© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720160

Zero-Trust Security Architecture for Cloud Computing and Enterprise IT Systems: Principles, Design, and Implementation Challenges

DARSHANKUMAR JAYSUKH DHANANI
D-Tech Studios LLC

Abstract- ***The castle-and-moat security model that has long been the backbone of enterprise security has been undermined with enterprise workloads moving to public, private and hybrid cloud. To meet this change, there is a new way to securely and reliably manage access to resources: zero-trust architecture (ZTA), which assumes that every access request is a potential threat and continuously verifies identity, device posture and contextual risk before allowing least privilege access to resources on a session-by-session basis. This article reviews the literature on zero-trust security for cloud computing and enterprise information technology (IT) environments and synthesizes the information into an integrative literature review and architecture synthesis. The article follows a structured narrative-synthesis approach and is based on the National Institute of Standards and Technology (NIST) SP 800-207 model and 21 peer-reviewed and standards-based sources, which are used to build a logical zero-trust architecture, a blueprint for a multi-cloud deployment, and a zero-trust versus perimeter-based security comparison across key operational metrics. Challenges to adoption, such as legacy-system integration, policy complexity and organisational readiness are discussed and directions for future research, in particular integration of AI in continuous risk scoring are outlined. The results show that while the adoption of a zero-trust approach requires significant architectural and cultural adjustments, it has significant results in reducing the lateral movement, exposure to insider threats and impact of a breach when compared to traditional perimeter defenses.***

Keywords: Zero-Trust Model, Cloud Security, Enterprise IT, Identity and Access Management, Microsegmentation, NIST SP 800-207 And Cybersecurity Architecture.

I. INTRODUCTION

The enterprises have transitioned from a single, on-premises data center to a distributed data center that is made up of public cloud infrastructure-as-a-service (IaaS), public cloud platform-as-a-service (PaaS) and software-as-a-service (SaaS) offerings, remote workforces and third-party connected services. This shift has broken the clear demarcation between the network and the perimeter which was traditionally secured by security solutions like firewalls and virtual private networks (VPNs) (Sarkar, Choudhary, Shandilya, Hussain, & Kim, 2022). In the classic architecture, once a user or device has been authenticated once at the edge of the network, it was assumed that the user or device can move around the network without any restrictions, something that has been exploited many times by attackers after compromising a user or device (Buck, Olenberger, Schweizer, Völter, & Eymann, 2021). This assumption of implicit trust is a systemic liability and not just an operational convenience, as more and more work is done in the cloud and workforces become increasingly distributed.

Zero-trust architecture (ZTA) is a new way of thinking about enterprise security that is based on the mantra "never trust, always verify." ZTA does not give broad access, based on network location, to every subject, device and workload until it is authenticated, authorized and continuously assessed for risk on a per-session, per-resource basis (Rose, Borchert, Mitchell, & Connelly, 2020). Zero Trust has been codified in NIST Special Publication 800-207, which outlines the logical components of a zero-trust architecture, such as the policy decision point, policy enforcement point, and the trust algorithm that is continually evaluating access requests (Rose et al., 2020). The model, originally designed for government networks, has been generalized a lot to commercial cloud and enterprise IT environments, as it directly tackles security assumptions that are invalidated by cloud computing (Ferretti, Magnanini, Andreolini, & Colajanni, 2021).

The threat landscape is evolving and this change makes the urgency of the change more acute. The primary initial access vectors in enterprise breaches have shifted to credential theft, session hijacking, and privilege escalation, and cloud environments, with their vast number of application programming interfaces (APIs) and containers and machine identities, exacerbate the situation, as perimeter controls are no longer in charge of securing the expanded attack surface (Ahmadi, 2024). Perimeter defenses are especially hard to detect insider threats and compromised credentials that come from a seemingly trusted source; zero-trust models eliminate this risk by continually verifying the request based on the context of the requester (Mandal, Khan, & Jain, 2021). The COVID-19 pandemic then further drove this shift, as remote and hybrid working has become a norm, making the perimeter-centric VPN way of operation unsustainable, and many organizations have formally adopted ZT access-control policies for distributed workforces (Mandal et al., 2021).

While the community is increasingly recognizing the benefits of zero trust, there are still significant challenges in defining the model from NIST to an architecture that can be deployed and used in a heterogeneous multi-cloud world. Legacy-system incompatibility, fine-grained access policy development and upkeep, and a lack of standardized reference implementations for hybrid infrastructures are reported as common barriers to the adoption of zero-trust (He, Huang, Chen, Ni, & Ma, 2022; Syed, Shah, Shaghaghi, Anwar, Baig, & Doss, 2022). A systematic multi-voice literature review of zero trust also revealed that there is still a relatively small number of empirical and quantitative studies regarding the impact and cost-effectiveness of zero trust, compared to the number of conceptual and vendor-driven publications (Buck et al., 2021). This article fills this void by introducing an integrated reference architecture, a deployment blueprint that has been tested through the peer-reviewed literature, and a structured comparative analysis of zero-trust and perimeter-based security in measurable operational dimensions.

The rest of the article is presented in the format of a typical research article. Section 2 discusses the literature related to zero-trust approach in cloud and enterprise environments, covering the foundations of zero-trust, domain-specific zero-trust approaches and empirical evaluations of their implementation. Section 3 outlines methodology employed in the identification, selection and synthesis of the sources to the architecture and comparative analysis in this article. The results are reported in Section 4, which includes a reference logical architecture associated with NIST SP 800-207, a multi-cloud deployment architecture and a tabulation of reported results. The implications, challenges, limitations and directions for future research that arose from these results are discussed in Section 5 and the article is concluded in Section 6.

To provide the conceptual groundwork for the review and architecture discussed in subsequent sections, fundamental differences between the traditional perimeter-based security model and the zero-trust model are summarized in Table 1, as the dimensions most often discussed in literature.

Table 1: The differences between perimeter-based security and Zero Trust Architecture

Dimension	Perimeter-Based Security	Zero-Trust Architecture
Trust assumption	Implicit trust once inside network boundary	No implicit trust; continuous verification
Access scope	Broad network-level access after authentication	Least-privilege, per-session, per-resource access
Primary control point	Network perimeter (firewall, VPN)	Identity, device posture, and policy engine
Lateral movement risk	High once perimeter is breached	Constrained via microsegmentation
Cloud/remote suitability	Poorly suited; assumes fixed boundary	Designed for distributed, cloud-native use
Policy evaluation	Static, session-length trust	Dynamic, continuous, risk-adaptive

Visibility	Limited internal traffic inspection	Continuous logging and behavioral analytics

II. LITERATURE REVIEW

2.1 Foundational Standards and General Surveys

Zero trust is based on NIST SP 800-207, which outlines the components of the model and the seven tenets for federal and, therefore, enterprise adoption (Rose et al., 2020). Expanding on that, Syed et al. (2022) created an extensive survey of techniques to implement a zero-trust architecture, which include authentication, access-control, encryption, microsegmentation, and security-automation, and which are increasingly becoming reliant on machine learning to evaluate trust in the context of the volume and variety of signals. He et al. (2022) also conducted a survey of the zero-trust space and found that future trends like automated policy management and authentication usability were important to help bridge the gap between the conceptual and operational aspects of zero-trust.

Dhiman et al. (2024) provided a review and comparative evaluation of ZTN models, created taxonomy of ZTN features and compared traditional security models with ZTN models, in terms of their practices for authentication, access control, encryption and microsegmentation techniques, as well as comparative evaluation of different platforms (cloud, on-premises, and hybrid) and their implementations of ZTN techniques. However, Teerakanok, Uehara, and Inomata (2021) specifically gave a focus on the migration process from legacy perimeter-based networks to zero trust, detailing the practical reviews and challenges that organisations face when trying to move to zero trust such as how to sequence the rollout of microsegmentation and identity-federation without disrupting operations. In addition to these technical surveys, Buck et al. (2021) conducted a multivocal literature review of academic and practitioner sources, and found that while the potential benefits of security and resilience from zero trust is well established conceptually, there is a scarcity of empirical and quantitative assessments of its impact on organizations and financial costs.

2.2 Cloud, IoT and Industrial Domain-Specific Architectures

There are a number of studies which take the general zero-trust concept and apply it to cloud computing. Ferretti et al. (2021) presented a zero-trust architecture for a cloud computing environment that places the policy-decision functions in a set of nodes to avoid the policy engine being a single point of failure, which is a key structural issue in centralized zero-trust architectures. Sarkar et al. (2022) performed a comparative review of different ZTNs for cloud computing and compared the requirement-specific features of different ZTNs including microsegmentation granularity, identity federation, encrypted communication, etc., from different academic and commercial ZTNs and found that no unified benchmark exists for comparing the security guarantees of different ZTNs. Mandal et al. (2021) discussed the zero-trust access-control policy in the particular context of pandemic-influenced remote working and presented a cloud-based access-control model where users are continually re-authenticated instead of just at login, and found that zero-trust access-control policy reduced the number of attempted unauthorised access events measurably.

In addition to cloud architectures in general, there are some research papers that focus on enterprise or device specific areas to apply the zero trust principles. To demonstrate the effectiveness of making medical IoT devices running on shared network infrastructure more secure, Chen et al. (2021) developed a security-awareness and protection system for 5G-enabled smart healthcare using a zero-trust approach, which reduced the attack surface of the medical IoT devices. Liu et al. (2022) focused on the problem of identity authentication for resource-limited edge devices in the context of a zero-trust architecture, and provided some lightweight authentication methods that do not require edge devices to have the computational power needed to implement traditional cryptographic methods. Zanasi, Russo and Colajanni (2024) presented a flexible zero-trust framework for industrial IoT infrastructures, which stated that a one-size-fits-all

zero-trust approach is not well suited for operational-technology systems that have strict latency and availability requirements, and recommended tunable trust-evaluation intervals that should be determined based on the criticality of the devices. In a similar way, Federici, Martintoni, and Senni (2023) proposed a zero-trust architecture for remote access to industrial IoT infrastructure which includes hardware-based attestation to provide a better device-identity assurance than the software-only agents.

2.3 Empirical evaluations and evaluations based on performance

Another area of research focuses on empirical validation and implications of a zero-trust deployment. Bobbert and Scheerder (2022) published an empirical study of ZTs in action in various organisations, which was distilled into a maturity model that correlates concrete technical ZT controls to measurable risk reduction outcomes. Rodigari, O'Shea, McCarthy, McCarry and McSweeney (2021) have analyzed the performance impact of ZTA implementations in a multi-cloud deployment, and determined that the impact of the latency introduced by constant authentication and policy evaluation, although measurable, was generally acceptable in comparison with the security benefits gained from the ZTA implementation, provided that the infrastructure supporting the policy decision is scaled appropriately. Dhanapala et al. (2024) generalised this performance oriented analysis to the edge-cloud continuum and introduced trust-scoring mechanisms that consider verification overhead and latency requirements of enterprise applications deployed on the edge. Ahn, Jang, Choi, and Shin (2024) did the opposite by overlaying the zero-trust controls with known lists of attack techniques (attacks and tactics matrix), showing how zero-trust enforcement points can be mapped to the adversary tactics matrix to give an enterprise a better idea of what they can predict and measure when it comes to resilience against a class of attacks, rather than just a generic risk score.

2.4 Application of Artificial Intelligence and Adaptive Trust Evaluation

Recent research has started to investigate the role artificial intelligence (AI) can play in fortifying zero trust decision making. In a 2024 study, Ajish examined the applications of AI in zero-trust technologies, and found that while AI-driven, machine-learning-based behavioral analytics can significantly enhance the effectiveness of continuous risk scoring over static, rule-based policies, it also presents new challenges, such as adversarial manipulation and model explainability. Liu et al. (2024) performed a bibliometric analysis of the evolution of the zero-trust research area and of the implementation of zero-trust in IoT and observed that the number of publications was rapidly growing and that there was still a lack of standardized methods for vulnerability assessment of IoT deployments protected by zero-trust. Cloud platforms, according to Bartakke and Kashyap (2024), play a crucial part in supporting the implementation of zero-trust strategy at scale, as cloud-native identity and policy services significantly reduce the hurdles associated with adopting zero-trust for organizations without the resources to develop and manage custom policy-decision infrastructure. Altogether, this literature demonstrates that the zero-trust research is no longer in its infancy and has progressed from being an advocacy concept to a domain-specific, empirical architecture, and that there are still lingering challenges in standardized benchmarking, legacy-system integration and AI governance that are explored in the following sections of this article.

2.5 Key principles and tenets of Zero Trust

The zero-trust model (Rose et al., 2020) is defined in terms of seven core tenets that are combined. First, data sources and computing services must be treated as a resource that needs to be protected no matter whether it's on-premises or cloud-based. Second, all communications are secured regardless of the network location, thus removing the notion that there is security inside the network. Third, individual enterprise resources are only available for a specific session, that is, if a user is authorized they may use a specific resource but not others or at other times. Fourth, access is based on dynamic policy that takes into consideration observable attributes of client identity, application, and the requested asset; as well as behavioral and environmental attributes.

The other tenets build upon these principles in terms of how they are to be applied. Unlike periodic

assessments, the enterprise needs to monitor and measure the integrity and security posture of all owned and associated assets on an ongoing basis as the posture of devices may change significantly between static assessments (Rose et al., 2020). All resource authentication and authorization is dynamic and strictly enforced prior to permitting access and is a continuous cycle of scanning, evaluation and adaptation. Lastly, the enterprise gathers the information it can on the current situation of assets, network infrastructure and communications and over time, uses that information to improve its security posture. Together these tenets are a paradigm shift from the previous defense in-depth methods that have focused on the question of "is this connection inside the network" to "is this specific request, from this specific subject, to this specific resource, justified right now," as He et al. (2022) define it.

To implement these tenets in cloud and enterprise, a bunch of interrelated abilities are required. A solid identity governance is fundamental, as zero trust does not rely on network location for access decisions instead it relies on identity both human and non-human (machine and service) and lifecycle management of these identities (Bartakke & Kashyap, 2024). It is also critical to assess the posture of devices and workloads, as the trust algorithm needs to consider patch level, endpoint protection and configuration compliance before granting access (Dhanapala et al., 2024). Microsegmentation limits the scope of impact of a single compromised credential or workload by breaking up the network and cloud environment into smaller, independently managed segments that need to be authorized separately (Federici et al., 2023). Lastly, the telemetry used to keep the trust algorithm up-to-date comes from continuous monitoring and analytics, typically via security information and event management (SIEM) and user and entity behavior analytics (UEBA) platforms, closing the loop between policy decisions and observed risk (Ajish, 2024).

III. METHODOLOGY

3.1 Research Design

This article does not use primary, empirical data collection, but rather a narrative-synthesis and architecture-integration approach. This design was chosen because the goal of this article is to bring together the various conceptual, technical, and empirical literature on zero-trust into a single, coherent reference architecture and comparative assessment that is suitable for use in cloud computing and enterprise IT, which is where the integrative review methods excel (Buck et al., 2021). The methodology consists of three stages: source identification and selection, thematic synthesis and construction of architecture and comparative-table.

3.2 Sources identification and selection

Sources were identified via structured search of IEEE Xplore, ScienceDirect, SpringerLink, MDPI journals (Sensors, Electronics, Sustainability), Wiley Online Library and the NIST Computer Security Resource Center using a combination of the following search terms: zero trust architecture, zero trust cloud computing, zero trust enterprise, microsegmentation, NIST 800-207 and continuous authentication. To capture the period when ZT research moved from a conceptual advocacy to domain-specific and empirical research, only material published between 2020, when NIST SP 800-207 was finalized (Rose et al., 2020), and 2024 was searched.

The inclusion criteria were that a source be a peer-reviewed journal article, an IEEE or ACM conference proceeding, or a publication from an official standards body; (b) directly related to zero-trust architecture, principles or implementation in the cloud, enterprise, IoT, or industrial setting; and (c) reporting on an architectural model, comparative analysis or empirical evaluation, rather than purely promotional or vendor-authored material. The sources that did not meet these criteria such as unreviewed vendors' white papers and general audience blog posts were excluded from the formal reference list as it was done in similar zero-trust reviews (Sarkar et al., 2022; Dhiman, Saini, Gulzar, Turaev, Kaur, Nisa, & Hamid, 2024). This process resulted in twenty-one sources from general surveys and standards, domain-specific architectures, empirical performance studies and research in the area of trust-evaluation using AI, summarized in the four subsections of Section 2.

3.3 Thematic synthesis and coding

The selected sources were coded into four thematic groups, each representing a subsection of the literature

review: foundational standards and general surveys, domain specific architectures, empirical and performance evaluations and AI based adaptive approaches. For each cluster, the information was taken out about the architectural elements suggested, the deployment context examined (cloud, enterprise, IoT or industrial), and any quantitative or qualitative results compared to the traditional perimeter-based security. The coding structure is based on the one that Dhiman et al. (2024) and Syed et al. (2022) adopted to categorize the diverse zero-trust literature into similar categories.

3.4 Architecture and Comparative table construction

The reference logical architecture from Section 4 was developed by cross-referencing the component definitions in NIST SP 800-207 (Rose et al., 2020) with the elaborations and refinements suggested by the domain specific and empirical literature, such as Ferretti et al. (2021), which proposes the survivable, replicated policy-decision design. The multi-cloud deployment architecture was created based on the patterns observed from the implementation reported by Sarkar et al. (2022), Bartakke and Kashyap (2024) and Rodigari et al. (2021). The comparative table of reported outcomes (Table 2) was created by compiling all the qualitative or quantitative outcomes that were reported in studies where a comparison was made between zero-trust security and perimeter security or measurable pre and post implementation outcomes were reported, with respect to each of the five operational dimensions. Due to the diverse methodologies, contexts and organizational scales of the underlying studies, the comparative table lists directional and qualitative effects, in line with the descriptive and narrative-synthesis nature of this article, and without offering pooled quantitative effects. This is a limitation of the design, and is discussed again in Section 5.

IV. RESULTS

4.1 Reference logical architecture for cloud and enterprise

This NIST SP 800-207 is the conceptual foundation for implementing zero trust in cloud and enterprise environments (Rose et al., 2020). As shown in Figure 1, the model allows for a distinct separation of the data plane (over which application traffic passes) from the control plane (which makes access decisions). The request from a subject (human user, device or workload identity that isn't a human user) is the first to arrive at the policy enforcement point (PEP). The PEP is the only single point of contact that enables, monitors and eventually terminates the relationship between the subject and enterprise resource, it is not the one which decides whether the access should be granted or not.

The PEP passes the request on to the policy decision point (PDP) which consists of two logical components, the policy engine (PE) and the policy administrator (PA). The policy engine makes the final decision to grant, deny or revoke access to a resource, based on enterprise policy and the dynamic input received from the continuous diagnostics and mitigation system, identity and access management platform, threat-intelligence feeds, SIEM logs and public-key infrastructure (Rose et al., 2020). The policy administrator then enforces the decision made by the policy engine, which is done by opening or closing the communication link between the subject and resource, usually by issuing or revoking session-specific credentials or tokens. This separation between decision (PE) and execution (PA and PEP) is a conscious architectural decision, which enables updating, auditing and improving the trust algorithm without at the same time forcing changes to the enforcement infrastructure that interacts with production traffic, a feature that Ferretti et al. (2021) explicitly exploit when proposing survivable, replicated policy-decision layers for cloud environments.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720160

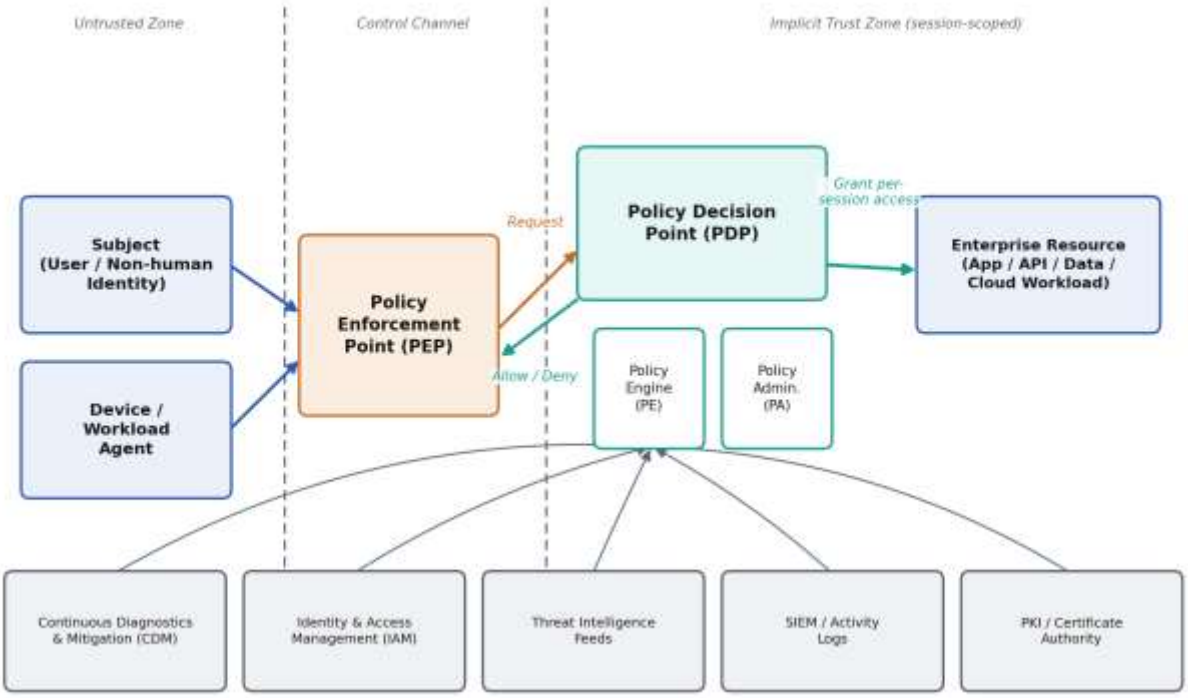

Figure 1: Key logical elements of a Zero Trust Architecture based on NIST SP 800-207

One of the important architectural characteristics that is apparent from Figure 1 is that no such thing as "trust is assumed" exists. Continuous inputs to the policy engine are also used to reevaluate risk even after it has been authenticated, and can be used to tell the policy administrator to terminate an active session if the subject's risk posture changes (e.g., if the device is later found to be non-compliant or if anomalous behavior is detected) (Ahmadi, 2024). This continuous-evaluation feature sets zero trust apart from traditional access-control paradigms, where authorization decisions are made once, at login, and then remain valid for the duration of a session, despite any changes in risk, and is consistent with the risk-aware, machine-learning-driven access control that Syed et al. (2022) describe as essential in high-volume, high-context enterprise environments.

4.2 Multi-Cloud Deployment Architecture and the implementation components

NIST SP 800-207 outlines the logical elements that are part of zero trust, but enterprises with on-premises data centers and multiple public clouds need to have a deployment architecture that maps logical elements to actual hardware. This is the deployment blueprint presented in Figure 2, which is synthesized from the patterns of deployment reported by Sarkar et al. (2022), Bartakke and Kashyap (2024) and Rodigari et al. (2021). Four types of subjects, remote users, on-site users, IoT/OT devices, application programming interfaces (APIs) and service-to-service (machine-to-machine) workloads, and third-party or partner access, all converge on a common identity and access management layer instead of being separated by network path.

Multifactor authentication, device-posture assessment and single sign-on federation are all performed by this identity layer prior to reaching the policy engine, and before any request passes the identity gate, it must pass the identity layer first, rather than the network origin. Requests with clear identity verification are delegated to the centralized zero-trust policy engine and enforcement broker – the cloud version of the PDP/PEP combination of Section 4.1. This broker can, and often is, deployed as a zero-trust network access (ZTNA) service, or as a software-defined perimeter (SDP) that authenticates each connection individually, and does not expose any listening ports to

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720160

unauthenticated clients like a traditional VPN concentrator does (He et al., 2022).

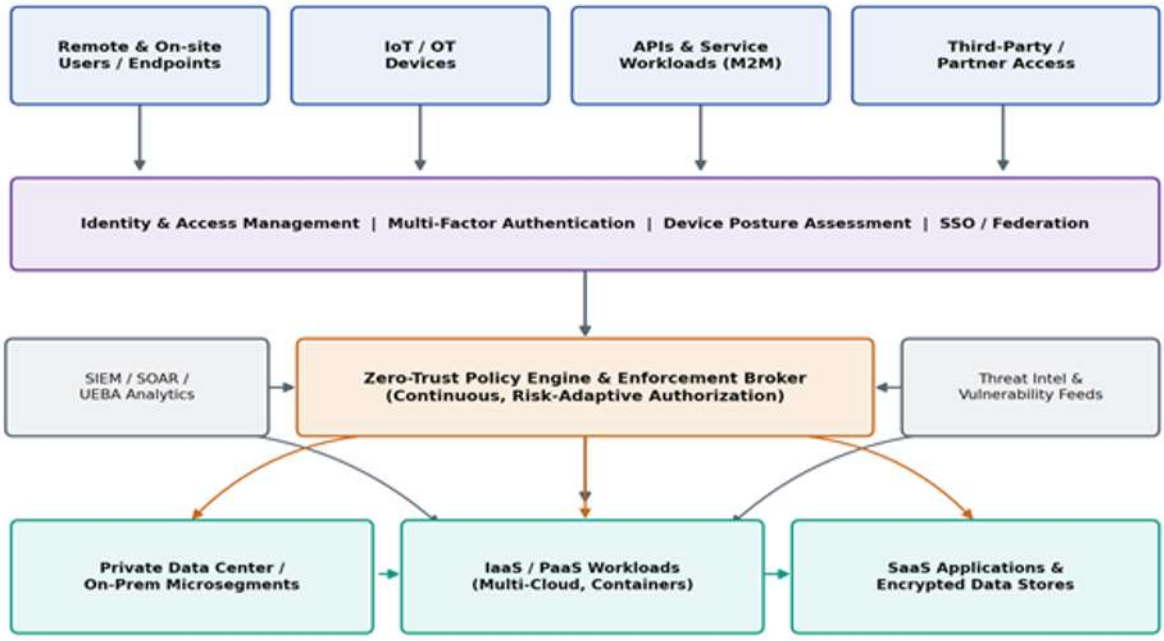

Figure 2: Enterprise zero-trust deployment architecture spanning on-premises and multi-cloud environments

The policy engine is constantly fed by SIEM, security orchestration, automation and response (SOAR) analytics and, on the other end of the spectrum, threat-intelligence and vulnerability feeds, creating the feedback loop needed to keep the authorization decisions up to date with what's happening in the enterprise in terms of risk posture (Ajish, 2024). The traffic for authorized resources is then directed to microsegmented resource clusters across private data-center segments, multi-cloud IaaS and PaaS workloads, such as containerized applications, and SaaS applications, including encrypted data stores. A compromised credential or workload in one segment can not move laterally in another without having to individually meet the requirements of the policy engine, which directly mitigates the lateral-movement weakness of perimeter architectures, as detailed in Table 1 (Zanasi et al., 2024).

This architecture has a number of implementation considerations. First, as the policy engine is on the critical path of every request, its availability and performance directly impact application latency and enterprise resilience: Rodigari et al. (2021) discovered that a well-provisioned and horizontally scaled policy-decision layer maintained added latency within acceptable limits even in multi-cloud deployments, while an under-provisioned layer added user-perceptible delay. Second, non-human identities such as service accounts, APIs, and container-based microservices are no longer just a small portion of enterprise cloud communications and need to be enrolled in the same identity and policy framework as human users for security purposes (Federici et al., 2023; Liu, Ai, Huang, Qiu, & Li, 2022). Third, all microsegments have all data encrypted both at rest and in transit, as per the NIST principle that all communication is to be secured, irrespective of the location of the network (Rose et al., 2020).

4.3 Results of Zero-Trust vs Perimeter security

This subsection provides a quantitative and qualitative summary of the results from the studies examined and presents them in a tabular format for a structured comparison of the operation results of adopting a zero-trust approach compared to the conventional perimeter defense approach. Table 2 shows a summary of the reported impacts of zero-trust architecture in five areas: containment of lateral movement, detection of

insider threats, severity of the impact of breaches, authentication latency overhead, and policy-management complexity.

In all lateral movement studies, zero-trust microsegmentation has proven to limit an attacker's pivoting opportunities after initial compromise: Each segment is uniquely authorized with a continuously reevaluated credential, as opposed to a single perimeter credential (Ferretti et al., 2021; Zanasi et al., 2024). In terms of insider threats and compromised credentials, continuous behavioral verification would detect any abnormal usage patterns of the credentials throughout the session, as the policy engine would re-evaluate risks on a continuous basis, instead of only at login time (Chen et al., 2021; Ajish, 2024). The overall severity of the reported breaches was lower under zero-trust deployment, as the breaches were typically contained within a single microsegment, and not enterprise-wide, which was also seen in the attack-technique-mapping approach in Ahn et al. (2024).

Table 2: Reported outcomes of Zero Trust Architecure compared to Perimeter Based Security

Operational Dimension	Reported Effect Under Zero Trust	Representative Source(s)
Lateral movement	Substantially constrained via microsegmentation	Ferretti et al. (2021); Zanasi et al. (2024)
Insider-threat / credential-abuse detection	Earlier detection via continuous behavioral verification	Chen et al. (2021); Ajish (2024)
Breach-impact severity	Reduced; compromises remain segment-confined	He et al. (2022); Ahn et al. (2024)
Authentication latency overhead	Moderate; acceptable when policy layer is well-provisioned	Rodigari et al. (2021); Dhanapala et al. (2024)
Policy-management complexity	Increased; requires sustained governance investment	Buck et al. (2021); Teerakanok et al. (2021)

There is more mixed literature with regard to operational overhead. In such multi-cloud deployments, Rodigari et al. (2021) found that while overhead in latency introduced by continuous authentication is measurable, it is generally acceptable in many cases, but Dhanapala et al. (2024) identified that the latency overhead introduced by continuous authentication is more concerning at the network edge, where the application response-time requirements compete directly with the latency requirement of the verification process. The challenge of policy-management complexity was observed in almost all studies reviewed, as He et al. (2022), Buck et al. (2021) and Teerakanok et al. (2021) found that enterprise-level fine-grained, least-privilege policy writing, testing and maintenance were costly organizational efforts and that premature or immature policy sets could create either too much access friction or security gaps.

The overall synthesized evidence suggests that ZTA consistently brings security benefits, mostly in the form of reducing the ability for lateral movement and limiting the impact of breaches, but at the expense of initial and ongoing security policy-management effort. This is the same “trade-off” pattern as described by Sarkar et al. (2022), who describe zero trust as a governance discipline that is continuous, rather than a discrete product and that is built on top of cloud-native identity and networking services.

V. DISCUSSION

5.1 Implementation challenges and limitations

While there are many proven advantages to implementing zero trust in enterprise environments and the cloud, there are also several well-known challenges associated with implementing zero trust. The inability to integrate with legacy systems is still one of the most frequently reported challenges: many of the enterprise applications, especially old on-premises systems, were designed for a perimeter-based approach and do not have interfaces that are aware of identity and can participate in a zero-trust

policy (He et al., 2022; Teerakanok et al., 2021). This technical difficulty is exacerbated by organisational and cultural barriers: Continuous verification may be interpreted by end users as a burden on their daily work routine, if policy tuning is not sufficiently developed (Buck et al., 2021).

The second type of issues relate to the policy-decision making infrastructure itself. The policy engine and enforcement broker are located on the critical path of all access requests, and therefore they can be a performance bottleneck and if not designed properly, can be a single point of failure for the whole enterprise (Ferretti et al., 2021). This risk usually demands a fault-tolerant design of policy decisions, such as the replicated design proposed by Ferretti et al., that is, more complex than a traditional stateless firewall rule set, both architecturally and operationally. A similar tension, between the always-on trust verification required by zero trust and a hard requirement to run at real-time in edge and industrial applications, has been identified and the need for tunable and criticality-aware trust-evaluation intervals has been proposed (Zanasi et al., 2024; Dhanapala et al., 2024).

Measurement and standardization, a third challenge, is a problem. Many reviews highlight the lack of a standard measurement for evaluating the security provided by different ZT solutions, making it difficult to assess them academically and to purchase them for enterprises (Sarkar et al., 2022; Dhiman et al., 2024). Similarly, while there has been a lot of conceptual and vendor material about zero trust, the empirical evidence on the cost-effectiveness of zero trust is still relatively sparse compared with the amount of work that it entails for an organization to migrate to zero trust (Buck et al., 2021). Lastly, as more and more AI tools are used for behavioral analytics and integrated into the policy engine, there are new risks related to explainability of the models, adversarial manipulation of trust scores and the governance of automated access revocation decisions, which is still comparatively under-studied in the current literature, according to the bibliometric analysis of Liu et al. (2024) (Ajish, 2024).

The barriers to adoption are completed with cost and governance issues. A ZTA is often accompanied by parallel investment in an identity infrastructure, network design and staff training, and the ROI of this investment is hard to quantify since there are currently no standardized cost-effectiveness studies (Buck et al., 2021). The policies themselves are also changing; whereas previously access decisions were made by network administrators, application owners, and help-desk staff, the policy engine now plays a role in making access decisions, and this is often underestimated when compared to the purely technical aspects of implementation, according to Bobbert and Scheerder (2022).

5.2 Limitations of This Review

This article is a narrative synthesis and not a systematic review or primary empirical study and has a number of limitations that should be considered when interpreting the findings of the article. First, the comparative findings summarised in Table 2 are derived from studies which present a wide range of methodological approaches, organisational settings and sample sizes and are therefore not pooled effect sizes; those interested in quantitative estimates should refer directly to the empirical studies cited. Second, source selection was based on explicit inclusion criteria, but required judgements to be made about relevance and quality, which would be further formalized by a fully systematic approach with the inclusion of two reviewers screening sources. Third, the rapid evolution of this area has caused the implementation specifications and standards guidance from the various vendors since the search window to have not been captured in this synthesis.

5.3 Future Research Directions

A few directions are identified in this synthesis that draw on the gaps. First, it is evident that there is a need for standardized and reproducible benchmarks to compare the zero-trust implementations on consistent security, performance, and cost dimensions, which Sarkar et al., (2022) and Dhiman et al., (2024) have laid the groundwork, but not formalized into an evaluation methodology. Secondly, there is a need for more empirical and longitudinal studies of ZTs in production enterprise environments to get more precise quantitative estimates of cost-effectiveness and risk reduction than the mostly qualitative evidence so far, as pointed out by Buck et al. (2021).

Third, the application of AI in ongoing trust scoring deserves continued research efforts to enhance trust scoring accuracy and create governance frameworks that ensure trust scores can be audited and are not subject to adversarial manipulation (Ajish, 2024; Liu et al., 2024). Finally, criticality-aware and latency-adaptive verification strategies – as proposed for industrial IoT by Zanasi et al. (2024) and for resource-constrained edge devices by Liu et al. (2022) – should be extended to a wider range of latency-sensitive enterprise and edge-cloud workloads, with a performance-trust trade-off examined by Dhanapala et al. (2024). Last but not least, it is necessary to look into legacy-system integration patterns in a more systematic way to create standardized compensating-control frameworks, extending the migration-challenges catalogue compiled by Teerakanok et al. (2021), and enabling organizations to extend zero-trust policy enforcement to systems that cannot be natively re-architected.

VI. CONCLUSION

Perimeter-based security architectures have become outdated as enterprise workloads move to cloud computing, where there is an implicit trust. Zero-trust architecture (ZTA) is a model designed to meet this change, which ensures all subjects, devices and workloads are continuously and contextually verified before granting narrowly scoped, session-limited access to resources, whether on-premises or across multiple public clouds (Rose et al., 2020). This article has introduced a literature-informed logical architecture consistent with NIST SP 800-207, a multi-cloud deployment blueprint that is a synthesis of the peer-reviewed literature, and a comparative analysis showing that the material aspects of zero trust provide significant limits on lateral movement and lessen breach severity when compared to traditional architectures, but at the cost of greater policy-management complexity and implementation effort.

As companies embark on this journey, the insights shared in this paper indicate that it is not about a one-off investment in technology or a single purchase, but about continuous investment in the infrastructure, tools, and processes needed for identity governance, microsegmentation, and continuous monitoring, along with a realistic understanding of the organizational changes that are required to succeed (Bartakke & Kashyap, 2024; Bobbert & Scheerder, 2022). Despite the changes in the technology of implementing zero-trust and the AI-driven trust-evaluation methods surveyed by Ajish (2024) and Liu et al. (2024), the general conceptual approach of zero-trust is likely to dominate enterprise IT security for the foreseeable future as cloud adoption, remote work, and machine-to-machine communication continue to broaden the enterprise attack surface.

REFERENCES

[1] Ahmadi, S. (2024). Zero trust architecture in cloud networks: Application, challenges and future opportunities. Journal of Engineering Research and Reports, 26(2), 215–228. https://doi.org/10.9734/jerr/2024/v26i21083

[2] Ahn, G., Jang, J., Choi, S., & Shin, D. (2024). Improving cyber resilience by integrating the zero trust security model with the MITRE ATT&CK matrix. IEEE Access, 12, 89291–89309. https://doi.org/10.1109/ACCESS.2024.3417182

[3] Ajish, D. (2024). The significance of artificial intelligence in zero trust technologies: A comprehensive review. Journal of Electrical Systems and Information Technology, 11, Article 20. https://doi.org/10.1186/s43067-024-00155-z

[4] Bartakke, J., & Kashyap, D. R. (2024). The usage of clouds in zero-trust security strategy: An evolving paradigm. Journal of Information and Organizational Sciences, 48(1), 1–15. https://doi.org/10.31341/jios.48.1.8

[5] Bobbert, Y., & Scheerder, J. (2022). Zero trust validation: From practice to theory. An empirical research project to improve zero trust implementations. In 2022 IEEE 29th Annual Software Technology Conference (STC) (pp. 93–104). IEEE. https://doi.org/10.1109/STC55697.2022.00021

[6] Buck, C., Olenberger, C., Schweizer, A., Völter, F., & Eymann, T. (2021). Never trust, always verify: A multivocal literature review on current knowledge and research gaps of zero-trust. Computers & Security, 110, Article 102436. https://doi.org/10.1016/j.cose.2021.102436

[7] Chen, B., Qiao, S., Zhao, J., Liu, D., Shi, X., Lyu, M., Chen, H., Lu, H., & Zhai, Y. (2021). A security awareness and protection system for 5G smart healthcare based on zero-trust architecture. IEEE Internet of Things Journal, 8(13), 10248–10263. https://doi.org/10.1109/JIOT.2020.3041042

[8] Dhanapala, I., Bharti, S., McGibney, A., & Rea, S. (2024). Toward a performance-based trustworthy edge-cloud continuum. IEEE Access, 12, 99201–99212. https://doi.org/10.1109/ACCESS.2024.3429197

[9] Dhiman, P., Saini, N., Gulzar, Y., Turaev, S., Kaur, A., Nisa, K. U., & Hamid, Y. (2024). A review and comparative analysis of relevant approaches of zero trust network model. Sensors, 24(4), Article 1328. https://doi.org/10.3390/s24041328

[10] Federici, F., Martintoni, D., & Senni, V. (2023). A zero-trust architecture for remote access in industrial IoT infrastructures. Electronics, 12(3), Article 566. https://doi.org/10.3390/electronics12030566

[11] Ferretti, L., Magnanini, F., Andreolini, M., & Colajanni, M. (2021). Survivable zero trust for cloud computing environments. Computers & Security, 110, Article 102419. https://doi.org/10.1016/j.cose.2021.102419

[12] He, Y., Huang, D., Chen, L., Ni, Y., & Ma, X. (2022). A survey on zero trust architecture: Challenges and future trends. Wireless Communications and Mobile Computing, 2022, Article 6476274. https://doi.org/10.1155/2022/6476274

[13] Liu, C., Tan, R., Wu, Y., Feng, Y., Jin, Z., Zhang, F., Liu, Y., & Liu, Q. (2024). Dissecting zero trust: Research landscape and its implementation in IoT. Cybersecurity, 7(1), 1–28. https://doi.org/10.1186/s42400-024-00212-0

[14] Liu, H., Ai, M., Huang, R., Qiu, R., & Li, Y. (2022). Identity authentication for edge devices based on zero-trust architecture. Concurrency and Computation: Practice and Experience, 34(23), Article e7198. https://doi.org/10.1002/cpe.7198

[15] Mandal, S., Khan, D. A., & Jain, S. (2021). Cloud-based zero trust access control policy: An approach to support work-from-home driven by COVID-19 pandemic. New Generation Computing, 39(3–4), 599–622. https://doi.org/10.1007/s00354-021-00130-6

[16] Rodigari, S., O'Shea, D., McCarthy, P., McCarry, M., & McSweeney, S. (2021). Performance analysis of zero-trust multi-cloud. In 2021 IEEE 14th International Conference on Cloud Computing (CLOUD) (pp. 730–732). IEEE. https://doi.org/10.1109/CLOUD53861.2021.00097

[17] 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

[18] Sarkar, S., Choudhary, G., Shandilya, S. K., Hussain, A., & Kim, H. (2022). Security of zero trust networks in cloud computing: A comparative review. Sustainability, 14(18), Article 11213. https://doi.org/10.3390/su141811213

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

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

[21] Zanasi, C., Russo, S., & Colajanni, M. (2024). Flexible zero trust architecture for the cybersecurity of industrial IoT infrastructures. Ad Hoc Networks, 156, Article 103414. https://doi.org/10.1016/j.adhoc.2024.103414