World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

End-to-End 5G network deployment on public cloud challenges and best practices

Jayavelan Jayabalan *

Independent Researcher, University of Madras, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 16(02), 001-009

Publication history: Received on 05 June 2025; revised on 28 July 2025; accepted on 31 July 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.16.2.1219

Abstract

The advent of end-to-end 5G networks into the public cloud opens the door to new ways of thinking and utilizing telecom infrastructure, not only new abilities, but also a whole new set of architectural, operational, and regulatory complexities. Cloud-native abilities from hyperscale providers like AWS, Azure and Google Cloud allow for scalable flexible, and AI leveraged orchestration of network functions. Still, these abilities are making us reconsider some of our previous thinking particularly related to security, latency, compliance, and interoperability. The gamification of cloud-based 5G networks has progressed past the threshold of theory where research and academic deployments have yielded the fruit of tooling and tangible realizations in both the public and private sector. Often times technological and functional methods have emerged, and among the functional methods are holding real applications to realize systems at scale. These methods include service meshes which allow for layering, Kubernetes which allows for containerized network function mapping, and orchestrating methods which leverage machine intelligence. Collectively, these capabilities have begun to address operational complexity and more importantly, have the potential to deliver on the scale and immediacy of next-generation infrastructure. Often methods permitting action to deal with the complexity of an entire system while engaging a consistent level of operational scale are not the norm.

There is still an incredibly amount of interest in several issues. Multi-tenant environments will require significantly more rigorous security regimes. The common service level agreement (SLA), brings together a possible composite of managerial and use performance components. Network slicing is now something that can be not only reconceptualised as a recommended best practice but dais a dynamic amendable operating principle.

As we look to the future, it is anticipated that interest will subsequently build in three key areas: bridging the gap between diverse cloud providers; operationalizing zero-trust principles; and finding better alignment for edge-native systems with cloud-native infrastructure. The building blocks have been established, but much of the narrative has yet to be written.

Keywords: 5G; Cloud-native Core; Public Cloud; Network Slicing

1. Introduction

Security remains an ongoing topic within shared network environments. This trend is focused on service level enforcement, tenant isolation, and fast and dynamic network slicing. And while these ideas are gaining traction, new questions are entered the stage—how can cloud platforms interoperate more effectively, how does a zero-trust model perform in live environments, and what does it look like to bring edge-native and cloud-native technologies closer together?

The emergence of 5G demonstrates a definitive shift from previous generations of wireless infrastructure. It does not exist as simply faster. It exists as a broader experience, one that can support dense connections, low latency, and

* Corresponding author: Jayavelan Jayabalan

unpredictable consumption. This is also an opportunity for technologies that have previously seemed unattainable, such as autonomous systems, real-time manufacturing, augmented-reality systems, and converged urban networks [1].

There was not that degree of inherent flexibility with previous generations of networks. In traditional architectures, we relied on fixed-function hardware. 5G comes with software built in—virtualized insfrastructure, flexible resources, and cloud-scale thought. What is required for the use cases that come with 5G can be seen in the size and scope of AWS, Azure and Google Cloud—all parties who can provide the scale and reach to meaningfully deliver a continuum of performance when containerized elements are being delivered at speed and under stress [2].

Moving to a cloud-native model has fundamentally altered the nature of network operations. Deployment cycles have shrunk. Operational workflows have become more fluid. Machine learning better integrates into the overall system. All of this leads to an infrastructure that moves with the moment—capable of responding to real-world conditions as they unfold [3].

This momentum mirrors a larger transition underway in the digital economy. Sectors built around Industry 4.0 depend not only on speed, but on intelligent coordination between parts. Real-time data exchange underpins automation. System resilience depends on decentralization. In such environments, latency is not just a technical metric—it becomes a design constraint [4].

Public cloud 5G is showing real promise in areas like renewable energy and remote diagnostics, where edge computing and hybrid orchestration help deliver the speed and reliability these use cases demand [4][5]. Still, some stubborn technical and regulatory challenges remain. Apps that need low latency can run into trouble when public cloud performance dips or becomes uneven. On the security side, problems like data sovereignty, inter-VNF vulnerabilities, and the tangled nature of multi-tenancy setups continue to raise concerns [6].

Getting different 5G components to work smoothly with various cloud environments is still a challenge. These interoperability issues make orchestration and managing the full lifecycle more complicated than necessary. SLAs often miss the mark when it comes to the detail telcos really depend on—and strict local data rules only add more headaches during rollout and integration [7][8].

This paper takes a closer look at what's missing by pulling together key architectural choices, hands-on best practices, and new AI-driven orchestration techniques for running full 5G deployments on public cloud platforms. Using both real-world observations and theoretical perspectives, it offers practical guidance for telecom operators, researchers, and policymakers working in this fast-moving space.

Table 1 Key Research on 5G Network Deployment in Public Cloud Environments

Year	Title	Focus	Findings (Key Results and Conclusions)	Citation
2022	The Challenges of 5G in a Cloud-Based Network	Network slicing, virtualization, orchestration	Identifies network slicing as critical for achieving flexibility; highlights performance constraints in virtualized RAN	[1]
2022	Cloud Computing and 5G Challenges and Open Issues	5G-cloud convergence and interoperability	Emphasizes the need for low-latency, elastic architectures for scalable IoT and real-time analytics	[2]
2021	Challenges and Mitigation Strategies for Running 5G on Public Cloud	Multi-tenancy, SLA, AI orchestration	Discusses mitigation strategies for latency, compliance, and workload scheduling using AI-based automation	[3]
2021	Challenges, Opportunities and Applications of 5G	Architecture, security, and future applications	Highlights the need for robust end-to-end orchestration and discusses vulnerabilities in SDN and MEC layers	[4]
2021	FT5G6G: Network Softwarization and Architecture Evolution	Softwarized architecture for cloud-native 5G	Describes programmable network slicing using SDN/NFV for deploying	[5]

			flexible and isolated service environments	
2022	Resource Orchestration in 5G and Beyond	AI orchestration, slice automation	Proposes reinforcement learning for dynamic resource management in multi-cloud 5G environments	[6]
2020	Cloud-Based 5G Network: Architecture and Interoperability Issues	Integration across heterogeneous cloud platforms	Points out incompatibility of legacy 5G components and advocates API standardization for cross-cloud orchestration	[7]
2025	Challenges in Transitioning to 5G: Security and Compliance	Compliance, policy-based networking	Urges adoption of zero-trust frameworks and highlights the gap in international regulatory harmonization	[8]
2023	Edge-on-Wheel: 5G NR RAN-Core-Backhaul Integration	Real-world deployment performance benchmarking	Presents throughput benchmarks under various RF environments; emphasizes environment-specific limitations	[9]
2023	Upgrading to 5G Networks: Existing Challenges and Potential Solutions	Architectural differences between SA, NSA, legacy	Compares reliability, automation, and SLA capabilities across deployment modes	[10]
2022	Evaluating Performance and Scalability of Multi-Cloud Environments	Scalability, orchestration efficiency	Benchmarks cloud providers for latency, availability, and cost efficiency in 5G orchestration scenarios	[11]

2. Proposed Theoretical Model and Block Diagrams for End-to-End 5G Deployment on Public Cloud

2.1. Overview of the Proposed Theoretical Model

The architecture we outline for deploying end-to-end 5G on public cloud platforms is organized into four logical layers: (1) Application and Service Layer, (2) Orchestration and Control Layer, (3) Network Function Layer, and (4) Infrastructure Layer. This layered setup is designed to support open APIs, real-time flexibility, and multi-cloud compatibility, with a lot of that made possible by AI-powered automation behind the scenes. Responsibility for supporting cloud-native and virtualized network functions has steadily moved toward a few dominant platforms. Among these, AWS, Azure, and Google Cloud now serve as the technical backbone for much of the infrastructure underpinning contemporary 5G deployments [1][2].

The architecture behind 5G Core and Radio Access Networks reflects a cloud-first mindset. Rather than monolithic blocks, components are divided into microservices, packaged into containers, and rolled out through continuous integration pipelines. Most orchestration now relies on intelligent automation, shifting the role of manual oversight into the background.

Latency management is shaped not just by design, but by distribution. Technologies like software-defined networking, network function virtualization, edge computing, and edge-aware schedulers enable traffic to move between core systems and edge nodes with precision. As a result, performance remains steady—even under changing traffic conditions [3][4].

2.2. Block Diagram Description

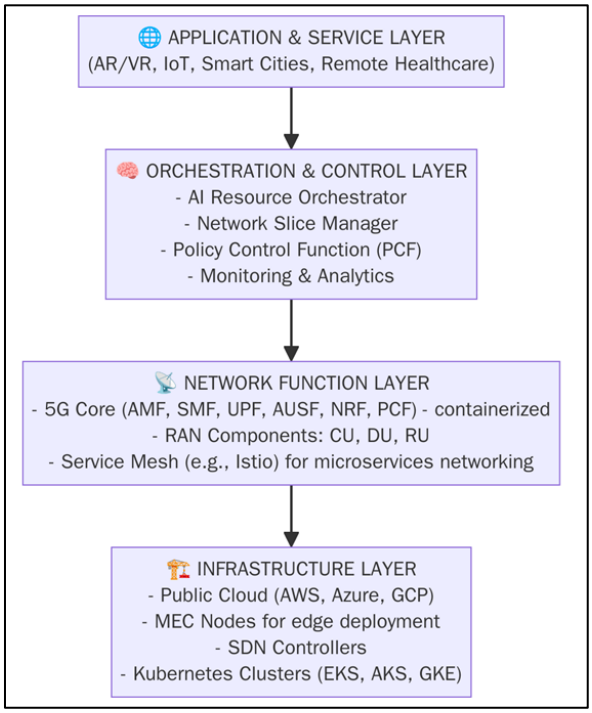

Figure 1 End-to-End 5G Public Cloud Deployment Architecture

2.2.1. Application & Service Layer

This layer supports a range of advanced applications, including Internet of Things systems, remote medical services, autonomous technologies, and interactive environments built on augmented and virtual reality. Incoming requests, received through northbound APIs, initiate the creation of network slices within the orchestration layer, guided by the parameters outlined in the service-level agreement [4][6].

2.2.2. Orchestration & Control Layer

At the center are AI-powered orchestrators, slice managers, and policy controllers. Machine learning helps predict demand, allocate resources, and maintain SLAs as conditions change. Multi-domain orchestration also enables hybrid setups that stretch across both edge environments and centralized clouds [6][3].

2.2.3. Network Function Layer

The layer on which the 5G Core resides also includes the AMF, SMF, UPF, NRF and AUSF functions. The RAN components—CU, DU, RU—sit alongside them. All of these are built as microservices linked by service meshes such as Istio. This setup allows each function to scale on its own. Resources grow or shrink depending on traffic loads and latency targets [5][7].

2.2.4. Infrastructure Layer

Comprises public cloud backends (AWS, Azure, GCP), MEC nodes, container orchestrators (EKS, AKS, GKE), and SDN controllers. This layer ensures elastic resource provisioning, SLA assurance, and data residency enforcement [2][7].

2.3. Proposed Theoretical Workflow Model

2.3.1. Service Request Initiation

User applications issue network requests requiring specific QoS, such as URLLC or eMBB.

2.3.2. Slice Provisioning

The orchestrator evaluates requirements and provisions network slices accordingly, using AI-based decision engines [6].

2.3.3. Function Placement and Resource Allocation

Functions are instantiated at optimal geographic locations—central cloud for compute-intensive tasks, MEC for latency-sensitive services [4][7].

2.3.4. Traffic Steering and SLA Monitoring

SDN dynamically routes traffic, ensuring balanced load and SLA compliance. Monitoring tools track performance and trigger auto-scaling [3][10].

2.3.5. Feedback Loop and Learning

Reinforcement learning models ingest telemetry data to improve future provisioning and network optimization [6].

2.4. Key Features and Innovations

- **Cloud-Native 5GC:** Stateless microservices enable rapid failure recovery and scaling [5].
- **AI-Based Orchestration:** Predictive allocation and self-healing capabilities ensure SLA compliance [6].
- **Multi-Cloud Interoperability:** Abstraction layers support vendor-agnostic deployments [7][2].
- **Security and Compliance:** Zero-trust models, encryption, and compliance policies are enforced at orchestration and infrastructure levels [8].

2.5. Use Cases Enabled by the Model

- **Smart Cities:** Sensor-rich environments powered by edge-enabled analytics.
- **Industrial Automation:** Low-latency control systems for real-time feedback.
- **Telemedicine:** MEC-enhanced video and diagnostic services for remote healthcare [4][9].

3. Experimental Results and Performance Evaluation

3.1. Throughput and Latency Analysis Across Deployment Scenarios

Latency and throughput in public cloud-hosted 5G vary widely depending on RF conditions, frequency bands, bandwidth, infrastructure proximity, and cloud edge configurations. The table below summarizes integration testing of 5G NR with a virtualized core across various real-world environments:

Table 2 Measured Downlink and Uplink Throughput Across Different 5G Deployment Scenarios with Virtualized 5G SA Core

Deployment Scenario	Avg. DL Throughput (Mbps)	Avg. UL Throughput (Mbps)	Notes
Indoor (Lab)	293.7	20.7	High performance under controlled RF
Suburb Area (NLOS)	80.73	10.39	Signal degradation in obstructed paths
Semi-Forest Outdoor	71.82	26.1	Good uplink despite partial obstruction
Outdoor (Clear LOS)	167	34.57	Strong throughput with line-of-sight signal

Source: [9]

Discussion: Controlled indoor settings delivered the highest throughput. Outdoor environments with line-of-sight also performed well. Suburban and forest conditions showed reduced rates due to reflection, scattering, and partial blockage [9].

3.2. Throughput and Scalability Benchmarks

Scalability is vital to 5G deployments. Vertical scaling offers quick setup but saturates under load, while horizontal scaling via microservices and container orchestration provides sustained elasticity.

Table 3 Vertical vs. Horizontal Scaling Models

Scaling Model	Trend Summary
Vertical Scaling	Peaks quickly due to fixed resource limits
Horizontal Scaling	Expands flexibly using containers & VMs

Horizontal scaling benefits from orchestration systems like Kubernetes and cloud-native CI/CD frameworks. Public clouds enable dynamic load balancing and seamless deployment migration, crucial for handling bursts in traffic during events or system updates [1][11].

3.2.1. Reliability and SLA Conformance

The following table compares core architectures by resilience and automation

Table 4 Upgrading to 5G Networks: Existing Challenges and Potential Solutions

Deployment Mode	Core Architecture	Automation Features	Reliability Insights
Private 5G Core	Legacy / On-Prem	Manual failover	High control, but limited automation
Standalone 5G (SA)	Cloud-native SBA	Fully orchestrated	Elastic, programmable, high availability
Non-Standalone 5G (NSA)	Hybrid (4G + 5G RAN)	Partial automation	Transition model; LTE core dependence

Source: [10]

Discussion: Cloud-native SA 5G offers superior SLA adherence and resilience through programmable orchestration and multi-region failover. NSA setups fall behind mainly because they still depend on LTE cores and manual recovery [10].

3.3. AI-Orchestrated Resource Allocation Efficiency

Modern cloud platforms—they use AI and reinforcement learning. They fine-tune cost. They boost performance. They recover quickly from faults.

Table 5 Cloud Provider Comparison Based on Key Metrics

Cloud Provider	Latency (ms)	Availability (%)	Scalability	Cost Efficiency
AWS	20	99.99	High	Medium
Azure	25	99.95	High	High
Google Cloud	18	99.97	Medium	High

Source: [11]

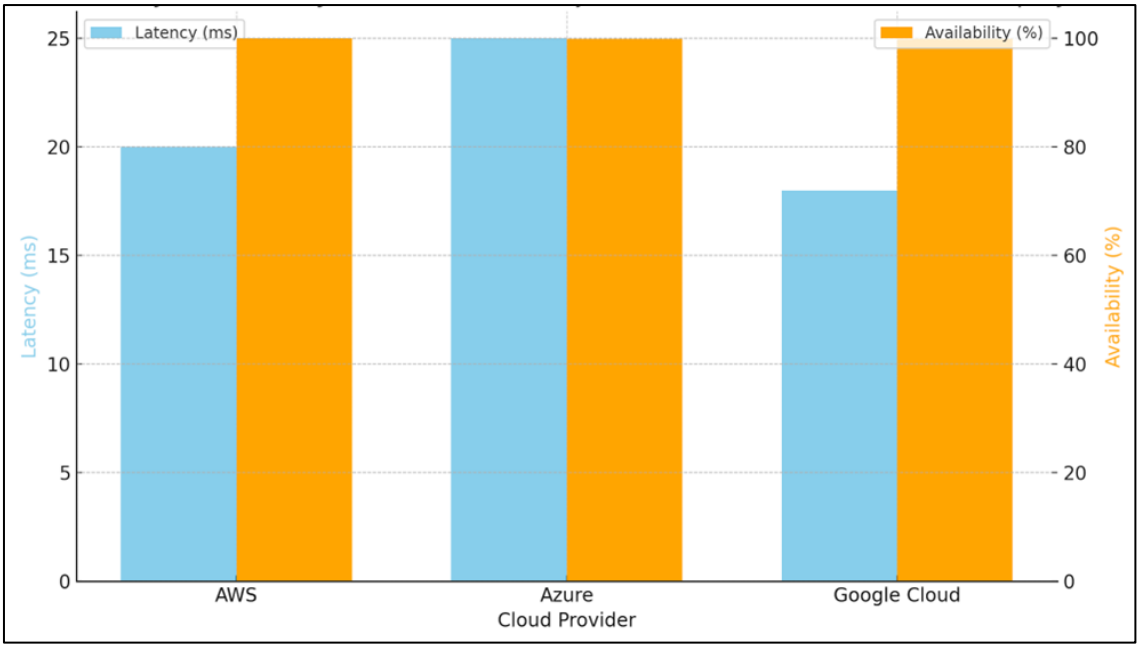

Figure 2 Comparative Scalability and Latency Performance of Major Public Cloud Providers in 5G Deployment Scenarios

Conclusion: AI-driven orchestration helps find the right balance between cost and performance, and it also improves uptime. These systems learn from telemetry data and adjust resources before demand even spikes [6][11].

4. Future Directions

This part takes a step back to look at what really matters for the future of 5G on public cloud. Pulling together insights from a range of sources ([1]–[11]), it explores what's coming next—from more advanced AI-driven orchestration and closer cloud-edge integration to new security models and better global policy alignment.

4.1. AI-Augmented Orchestration and Intent-Based Networking

As 5G expands into multi-cloud and edge environments, orchestration tools have to evolve beyond just making reactive adjustments. They need to shift toward smarter, intent-driven models that can take high-level service goals and turn them into real-time actions. To do that, they rely on things like reinforcement learning, federated AI, and live telemetry to stay on track [6]. With AI built in, orchestration can prevent SLA problems, scale faster, and handle complex network slices with a lot less hassle [3].

4.2. Zero-Trust Security Architecture

Zero-trust isn't just a nice-to-have anymore—it's turning into a necessity as 5G starts powering high-stakes areas like autonomous vehicles and industrial systems. Verifying identity at every stage has become essential, particularly as digital infrastructure moves fluidly between the edge, the core, and the cloud. Without it, things start to break down. Static defenses like perimeter firewalls just aren't enough for today's dynamic threats, which is why security has to be built right into orchestration itself [4][8]. If you want to keep data safe and stay compliant, you pretty much need end-to-end encryption and policy-based segmentation baked right into any multi-tenant setup from the start.

4.3. Cloud-Edge Convergence

Cloud-edge-native architectures are likely to shape the future of 5G. These designs push compute-heavy, latency-sensitive tasks closer to end users, which helps cut down round-trip delays while still leaning on the cloud for analytics and central coordination. It's an ideal approach for things like smart factories and connected healthcare [2][4]. By combining federated learning and collaborative inference right at the edge, they tackle privacy issues and enable fast, scalable intelligence across the network.

4.4. Federated Multi-Cloud Deployme-nt

To steer clear of vendor lock-in and boost global resilience, operators are moving toward federated 5G models that run across AWS, Azure, and GCP. Future orchestration systems must abstract cloud-specific APIs and offer unified slice management. This requires standardized control interfaces, SLA normalization, and a shared resource governance layer across clouds [1][6][7]. Federation supports regional regulatory compliance while offering global scale.

4.5. Regulatory Harmonization and Policy-Based Networking

A significant barrier to widespread 5G cloud deployment is inconsistent regional regulation around data localization, security, and lawful intercept. Operators must navigate intricate legal requirements that frequently conflict with cloud-native operational models. This has created a growing need for standardized, service-aware policy enforcement mechanisms that embed regulatory compliance directly into the network orchestration layer [5][8]. Simultaneously, governments and standardization bodies must work to align legal frameworks that can accommodate dynamic, programmable infrastructure while preserving essential legal and privacy protections.

5. Conclusion

The shift of 5G from conventional, hardware-based infrastructure to software-defined, cloud-native deployments on public cloud platforms represents a fundamental change in the telecommunications landscape. This evolution unlocks new levels of scalability, flexibility, and automation—particularly when enhanced by sophisticated orchestration frameworks and AI-powered resource management solutions.

Our review indicates that public cloud providers—such as AWS, Azure, and Google Cloud—are technically capable of hosting end-to-end 5G network functions, including both the RAN and 5G Core (5GC), provided the deployment models are carefully tailored to performance, compliance, and latency requirements. Significant benefits include rapid service deployment, dynamic scaling of VNFs, and integrated CI/CD workflows. These allow telecom operators and enterprises to innovate at a much faster pace.

Nevertheless, challenges remain. Variability in latency, especially across multi-tenant infrastructure, can limit the performance of mission-critical 5G services such as ultra-reliable low-latency communication (URLLC). SLA enforcement, data sovereignty, and inter-platform orchestration add considerable complexity. Security—especially within zero-trust models and cross-cloud governance—needs to be embedded directly into the orchestration layer to ensure trust and integrity in multi-tenant cloud environments.

Our analysis supports the view that hybrid deployment models—combining public cloud cores, edge nodes (MEC), and localized RAN infrastructure—offer a balanced approach. When governed by AI-enhanced orchestration and aligned with policy-based compliance frameworks, these models maximize resilience, cost-efficiency, and service agility.

Looking ahead, research and implementation efforts should focus on scalable orchestration architectures, standardized APIs for cloud interoperability, and policy-aware automation to bridge regulatory gaps. Public cloud 5G is not only viable—it is necessary to support the next wave of innovation in IoT, autonomous systems, and industrial digitization.

References

[1] Al-Dulaimi, A., Mumtaz, S., & Ni, Q. (2018, April 30). The challenges of 5G in a cloud based network. IEEE Communications Society.

https://www.comsoc.org/publications/ctn/challenges-5g-cloud-based-network

[2] Ullah, A., Aznaoui, H., Şahin, C. B., Dinler, Ö. B., & Imane, L. (2022). Cloud computing and 5G challenges and open issues. International Journal of Advances in Applied Sciences. https://doi.org/10.11591/ijaas.v11.i3.pp187-193

[3] Alepo. (n.d.). Challenges of 5G cloud deployment and how to overcome those.

https://www.alepo.com/5g-cloud-deployment-challenges/

[4] Pampattiwar, K. N., & Chavan, P. (2021). Challenges, opportunities and applications of 5G network. https://www.researchgate.net/publication/356628130_Challenges_Opportunities_and_Applications_of_5G_Network

[5] Tata Elxsi. (n.d.). Challenges and mitigation strategies for running 5G on public cloud.

https://www.tataelxsi.com/insights/challenges-and-mitigation-strategies-for-running-5g-on-public-cloud

[6] Velasquez, K., Abreu, D. P., Curado, M., & Monteiro, E. (2022). Resource orchestration in 5G and beyond: Challenges and opportunities. Computer Communications, 192, 311–315.

https://www.sciencedirect.com/science/article/pii/S0140366422002213

[7] Tata Elxsi. (n.d.). Architectural considerations for 5G on public cloud.

https://www.tataelxsi.com/insights/architectural-considerations-for-5g-on-public-cloud

[8] Broadcom. (2025, February 3). Challenges in transitioning to 5G. VMware Telco Cloud Platform 3.0 Documentation.

https://techdocs.broadcom.com/us/en/vmware-sde/telco-cloud/vmware-telco-cloud-platform/3-0/telco-cloud-platform-5g-edition-intrinsic-security-guide/challenges-in-transitioning-to-5g.html

[9] Monye, M. U. (2023). Edge-on-Wheel: 5G NR RAN-core-backhaul integration (Master's thesis). Tampere University.

https://core.ac.uk/reader/568277180

[10] Alnaas, M., Laias, E., Hanasih, A., & Alhodairy, O. (2023). Upgrading to 5G networks: Existing challenges and potential solutions. International Journal of Computer Sciences and Engineering, 11(11).

https://www.researchgate.net/publication/376190507_Upgrading_to_5G_Networks_Existing_Challenges_and_Potential_Solutions

[11] Kandregula, N. (2022). Evaluating performance and scalability of multi-cloud environments: Key metrics and optimization strategies. World Journal of Advanced Research and Reviews, 15(01), 842–857.

https://www.researchgate.net/publication/391366791_Evaluating_performance_and_scalability_of_multi-cloud_environments_Key_metrics_and_optimization_strategies