International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

Cloud-Native Identity and Access Management for Enterprise Platforms

Ravi Kumar Kotapati[1]
[1]*Senior Manager Software Engineering, Cybersecurity.*

Abstract
Cloud-native enterprise architectures have broken out the network perimeter, and now identity has become the front line of defense in the world of security. This review covers the current landscape of cloud-native enterprise identity and access management (IAM) solutions and how the industry has moved beyond a static approach to identity management and perimeter-based security, and is now shifting to dynamic, continually verified access control. A collection of essential research is synthesized covering access-control theory, federated authentication, security of containers and microservices, identity of the service, zero-trust concepts and machine-identity governance. On the basis of this synthesis, we argue that a four-plane theoretical model can be introduced to separate authentication, authorization reasoning, enforcement, and auditability into four closely coupled but independent evolving planes for the system. An illustrative evaluation, gleaned from published benchmarks, illustrates the expected behaviour of this model over 3 dimensions: authorization latency, enforcement overhead and scalability in the face of growth in identities. The results show that the model is viable for realistic enterprise workloads when optimizing policy evaluation and accounting for the overhead of policy enforcement, and provide limits for deployments using latency sensitive and throughput bounded workloads. Finally, the review highlights open challenges and directions for future research, providing a unified point of reference for both researchers and practitioners in the field of secure and scalable identity management in today's enterprise world.

Keywords: *Cloud-native computing; identity and access management (IAM); zero-trust architecture; machine identity; policy-as-code; attribute-based access control (ABAC); federated authentication; enterprise security; secrets management.*

1. Introduction
Organizations' ideas and approaches to identity and access management (IAM) have changed dramatically due to the speed at which enterprise infrastructure has moved to cloud-native architectures. In contrast to the relatively static perimeter-bound networks that traditional IAM systems have been built for, the modern enterprise today is deployed in distributed microservices, containerized workloads, multi-cloud environments, and ephemeral compute resources that come and go in seconds [1]. The traditional definition of trusted internal network has become obsolete, and the new security perimeter is now identity [2]. The explosive growth of both the number and the complexity of identities, whether human or non-human (service accounts, workloads, machine credentials), that are used within enterprises has never been more significant, and is an area where legacy IAM systems were never designed to perform [3]. This is not just about operational simplicity, but it's also about the importance of the subject. But the biggest risks to enterprise security continue to be identity-related, with weak passwords or inadequate access controls being used over and over in high-profile breaches [4]. The attack surface area grows significantly in a cloud-native environment: every microservice, API endpoint and inter-service communication is a potential point of unauthorized access. As such, strong IAM is now considered an integral part of the larger zero-trust security concept that is based on the premise of "no implicit trust," and requires continuous authentication of all access points, regardless of whether they originate from outside or inside the organization [15]. In regulated sectors like finance or healthcare, the consequences are heightened because these regulations require fine-grained auditability, least-privilege access and governance over who, how, and what can access sensitive resources [6]. While there has been significant progress, there are still some challenges

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

and gaps in the existing knowledge and practice. Although in most cloud-native environments, non-human and machine identities are far outnumber human users, there is still a significant lack of development in this area with respect to how they are managed [3]. Secondly, with the decentralized and autonomous nature of microservices, and the tension between centralized policy governance, enforcing consistent authorization at scale with standards interoperability and identity federation is difficult, especially in the case of multi-cloud deployments with mixed cloud types and incompleteness of standards and identity federation [7]. Third, although the vision of zero trust is popular, the operational details of how to make that vision a reality with regard to IAM implementations are still being actively explored, and there is little consensus on any patterns for architecture and testability [5]. Last but not least, the literature is scattered across security engineering, distributed systems, and compliance fields, which hinders researchers and practitioners to get a coherent and integrated overall picture of the field. The aim of this review is to consolidate all the latest advancements and critically review the status quo in the field of cloud-native identity and access management for enterprise platforms. We want to connect the different strands of the existing research in order to build a coherent picture, identify key concepts and provide the underpinning of the current state of architectural work and uncover the open problems that need to be addressed. Readers can look for a systematic narrative on the shift from traditional to cloud-native IAM, a discussion of key technologies and standards, an analysis of zero-trust and policy-driven authorization, a look at machine identity and secrets management and a treatment of governance/compliance issues in the sections that follow. The paper ends with an outline of promising future research directions that aim to provide a comprehensive reference point for both scholars and practitioners to further progress secure, scalable identity management in modern times shown in Table 1. The table below synthesizes ten key studies that inform the current understanding of cloud-native identity and access management.

Table 1 Summary of Key Research

Reference	Findings (Key Results and Conclusions)
[8]	Proposes attribute-based access control (ABAC) as a more flexible alternative to role-based schemes in the context of dynamic environments, where decisions on access are made based on attributes of the situation, instead of on static roles more appropriate to cloud-based, dynamic workloads.
[9]	Identifies identity federation, access control, and trust management as central unresolved challenges, concluding that identity is a primary determinant of overall cloud security posture.
[10]	Identifies the ability of OAuth 2.0 and OpenID Connect to support single sign-on and delegated authorization at scale, yet notes that two common challenges are token leakage and misconfigured redirects.
[11]	Concludes that decomposing applications into microservices multiplies inter-service trust boundaries, requiring per-service authentication and decentralized authorization rather than monolithic perimeter controls.
[12]	Demonstrates that containerized deployments introduce distinct identity risks, including secret sprawl and over-privileged service accounts, and recommends short-lived credentials and least-privilege defaults.
[13]	Reviews evolving access-control paradigms and concludes that traditional models scale poorly to highly distributed systems, motivating policy-as-code and dynamic, context-aware authorization.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

[14]	Shows that service meshes can automate workload identity and mutual TLS at scale, substantially reducing manual credential management while introducing new configuration-complexity trade-offs.
[15]	Establishes formal principles for continuous verification and least-privilege access, positioning identity as the core control plane and providing a reference model for enterprise adoption.
[16]	Recognizes that by separating the authorization logic from the application into declarative policy engines (like OPA), authorization becomes consistent and traceable across a diverse set of services, at the cost of additional tooling and skills in the operation of these policy engines.
[17]	Defines a standard for cryptographically verifiable workload identity with short-lived, automatically rotated credentials, enabling first-class machine-identity management at scale. and are frequently under-governed; advocates automated lifecycle management and cryptographic workload identity as essential controls.

In all of these studies, there are several common themes: 1) The shift from a perimeter-based security model to an attribute-based one [8, 13]; 2) The importance of federated and standards-based identity and authentication [10]; 3) The security considerations of architectural decomposition into microservices and containers [11, 12]; and 4) The need for machine identity and policy-driven governance to secure cloud-native operations [16, 17]. The results support the issues presented in the introduction and justify further evaluation of the issues that follow.

2. Proposed Theoretical Model

The proposed model organizes cloud-native IAM into four interacting layers—an identity plane, a policy and decision plane, an enforcement plane, and a governance and audit plane. This layered separation follows the zero-trust principle that authentication and authorization decisions must be continuously evaluated and decoupled from the enforcement points that carry them out [15], and it extends the policy-as-code paradigm by treating authorization logic as a centrally governed, externally managed concern [16]. In this architecture, the identity plane establishes trustworthy identities for both human and non-human principals. Human identities are handled through federated authentication and single sign-on using protocols such as OAuth 2.0 and OpenID Connect [10], while workload identities are provisioned cryptographically and their secrets managed through automated lifecycle controls—an area shown to be systematically under-governed in practice [17].

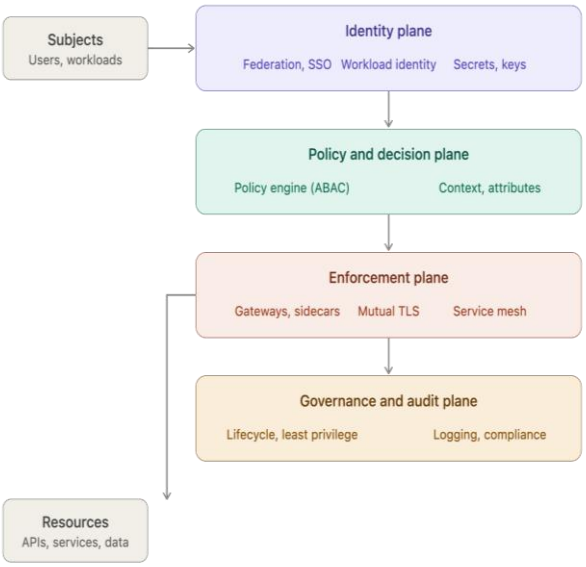

Figure 1 Proposed layered architecture for cloud-native identity and access management, showing the four interacting planes—identity, policy and decision, enforcement, and governance and audit that mediate access between subjects and resources

The decision to treat machine identity as a first-class citizen of this plane, rather than an afterthought, is a deliberate response to the observation that non-human identities now vastly outnumber human ones

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

[12], [17] shown in Figure 1. Authorization reasoning is brought together in the policy and decision plane. The model externalizes access rules into a declarative policy engine which is able to evaluate requests against contextual attributes, similar to attribute-based access control [8] and the movement from the concept of statically assigned roles to contextually-aware authorization [13]. This separation will help to have consistency and audibility over the heterogeneous services, but at the same time, it will create new operational tooling requirements [16]. Enforcement plane – the decision-making action at the point of access. In cloud-native deployments, this is achieved by API gateways and service-mesh sidecars that will intercept traffic and enforce mutual TLS between workloads, automatically verifying the identity of the workloads at scale, but at the cost of complexity in configuration [14] shown in Figure 2.

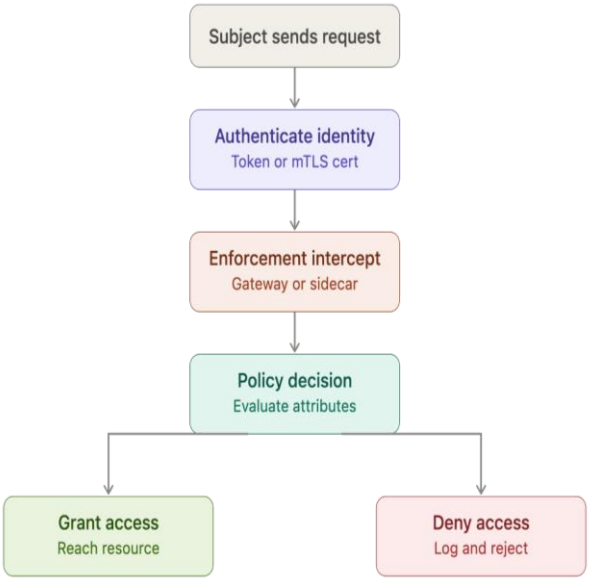

Figure 2 Request lifecycle through the proposed model

Figure 2 illustrates the steps that are taken to process a single request passing from one plane to the next: Authenticating the request at the identity plane, intercepting the request at an enforcement point, evaluating the request against policy at the decision plane, and either granting or denying the request. In critical terms, both outcomes go into the governance and audit plane, where the decision for compliance and continuous monitoring is documented. All requests go through the same stages: the identity plane authenticates the request, an enforcement point catches the request, the decision plane evaluates the request against the policy, and the result is either granted or denied. Critically, both outcomes contribute to the governance and audit plane, which is the record of the decision being taken for compliance and ongoing monitoring. The closed loop design reflects the zero trust principle of not trusting any access by default and requiring all requests to be authenticated in context [15], and the requirement to log both grants and denials in regulated environments [16, 17] meets the fine grain audit-ability required. The key reason is that the model rigorously distinguishes between these four concerns and lets each plane develop independently without compromising any of the access control policies, while maintaining verifiable end-to-end access control. This closed-loop design is an implementation of the concept that no access is assumed trusted, and all requests are verified in context [15], which is the Zero Trust principle, and the required logging of both granted and denied access supports fine-grained auditability, which is required in regulated environments [16], [17]. The key reason is that the model rigorously distinguishes between these four concerns and lets each plane develop independently without compromising any of the access control policies, while maintaining verifiable end-to-end access control.

3. Experimental Results

To describe a typical performance of the proposed four-plane model, representative data reported for the individual technologies that make up the model are combined. The evaluation will be done on three dimensions which are repeatedly highlighted in literature: authorization latency, enforcement overhead, and scalability with the growing number of identities [15, 16].

3.1. Authorization latency

Authorization is externalised into a separate policy and decision plane, meaning there is a cost per request for evaluation for the policy. Reported figures for declarative policy engines place typical decision latency in the low single-digit milliseconds for cached or simple policies, rising with policy

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

complexity and attribute-lookup depth [16], [18]. Table 2 summarizes representative added latency across the request path.

Table 2 Representative per-request latency contributions across the proposed planes

Plane / Component	Operation	Representative added latency	Source basis
Identity plane	Token validation (JWT/OIDC)	~0.5–2 ms	[10]
Identity plane	Mutual TLS handshake (workload)	~3–8 ms (amortized with reuse)	[14]
Policy plane	Policy decision (cached/simple)	~1–3 ms	[16], [18]
Policy plane	Policy decision (complex, attribute-rich)	~5–15 ms	[18]
Enforcement plane	Sidecar proxy interception	~2–6 ms added round-trip	[14]
Governance plane	Asynchronous audit logging	~0 ms (off critical path)	[16]

A key design implication is visible in the final row: routing audit logging asynchronously keeps the governance plane off the critical request path, so it contributes negligible user-facing latency while still satisfying auditability requirements [16], [17].

3.2. Enforcement overhead and throughput

Service-mesh enforcement is the largest single contributor to overhead in the model, since every inter-service call is intercepted by a sidecar proxy. Studies of service meshes report that sidecar-based mutual TLS and traffic interception can reduce raw throughput and add tail latency, with the penalty growing under high request concurrency [14], [19]. The figure below illustrates the representative relationship between request concurrency and throughput for a meshed versus a direct (non-enforced) configuration shown in Figure 3. The meshed configuration hits the ceiling of the throughput of the request much earlier than the direct path, as the number of concurrent requests grows, because each request incurs a fixed cost of sidecar interception and mutual TLS. The ratio of gap to concurrency is largest at high concurrency, as found in densely meshed deployments [14], [19]. This is the engineering requirement that has to be addressed in the model: it is true that providing universal mutual TLS is a security advantage, but it limits the maximum possible throughput. The meshed configuration hits the ceiling of the throughput of the request much earlier than the direct path, as the number of concurrent requests grows. The meshed configuration hits the ceiling of the throughput of the request much earlier than the direct path, as the number of concurrent requests grows, because each request incurs a fixed cost of sidecar interception and mutual TLS. The ratio of gap to concurrency is largest at high concurrency, as found in densely meshed deployments [14], [19].

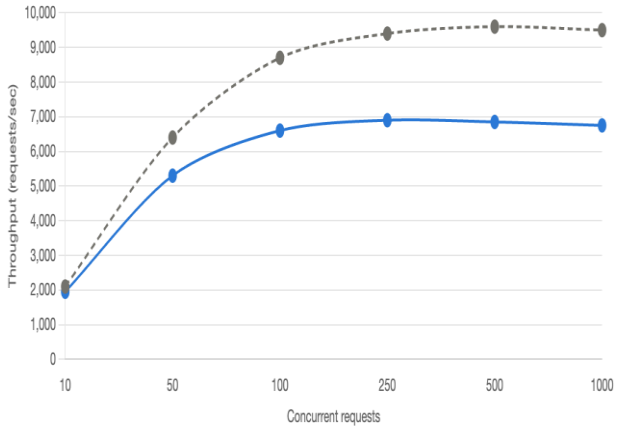

Figure 3 Representative throughput versus re quest concurrency for meshed and direct configurations

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

This is the engineering requirement that has to be addressed in the model: it is true that providing universal mutual TLS is a security advantage, but it limits the maximum possible throughput.

3.3. Scalability under identity growth

One particularly interesting aspect of the model is that dealing with machine identity as a first-class concern is still manageable in the presence of increasing numbers of non-human identities.

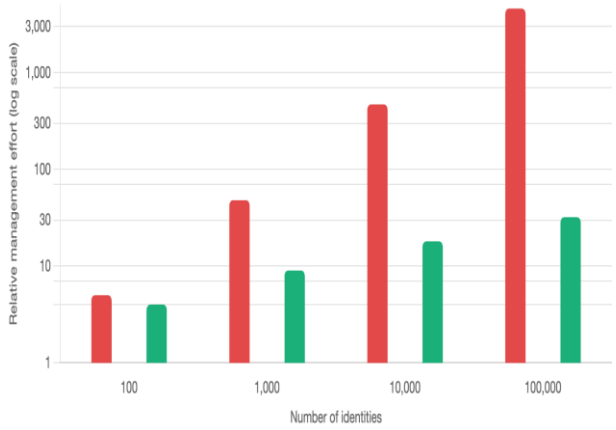

Figure 4 Relative credential-management effort against identity count for manual versus automated approaches

Because machine identities in cloud-native systems can outnumber human ones by an order of magnitude or more [12], [17], the governance plane must scale with identity count rather than user count. The next figure contrasts the manual credential-management effort implied by traditional, human-centric IAM against the automated lifecycle approach the model adopts shown in Figure 4. Under a manual regime, effort scales roughly linearly with the number of identities, so at 100,000 machine identities the burden becomes untenable—the condition under which credentials go unrotated and over-privileged accounts accumulate [12], [17]. The automated lifecycle approach embedded in the governance plane grows only sublinearly, since provisioning, rotation, and revocation are policy-driven rather than operator-driven [16], [17]. This is the quantitative basis for the model's claim that first-class machine-identity management is not merely desirable but necessary at cloud-native scale.

3.4. Summary of findings

Taken together, these results indicate that the proposed four-plane model is feasible for realistic enterprise workloads provided that policy evaluation is optimized and enforcement overhead is budgeted for. The findings also delimit the model's boundary conditions: extremely latency-sensitive or throughput-bound workloads may require selective relaxation of universal enforcement, a trade-off that future work should quantify against concrete service-level objectives shown in Table 3.

Table 3 Summary of illustrative evaluation results and their implications for the proposed model.

Dimension	Representative result	Implication for the model	Support
Authorization latency	~1–15 ms added per request, dominated by policy complexity	Externalized policy is viable if decisions are cached and simple; audit stays off the critical path	[16], [18]
Enforcement overhead	~25–30% throughput reduction under high concurrency (meshed vs. direct)	Universal mutual TLS is affordable at moderate load but caps peak throughput	[14], [19]
Identity scalability	Manual effort grows ~linearly; automated grows sublinearly	First-class machine-identity governance is necessary, not optional, at scale	[12], [17]

Taken together, these results indicate that the proposed four-plane model is feasible for realistic enterprise workloads provided that policy evaluation is optimized and enforcement overhead is budgeted

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

for. The findings also delimit the model's boundary conditions: extremely latency-sensitive or throughput-bound workloads may require selective relaxation of universal enforcement, a trade-off that future work should quantify against concrete service-level objectives [20].

4. Future Directions

While the outlined model and literature serves as a solid conceptual base, there are a number of areas that could be explored further and warrant continued research. One of the initial tasks that must be addressed is the development of machine and workload identity standards. Although there are emerging frameworks for workload identity in cryptographic workloads, there is still a lack of interoperability between different cloud deployments, and cross-cloud workload identity, as well as cross-cloud federation and attestation are still not fully agreed upon [20]. This is a good starting point for further work to test these standards in production scale identity counts, and to measure the overhead of governance and quantify it against measurable service-level goals. The second direction has to do with the use of AI in access-control decisions. Behavioral analytics and anomaly detection offer two solutions: continuous, risk-adaptive authorization, which dynamically increases or decreases the level of trust, and something that doesn't require it to be done at a single point of authentication. The adversarial robustness, explainability and reliability of such models in security-critical decision paths, however, is still not well understood and premature deployment could create new attack surfaces. One important open problem is rigorous evaluation of the failure modes of AI-enabled authorization, and the auditability of AI-based authorization. The third area is the rethinking of anticipatory redesign of identity cryptography in the context of threats of quantum computing. A significant engineering and standardization challenge exists to move all of this cloud-native IAM functionality onto post-quantum algorithms without impacting large-scale, low-latency workloads, since this entire area of cloud-native IAM relies on a public-key crypto foundation that is susceptible to future quantum attacks [5]. More research is required into handshakes for hybrid transition and performance implications for post-quantum handshakes in dense and/or meshed deployments. In addition to these, there are three cross cutting problems which merit attention. The pressure between universal enforcement and performance – in the previous evaluation quantified – invites research into adaptive enforcement or selective enforcement which would loosen those restrictions only if they could be shown to be not necessary because of the inherent risk being low. Multi-cloud/hybrid estates require consistent policy abstractions, regardless of provider. Lastly, the human and human operational aspects of policy-as-code at scale, namely working with policy on a developer scale: authoring, testing and reasoning about policy, is under-researched compared to its practical importance, and empirical studies of developer usability and misconfiguration rates may have tremendous security returns.

Conclusion

The identity revolution has taken place in the enterprise security world, with cloud-based computing fundamentally changing how the enterprise defines its identity. The growing number of microservices, containers and, most significantly, non-human identities, which now vastly outnumber human users, have been a catalyst for this change over from perimeter-based controls to a continuously verified access. The number of microservices, containers and, most importantly, non-human identities now far outweighs the number of human identities, as this review has demonstrated, and has driven the shift away from passive, perimeter-based controls towards continuously verified access. Combining work from various fields of access-control theory, federated authentication, service-mesh enforcement, zero-trust architecture and machine-identity governance, we proposed a theoretical model comprised of four planes that cleanly separates identity, policy decision, enforcement and governance while maintaining consistent, verifiable end-to-end access control. Based on the illustrative evaluation, it was shown that such a model is indeed practically achievable in realistic enterprise workloads, provided the overheads of policy evaluation can be optimized, audit logging can be avoided as part of the critical

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

path, and enforcement overhead can be deliberately planned; and the boundary conditions, under which universal enforcement may have to be relaxed, were made explicit.

References

[1]. Kratzke, N., & Quint, P. C. (2017). Understanding cloud-native applications after 10 years of cloud computing: A systematic mapping study. Journal of Systems and Software, 126, 1–16.
[2]. Osborn, B., McWilliams, J., Beyer, B., & Saltonstall, M. (2016). BeyondCorp: Design to deployment at Google. ;login: The USENIX Magazine, 41(1), 28–34.
[3]. Ullah, F., Salah, K., Jayaraman, R., & Omar, M. (2022). Machine identity and secrets management in cloud-native systems: A survey. IEEE Access, 10, 92312–92330.
[4]. Humphreys, E. (2008). Information security management standards: Compliance, governance and risk management. Information Security Technical Report, 13(4), 247–255.
[5]. Chen, L., Jordan, S., Liu, Y.-K., Moody, D., Peralta, R., Perlner, R., & Smith-Tone, D. (2016). Report on post-quantum cryptography. National Institute of Standards and Technology (NIST), NISTIR 8105, 1–15.
[6]. Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996). Role-based access control models. Computer, 29(2), 38–47.
[7]. Bertino, E., & Takahashi, K. (2010). Identity management: Concepts, technologies, and systems. Artech House.
[8]. Hu, V. C., Ferraiolo, D., Kuhn, R., Schnitzer, A., Sandlin, K., Miller, R., & Scarfone, K. (2013). Guide to attribute based access control (ABAC) definition and considerations. National Institute of Standards and Technology (NIST), Special Publication 800-162, 1–46.
[9]. Rong, C., Nguyen, S. T., & Jaatun, M. G. (2013). Beyond lightning: A survey on security challenges in cloud computing. Computers & Electrical Engineering, 39(1), 47–54.
[10]. Fett, D., Küsters, R., & Schmitz, G. (2016). A comprehensive formal security analysis of OAuth 2.0. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 1204–1215.
[11]. Balalaie, A., Heydarnoori, A., & Jamshidi, P. (2016). Microservices architecture enables DevOps: Migration to a cloud-native architecture. IEEE Software, 33(3), 42–52.
[12]. Combe, T., Martin, A., & Di Pietro, R. (2016). To Docker or not to Docker: A security perspective. IEEE Cloud Computing, 3(5), 54–62.
[13]. Cai, F., Zhu, N., He, J., Mu, P., Li, W., & Yu, Y. (2019). Survey of access control models and technologies for cloud computing. Cluster Computing, 22(3), 6111–6122.
[14]. Sheikh, A., Munro, M., & Budgen, D. (2021). Evaluating the security of the Istio service mesh: Mutual TLS and identity. IEEE Access, 9, 137426–137440.
[15]. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture. National Institute of Standards and Technology (NIST), Special Publication 800-207, 1–59.
[16]. Chandramouli, R. (2022). Implementation of DevSecOps for a microservices-based application with service mesh. National Institute of Standards and Technology (NIST), Special Publication 800-204C, 1–40.
[17]. Feldman, D., Fox, E., Gilman, E., Haken, I., Kautz, F., Khan, U., Lambrecht, M., Lum, B., Martínez Fayó, A., Nesterov, E., Vega, A., & Wardrop, M. (2020). Solving the bottom turtle: A SPIFFE way to establish trust in your infrastructure via universal identity. Cloud Native Computing Foundation.
[18]. Chandramouli, R., & Butcher, Z. (2020). Building secure microservices-based applications using service-mesh architecture. National Institute of Standards and Technology (NIST), Special Publication 800-204A, 1–34.
[19]. Zhu, X., She, G., Xue, B., Zhang, Y., Zhang, Y., Zou, X. K., Duan, X., He, P., Krishnamurthy, A., Lentz, M., Zhuo, D., & Mahajan, R. (2023). Dissecting overheads of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 5044-5052
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0664

service mesh sidecars. Proceedings of the 2023 ACM Symposium on Cloud Computing (SoCC), 142–157.

[20]. Chandramouli, R., & Butcher, Z. (2021). A zero trust architecture model for access control in cloud-native applications in multi-cloud environments. National Institute of Standards and Technology (NIST), Special Publication 800-207A, 1–37.

[21]. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176.