International Journal of Sustainable Development Through AI, ML and IoT

Volume 4 | Issue2 | 2025
https://ijsdai.com/index.php/IJSDAI/index
ISSN (Online): 2584-0827

Cloud Computing: Architectures, Service Models, Security, and Emerging Trends

Anupriya Jain [1]
MRIIRS, India
Corresponding author: ***Dr. Anupriya jain***

ARTICLE INFO

Received: 31 May 2025
Revised: 20 Nov 2025
Accepted: 31 Dec 2025

ABSTRACT

Cloud computing has become the default infrastructure for modern software delivery, with the large majority of new digital workloads now built on cloud-native platforms rather than on-premises hardware. This paper surveys the current state of cloud computing: its foundational service models (Infrastructure-, Platform-, and Software-as-a-Service), deployment models (public, private, hybrid, and multi-cloud), and the architectural shift toward containerization and serverless computing. It reviews the security landscape, where misconfiguration and identity-related failures, not exotic exploits, remain the dominant cause of breaches, and surveys the shift toward zero-trust architectures and cloud-native application protection platforms. It further examines the growing entanglement of cloud computing with artificial intelligence, both as a workload that is reshaping infrastructure demand and as a tool used for security and operations, and reviews the environmental costs of cloud growth alongside emerging sustainability practices such as carbon-aware scheduling and energy-efficient serverless design. The paper closes with a discussion of open challenges, including multi-cloud complexity, the widening skills gap, and the tension between rapid AI-driven infrastructure growth and sustainability commitments.

Keywords: *autonomous agents; agent evaluation benchmarks; AI safety and governance*.

1. Introduction

Cloud computing, the on-demand delivery of computing power, storage, and applications over the internet on a pay-as-you-go basis, has moved from a cost-saving alternative to on-premises infrastructure to the default substrate of modern computing. Global public cloud end-user spending has been forecast to exceed $700 billion in a single year, and industry analysts project that the overwhelming majority of new digital workloads will be built on cloud-native platforms rather than deployed to on-site servers, a dramatic shift from a decade earlier when cloud-native development was still a minority practice.

This growth has been driven by several converging forces: the maturation of virtualization and containerization technology, the rise of serverless architectures that abstract away infrastructure management entirely, and, most recently, the enormous compute demands of generative AI, which has made GPU-as-a-Service one of the fastest-growing categories in the industry. At the same time, this growth has amplified long-standing challenges around security, cost management, and environmental impact. This paper surveys the technical foundations of cloud computing, its evolving architecture, its security landscape, and the sustainability and governance questions that are shaping the field going into the second half of the 2020s.

2. Service Models: IaaS, PaaS, and SaaS

Cloud offerings are conventionally organized into three service models that differ in how much of the technology stack the customer manages versus the provider. Infrastructure-as-a-Service (IaaS) provides raw compute, storage, and networking resources, leaving the customer responsible for the operating system, middleware, runtime, and application layers; Platform-as-a-Service (PaaS) adds a managed development and runtime environment on top of that infrastructure, letting developers focus on application code rather than the underlying servers; and Software-as-a-Service (SaaS) delivers a complete, ready-to-use application, with the provider managing everything below the user's own data and configuration.

SaaS remains the largest segment by revenue, with 2025 projections placing it around $390 billion, ahead of PaaS at roughly $209 billion and IaaS at roughly $180 billion, though IaaS has recently been described as the fastest-growing segment of the three, propelled largely by AI training and inference compute demand. Table 1 summarizes the core distinctions between the three models, alongside a fourth category, Containers-as-a-Service (CaaS), that has grown in relevance as organizations modernize legacy applications through containerization.

Table 1. Comparison of core cloud service models.

Attribute	**IaaS**	**PaaS**	**SaaS**
What is provided	Raw compute, storage, and networking	Managed development and runtime environment	Complete, ready-to-use application
Customer manages	OS, middleware, runtime, applications, data	Applications and data only	Configuration and user data only
Typical users	Infrastructure and DevOps teams	Application developers	End users and business teams
Examples	AWS EC2, Azure VMs, Google Compute Engine	Azure AI Foundry, Google Vertex AI, Amazon Bedrock, Heroku	Salesforce, Microsoft 365, HubSpot, Dropbox
2025 market share signal	Fastest-growing segment, driven by AI training and inference demand	Roughly $209B in projected 2025 revenue	Largest segment by revenue, roughly $390B projected for 2025

In practice, most enterprises operate across all three models simultaneously: legacy systems on IaaS, new application development on PaaS, and standard business functions delivered through SaaS. This has produced a further challenge of its own, SaaS sprawl, where large enterprises now manage several hundred SaaS applications on average, often with limited centralized visibility into which tools are actually in use across the organization.

3. Deployment Models and Architectural Trends

3.1 Public, Private, Hybrid, and Multi-Cloud

Cloud deployments are also classified by ownership and location: public clouds are operated by third-party providers and shared across customers; private clouds are dedicated to a single organization, whether hosted on-premises or by a provider; and hybrid clouds combine the two, allowing organizations to burst workloads into public cloud capacity without fully migrating sensitive systems. The large majority of enterprises now report prioritizing hybrid cloud architectures, and a comparable share describe multi-cloud strategies, using more than one public cloud provider, as core to their operational resilience and to avoiding vendor lock-in.

3.2 Containerization and Serverless Computing

Two architectural shifts have reshaped how applications are built and deployed on cloud infrastructure. Containerization, most commonly implemented through Docker and orchestrated with Kubernetes, packages an application with its dependencies into a portable unit that runs consistently across environments, which has made it a foundational technology for hybrid and multi-cloud strategies. Serverless computing, and specifically the Function-as-a-

Service (FaaS) model exemplified by AWS Lambda, goes a step further by eliminating server management entirely: developers deploy individual functions that the provider automatically scales and bills only for actual execution time, removing idle-resource costs altogether.

Serverless adoption has been reinforced by its natural fit with AI and IoT workloads that involve bursty, event-driven processing, such as real-time image analysis or sensor data ingestion at the edge. However, serverless computing carries its own trade-offs, including cold-start latency when a function is invoked after being idle, and, as discussed in Section 6, a more complex and sometimes counterintuitive energy profile than its pay-per-use billing model might suggest.

4. Security Challenges

Security remains the most consistently cited concern in cloud adoption research, and the data paints a fairly consistent picture across independent surveys: the dominant cause of cloud breaches is not sophisticated attack techniques but basic misconfiguration and identity failures. Industry analysis attributes the substantial majority of cloud security failures to the customer's side of the shared-responsibility model, primarily through identity and configuration gaps, and a comparable share of reported cloud data breaches has been linked to human error such as credential misuse or overly permissive access settings.

Multi-cloud adoption has compounded this problem. With most enterprises now running workloads across two or more cloud providers, security teams must maintain consistent identity governance, logging, and policy enforcement across platforms that were not designed to be managed as a single system, a challenge often described as tool sprawl and identity sprawl. Infrastructure-as-code practices, while valuable for deployment speed, have introduced a further wrinkle: misconfigurations are no longer isolated human errors but can be encoded into templates and replicated automatically across every environment that inherits them.

4.1 Zero Trust and Cloud-Native Protection

In response, organizations are converging on zero-trust architecture as the dominant security model for cloud environments, replacing the assumption of a trusted internal network with continuous verification of every access request regardless of origin. A substantial share of enterprises report having implemented or piloted zero-trust principles for cloud workloads, alongside growing adoption of Cloud-Native Application Protection Platforms (CNAPP) that consolidate posture management, workload protection, and identity entitlement management into a single, unified tool rather than a patchwork of point solutions. DevSecOps practices, embedding security scanning directly into CI/CD pipelines so vulnerabilities are caught during development rather than after deployment, have also become a standard component of this shift.

Artificial intelligence plays a role on both sides of this contest. Security teams increasingly rely on AI-driven anomaly detection to learn normal identity and workload behavior and flag deviations across cloud environments, meaningfully reducing average detection and response times in mature programs. At the same time, attackers are using AI to automate credential theft and to map permission paths more quickly than manual reconnaissance would allow, meaning the AI-driven security arms race is itself becoming a defining feature of the cloud threat landscape.

5. The Convergence of Cloud Computing and Artificial Intelligence

Generative AI has become deeply entangled with cloud infrastructure in two directions simultaneously. First, AI is now one of the largest and fastest-growing categories of cloud spending, with generative AI usage reported among the majority of surveyed organizations and analyst projections placing generative AI's share of cloud spending in the hundreds of billions of dollars by the end of the decade. GPU-as-a-Service offerings have grown especially quickly as hyperscalers race to meet training and inference demand, and managed AI platforms, such as Azure AI Foundry, Google Vertex AI, and Amazon Bedrock, have effectively become a new sub-layer of the PaaS model.

Second, AI is increasingly embedded within the cloud services themselves, from AI-assisted resource provisioning and automated scaling in IaaS offerings to AI features built directly into SaaS products such as productivity suites and CRM platforms. This dual role, AI as both a major cloud workload and a feature embedded across the cloud stack, is reshaping cost structures: AI and machine learning workloads now represent a meaningful share of cloud costs at SaaS and technology companies, and these costs have proven harder to forecast than traditional infrastructure spending, contributing to rising cloud waste even as organizations improve their financial operations (FinOps) practices.

6. Sustainability and Environmental Impact

Cloud computing's rapid growth has made data center energy consumption and carbon emissions a significant and growing environmental concern, and cloud providers have responded with a mix of infrastructure and software-level interventions. At the infrastructure level, hyperscalers have invested in renewable-powered data centers, advanced liquid and underwater cooling designs, and AI-driven cooling optimization to reduce the power usage effectiveness of their facilities.

At the software level, serverless computing is frequently promoted as inherently more sustainable, since its pay-per-use model eliminates idle server capacity, and some empirical studies report substantial energy and cost reductions when workloads are moved to serverless platforms. This picture, however, is contested: other measurement studies have found that serverless execution can be considerably more energy-intensive per unit of useful work than native execution, largely due to the overhead of containerization, cold starts, and platform abstraction layers, indicating that the sustainability benefit depends heavily on workload type and is not a settled question in the literature. Emerging research directions include carbon-aware function scheduling, which shifts workload execution to times or regions with cleaner electricity grids, and machine-learning-based resource modeling intended to close the gap between the theoretical and observed energy efficiency of serverless systems.

7. Open Challenges

7.1 Multi-Cloud Complexity and Vendor Lock-In

While multi-cloud strategies are widely adopted for resilience and negotiating leverage, they introduce real operational cost: inconsistent APIs, differing security models, and duplicated tooling across providers make unified visibility and governance difficult to achieve in practice, and organizations that have not invested in cross-cloud identity governance report the most severe security gaps.

7.2 Cost Management and Cloud Waste

Despite improvements in financial operations tooling, a meaningful share of cloud spending continues to be classified as waste by the organizations themselves, a problem compounded by the harder-to-forecast, bursty nature of AI workload costs relative to traditional, steady-state infrastructure spending.

7.3 The Sustainability-Growth Tension

The infrastructure buildout required to meet AI compute demand sits in direct tension with sustainability commitments, since GPU-intensive training and inference workloads are energy-dense even as providers pursue carbon-neutral operations, meaning the net environmental trajectory of the industry depends on whether efficiency gains can outpace the sheer growth in AI-driven demand.

7.4 Skills and Governance Gaps

Finally, the pace of cloud innovation continues to outstrip organizational capacity to secure and govern it; new features and services expand the attack surface faster than security teams can develop corresponding expertise and controls, a gap that industry researchers consistently point to as a root cause behind the persistence of basic, preventable misconfiguration failures even as security tooling matures.

8. Conclusion

Cloud computing has completed its transition from an alternative to on-premises infrastructure to the default environment for new software development, a shift now being accelerated further by the compute demands of generative AI. Its service and deployment models have matured into a well-understood taxonomy, but the operational challenges around security, cost, and sustainability have not been solved so much as displaced onto increasingly complex multi-cloud environments. The evidence reviewed here suggests that the industry's most persistent problems, misconfiguration-driven breaches, unpredictable AI-driven costs, and an uncertain sustainability trajectory, are not primarily technology gaps but governance and visibility gaps. Closing them is likely to depend less on any single new technology than on organizations' ability to operationalize consistent identity, security, and cost governance across increasingly distributed and heterogeneous cloud environments.

References

Akour, M., & Alenezi, M. (2025). Reducing environmental impact with sustainable serverless computing. Sustainability, 17(7), 2999. https://doi.org/10.3390/su17072999

Axis Intelligence. (2026). Cloud computing statistics 2026: Market, adoption & spending. Retrieved from https://axis-intelligence.com/cloud-computing-statistics/

CloudZero. (2026). 100+ cloud computing statistics: A 2026 market snapshot. Retrieved from https://www.cloudzero.com/blog/cloud-computing-statistics/

Cyble. (2026). Top cloud security challenges in 2025 & insights for 2026. Retrieved from https://cyble.com/knowledge-hub/top-cloud-security-challenges/

DataStackHub. (2025). 50 cloud data security statistics for 2025-2026. Retrieved from https://www.datastackhub.com/insights/cloud-data-security-statistics/

DataStackHub. (2026). 50 cloud security statistics for 2025-2026. Retrieved from https://www.datastackhub.com/insights/cloud-security-statistics/

Faddom. (2026). 42 cloud computing statistics you must know in 2026. Retrieved from https://faddom.com/29-cloud-computing-statistics-you-must-know-in-2025/

Finout. (2026). 49 cloud computing statistics you need to know in 2026. Retrieved from https://www.finout.io/blog/49-cloud-computing-statistics-in-2026

IBM. (2026). IaaS, PaaS, SaaS: What's the difference? Retrieved from https://www.ibm.com/think/topics/iaas-paas-saas

IEEE. (2026). A comprehensive analysis of cloud service models: IaaS, PaaS, and SaaS in the context of emerging technologies and trend. IEEE Conference Publication. https://ieeexplore.ieee.org/document/10823401/

Jack Henry. (2026). Top cybersecurity trends for 2026 every financial leader must know. Retrieved from https://www.jackhenry.com/fintalk/top-cybersecurity-trends-for-2026-every-financial-leader-must-know

Java Code Geeks. (2025). Green cloud computing practices 2025: Measuring & reducing energy consumption. Retrieved from https://www.javacodegeeks.com/2025/04/green-cloud-computing-practices-2025-measuring-reducing-energy-consumption.html

Lazreg, Y., Sadoghi, M., & Chen, L. (2025). Sustainable serverless computing: A taxonomy of challenges and solutions. Proceedings of the 11th International Workshop on Serverless Computing. https://doi.org/10.1145/3774899.3775019

Mitiga. (2026). IaaS vs PaaS vs SaaS vs CaaS: Cloud models compared. Retrieved from https://www.mitiga.io/blog/iaas-vs-paas-vs-saas-what-are-the-differences

PIT Journal. (2026). Sustainable scalability: Green cloud computing and the future of digital growth. University of North Carolina. Retrieved from https://pitjournal.unc.edu/2026/05/16/sustainable-scalability-green-cloud-computing-and-the-future-of-digital-growth/

Perception System. (2025). Serverless: The future of cloud computing in 2025. Retrieved from https://www.perceptionsystem.com/blog/serverless-future-cloud-computing-2025/

Sennovate. (2026). Top multi-cloud security risks to know in 2026. Retrieved from https://sennovate.com/blog/multi-cloud-security-challenges-in-2026/

SentinelOne. (2026). 50+ cloud security statistics in 2026. Retrieved from https://www.sentinelone.com/cybersecurity-101/cloud-security/cloud-security-statistics/

Softjourn. (2026). 100+ cloud computing statistics for 2026: Complete report. Retrieved from https://softjourn.com/insights/cloud-computing-stats

TierPoint. (2026). Top cloud security trends in 2026: Key strategies & risks. Retrieved from https://www.tierpoint.com/blog/cloud/cloud-security-trends/

Wang, J., Gupta, U., & Sriraman, A. (2024). The hidden carbon footprint of serverless computing. Proceedings of the 2024 ACM Symposium on Cloud Computing. https://doi.org/10.1145/3698038.3698546

Wickramarachchi, T. (2026). Cloud security challenges in a multi-cloud world (2026). Medium. Retrieved from https://medium.com/@thinuridulsini/cloud-security-challenges-in-a-multi-cloud-world-2026-41de572dec44