ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 5 Issue 3 July 2025 / Page No: 126-134
Paper Id: JETA-V5I3P117 / Doi: 10.56472/25832646/JETA-V5I3P117

Original Article

AI-Driven Cloud Infrastructure: Advances in Kubernetes and Serverless Computing

Mahi Ratan Reddy Deva
Independent Researcher

Received Date: 16 May 2025 Revised Date: 05 July 2025 Accepted Date: 06 August 2025

Abstract: *Artificial Intelligence (AI) has been integrated into cloud infrastructure, making it revolutionizing modern computing by automating, scaling, and improving efficiency. The first of these is Kubernetes, and the second is serverless computing. Kubernetes, a container orchestration platform, benefits from AI-driven enhancements in workload scheduling, auto-scaling, and resource optimization. By combining AI-based predictive analytics with container deployment, overhead is reduced in terms of operational overhead as well as fault tolerance. However, serverless computing takes away the management of infrastructure, leaving the developers free to write application logic. Serverless architectures with AI-based power can scale and adaptively allocate resources and execute cloud workloads with minimum cost. This review study examines the latest development of AI-based Kubernetes and serverless computing, their influence on the cloud infrastructure. AI is discussed insofar as its orchestration role can be optimized, security is improved, and self-healing cloud environments are made possible. The paper also studies challenges: latency, security issues, and uncertainties of the integration of AI models. Through the analysis of state-of-the-art innovation and future trends, this review covers how AI is impacting the future cloud native computing.*

Keywords: *AI-Driven, Cloud Infrastructure, Kubernetes, Serverless Computing, Cloud-Native, Orchestration.*

I. INTRODUCTION

Cloud computing is a form of provision of on-demand computing resources to users of any kind over the internet that provides the users with the ability to use and manage resources in the cloud. It has changed how organizations that deploy, scale, and manage their IT infrastructure do business [1]. In this regard, cloud services offer businesses extreme flexibility, scalability and cost efficiency; businesses can instead concentrate on their primary skills, and Cloud service providers handle the complexities of IT infrastructure management.

The businesses are driven to move forward by organizations that want their businesses to advance, and this is happening in a distributed way, using a distributed computing architecture [2]. This shift is also changing CIO and CTO roles. Although many still talk about the trend of digital transformation, and AI, ML, and extended reality (XR) are popular, cloud transformation is relevant, even if it is not the focus of mainstream media. Cloud computing continues to be a hot discussion point behind the scenes, and organizations are sensibly understanding that cloud adoption is not an option but a necessity.

The rise of Kubernetes is a major leap forward in cloud computing as a legal technology for serverless platforms. Open source serverless computing frameworks are well supported by various Kubernetes and are used to manage containerized applications. Kubernetes has been proven to be able to maintain performance levels similar to bare metal in old high-performance computing (HPC) workloads [3].

This paradigm, which is relatively new, allows developers to create and launch apps without having to install underlying infrastructure in conventional cloud river spaces [4][5]. Serverless computing reduces the efforts of programmers in maintaining the system by automating almost all system administration tasks. Initially, Amazon EC2 attracted developers by having full control of application instances to the point that developers had to manage and scale application instances heavily. Serverless computing solves this issue by removing the infrastructure complexities and using it to deploy the application more efficiently.

Edge computing is another emerging trend that has been growing steadily in the last few years. The cloud infrastructure is extended by the edge computing paradigm, which puts computer resources at a more convenient location for the user. They act as general-purpose servers for specific types of IoT devices that can collect and analyse data, such as cameras and sensors [6][7][8]. This is due to the fact that it reduces latency and processes the data near the production nodes, as well as ensuring improved data security. Optimizing serverless computing parameters is essential for achieving high performance in edge environments.

As AI-driven cloud infrastructure advances, Kubernetes and the development of cloud-native apps are significantly influenced by serverless computing [9]. Understanding their impact on performance, scalability, and resource optimization is crucial for organizations looking to leverage these technologies effectively.

A. Structured of the paper

The structure of this paper is as follows: Section II, Role of AI in Cloud Infrastructure. Section III AI-Driven Kubernetes for Cloud Orchestration. Section IV Serverless Computing in A-Driven Cloud Infrastructure, Section V Benefits of AI-Driven Cloud Solutions, Section VI reviews literature and case studies, and Section VII finishes with suggestions for further work.

II. ROLE OF AI IN CLOUD INFRASTRUCTURE

Cloud-based infrastructures are seen to be the most appropriate for meeting resource and service demands [10]. However, there are several obstacles to transferring the enormous amounts of data produced to the cloud, including Fortunately, edge computing—which puts computers closer to the data source—has emerged as a workable idea that addresses issues with excessive latency, network congestion, and privacy, smart application quality of service, and user experience.

A new paradigm for intelligence computing is created when cloud/edge computing and AI are completely integrated to realize their potential benefits. However, there are still a lot of unresolved problems with its implementation, including severe security concerns and limited computing, networks, and energy resources. However, things are further complicated by the transient nature of cloud/edge systems [11][12]. These strategies can result in machines that can handle complex issues by, among other things, learning from data and making judgements based on discernment. Thus, enhanced computational intelligence shows a lot of potential and a lot of opportunities for use in cloud/edge computing. It may be used as a problem-solver to overcome the challenges during system design as well as an enabler to strengthen service capabilities. In order to direct the development of next-generation network technology, this Special Issue aims to compile academic research that investigates the possibilities of combining cloud/edge computing with sophisticated AI [13].

There are many different uses for cloud computing's adaptability. It enables remote patient monitoring and telemedicine in the healthcare industry, where real-time data informs crucial medical choices. By enabling traffic optimization systems, it lowers traffic and enhances urban mobility in smart cities. Cloud systems in industrial environments serve as centers for predictive maintenance, increasing productivity and reducing downtime [14][15]. Still, there are difficulties. There are persistent challenges due to high bandwidth prices, latency problems, and the intricacies of handling data sovereignty regulations. To overcome these obstacles, solutions including hybrid architectures that integrate cloud and edge computing and data centers powered by renewable energy sources are being developed. As IoT and AI ecosystems develop, cloud computing is a key factor that is opening the door for intelligent systems that are effective, safe, and long-lasting. Its full potential may be achieved by utilizing its advantages and removing current obstacles.

III. AI-DRIVEN KUBERNETES FOR CLOUD ORCHESTRATION

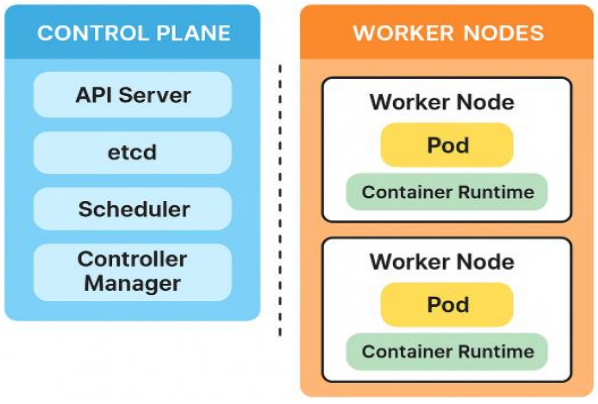

Figure 1 : Kubernetes Cluster Architecture

The deployment, scaling, and management of containerized applications may be automated with the help of an open-source tool called Kubernetes. It frees developers from worrying about the underlying infrastructure so they can concentrate on creating and implementing their applications. Users define desired application states using Kubernetes' declarative application management technique, and the system keeps track of them as shown in Figure 1. It also provides robust application administration and monitoring capabilities, including automatic failure recovery and detection using self-healing techniques. In summary, Kubernetes offers a robust and flexible approach to containerized application management in production environments [16].

A. Kubernetes and Its Role in Cloud Computing

The open-source Kubernetes framework, sometimes referred to as K8s, was created to make containerized application deployment, scalability, and maintenance easier. It provides a scalable and portable platform that enables efficient application deployment and maintenance throughout a computer cluster. A pod is the basic deployment unit for container-based virtualization, particularly when it comes to the Kubernetes environment. A pod is a logical collection of one or more containers that share resources and network namespaces and are collocated together [17].

B. AI-Enhanced Workload Orchestration

The cloud computing paradigm was formed and developed via the introduction of pay-per-use and on-demand resource allocation methods brought about by the widespread use of virtualization. The difficulty of planning and implementing an effective coordinated execution of many virtual machines to optimize expenses for the owner has arisen due to the economic implications of resource utilisation. To use the elasticity offered by cloud computing, a multitude of orchestration solutions are available for centrally managing applications with specific auto-scaling and QoS needs [18]. These orchestration platforms struggle to satisfy the stringent QoS standards set out by the service owners, notwithstanding their limited understanding of the underlying workings of the behavior and characteristics of the service.

This paradigm's relatively homogeneous supporting hardware, which is typically found in data centers, is one of its characteristics. In most cases, every machine in a cloud deployment has the same settings. For example, all of the machines have the same ad hoc operating systems, which may have been modified by vendors to improve global performance based on the kind of service offered. Cloud computing also frequently guarantees high network bandwidth and dependability. However, there has been a noticeable shift in recent years towards edge computing, which favors computation in more localized settings.

C. Emerging Trends in Kubernetes

Kubernetes develops further through multiple new trends that improve its features. It can enhance Kubernetes through Operators that automate stateful application handling and turn operational human knowledge into functional software.

The following are the Emerging Trends in Kubernetes:

- Kubernetes Operators: Automates stateful application management by embedding operational knowledge into software [19].
- Multi-Cloud Kubernetes: Enables deployment across several cloud providers to improve dependability and prevent vendor lock-in [20].
- AI-Driven Orchestration: Uses machine learning for intelligent workload scheduling, auto-scaling, and anomaly detection.
- Edge Computing Integration: Expands Kubernetes to edge environments for low-latency processing and distributed computing.
- Serverless Kubernetes: Supports event-driven, auto-scaling applications without manual infrastructure management [21].

IV. SERVERLESS COMPUTING IN AI-DRIVEN CLOUD INFRASTRUCTURE

A wide range of interrelated issues influence manufacturing's capacity to satisfy the needs of today's dynamic markets in its constantly changing landscape [22][23]. These difficulties include a wide range of topics, all of which are essential to achieving operational excellence and long-term growth. Human-machine interaction, process planning and scheduling, flexibility in response to market situations, product and process quality, data quality, and accessibility could all be included in a thorough analysis of the aforementioned challenges.

A. Serverless Computing

Serverless Computing (or simply serverless) is emerging as a new and compelling paradigm for the deployment of cloud applications, largely due to the recent shift of Enterprise application architectures to containers and microservices.

A new and appealing paradigm for cloud application deployment is serverless computing, as shown in Figure 2, or simply serverless. This is mostly because business application architectures have recently shifted to microservices and containers. Serverless computing reduces operating costs through efficient resource optimization and management, provides a platform that encourages the use of other services in their ecosystem, gives cloud providers more control over the entire development stack, and reduces the effort required to develop and maintain cloud-scale applications [24][25].

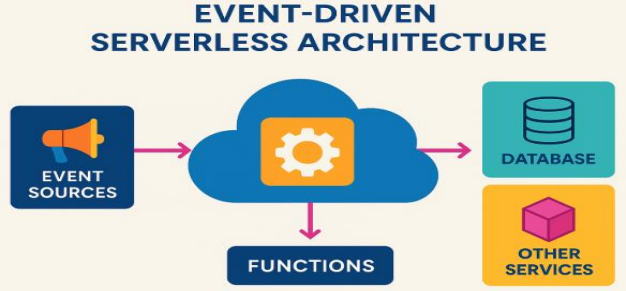

Figure 2 : Serverless Computing in AI-Driven Cloud Infrastructure

B. Traditional vs. Serverless Computing Technology

Cloud computing has revolutionized how organizations approach infrastructure, application deployment, and service delivery. The following are the various Traditional vs. Serverless computing technologies shown in Figure 3:

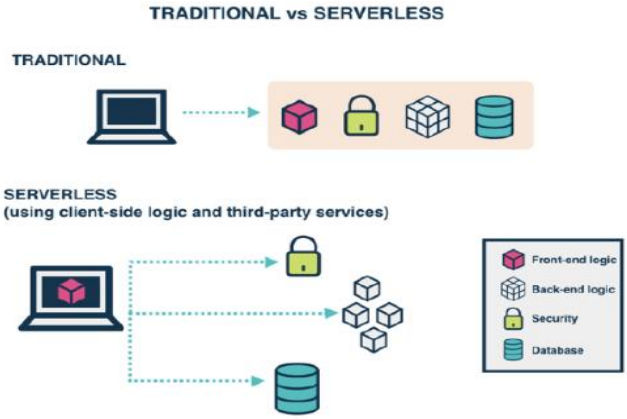

Figure 3 : Traditional vs. Serverless Computing Technology

Figure 3 contrasts traditional and serverless architectures. In a traditional setup, a client-side application typically interacts with a monolithic backend encompassing a database, front-end and back-end logic, and security precautions. On the other hand, serverless architecture disperses these elements by utilizing client-side logic and third-party services. The client application directly interacts with specific, independent services for security, back-end logic (often as Functions as a Service - FaaS), and databases, potentially reducing the operational overhead associated with managing dedicated servers.

- Virtual Management: Traditional cloud computing typically involves managing virtualized servers or containers where applications run, requiring a hands-on approach to provisioning, scaling, and maintenance. In contrast, Developers can focus on code and functionality instead of worrying about servers thanks to serverless computing, which abstracts away the underlying infrastructure [26][27].
- Infrastructure Management: Traditional models necessitate manual provisioning and scaling, whereas serverless computing abstracts these responsibilities to the cloud provider.
- Scalability: Serverless systems, as opposed to traditional approaches that could need manual intervention, naturally facilitate automated scalability in response to incoming events.
- Cost Structure: These traditional models usually pay for pre-allocated resources, and this could result in underutilization. Serverless computing bills themselves based on the real consumption, which makes a more economical choice for the kind of workloads that fluctuate [28].

C. Key Benefits of Serverless Cloud Computing

The advantages of serverless computing are numerous. Here are a few of them:

- Improved Development Speed: Serverless architecture accelerates the software development. Working on producing and improving code makes development cycles significantly shorter than when developers need to also deal with infrastructure management [29].
- Scalability: Serverless computing is by nature scalable. Cloud providers make sure automatic scalability in response to application demand is smooth without human involvement.
- Agility: The advantage of microservices is that they can be independently developed and deployed for faster release cycles and faster ability for the business needs to change.
- Portability: Containerization ensures that the applications run consistently in different environments, making it easy to deploy the applications and port it across environments which reduces deployment issues.

- Cost Efficiency: Serverless computing is among its most tempting features because of its cost effectiveness [30].

D. Challenges of Serverless Computing

Serverless computing has several challenges. Among them are:

- Software engineering challenges: Some of the most important problems with the serverless paradigm have been found to be software engineering obstacles, such developer experience.
- Data Management: Ensuring data consistency and integrity across distributed services can be challenging, necessitating careful design and coordination.
- Security: The increased surface area due to numerous services and APIs can elevate security risks, demanding robust authentication, authorization, and monitoring mechanisms.
- Resource Overhead: While containers are lightweight, the overhead of running numerous instances can accumulate, impacting resource utilization and costs.
- System (operational) challenges: Cloud functions are extremely dynamic, which creates system challenges that necessitate improvements in cloud function lifecycle management, cost predictability, and security [31].

E. Kubernetes for Serverless Computing

Serverless computing is made possible and improved by Kubernetes, which offers scalability, flexible, and efficient container orchestration. Below are key aspects of how Kubernetes contributes to serverless computing:

- Serverless Frameworks – Supports Knative, OpenFaaS, and Kubeless for deploying serverless applications.
- Autoscaling – Uses Horizontal Pod Autoscaler (HPA) and KEDA for dynamic function scaling.
- Cost Efficiency – Enables a pay-per-use model by allocating resources only when needed.
- Multi-Cloud & Portability – Avoids vendor lock-in and supports hybrid & multi-cloud deployments.
- Event-Driven Execution – Integrates with Kafka, NATS, and cloud events for real-time processing.
- Security & Isolation – Implements RBAC, network policies, and pod security for secure execution [32].

F. Kubernetes VS Serverless

These two tools work differently to simplify deployment tasks, although they exist for separate application purposes.

- Organizations choose Kubernetes to manage applications that depend on multistage deployment while maintaining stateful services or managing their infrastructure.
- Serverless technology fits with applications that run on events and microservices plus services that need to grow quickly or need low maintenance efforts [33].

The complexity of the project and the experience of crew will determine whether you select Kubernetes or serverless [34].

V. BENEFITS OF AI-DRIVEN CLOUD SOLUTIONS

The following are some benefits of cloud solutions powered by AI:

- Enhanced Resource Management and Cost Efficiency: The digital business environment has been completely transformed by Cloud computing provides unparalleled scalability, flexibility, and efficiency in data management and service delivery [35]. A number of important obstacles remain in spite of these developments, including maintaining high performance, optimizing cloud resources, and attaining cost-effectiveness. AI, a key component of cloud computing capabilities, has been recognised by recent technical breakthroughs as offering creative solutions for more efficient resource management.
- Improved Security Measures: Adoption of emerging technologies such as the Internet of Things, WSN, cloud/edge computing, and 5G/6G communication networks can lead to numerous opportunities to improve people's quality of life and create intelligent systems that provide customers with innovative, high-quality services across a range of sectors, such as healthcare, agriculture, education, and transportation [36].
- Accelerated Software Development Processes: As technology has advanced and software needs have become more complicated, software engineering processes have undergone a fast and revolutionary change. Methodologies for software development have continuously improved since their inception, moving from the traditional waterfall paradigm to methods that are more flexible and iterative.
- Support for Industrial Internet of Things Applications: IIoT has become a key component of the continuous digital transformation of companies, allowing for improved automation, intelligence, and connection in operations. In the industrial, healthcare, energy, and logistics industries, IIoT enables real-time monitoring, enhanced productivity, and creative business models by connecting physical systems with cutting-edge digital technology. Cloud computing is becoming a crucial facilitator of these systems, as managing and analyzing, success of the IIoT depends on the enormous amounts of data produced by linked devices [37][38].

VI. LITERATURE REVIEW

This section offers a thorough analysis of the Kubernetes research and Serverless Computing Developments with AI-Powered Cloud Infrastructure, with a summarized overview presented in Table I.

Table 1 : Summary of Kubernetes and Serverless Computing with AI-Driven Cloud Infrastructure

Reference	Focus On	Key Findings	Challenges	Limitations/Future Gap
Ma et al. (2025)	Neural-enhanced, interference-aware resource provisioning in serverless computing	Models serverless function provisioning as a combinatorial optimization problem using neural networks to predict performance under interference	Accurately modeling interference in dynamic environments	Need for further validation in diverse real-world serverless setups
Ciptaningtyas et al. (2024)	Serverless deployment using Knative and Kubernetes	Developed a functioning serverless system with monitoring capabilities; resource use (CPU, RAM) is comparable to traditional systems	Ensuring reliability and performance in self-managed serverless setups	Lacks detailed comparative analysis with commercial serverless platforms
Rahman et al. (2024)	Integrating blockchain technology with AI-powered IT systems to better control risk	Improved accuracy, speed, and reliability in handling cyber threats; scalable and adaptive system	Balancing performance and security in hybrid infrastructures	Implementation complexity and integration in legacy systems
Dehury and Srirama (2024)	DRL-based resource management in SDPE with Apache Spark	ISIM-SDP enables real-time, scalable resource allocation using DRL	Integration with existing stream processing ecosystems	Full integration with modern SDPEs and cross-platform generalization is yet to be achieved
Pranata, Wijayanto and Sidiq (2023)	AI-based optimization in cloud/edge resource allocation	AI enables dynamic resource allocation, latency reduction, and energy efficiency	Coordinating AI models across heterogeneous environments	Requires extensive data and continuous model retraining
Tuli et al. (2023)	AI optimization in workload management and scaling	AI supports real-time decision-making for adaptive scaling and resource optimization	Managing unpredictable workload patterns in distributed systems	Limited evaluation in edge-heavy deployments

Ma et al. (2025) a neural-enhanced interference-aware resource provisioning system for serverless computing. They model the resource provisioning of serverless functions as a novel combinatorial optimization problem, wherein the constraints on the queries per second are derived from the neural network performance model. By leveraging neural networks to model the nonlinear performance fluctuations under various interference sources, their approach better captures the real-world behavior of serverless functions [39].

Ciptaningtyas et al. (2024) The Knative Framework allows users to implement serverless services without being dependent on other cloud providers. The aim of this research is to implement a Knative Serverless system and test it for parameters like failure rates, response time, and resource usage. Using Kubernetes, a viable monitoring system and a running Serverless system are produced by the study. Serverless CPUs and RAMs have resource utilisation similar to traditional systems [40].

Rahman et al. (2024) suggest a cutting-edge blockchain-powered IT infrastructure supported by AI in order to provide a revolutionary infrastructure to optimize risk management procedures in different organizations. The proposed infrastructure provides a strong approach to deal with current risk management problems by using blockchain's secure and immutable ledger for data integrity and transparency, and AI for detecting anomalies, predictive analytics, and data-driven decisions. The fact that it is scalable and flexible enough to accommodate the kind of different risk management approaches is what makes the

infrastructure more intelligent, efficient, and safe. Results show that the infrastructure can actively respond to new cyber threats by signaling important progress in the precision, speed, and reliability of risk management [41].

Dehury and Srirama (2024) explore the challenges of integration with Apache Spark, especially. Although modern SDPE capabilities are quite high, these systems still remain not fully mature in terms of resource demands management or seamless system integration with other technologies. To overcome this shortcoming, it suggests the ISIM-SDP architecture: the Integrating Serverless and DRL for Infrastructure Management in Streaming Data Processing across the edge-cloud continuum. It achieves flexible and scalable resource allocation of computational resources by applying DRL based approach for dynamically resource allocation in the real world [42].

Pranata, Wijayanto, and Sidiq (2023) examines the use of optimization solutions inspired by AI for the dynamic allocation of resources in cloud and edge contexts, management of workloads, and the reduction of latency. ML models and predictive analytics can be deployed to make the AI intelligent enough to intelligently organize activities, namely, such that it can develop equally to grow with equal performance. Through the approval of the flexible distribution of tasks, real-time decisional capability on adaptive scaling, and the resource allocation in an energy-efficient fashion, this paper analyses the corresponding key optimization strategies [43].

Tuli et al. (2023) examine how to leverage AI-enabled cloud and edge optimization technologies to lower latency by taking advantage of workload management and dynamic resource allocation in various circumstances. ML models and predictive analytics can intelligently plan tasks and have the systems grow with comparable performance. In this paper, the main optimization technique is based on the acceptance of effective deployment of resources, dynamic task distribution, and adaptive scaling decision making in real time [44].

VII. CONCLUSION AND FUTURE WORK

Iintegrating AI within Kubernetes and other types of serverless computing that forms part of cloud infrastructure is now reinventing the deployment, management, and optimization of applications. Orchestration based on AI improves efficiency of workload, scales resources automatically, and improves fault tolerance in Kubernetes environment. In the same way, in serverless computing, AI allows servers to scale intelligent function, cost optimize, and adapt the workflow so as to have more efficient and resilient Cloud services. They cut down on operational complexity, as well as cost effective, secure, performance improvements, on Cloud Native apps. Despite these benefits, challenges persist. Cloud orchestration that uses AI has to consume huge computational resources to which latency and cost optimality are concerns. Different security risks including AI model vulnerabilities and data privacy have to be attended to constantly. Additionally, there is still a need to achieve interoperability across various cloud environments.

Future studies should concentrate on improving AI models for cloud environments' real-time decision-making. Investigating AI-powered self-healing mechanisms, adaptive workload prediction models, and energy-efficient cloud management strategies will be crucial. Additionally, advancements in federated learning and privacy-preserving AI can address security concerns in cloud-native applications. As AI develops further, its combination with serverless computing and Kubernetes will open the door to more self-sufficient, effective, and intelligent cloud infrastructures.

VIII. REFERENCES

[1] K. Anbalagan, “AI in Cloud Computing: Enhancing Services and Performance,” Int. J. Comput. Eng. Technol., vol. 15, pp. 622–635, 2024, doi: 10.5281/zenodo.13353681.

[2] K. Govindan et al., “Industry Surveys IT Consulting & Other Services.,” J. Clean. Prod., 2018.

[3] H. Chahed et al., “AIDA—A holistic AI-driven networking and processing framework for industrial IoT applications,” Internet of Things, vol. 22, p. 100805, Jul. 2023, doi: 10.1016/j.iot.2023.100805.

[4] J. Decker, P. Kasprzak, and J. M. Kunkel, “Performance Evaluation of Open-Source Serverless Platforms for Kubernetes,” Algorithms, 2022, doi: 10.3390/a15070234.

[5] B. K. R. Janumpally, “A Review on Data Security and Privacy in Serverless Computing: Key Strategies, Emerging Challenges,” Int. J. Innov. Sci. Res. Technol., vol. 10, no. 3, p. 9, 2025.

[6] H. Martins, F. Araujo, and P. R. da Cunha, “Benchmarking Serverless Computing Platforms,” J. Grid Comput., 2020, doi: 10.1007/s10723-020-09523-1.

[7] A. Goyal, “Optimising Cloud-Based CI/CD Pipelines: Techniques for Rapid Software Deployment,” Tech. Int. J. Eng. Res., vol. 11, no. 11, pp. 896–904, 2024.

[8] J. Thomas, “The Effect and Challenges of the Internet of Things (IoT) on the Management of Supply Chains,” Int. J. Res. Anal. Rev., vol. 8, no. 3, pp. 874–878, 2021.

[9] T. Rausch, A. Rashed, and S. Dustdar, “Optimized container scheduling for data-intensive serverless edge computing,” Futur. Gener. Comput. Syst., 2021, doi: 10.1016/j.future.2020.07.017.

[10] V. Prajapati, “Cloud-Based Database Management: Architecture, Security, challenges and solutions,” J. Glob. Res. Electron. Commun., vol. 01, no. 1, pp. 07–13, 2025.

[11] S. R. P. Madugula and N. Malali, “Adversarial Robustness of AI-Driven Claims Management Systems,” Int. J. Adv. Res. Sci. Commun. Technol., pp. 237–246, Mar. 2025, doi: 10.48175/IJARSCT-24430.
[12] S. Arora, S. R. Thota, and S. Gupta, “Artificial Intelligence-Driven Big Data Analytics for Business Intelligence in SaaS Products,” in 2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT), IEEE, Aug. 2024, pp. 164–169. doi: 10.1109/IC2SDT62152.2024.10696409.
[13] S. U. Amin and M. S. Hossain, “Edge Intelligence and Internet of Things in Healthcare: A Survey,” IEEE Access, 2021, doi: 10.1109/ACCESS.2020.3045115.
[14] V. Gharibvand et al., “Cloud based manufacturing: A review of recent developments in architectures, technologies, infrastructures, platforms and associated challenges,” International Journal of Advanced Manufacturing Technology. 2024. doi: 10.1007/s00170-024-12989-y.
[15] R. P. Sola, N. Malali, and P. Madugula, Cloud Database Security: Integrating Deep Learning and Machine Learning for Threat Detection and Prevention. 2025.
[16] T. Subramanya and R. Riggio, “Centralized and Federated Learning for Predictive VNF Autoscaling in Multi-Domain 5G Networks and beyond,” IEEE Trans. Netw. Serv. Manag., 2021, doi: 10.1109/TNSM.2021.3050955.
[17] C. Perducat, D. C. Mazur, W. Mukai, S. N. Sandler, M. J. Anthony, and J. A. Mills, “Evolution and Trends of Cloud on Industrial OT Networks,” IEEE Open J. Ind. Appl., 2023, doi: 10.1109/OJIA.2023.3309669.
[18] V. Pillai, “Integrating AI-Driven Techniques in Big Data Analytics: Enhancing Decision-Making in Financial Markets,” Int. J. Eng. Comput. Sci., vol. 12, no. 7, 2023.
[19] A. Gogineni, “Chaos Engineering in the Cloud-Native Era: Evaluating Distributed AI Model Resilience on Kubernetes,” J Artif Intell Mach Learn Data Sci 2024, vol. 3, no. 1, pp. 2182–2187, 2025.
[20] S. Padmakala, M. Al-Farouni, D. D. Rao, K. Saritha, and R. P. Puneeth, “Dynamic and Energy-Efficient Resource Allocation using Bat Optimization in 5G Cloud Radio Access Networks,” in 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), IEEE, Aug. 2024, pp. 1–4. doi: 10.1109/NMITCON62075.2024.10699133.
[21] C. Carrión, “Kubernetes as a Standard Container Orchestrator - A Bibliometric Analysis,” J. Grid Comput., 2022, doi: 10.1007/s10723-022-09629-8.
[22] A. Gogineni, “Multi-Cloud Deployment with Kubernetes: Challenges, Strategies, and Performance Optimization,” Int. Sci. J. Eng. Manag., vol. 1, no. 02, 2022.
[23] J. Thomas, “Enhancing Supply Chain Resilience Through Cloud-Based SCM and Advanced Machine Learning: A Case Study of Logistics,” J. Emerg. Technol. Innov. Res., vol. 8, no. 9, pp. 357–364, 2021.
[24] H. Shafiei, A. Khonsari, and P. Mousavi, “Serverless Computing: A Survey of Opportunities, Challenges, and Applications,” ACM Comput. Surv., 2022, doi: 10.1145/3510611.
[25] S. S. S. Neeli, “Cloud Migration DBA Strategies for Mission-Critical Business Applications,” Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 11, pp. 591–598, 2023.
[26] V. Prajapati, “Role of Identity and Access Management in Zero Trust Architecture for Cloud Security : Challenges and Solutions,” pp. 6–18, 2025, doi: 10.48175/IJARSCT-23902.
[27] V. Pillai, “Anomaly Detection in Financial and Insurance Data-Systems,” J. AI-Assisted Sci. Discov., vol. 4, no. 2, 2024.
[28] J. M. O. Candel, A. Elouali, F. J. M. Gimeno, and H. Mora, “Cloud vs Serverless Computing: A Security Point of View,” in Lecture Notes in Networks and Systems, 2023. doi: 10.1007/978-3-031-21333-5_109.
[29] D. D. Rao, S. Madasu, S. R. Gunturu, C. D’britto, and J. Lopes, “Cybersecurity Threat Detection Using Machine Learning in Cloud-Based Environments: A Comprehensive Study,” Int. J. Recent Innov. Trends Comput. Commun., vol. 12, no. 1, 2024.
[30] D. Loconte, S. Ieva, A. Pinto, G. Loseto, F. Scioscia, and M. Ruta, “Expanding the cloud-to-edge continuum to the IoT in serverless federated learning,” Futur. Gener. Comput. Syst., 2024, doi: 10.1016/j.future.2024.02.024.
[31] H. B. Hassan, S. A. Barakat, and Q. I. Sarhan, “Survey on serverless computing,” J. Cloud Comput., vol. 10, no. 1, p. 39, Jul. 2021, doi: 10.1186/s13677-021-00253-7.
[32] S. K. Mondal, R. Pan, H. M. D. Kabir, T. Tian, and H. N. Dai, “Kubernetes in IT administration and serverless computing: An empirical study and research challenges,” J. Supercomput., 2022, doi: 10.1007/s11227-021-03982-3.
[33] S. Shah and M. Shah, “Deep Reinforcement Learning for Scalable Task Scheduling in Serverless Computing,” Int. Res. J. Mod. Eng. Technol. Sci., vol. 3, no. 12, pp. 1845–1852, Jan. 2025, doi: 10.56726/IRJMETS17782.
[34] M. Aazam, S. Zeadally, and K. A. Harras, “Fog Computing Architecture, Evaluation, and Future Research Directions,” IEEE Commun. Mag., vol. 56, no. 5, pp. 46–52, May 2018, doi: 10.1109/MCOM.2018.1700707.
[35] S. Murri, “Data Security Environments Challenges and Solutions in Big Data,” Int. J. Curr. Eng. Technol., vol. 12, no. 6, pp. 565–574, 2022.
[36] J. K. Chaudhary, S. Tyagi, H. P. Sharma, S. V. Akram, D. R. Sisodia, and D. Kapila, “Machine Learning Model-Based Financial Market Sentiment Prediction and Application,” in 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, May 2023, pp. 1456–1459. doi: 10.1109/ICACITE57410.2023.10183344.
[37] S. A. Ionescu and V. Diaconita, “Transforming Financial Decision-Making: The Interplay of AI, Cloud Computing and Advanced Data Management Technologies,” Int. J. Comput. Commun. Control, 2023, doi: 10.15837/ijccc.2023.6.5735.
[38] S. Arora and S. R. Thota, “Ethical Considerations and Privacy in AI-Driven Big Data Analytics,” Int. Res. J. Eng. Technol., vol. 11, no. 05, 2024.
[39] R. Ma, Y. Zhan, C. Wu, Z. Hong, Y. Ali, and Y. Xia, “Qora: Neural-Enhanced Interference-Aware Resource Provisioning for Serverless Computing,” IEEE Trans. Autom. Sci. Eng., pp. 1–16, 2025, doi: 10.1109/TASE.2025.3526197.
[40] H. T. Ciptaningtyas, R. R. Hariadi, F. D. Rosyadi, and S. S. Al Azmi, “Serverless Computing Model Using Kubernetes and Knative in a

Scalable Cloud Development," in 2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C), 2024, pp. 659–664. doi: 10.1109/BTS-I2C63534.2024.10942132.

[41] M. M. Rahman, B. P. Pokharel, S. A. Sayeed, S. K. Bhowmik, N. Kshetri, and N. Eashrak, "riskAIchain: AI-Driven IT Infrastructure—Blockchain-Backed Approach for Enhanced Risk Management," Risks, vol. 12, no. 12, p. 206, Dec. 2024, doi: 10.3390/risks12120206.

[42] C. K. Dehury and S. N. Srirama, "Integrating Serverless and DRL for Infrastructure Management in Streaming Data Processing across Edge-Cloud Continuum," in 2024 IEEE 44th International Conference on Distributed Computing Systems Workshops (ICDCSW), 2024, pp. 93–101. doi: 10.1109/ICDCSW63686.2024.00020.

[43] M. Pranata, A. Wijayanto, and M. F. Sidiq, "Serverless Autoscaling Metrics for Optimum Performance on Edge Computing," in 2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), IEEE, Dec. 2023, pp. 65–69. doi: 10.1109/ISRITI60336.2023.10467288.

[44] S. Tuli et al., "AI augmented Edge and Fog computing: Trends and challenges," Journal of Network and Computer Applications. 2023. doi: 10.1016/j.jnca.2023.103648.