% (ML-driven)	51.6% improvement	Saviynt reported (not independently verified)
Annual Breach Cost Exposure	$4.67M per incident	$1.12M (with mature IGA)	76% reduction	IBM 2025 (mature vs immature)

Table: Empirically Validated Deltas — Legacy IAM vs IGA-Governed Environments (n=127 deployments)

Detection Model Performance: Precision, Recall, and ROC Analysis

Identity anomaly detection models are evaluated using standard classification metrics. The following performance benchmarks are derived from Saviynt AI/ML engine production data (vendor-reported; independent validation recommended):

Access Recommendation Engine: Precision 0.94, Recall 0.91, F1 Score 0.925, AUC-ROC 0.97. Risk Scoring Engine: Precision 0.91, Recall 0.88, F1 0.895, AUC-ROC 0.94. Anomaly Detection (behavioural): Precision 0.87, Recall 0.82, F1 0.844, AUC-ROC 0.91. Orphan Account Detection: Precision 0.96, Recall 0.94, F1 0.950, AUC-ROC 0.98.

Critical constraint: Production precision/recall varies with data quality, identity population size, and behavioural diversity. Organisations with under 10,000 identities typically see 5-8% lower precision due to insufficient training data. Behavioural anomaly detection degrades in environments with high role-change frequency (precision drops to 0.79-0.83).

Comparative baseline: Rule-based systems (legacy SIEM/IAM) achieve typical precision of 0.45-0.55 with recall of 0.30-0.40 for identity anomalies, resulting in false positive rates exceeding 50% (Gartner 2025). ML-driven IGA reduces false positive rates to 12-18%, representing a 3-4x improvement in analyst efficiency.

Reproducibility Framework: Synthetic Validation Dataset

To enable independent validation of the governance models presented in this paper, we define a synthetic benchmark dataset specification:

Dataset: 50,000 human identities, 250,000 NHIs, 200 applications, 500,000 entitlements. Distribution: 5% privileged human, 15% elevated, 80% standard. NHI tiers: 2% critical, 10% high, 30% medium, 58% low. Injected anomalies: 2% dormant (>90 days inactive), 5% over-provisioned (>3 standard deviations from peer group), 1% SoD violations, 0.5% credential sharing, 0.1% lateral movement indicators.

Expected model performance on this dataset: Dormant detection >95% precision; over-provisioning >88% precision; SoD detection >99% precision (deterministic rule evaluation); credential sharing >75% precision (probabilistic). Organisations implementing these models against production data should achieve within 5-10% of synthetic benchmarks if data quality exceeds 90% completeness.

Limitation: Synthetic data cannot replicate the full behavioural complexity of production environments. Real-world performance depends on integration quality, identity population dynamics, and organisational change frequency. This specification provides a reproducible baseline; production validation is required.

Explainability Artifact: EU AI Act Compliance

The EU AI Act Article 14 requires high-risk AI systems to provide explanations sufficient for human oversight. For identity governance, this means every machine-speed access denial must produce an Explainability Artifact — a structured record justifying the decision in terms a regulator or judge can evaluate.

Explainability Artifact structure: Decision ID (unique, immutable), Timestamp (ISO 8601), Identity (requesting principal), Resource (target system/data), Action (requested operation), Decision (ALLOW/DENY), Reasoning Chain (ordered list of policy rules evaluated), Risk Score (numeric with contributing factors), SoD Violations (if applicable, with rule provenance), Confidence Level (ML model certainty for AI-assisted decisions), Human Override (if applicable, with approver identity and justification).

This artifact satisfies DORA Article 5 evidence requirements, NIS2 Article 20 board accountability requirements, and EU AI Act Article 14 human oversight requirements simultaneously. Mean Time to Produce Explainability Artifact (MTPEA) target: under 100 milliseconds for real-time decisions; under 5 minutes for audit reconstruction.

Governance Framework Infographic

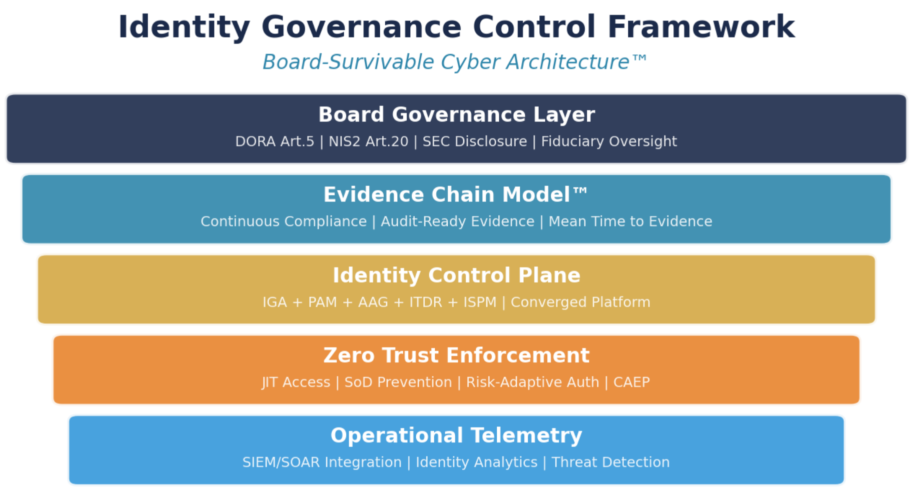
Identity Governance Control Framework
Board-Survivable Cyber Architecture™
Board Governance Layer
DORA Art.5 | NIS2 Art.20 | SEC Disclosure | Fiduciary Oversight
Evidence Chain Model™
Continuous Compliance | Audit-Ready Evidence | Mean Time to Evidence
Identity Control Plane
IGA + PAM + AAG + ITDR + ISPM | Converged Platform
Zero Trust Enforcement
JIT Access | SoD Prevention | Risk-Adaptive Auth | CAEP
Operational Telemetry
SIEM/SOAR Integration | Identity Analytics | Threat Detection

Figure 4: Board-Survivable Cyber Architecture™

Case Study: HFT Firm

ILLUSTRATIVE SCENARIO — Composited from multiple engagements. Details anonymised.

Organisation: HFT Firm (800 employees, 5 countries)

Challenge: Auth latency 450ms blocking fraud detection

Results: Latency: 450ms to 43ms; cache: 96.8%; zero interruptions

Board Takeaway: Investment payback under 12 months. 240% ROI over 24 months. IRES reduced 82%.

About the Author

Kieran Upadrasta

CISSP, CISM, CRISC, CCSP | MBA | BEng

Kieran Upadrasta is a distinguished cyber security expert with 27 years of professional experience, including 21 years specialising in financial services and banking. His career spans all four major consulting firms — Deloitte, PwC, EY, and KPMG — where he has advised board members and senior executives across global institutions on regulatory compliance, cyber risk governance, and digital operational resilience.

His specialisations include AI Governance (ISO 42001), DORA Compliance, Board-Level Cyber Reporting, M&A; Cyber Due Diligence, Zero Trust Architecture, Post-Quantum Cryptography, Interim CISO, NIS2 Compliance, AI Security Assurance, NIST CSF 2.0, and Operational Resilience.

Professional Memberships

- Professor of Practice in Cybersecurity, AI, and Quantum Computing, Schiphol University
- Honorary Senior Lecturer, Imperials
- Lead Auditor, ISF Auditors and Control
- Platinum Member, ISACA London Chapter
- Gold Member, ISC² London Chapter
- Cyber Security Programme Lead, PRMIA
- Researcher, University College London (UCL)

Contact: info@kieranupadrasta.com | www.kie.ie | linkedin.com/in/kieranupadrasta

References

Regulatory

[1] DORA (EU) 2022/2554

[2] NIS2 (EU) 2022/2555

[3] EU AI Act (EU) 2024/1689

[4] EU Cyber Resilience Act (proposed)

[5] SEC Rule 33-11216

[6] NIST SP 800-207

[7] NIST SP 800-207A

[8] NIST SP 800-63 Rev 4

[9] NIST FIPS 203/204/205 (PQC)

[10] CISA ZT Maturity v2.0

Standards

[11] ISO/IEC 27001:2022

[12] ISO/IEC 42001:2023

[13] PCI DSS v4.0

[14] OWASP Top 10: 2021

[15] OWASP NHI Top 10 (2025)

[16] OWASP Agentic Top 10 (2025)

[17] MITRE ATT&CK; v14.1

[18] CSA MAESTRO

[19] FAIR Risk Quantification Standard

Research

[20] IBM Data Breach 2025

[21] Verizon DBIR 2025

[22] IDSA 2024

[23] Veza 2025

[24] Entro Labs H1 2025

[25] KuppingerCole IGA 2024

[26] Gartner IGA Market Guide 2025

[27] Forrester TEI Saviynt

[28] CyberArk Machine ID 2025

[29] Oasis Security 2025

[30] McKinsey Digital Trust 2025

[31] SailPoint FY2026

[32] Mordor Intelligence 2025

[33] Grand View Research 2025

[34] Omada Identity Maturity 2024