A Brief on 6G Networks Resilience and Service Availability Metric*

Revisiting standard ITU network availability and number of 9s in the context of emerging software-based core telco cloud networks 3CN†

Hakima CHAOUCHI
Institut Polytechnique de Paris
Laboratoire SAMOVAR
91123 Palaiseau Cedex, France
hakima.chaouchi@imt.fr

Bilal GHANI
Institut Mines-Télécom
Télécom SudParis, Institut Polytechnique de Paris
91123 Palaiseau Cedex, France
bilal.ghani@telecom-sudparis.eu

***Abstract*—The design of network architectures and standard communication protocols follow requirements and specifications from the International Telecommunication Union (ITU) to establish the corresponding Quality of Service (QoS) requirements and ensure the interoperability and continuity of services. Availability being one of the ITU QoS metrics, it is revisited in this paper to explain the need to clarify the 6G architecture design principles such as centralized or distributed approaches, the communication concept as point-to-point of the physical layer and end-to-end of the network and transport layers and their effect on the expected performance in 6G networks, particularly in the emerging 3CN software and programmable network approach. The SimpleRAN architecture is proposed to showcase the need to distribute the core network to meet or closely approach the required QoS of the future complex 6G architecture.**

***Index Terms*—6G Core and RAN, ITU Network Availability Metric, Programmable Networks, Network Functions.**

I. Introduction

Network availability is part of a Service Level Agreement (SLA) represented by the ITU QoS metric often referred to as the "*number of nines*" where, for example, five-nines correspond to 99.999% of availability, meaning a network that offers always access but only 5 minutes of downtime per year, which is usually ensured by a highly critical communication network. The old telephony circuit-switched network is based on a five-nines availability guarantee, whereas the packet-switched model used by Internet Service Providers (ISPs) generally aims for an SLA guarantee with a minimum of three-nines (99.9%) for residential and basic business services corresponding to 8 hours per year downtime.

5G technology features make it capable of supporting the requirements of factory control systems for real-time determinism and six-nines (99.9999%) availability. Yet the real-world experience of most mobile handset users accessing 2G, 3G, or 4G networks still involves black spots where coverage is weak or nonexistent, and of occasional and unpredictable dropped connections.

So, is there a realistic prospect that mobile phone technology will be used to connect mission-critical, time-sensitive industrial machines? In fact, the recent installation of 5G networks known as Industrial 5G and future 6G, marks the first time that a new generation of mobile technology has been built around the needs of machines and systems rather than handset users [1]. However, meeting the stringent reliability and latency requirements of such applications cannot be accomplished by high air-interface QoS alone.

Therefore, in addition to the high QoS at the radio air interface, different key architectural components are necessary to ensure service continuity. These include the redundancy of hardware/equipment, software replication enabled through virtualization, and the distribution of network functions [2].

For instance, in the context of future 6G networks, a distributed 6G core network will be a paradigm shift compared to the 5G core that promises to significantly improve network resilience by eliminating single points of failure, which are inherent in centralized designs [3]. Critical control functions such as Access and Mobility Management Function (AMF), Session Management Function (SMF), etc., often exist in a centralized node, which makes them prone to network outages caused by hardware failures, connectivity disruptions, cyberattacks, or natural disasters. On the other hand, a distributed network elevates the redundancy by geographic distribution of its Network Functions (NFs). The same NF can have multiple instances deployed at different sites, to place them close to users to reduce latency and enhance both performance and resilience [4].

The remainder of the paper reviews the key architectural components that support network access and service availability. We then present network architectural design considerations, including the decentralization of network functions, redundancy, and the need for end-to-

*This work is part of the French-funded BPI France 2030 research project SIMPLERAN.

† 3CN: Collaborative, Computing, Communication Networks.

979-8-3315-7111-5/26/$31.00 ©2026 IEEE

end network availability metrics, beyond point-to-point air interface availability, to maximize user experience. Furthermore, the paper discusses the control plane and resilience metrics, and finally illustrates the implementation of distributed core functions through a 5G advanced scenario in an industrial innovation project to enhance network resilience.

II. 6G IMT 2030 QoS Requirements Wheel

The ITU International Mobile Telecommunications (IMT)-2030 QoS requirement document provides the network service capabilities as depicted in the Figure 1 below:

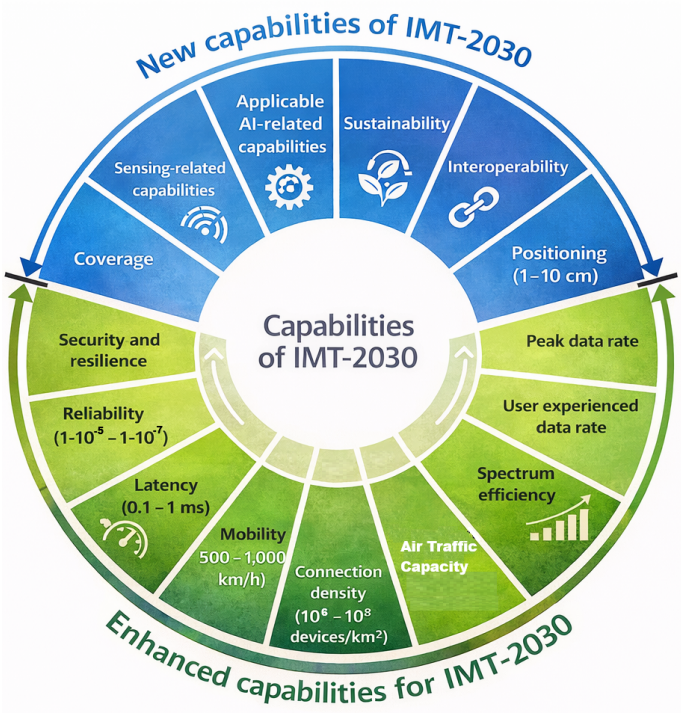

Fig. 1: 6G IMT 2030 network service capabilities

The proposed capability is lacking in the 6G network availability metric compared to the 5G and previous mobile network generations. Network availability is one of the monitoring metrics for assessing the resilience of the network [5].

III. ITU Availability Requirement And Network Functions Redundancy

The "number of nines" is a common industry shorthand, often adopted for SLAs, to express the availability percentage. It explicitly quantifies the maximum permissible network downtime within a defined period, typically a year. The availability percentage is represented by the number of nines; for example, "five-nines" is 99.999% and measures the network and service up-time per year. The general formula to find the maximum annual downtime for an availability percentage A is [6]:

$$\text{Maximum Downtime} = (1 - A) \times 8760 \text{ hours/year} \quad (1)$$

The table below illustrates the relationship between the number of nines, the availability percentage, and the maximum allowed downtime in a year (365 days): Achieving a higher number of nines (e.g., four or five nines) is significantly more complex and costly because system architecture must incorporate a higher level of redundancy, fault tolerance, and fast recovery mechanisms (e.g., automated fail-over, geographically dispersed data centers, redundant power supplies, etc.). In telecommunications, high network availability is crucial for services such as Mission-Critical Services (emergency services, financial transactions, and real-time control systems).

TABLE I: Network Availability, "9s" Term, and Downtime per Year

Availability (%)	"9s" Term	Unavailability (%)	Downtime per Year (8760 hrs)
99.9%	Three-nines	0.1%	8.76 hours (525.6 minutes)
99.99%	Four-nines	0.01%	52.6 minutes
99.999%	Five-nines	0.001%	5.26 minutes
99.9999%	Six-nines	0.0001%	31.5 seconds

Traditional Telecom (Hardware-Based) offers typically five-nines (99.999%) or higher availability [7]. This is the default rate because the physical components (router, switch, base station) are built with internal hardware redundancy, duplicated power supplies, and fast failover mechanisms built into the chassis and network card architecture.

Basic Cloud (Software-Based) offers typically three to four-nines (99.9% to 99.99%) network availability [8]. This is the default rate for a single Virtual Machine (VM) or a container instance. A single server rack failure, maintenance window, or software crash can take this instance down, hence lowering the guarantee. Achieving high network availability with the software cloud approach used in telco architecture requires a software architectural difference. Indeed, a cloud architecture can achieve the required telco 99.999% (Five-nines) or even higher, but it requires the application to be deployed across a redundant, distributed infrastructure, which the telecom network achieves internally within its dedicated hardware boxes. Thus, the internal redundancy and decoupling can be achieved by designing the NFs (like AMF, SMF) to be stateless and distributed as microservices across multiple fault domains. The failure of one instance is immediately and automatically replaced by another. This minimizes the service outage to milliseconds, maintaining the high availability of the end-to-end path (ITU standard).

IV. Brief On Key Architectural Components For Network Access And Service Availability

Telco mobile architectures evolved in their design of NFs from proprietary dedicated telco equipment implementa-

tions to emerging virtualized software-based telco equipment functions. The hardware telco equipment approach is designed to guarantee a high level of availability, as explained in the previous section. Other key architectural components are used for this same objective and are explained in the following.

Traditional Hardware equipment redundancy: The biggest hurdle for redundancy in traditional telecom networks was the tight coupling of network functions (like a firewall, router, or load balancer) to a proprietary, vendor-specific hardware appliance. If the hardware failed, the function failed.

Software redundancy: Specifically, through Network Functions Virtualization (NFV)—enabled software redundancy by completely decoupling network functions from dedicated hardware, allowing them to be managed like any other cloud application.

Modularity: 5G Core is built as a Service-Based Architecture (SBA), where NFs are implemented as virtualized, loosely coupled microservices (e.g., AMF, SMF, Policy Control Function (PCF), etc.). These NFs interact via standardized Representational State Transfer (REST) APIs, typically over HTTP/2. This modularity allows for individual functions to be deployed, scaled, and upgraded independently.

Load Balancing: Functions like AMF and SMF handle signaling and session management. While they remain logically centralized or regional, their microservice-based deployment facilitates load balancing and fault tolerance across distributed cloud infrastructure.

Distribution / The Decentralized Edge: The User Plane Function (UPF) is the primary component enabling core network decentralization. The UPF is responsible for packet routing, forwarding, and QoS enforcement. UPFs are designed to be deployed close to the edge of the network, often in regional data centers or at the Mobile Edge Compute (MEC) platform, far from the central core data center. The technical benefits of decentralization can be sensed in Ultra-Reliable Low Latency Communications (URLLC) services, where local data offloading via distributed UPFs minimizes Round-Trip Time (RTT) (URLLC services target RTTs < 1 ms). It can also be sensed in Massive Machine-Type Communications (mMTC), where scalability is needed; it is achieved by the distributed Cloud-native microservices, which allow NFs to be instantiated and scaled elastically based on demand needed in mMTC.

Network Slicing: Decentralization supports the creation of isolated, customized network slices, each with its own set of UPFs optimized for specific service requirements (e.g., a slice for Enhanced Mobile Broadband (eMBB) demanding high throughput vs. a slice for URLLC demanding high availability).

Local Data Offload: By placing the UPF near the User Equipment (UE), traffic intended for local services (e.g., smart factory control, autonomous driving) can be locally offloaded to the Data Network (DN) without traversing the central core. This dramatically reduces end-to-end latency.

V. User And Control Plane Resilience Related Aspects

A. User Plane

The decentralized core, especially the distribution of the UPF, directly impacts the ability of 5G networks to meet the most demanding ITU availability objectives (e.g., "six-nines" or 99.9999%) as required in connected factories for critical communications. In fact, the ITU-T Recommendations (G.827.1) set stringent objectives for availability performance for end-to-end digital path; for 5G URLLC services, the availability objective can be as high as 99.9999% (Six-nines), which permits a total downtime of only 31.5 seconds per year.

Decentralization enhances availability by addressing the key sources of downtime, which is the elimination of a single point of failure. In a centralized core, a failure of the core gateway (like the 4G Packet Data Network Gateway (PGW)) impacts all users served by that central site. In the 5G Core (5GC), the distributed UPFs mean that a fault at one edge location only affects the localized area (e.g., a single factory). The failure domain is drastically reduced. This localized failure does not trigger a widespread "Unavailable State" as defined by the ITU standards (Rec. G.826/G.827) [9], [10].

One can argue that the software-based approach, as cloud-native functions allow for rapid spin-up and replacement of failed NFs (e.g., containerized UPFs) in minutes, or even seconds, rather than hours. In addition, geo-redundancy can be achieved by deploying UPFs in multiple physically separate edge clouds. If one site fails, traffic can be rerouted to an adjacent UPF without significant service interruption, ensuring the link remains in the available state.

For the most critical services, the 5GC and future 6G Core (6GC) architecture can support the establishment of dual Packet Data Unit (PDU) sessions or redundant user-plane paths from the UE to the DN through different UPFs. This mechanism, sometimes called end-to-end user plane redundancy, is a direct strategy to achieve the extremely high reliability levels (six-nines: 99.9999%) required by URLLC, as a single failure in one path is masked by the operational status of the parallel paths.

B. Control Plane

While the UPF is decentralized, the Control Plane (CP) NFs (AMF, SMF) still maintain session state following a centralized approach. The use of microservices and the SBA ensures that control plane NFs can be horizontally scaled across multiple instances, often using Kubernetes [11], which provides automatic monitoring and rapid failure recovery, further minimizing the duration of any downtime that could affect the overall service. This means that

the 5G decentralized core shifts availability assurance from a single, high-capacity core to a mesh of highly redundant, rapidly recoverable, and geographically distributed UPFs. This architectural resilience is the primary mechanism by which 5G aims to fulfill the ITU's most stringent "six-nines" availability requirements for next-generation critical services. In the future 6G networks, function distribution in both the user and control planes needs to be further investigated for better network resilience.

In fact, while the UPF is the most geographically distributed core element, the CP functions themselves are logically separated and implemented as a collection of independent microservices. This separation allows for granular distribution, independent scaling, and high resilience across multiple data centers (regional and central). Some of the key CP NFs in 5G/6G, which interact via the Service-Based Interface (SBI), are:

- AMF handles UE registration, connection management, and mobility. Multiple AMF instances are deployed regionally. A single UE is typically registered with one AMF, but the AMFs in a pool share a load, and a failure in one AMF is contained to its connected UEs, which can swiftly re-register with another.
- SMF manages the life cycle of PDU sessions (data connectivity), assigning IP addresses, and selecting/-controlling the UPF. Instances can be regional to support local services, such as MEC, or centrally to handle general internet traffic. The key requirement is the ability to independently scale based on session load.
- Authentication Server Function (AUSF) performs authentication of the UE. It can be centralized or regionally distributed, leveraging a common backend data layer.
- PCF provides policy rules (QoS, charging) to the AMF and SMF. It is often centralized to maintain a unified network policy, but edge instances can exist for specific local services.
- Unified Data Management (UDM) stores subscriber data, subscription profiles, and security credentials. Although the function itself might be accessible regionally, the actual data (HSS/HLR equivalent) is typically synchronized across a highly resilient, geo-redundant database layer.
- Network Repository Function (NRF) acts as a service registry, allowing CP NFs to discover and communicate with each other. It is essential for SBA; its availability is paramount. It is deployed with strong redundancy and often distributed to serve regional NF pools.

VI. Control Plane Resilience Mechanisms

The inherent design of the 5GC CP using cloud-native principles is what guarantees its high resilience, thereby allowing the network to meet demanding ITU availability targets.

A. SBA and Microservices

Stateless Design (Compute/Storage Separation): Most CP NFs, such as the AMF and SMF, are designed to be stateless (or soft-state), meaning that persistent user and session information is stored in a separate, centralized, and highly redundant data layer (e.g., the UDM/UDR/AUSF/PCF).

Benefit of Resilience: If a CP NF instance fails (a container crash, a server outage), the compute instance can be instantly terminated and replaced by a new instance in the cloud without losing the critical session state. The new instance simply retrieves the state from the dedicated data layer, resulting in near-zero service interruption time and minimal impact on the ITU availability metric.

Loose Coupling and Failure Isolation: Each NF is a loosely coupled microservice. The failure of one NF (e.g., a software bug in an SMF instance) only affects that specific function and does not cascade to the entire CP. This compartmentalization minimizes the scope and duration of any outage, drastically reducing the Unavailable Time defined by ITU-T Rec. G.827 [10].

High Availability and Redundancy: Extensive redundancy and intelligent traffic distribution of NFs in cloud-native realization of 5GC CP offers intrinsic high network availability.

Horizontal Scaling and Redundancy Pools: CP NFs are deployed as multiple identical instances across virtualized infrastructure (VMs/Containers). Standard cloud orchestration tools (like Kubernetes) ensure that if an instance fails, the orchestrator automatically reroutes traffic and launches a replacement. This is an $N + K$ redundancy model (where $K > 1$ spare instances handle failures of N working instances), which is far more efficient than the traditional $N + 1$ hardware-based redundancy.

Service Communication Proxy (SCP): The SCP acts as a mandatory intermediary for all SBI traffic in some deployments. It is a critical enabler for resilience.

Load Balancing: The SCP distributes signaling traffic (HTTP/2 requests) across multiple available NF instances, preventing any single instance from becoming overloaded and failing.

Alternate Routing: If an SCP detects that a specific NF instance is unavailable or slow (outlier detection), it can immediately reroute subsequent requests to an alternate, healthy instance. This mechanism ensures service continuity, which is essential for meeting the continuous availability criteria.

B. Enhanced Recovery and Management

Automated Self-Healing: The entire cloud-native CP is managed by a closed-loop automation system. Performance degradation or failure detection triggers an automated response, such as auto-scaling (creating new instances), auto-healing (restarting failed instances), or

rerouting traffic. This significantly reduces the Mean Time To Recovery (MTTR), a key factor in maximizing the "number of nines" availability.

Canary Deployment and A/B Testing: New software versions or patches for an NF can be rolled out to a small subset of instances (canary). If issues arise, traffic can be instantly rolled back or rerouted to the old, stable version. This ensures updates do not cause a full-network outage, maintaining high availability during maintenance windows.

In summary, the 5G/6G CP's resilience is a direct outcome of its distributed, cloud-native architecture, where the stateless design, microservice isolation, and automated redundancy via the NRF and SCP ensure that any localized failure is detected and mitigated within seconds, maintaining the high availability required by modern ITU standards.

VII. SimpleRAN Project Scenario and Distributed Core

As part of the SimpleRAN project [12], we are working on the Open Radio Station (ORS) [13] provided by RapidSpace [14]. Each ORS has one core network for a single base station. The distributed architecture shown in Figure 2 ensures resilience to disasters affecting the backhaul network or any ORS base station, eliminating a single point of failure.

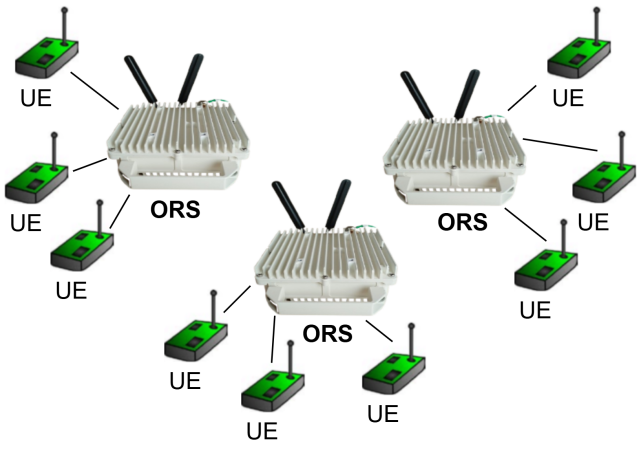

Fig. 2: Distributed Core Network Using ORS

A. Challenge and Consideration: Handover

A central core network can be used to achieve traditional handover. Furthermore, to increase network resilience, the central core can be deployed across multiple hosts. This approach has been validated to ensure that when a UE moves from one ORS to another, its Transmission Control Protocol (TCP) session remains uninterrupted. A logically centralized core, even if physically distributed, can still cause total ORS connectivity loss if the central core fails—making it unsuitable for applications requiring resilience. We are looking to improve the handover using a routing approach in a distributed ORS core network without a centralized core.

1) sroamd

We have tested seamless handover using Simple Roaming Daemon (***sroamd***) [15] for an untrusted Non-3GPP (WiFi) access using free5GC [16].

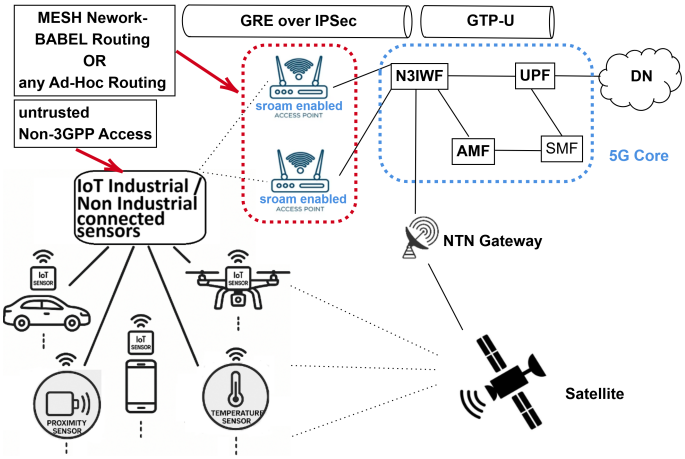

Fig. 3: Seamless Handover using sroamd

In Figure 3, every UE is allocated a unique IPv6 address, which does not change regardless of the Access Point (AP) to which it is attached. We run the Babel Routing protocol on the APs to distribute routes to the roaming clients. For the sake of simplicity, we consider two APs connected to a Non-3GPP Inter-Working Function (N3IWF). Once the Babel routing protocol is up on both APs (AP1 and AP2), we configure ***sroamd*** on each AP to enable seamless roaming for the UE between them.

VIII. Conclusion

Within untrusted Non-3GPP access, sroamd collaborates with ***hostapd*** to identify UE association and disassociation events with an access point.

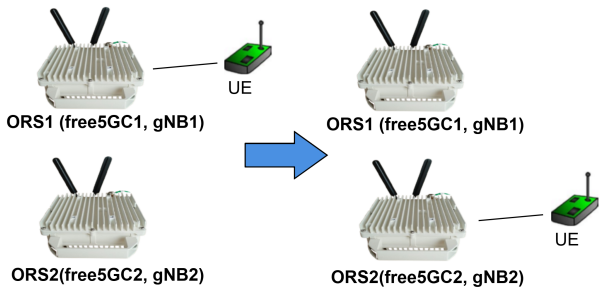

Fig. 4: UE connects with same IPv6 to ORS2

Similarly to this, we plan to modify the free5GC so that UEs always receive the same IPv6 address, regardless of which ORS they are connected to. When the signal becomes too weak, the UE detaches from one gNB and

performs handover through the NG interface, since ORSs are assumed to be independent in a truly distributed setup. The UE then attaches to a new gNB connected to a different 5GC as shown in Figure 4, but continues to use the same IPv6 address as before. This ensures that the UE maintains the same IPv6 address throughout the handover process. Including a smooth handover in 3D-type communication, considering also non-terrestrial networks, will complete the distributed approach to ensure resilience with the extensive connectivity of Non-Terrestrial Networks (NTN) in 6G.

References

[1] M. Khoshnevisan, V. Joseph, P. Gupta, F. Meshkati, R. Prakash, and P. Tinnakornsrisuphap, "5g industrial networks with comp for urllc and time sensitive network architecture," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 947–959, 2019.

[2] X. Zhang and Q. Zhu, "Ai-enabled network-functions virtualization and software-defined architectures for customized statistical qos over 6g massive mimo mobile wireless networks," *IEEE Network*, vol. 37, no. 2, pp. 30–37, 2023.

[3] I. Ahmad, F. Rodriguez, J. Huusko, and K. Seppänen, "On the dependability of 6g networks," *Electronics*, vol. 12, no. 6, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/6/1472

[4] J. Kong, I. Kim, X. Wang, Q. Zhang, H. C. Cankaya, W. Xie, T. Ikeuchi, and J. P. Jue, "Guaranteed-availability network function virtualization with network protection and vnf replication," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[5] P. Smith, D. Hutchison, J. P. Sterbenz, M. Schöller, A. Fessi, M. Karaliopoulos, C. Lac, and B. Plattner, "Network resilience: a systematic approach," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 88–97, 2011.

[6] G. L. Santos, P. T. Endo, G. Gonçalves, D. Rosendo, D. Gomes, J. Kelner, D. Sadok, and M. Mahloo, "Analyzing the it subsystem failure impact on availability of cloud services," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, 2017, pp. 717–723.

[7] S. Sharma, D. Staessens, D. Colle, M. Pickavet, and P. Demeester, "Openflow: Meeting carrier-grade recovery requirements," *Computer Communications*, vol. 36, no. 6, pp. 656–665, 2013, reliable Network-based Services.

[8] M. Aldossary, "A review of dynamic resource management in cloud computing environments," *Computer Systems Science and Engineering*, vol. 36, pp. 461–476, 01 2021.

[9] "ITU-T Recommendation G.826: End-to-end error performance parameters and objectives for international constant bit-rate digital paths and connections," International Telecommunication Union - Telecommunications Standardization Sector (ITU-T), 2002, available: https://www.itu.int/rec/T-REC-G.826.

[10] "ITU-T Recommendation G.827: Availability performance parameters and objectives for end-to-end international constant bit-rate digital paths," International Telecommunication Union - Telecommunications Standardization Sector (ITU-T), 2003, available: https://www.itu.int/rec/T-REC-G.827.

[11] S. D. A. Shah, M. A. Gregory, and S. Li, "Cloud-native network slicing using software defined networking based multi-access edge computing: A survey," *IEEE Access*, vol. 9, pp. 10 903–10 924, 2021.

[12] "Simpleran," https://www.simpleran.org/, SimpleRAN Initiative, 2025, accessed: 2025-12-06.

[13] Rapid.Space, "Rapid.space ran – radio access networks," https://www.rapid.space/products/ran, accessed: 2025-09-17.

[14] ——, "Rapid.space: Fully open 5g edge cloud," https://www.rapid.space/, accessed: 2025-09-17.

[15] J. Chroboczek, A. Décimo, and Joanne, "sroamd: Simple roaming daemon," https://github.com/jech/sroamd, 2025, accessed: 2025-09-17.

[16] free5GC Developers, "free5gc: 5g core network implementation in go," https://github.com/free5gc/free5gc, 2025, open-source project. [Online]. Available: https://github.com/free5gc/free5gc