Pertanika J. Sci. & Technol. 34 (2): 913 - 950 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

*Review Article*

# A Critical Review of User-centric QoS Provisioning Frameworks in 5G Networks

**Wai Leong Pang[1], Wen Hao Anselm Chow[1], Hui Hwang Goh[1], Swee King Phang[1], and Kah Yoong Chan[2,3]***

[1]*School of Engineering, Faculty of Innovation and Technology, Centre for Sustainable Societies, Taylor's University, 47500 Subang Jaya, Selangor, Malaysia*
[2]*Faculty of Artificial Intelligence and Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia*
[3]*Centre for Advanced Devices and Systems, COE for Robotics and Sensing Technologies, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia*

## ABSTRACT

5G networks provide the Quality of Service (QoS) to various services, including Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC) with different requirements. However, the existing 5G network deploys the static 5G QoS Identifiers (5QI) framework, which provides a basic level of services differentiation and supports the class-based prioritisation. The static 5QI fails to support the dynamic and user-centric QoS requirements. This paper reviews user-centric QoS provisioning schemes that focus on the Software-Defined Networks (SDN), Machine Learning (ML), and network slicing schemes. A structured evaluation methodology is used to analyse the existing works in terms of architecture, mechanisms, QoS metrics, scalability, and real-world feasibility. The reviews discover that SDN improves programmability and policy-based control, ML predicts the traffic adaptively, and network slicing provides service isolation. However, user-centric schemes can increase computational load, raising controller overhead and latency. This paper identifies the research gaps in the user-centric schemes related to scheme readiness, scalability, security, and heterogeneity. The paper also proposed a direction for deploying user-centric QoS schemes in real 5G networks that support future 5G enhancement that covers the adaptive, efficient, and personalised QoS provisioning.

*Keywords:* 5G, quality of service, software-defined networking, user-centric

ARTICLE INFO

*Article history:*
Received: 26 February 2025
Accepted: 25 March 2026
Published: 17 April 2026

DOI: https://doi.org/10.47836/pjst.34.2.14

*E-mail addresses*:
Waileong.pang@taylors.edu.my (Wai Leong Pang)
0361738@sd.taylors.edu.my (Wen Hao Anselm Chow)
jonathan.goh@taylors.edu.my (Hui Hwang Goh)
sweeking.phang@taylors.edu.my (Swee King Phang)
kychan@mmu.edu.my (Kah Yoong Chan)
* Corresponding author

e-ISSN: 2231-8526

## INTRODUCTION

The environments, network speed, reliability, and service consistency play a crucial role in ensuring user satisfaction across different applications in existing mobile networks. The latest 5G wireless networks improve network speed and capacity compared with previous generations. Consequently, 5G can support various user-centric and mission-critical applications, including Virtual Reality (VR) applications, autonomous vehicles, smart manufacturing, and real-time applications. However, there is still a challenging task to ensure a stable QoS provisioning for various types of services.

The previous cellular networks, with static configurations and specialised hardware, fail to support the current high complexity and scalability of network conditions (Haji et al., 2021). The current mobile networks fail to provide QoS to all users, especially premium customers, when the networks are overloaded. Conventionally, the 5G network relies on fixed QoS levels for traffic management, which are pre-defined and cannot be changed quickly in the event of varying network conditions. The current design of QoS allows for up to nine levels of priority for varying telco service provided, designed to prioritise traffic across different service types, including voice, video, and best-effort data transfers. Although this approach can work well under normal conditions, it may break during congestion events, resulting in higher latency, packet loss, and general degradation of user experience. Therefore, the key technical challenge is not just to define QoS classes, but also to ensure QoS enforcement stability and efficiency under dynamic load conditions.

As summarised in Table 1, the existing QoS framework for 5G networks allocates bandwidth resources according to the predefined priorities of the services. Although the QoS framework based on priorities can provide a certain level of service differentiation, it is not fully capable of satisfying the new requirement for personalised QoS provisioning, especially for the premium segment that requires high availability, reliability, and integrity of the services provided. Regarding the QoS in the 5G network architecture, the framework specifically supports a flexible configuration of 64 QoS Identifiers known as 5QI (5G QoS Identifier) for differentiating the services according to utilisation demands (Debnath et al., 2023). The 5G QoS architecture essentially supports a completely different configuration than the existing 4G architecture. It facilitates a total of nine Quality Class Identifiers (QCIs) of resources for the services, resembling the nine levels of priorities. The levels of priorities in the 5G architecture are essentially determined by the 3rd Generation Partnership Project (3GPP) for standardising the overall 5G communication framework (Hussain et al., 2020; Shrivastava et al., 2022). As shown in Table 1, delay-sensitive services such as conversational voice and real-time gaming are assigned higher priority and stricter delay constraints to preserve interactivity, while less delay-sensitive services such as non-conversational video are allocated comparatively relaxed QoS budgets. This structure design allows for differentiation in service level establishment. However, the issue

Table 1
*Standardised 5QI mappings*

| 5QI | Service type | Resource type | Priority | Packet Delay Budget | Packet Error Rate |
|---|---|---|---|---|---|
| 1 | Conversational Voice (VoIP) | Guaranteed Bit Rate (GBR) | Highest | 100ms | $10^{-2}$ |
| 2 | Conversational video | GBR | High | 150ms | $10^{-3}$ |
| 3 | Real-time gaming | GBR | High | 50ms | $10^{-3}$ |
| 4 | Non-conversational video streaming | GBR | Medium | 300ms | $10^{-6}$ |
| 5 | IMS signalling | Non-GBR | Medium | 100ms | $10^{-6}$ |
| 6 | Voice/video signalling | Non-GBR | Low | 100ms | $10^{-6}$ |
| 7 | Live streaming | Non-GBR | Medium | 100ms | $10^{-6}$ |
| 8 | Low-latency data (Interactive gaming/ Buffered streaming) | GBR | High | 10ms | $10^{-6}$ |
| 9-63 | Mission-critical applications | Non-GBR/GBR | Variable | Variable | Variable |

*Note.* Adapted from Al-Shammari et al. (2018); Hoyhtya et al. (2018); Hussain et al. (2020); and Trifan et al. (2015)

now is how to ensure optimum bandwidth utilisation with optimum QoS performance in a real-world scenario, especially in a network with unbalanced network utilisation (Khan, 2022). It is important to note that the static QoS mapping may not consider the dynamic user behaviour or traffic patterns. Especially in cases when premium users who are guaranteed the highest priority and bandwidth do not utilise their granted resources. The network bandwidth will be underutilised, while the performance of non-premium users will be low due to bandwidth constraints. This highlights the need to shift QoS provisioning from strictly service-class differentiation toward more adaptive and user-centric resource allocation strategies.

The 5G QoS Identifier (5QI) is used to provide service differentiation in the existing 5G framework. The priority and resource allocation are assigned according to the service class (eMBB, URLLC, mMTC). All the users receive the same QoS for the same service class, and this is clearly a service-centric framework. However, the QoS for the user-centric framework is delivered to the individual users according to the users' priority, subscription tier, real-time user experience, etc. The user-centric QoS can be deployed on the programmable SDN with the ML-powered prediction model to adaptively manage the limited resources.

In this review paper, various algorithm-based and Machine Learning (ML)-assisted approaches have been explored to analyse network traffic and improve bandwidth allocation under both overloaded and non-overloaded scenarios. These approaches are generally integrated with SDN, which allows for programmatic control to dynamically control traffic. For instance, the closed-cycle process for traffic management, where traffic can be monitored in real time to adapt policies to traffic conditions, has already been validated in research using machine learning to analyse traffic patterns to determine overload conditions in traffic (Hyder & Lung, 2018). As opposed to the previous reliance, the method aims to facilitate the adaptation in the allocation of prioritising the bandwidth based on the time-varying demands.

Several works further indicate that dynamic allocation schemes have been reported to enhance resource utilisation by redistributing unused high-priority resources when demand is lower and still granting service guarantees at increased demands. For example, Nikolaidis et al. (2023) discuss per-user QoS guarantees in network slicing scenarios, illustrating how more flexible provisioning of resources across users can take place. The literature indicates that the combination of SDN and ML enhances QoS flexibility and bandwidth efficiency but also introduces practical constraints, which include control-plane overhead, inference latency, and scalability challenges in dense 5G environments (Amin et al., 2021). Conveying the rapidly changing demands of various modern network services, 5G networks will have to be highly flexible, scalable, and dynamic. As compared to traditional networks, which are bound to their configuration and are only supported through hardware-based proprietary solutions such as routers and firewalls. SDN is a relatively new paradigm that has come to offer a novel approach in which the control plane is decoupled from the data plane (Etxezarreta et al., 2023). The integration of SDN reduces the Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) in 5G networks. There are also opportunities for managing dynamic resources and services in response to real-time conditions.

The evolution from Long Term Evolution (LTE) to 5G improved the service quality in mobile communications. Resource allocation is a promising solution to support the surge in mobile data traffic under limited licensed spectrum. The coexistence of LTE and Wi-Fi has a significant impact on performance in terms of throughput degradation and energy efficiency loss under various modulation and coding schemes (Yaqoob et al., 2020). A robust mobile network system should protect the QoS of the real-time streaming applications (Yaqoob et al., 2019).

Beyond technical performance, the QoS stability problem has a growing global impact on how 5G services are perceived and monetised. As increased deployments of 5G technology within urban cities and industrial sites are underway, the deterioration of service within such scenarios of increased congestion stands to have an immediate impact on the end consumer with respect to the services they can access within the 5G environment,

including more robust applications such as smart manufacturing, the Internet of Vehicles, and smart cities. As such, the issue of scalable QoS within the 5G environment continues to become a determining aspect in the successful rollout of 5G technology across the globe.

In addition to QoS classification challenges, practical QoS performance is also influenced by coexistence effects and resource allocation efficiency in multi-access and multi-user environments. Furthermore, the involvement of trade-offs in resource allocation problems of Non-Orthogonal Multiple Access (NOMA) systems for 5G networks is focused on where researchers investigate how a balance between throughput maximisation and fairness can result in optimally allocating user data rates by using sophisticated algorithms like Integer Linear Programming (ILP) and Particle Swarm Optimisation (PSO) (Abuajwa et al., 2022). A fair bandwidth distribution to the users through the resource allocation techniques in poor channel conditions, regardless of the types of services. The enhanced exponential rule-based scheduling algorithms are proposed to enhance the performance of the real-time traffic for the LTE systems (Yaqoob et al., 2020). The latency and QoS guarantees of the real-time data services are improved to ensure a stable performance for these services. The effective QoS provisioning for real-time services depends on QoS classification, practical scheduling and resource allocation strategies under real network constraints.

5G and beyond systems improve the efficiency of QoS management and resource allocation amid multi-access coexistence. Bandwidth-optimisation techniques can improve reliability but introduce trade-offs. In this review, SDN, ML and adaptive QoS-driven frameworks are examined with a focus on their strengths, limitations, and outstanding research challenges. Prior studies report theoretical QoS improvements that depend on deployment feasibility and operational constraints.

The limitations observed in 5G QoS deployment are largely associated with maintaining reliable service guarantees under fluctuating traffic conditions, rather than with the absence of formal QoS definitions. Static service class prioritisation offers an initial structure, but real-world performance varies according to user behaviour, congestion, mobility, and competing resource demands among services. Hence, the adaptive QoS provisioning needs to be considered in the context of the implementation challenges, such as controller scalability, signalling overheads, and decision latency, apart from the reported improvements.

The large-scale commercial 5G network deployed the static 5QI-based management system, but the existing works have shown that the performance of the dynamic QoS provisioning schemes is better than the static 5QI-based approaches. The experimental test shows that the dynamic scheme provides a lower packet loss rate and a more stable bitrate for the video streaming in an outdoor environment compared to the static 5QI scheme (Ito et al., 2025). Another work that incorporated adaptive network management

has shown a significant improvement in throughput and packet drop rate compared to the static approach (Yin et al., 2020). The dynamic SDN resource allocation scheme also demonstrated a significant improvement in bandwidth utilisation compared to the static QoS approaches (Yang & Tsai, 2024). These show that the dynamic approaches have significant advantages over the static 5QI-based schemes in the 5G network.

**Addressing QoS Challenges in 5G**

The motivation for this review results from the growing challenge of providing consistent QoS service delivery within modern 5G networks. As 5G infrastructure is used to deliver mission-critical services like autonomous transportation systems and immersive virtual environments, the tolerance for service delivery instability is much lower. Yet, providing consistent QoS service delivery remains an ongoing challenge within operational 5G networks, especially against dynamic fluctuations, congestion due to mobility, and burst-based demand patterns.

One of the major drawbacks of the conventional QoS approaches is that they are based on a set of static rules of prioritisation that are not effective enough to deal with real-time dynamic changes in traffic and user behaviours. In fact, during congestion, inflexibility in priority rules can cause less-than-optimal use of available system resources and service performance, especially when there are a variety of applications competing for limited radio and core network resources. The issue is therefore not just about service classification anymore. It is also about service allocation strategies that are not only adaptive and context-aware but also provide a certain level of fairness and reliability between users. To address these limitations, prior studies have investigated adaptive and smart QoS management approaches that integrate SDN and ML. SDN has depicted its merits in terms of making flexible programmable traffic management possible for enforcing dynamic policies according to the real-time evolving states of networks. In parallel, ML techniques have been applied to enhance traffic awareness, enabling congestion prediction, traffic classification, and decision support for prioritisation strategies. Although these existing solutions demonstrated strengthened adaptive capabilities over traditional static management techniques, their own constraints on the control plane delay over SDN, the processing latency during ML-assisted decision-making processes, as well as scalability in high traffic density, are depicted.

Consequently, this review critically examines SDN, ML, and dynamic QoS-based frameworks for user-centric QoS provisioning in support of 5G networks through the critical assessment of mechanisms, strength and gaps associated with various approaches. The objective is to determine the suitability of different frameworks under service requirements and network scenarios, along with the key trade-offs and open issues to be addressed for effective QoS assurance in 5G and beyond wireless communication systems.

Existing works that simulated ML-powered SDN and network slicing in 5G have reported the effectiveness of the proposed solution. These works provide service-centric QoS provisioning to the services without considering the users' privileges and priorities. There are limited studies on the impact of the user-level behaviour on the 5G QoS provisioning. A comprehensive review is carried out in this paper to evaluate the potential of the SDN, ML and slicing-based framework from a user-centric perspective. The review reported the ability of these solutions and the challenges in supporting the user-centric driven framework for the 5G network.

**Paper Outline**

This review paper adopts a structured and comparative approach to analyse existing research on user-centric QoS provisioning in 5G networks. Generally, the review is focused on SDN-based frameworks, ML-assisted QoS intelligence, and dynamic QoS and resource management mechanisms, as these represent the most common and promising directions for enhancing QoS adaptability under real-world traffic variability.

The paper is organised as follows. Firstly, the background and motivation for user-centric QoS provisioning in 5G are introduced to highlight the limitations associated with the traditional QoS support under congested and heterogeneous environments. Next, the review examines the SDN-enabled QoS control techniques, emphasising how programmability and centralised orchestration can support flexible bandwidth management. Then, this is followed by an observation on solutions based on ML, and their significance associated with enhanced traffic knowledge and prediction-based QoS management is explained. Exploring dynamic QoS solutions is also one of the steps taken in this review.

Lastly, a comparative synthesis is presented to summarise significant technical trade-offs among all the studied solutions in terms of performance enhancement, scaling limits, control-plane overhead, inference delay, and deployment simplicity. The rest of this paper is concluded with an overview of existing research gaps and a discussion of future studies related to deployable and user-oriented QoS in 5G and future communication systems.

A structured screening process was applied to ensure the review is systematic instead of descriptive. The review covered the works reported since 2017 in the areas of SDN, ML-powered, network-slicing, adaptive scheduling and user-centric QoS provisioning solutions. The QoS-related solution that is deployable on 5G or beyond is reported for each study. The mode of evaluation, either through simulation or experiment, is provided. The QoS metrics and the significant performance likes delay, throughput, fairness, and packet loss are analysed. The properties of the 5G network, like model architecture (SDN, ML, network slicing), scalability, adaptability, feasibility and user-centric QoS are evaluated for all the reviewed papers.

## Trends and Evolutions of User-Centric QoS Research

The development of user-centric QoS provisioning in wireless networks has witnessed colossal evolution from 2G to 5G and is now expanding into the realm of beyond 5G (B5G) technologies. The early mobile networks, such as Global System for Mobile Communications (GSM) and Wideband Code Division Multiple Access (WCDMA), were voice-oriented with less support for data. In these networks, QoS was static and primarily preconfigured, with static bandwidth reservations and minimal differentiation by user or application. The introduction of LTE and LTE-Advanced brought about QoS provisioning as a concept. As explained in (Toor et al., 2019) Improvements to the Random-Access Procedure (RAP) in LTE allowed devices to transmit connection requests using contention-based and contention-free procedures. Access Class Barring (ACB) and Extended Access Barring (EAB) were added to control network congestion for Human-Type Communications (HTC) and Machine-Type Communications (MTC). There is better control of device access and a reduction in delay.

In 5G networks, the transition is more tense; technologies such as the grant-free transmission, NOMA, and Pattern Division Multiple Access (PDMA) were employed to facilitate massive connectivity and ultra-low latency, as demonstrated in simulation evaluations described in (Toor et al., 2019). The protocols allow devices to transmit data without the need for a grant from the base station (gNodeB), which greatly reduces the overhead while making the protocols more effective for use in environments of high density. As explained in (Sufyan et al., 2023), 5G leverages an extremely programmable and flexible system design that supports context-dependent, user-preference, and type-of-service-sensitive dynamic QoS adaptation. Edge computing, massive Multiple-Input Multiple-Output (MIMO), and network slicing complement support 5G to ensure real-time, user-level quality guarantee. However, with the sudden increase in demands from Internet of Things (IoT) ecosystems and immersive applications like the Metaverse, VR and Augmented Reality (AR), even 5G is insufficient. As a result, the research strategy is progressively shifting towards B5G technology with ultra-dense networks, an AI-driven strategy, and edge-cloud coordination, having the capability to further improve the user-centric QoS as described in (Sufyan et al., 2023). Besides that, Mahmood et al. (2021) indicate that mission-critical services based on industrial applications, such as Factory 5G, necessitate deterministic QoS, with a tendency towards vertical-specific QoS design for manufacturing and automation sectors. Moreover, the evolution of mobile communication networks and how 5G technology paves the way for B5G use cases with even more stringent latency, reliability, and flexibility requirements. In general, the evolution of user-centric QoS is a shift from fixed best-effort delivery to intelligent, adaptive, and service-aware provisioning. The current trend is centred on the use of AI/ML in prediction and real-time decision-making, adaptive access methods, and architectural shifts that extend the decision

logic to the network edge. The goal is to provide reliability, low latency, and optimal user experience in highly heterogeneous environments.

Despite the reviewed evolution, it indicates that the QoS-related problems have changed rather than disappeared. Although techniques such as grant-free access and advanced multiple access techniques reduce the overhead of scheduling and connectivity, new problems arise in contention resolution, fairness, and QoS guarantee. Thus, a major research gap that remains to be addressed is the design of user-centric QoS frameworks that can meet strict latency and reliability requirements while being scalable and computationally tractable.

**Software-defined Network**

Literature on the QoS provisioning for 5G networks highlights the need for SDN, Network Function Virtualisation (NFV), and ML to facilitate the adaptation of resource management to satisfy the heterogeneous QoS demands of the applications. The 5G network is equipped with various applications such as URLLC, eMBB, and mMTC, which have different QoS requirements, such as latency, throughput, and reliability (Khan, 2022; Kunasegran et al., 2025a). Thus, the QoS paradigms may not work, and there is a need to implement a flexible control plane architecture to satisfy the applications' demands.

SDN revolutionises conventional networking by introducing a separation between the control and data planes of communication to allow for centralised programmatic controls for managing network traffic according to predefined policies. As shown in Figure 1, network administrators can leverage intelligent network management techniques to dynamically set up QoS policies, rather than depending upon conventional static configurations of networks (Kunasegran et al., 2025b). In a typical 5G network setup, due to dynamic changes in user traffic requirements for bandwidth over short intervals of time, it is necessary to integrate SDN-based network control functionalities. For instance, bandwidth-intensive applications such as video-on-demand services, interactive applications such as online gaming that require lower latency or applications which requires forestalling over other applications. These can take advantage of the network monitoring provided by SDN in supporting QoS depending on the network loads (Akinola et al., 2024). Figure 2 shows the traffic classifications of SDN. It highlights the importance of dynamic QoS rearrangement. SDN was applied to optimise the bandwidth allocation of the 5G network. Similarly, SDN controllers have been interfaced with optical networks to improve bandwidth allocation efficiency and adaptability of traffic for hybrid environments (Yang & Tsai, 2024).

Within the SDN architectures, communication across network layers is enabled through Northbound and Southbound Interfaces (NBI and SBI), which connect applications, controllers, and forwarding elements. The NBI in SDN enables applications and management systems to specify service-level QoS requirements,

allowing network configurations to adapt to real-time operating conditions (Alabarce et al., 2020). These requirements are then mapped by the SBI into device-level instructions for flow-level policy enforcement. Although this interface design improves programmability, it also increases implementation complexity, especially in dense networks that require frequent reconfiguration. Moreover, SDN deployment itself presents practical challenges. As observed by Jiang et al. (2023), configuring appropriate SDN functions can be cumbersome, prompting the use of ML-based performance prediction to estimate metrics such as Round-Trip Time (RTT), Switch-to-Controller (S2C) Traffic, and Controller-to-Controller (C2C) traffic before deployment.

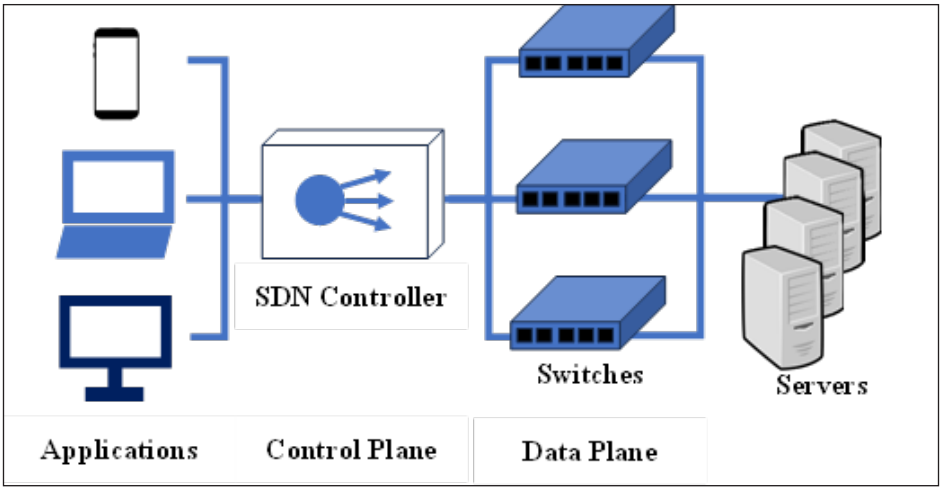


*Figure 1.* Network architecture of SDN

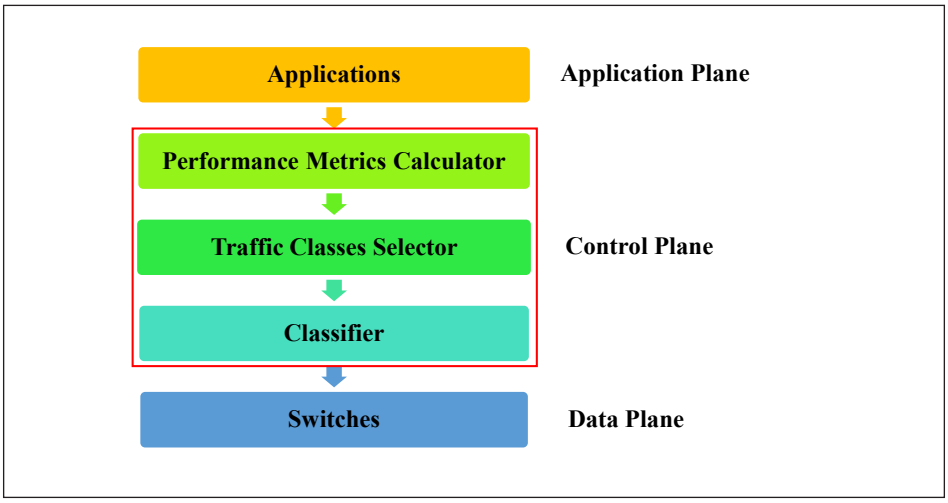


*Figure 2.* Traffic classifications of SDN

Their findings show that a neural network boosting regression model (NNBoost) achieves superior prediction accuracy compared with baseline ML and deep learning models such as RF, AdaBoost, XGBoost, LightGBM, DNN, and LSTM. This suggests that QoS provisioning in SDN networks needs to consider not only the optimisation during runtime, but also the assessment before deployment to avoid configuration risk, enhance scalability, and reduce trial and error in practice (Jiang et al., 2023).

Overall, SDN-based QoS frameworks provide a solid foundation for user-centric and dynamic QoS provisioning in 5G networks due to the centralised control, programmability, and policy-driven adaptability they offer. However, the reviewed analysis identified the critical trade-offs involved in the integration of SDN, which invariably include an increase in the control plane overhead, interface complexities, and scalability challenges, especially when the density of devices is high. All these challenges suggest that the QoS solutions based on SDN need to be critically evaluated based on the enhancements being offered for QoS and their operational feasibility, readiness, and stability.

SDN can support dynamic QoS provisioning and is feasible for the dense 5G deployment. Many works show the ability of the SDN framework in enhancing the network performance. However, the increased amount of traffic overloaded the QoS provisioning scheduler and potentially shifted the bottleneck from data plane to control plane (Shameli & Rajkumar, 2026). The experimental test result shows that SDN does not shift the complexity, but it provides QoS to support the controllers (Yang & Tsai, 2024). The conditional feasibility of the SDN supports the QoS provisioning in 5G, but a control mechanism should be deployed to manage the traffic from the dense network. A lightweight model is preferable for the edge deployment, and the lightweight SDN controller is suitable for handling the QoS at the edge. The solution can be strengthened by deploying an ML model to provide local prediction at the edge. The system can be managed and coordinated at a high level to maintain consistency among the vendors. This architecture avoids the bottleneck at the server and still provides QoS adaptively.

## User-centric Quality of Service (QOS) Management

In the context of 5G networks, user-centric QoS management is seen to go beyond strict service class prioritisation and to better align QoS service delivery with individual user requirements and context. This direction has become even more important with the support of highly heterogeneous applications with different sensitivities to service performance, for which QoS information alone is not sufficient to ensure service stability with frequently changing network states. One of the means to achieve user-centric QoS optimisation is to introduce Radio Access Network (RAN) flexibility using Functional Splitting (FS), where different functions are split between central and edge entities, depending on latency and bandwidth limitations.

This flexibility can increase reliance on QoS for latency-sensitive applications by efficiently determining processing offload and enhancing system responsiveness to changing load conditions. At the same time, FS adds complexity to coordination because of the requirement to ensure persistent splitting strategies in mobility and traffic variations to exclude overhead costs of reconfiguration that might impede performance gains. Thus, various implementations of using FS have proven to make RAN more flexible for performance improvement (Ciceri et al., 2024).

In addition to RAN-level enhancements, the provision of user-centric QoS has also been explored for the integration of the transport network over optical access systems, such as Passive Optical Networks (PON), that provide more efficient bandwidth allocation for a hybrid 5G setting. Optical-based technologies have the potential to deliver better QoS assurance for supporting bandwidth-demanding and delay-sensitive services through the provision of additional network transport capacity, especially with the integration of edge computing technology. Other than the RAN layer, QoS support from a user perspective has also been explored with the integration of the 5G transport segments and optical access platforms, such as Passive Optical Networks, for improving bandwidth distribution and transport efficiency, especially with the potential to do so for high-throughput and delay-sensitive services with the integration of edge computing technology. On the other hand, the extent of the possible gains in QoS is highly dependent on the degree of interoperability between optical access networks and the existing network infrastructure. Legacy issues imply that the performance gains are more likely to be deployment-dependent rather than general (Dias et al., 2023). Moreover, another important set of issues in user-centric QoS management is concerned with adaptive scheduling and real-time prioritisation, specifically for dense and IoT-congested environments. The controlled scheduling is critical to help mitigate the situation, mainly by prioritising real-time traffic. Although the approaches are suitable in dynamic traffic environments, certain trade-offs are encountered with signal overhead, computational costs, and service classification. These limitations become increasingly important under high traffic density, where frequent adjustments within scheduling may affect scalability (Anjum et al., 2024; Albekairi, 2025).

The QoS and QoE are closely related but cannot be treated interchangeably. QoS evaluates the performance of the network by measuring the performance metrics of throughput, jitter, delay and packet loss rate. However, QoE focuses on evaluating the users' performance. The performance metrics for evaluating the user experience are startup delay of the application, bitrate stability, smoothness of the video streaming, and continuity of the application during handover. This review separated the QoE-based solutions that focused on 5G user experience and the QoS-based solutions that focused on protecting real-time services.

Furthermore, the concept of mobility-aware user-centric QoS management has been explored to enhance the QoS in the scenario of dynamic user mobility. Mode-switching schemes, for instance, need to be developed in such a way that they maintain the level of QoS while making decisions based on the anticipated effects of user mobility, which is essential in the scenario of applications where continuous service reliability is required despite the constant switching between modes (Baz et al., 2024). The comparative analysis reveals that the QoS management using deep learning prediction in mobility scenarios has the potential to surpass the existing modes of mobility management while making decisions regarding handover to avoid the loss of packets in real-time, despite generating overhead in the process.

The user-centric QoS approaches consistently enhance differentiation and adaptability in heterogeneous service environments through mechanisms such as RAN flexibility, hybrid architectures, and priority-driven scheduling. Nevertheless, scalability and deployment feasibility remain insufficiently addressed in many existing frameworks. Control overhead, reconfiguration volatility, and fairness trade-offs continue to limit operational adoption. Further work is therefore required to translate user-centric QoS concepts into deployable, implementation-level solutions.

The existing 5G system evaluates the QoS through the following performance metrics, like throughput, packet loss and packet delay. More efforts are needed to consider the users' fairness, mobility and network stability to support the user-centric solution. The user-level QoS provisioning metrics are incorporated in the following analyses that have high potential to improve the user experience.

**Bandwidth Allocation**

ML-assisted bandwidth allocation has increasingly been explored in SDN-based QoS management frameworks, where traffic flow detection supports dynamic prioritisation under varying traffic conditions (Khairi et al., 2021). It applies a hybrid Decision Tree-Support Vector Machine (DT-SVM) approach to improve bandwidth allocation, ensuring that higher-priority flows receive sufficient resources. Due to the implementation of ML into the SDN system, it lessens the need for manual configuration of network updates and network policies. This also allows better real-time analysis and real-time adjustments based on the network traffic conditions (Chia et al., 2025). Meanwhile, flexible bandwidth partitioning has also been discussed, where higher-priority bandwidth is split into smaller functional allocations to satisfy both latency and throughput requirements (Ciceri et al., 2024). Besides that, the usage of QoS management with the help of ML is evaluated (Raeisi & Sesay, 2023). This study applied a K-Means algorithm, which differentiates between the high-speed and low-speed users in the variable frame rate (VFR) scheme. Although the work targets autonomous vehicle QoS enhancement, its main contribution is mobility-aware bandwidth control by adjusting handover thresholds to prioritise high-speed users.

This supports more adaptive bandwidth allocation across channels, improving overall QoS consistency under varying mobility conditions. Besides, an efficient resource allocation scheme is demonstrated according to the service prioritisation, and this scheme is evaluated on a 100Gb/s Ethernet passive optical network (NG-EPON) upstream (Rawshan et al., 2024). Proportional Fairness allocation has also been proposed to distribute bandwidth according to users' real-time traffic rates (Tanuja et al., 2023). This gives users with high demands for data to receive a share of bandwidth proportionally and not completely sideline the users with lower priorities. However, while ML-driven prioritisation improves adaptability and Proportional Fairness improves multi-user stability, there are certain trade-offs with respect to responsiveness, fairness enforcement, and scalability for a highly trafficked network.

The existing works show that although the user priority and fairness are evaluated, these cannot guarantee the user experience. Hybrid solutions are proposed to reserve bandwidth for the premium users and provide best effort services to the basic users. However, the basic users may be starving when the network is overloaded, and most of the resources are utilised to support the premium users. Various approaches, such as SDN and ML-powered schedulers, are used to improve the premium users' experience. The user-centric QoS provisioning scheme must have a balance between the services and users to optimise the resource allocation.

Resource allocation can be improved using the Max-Min fairness algorithm and Weighted Fair Queuing with a preference for heavy users, besides allocating base bandwidth for other users as well. The adaptive bandwidth allocation requires a dynamic adjustment of resource distribution in response to network congestion or low-demand periods. In addition, dynamic bandwidth allocation frameworks in MIMO-based 5G systems are designed to optimise bandwidth allocation between antenna elements. As a result, spectral efficiency enhancement as well as simultaneous transmission to multiple users are achieved (Prodromos et al., 2024). Deep learning methods have been employed for estimating traffic volumes and allocating bandwidth dynamically for base station antennas. Algorithms such as the Hungarian algorithm and the Minimum Cost Flow algorithm were used to optimise bandwidth distribution. Though such strategies enhance dynamic bandwidth allocation as well as real-time traffic control, their viability can be limited by computation costs and real-time constraints, even in dense deployment. Meanwhile, a system that gives priority to the URLLC traffic in terms of bandwidth is proposed to achieve ultra-low latency, even under high traffic load (Wu et al., 2024). This dynamic allocation prevents under-provisioning for delay-sensitive applications. Regardless of that, the theoretical model proposed by the authors helps in determining the optimal amount of bandwidth required for each service type under varying conditions. Furthermore, efficient handling of the bandwidth is one of the key requirements to fulfil the requirements of the QoS for diverse classes of traffic.

The problem is more prominent when the topic is touched in the field of vehicular or high-mobility scenarios, where maintaining throughput while keeping packet loss under control is naturally difficult. To address this, Adaptive Network Coding (ANC) approaches have been investigated, where Hidden Markov Models are used to estimate packet loss behaviour and adjust coding redundancy in real time (Yin et al., 2020). By adjusting redundancy levels according to observed link conditions, these mechanisms help reduce avoidable retransmissions and improve effective throughput. In network slicing contexts, bandwidth control is often coupled with Virtual Network Function (VNF) migration to maintain service continuity through adaptive resource reassignment and reduced overprovisioning (Vidhya et al., 2025). ML-based control mechanisms react more effectively to traffic variation, while fairness-based schemes stabilise multi-user performance. URLLC-oriented approaches, in turn, are best suited to high-load scenarios that impose strict latency requirements. SDN-ML facilitates rapid policy updates and fairness-based scheduling to improve multi-user stability. The URLLC-focused allocation ensures low-latency performance for delay-critical services. These observations point to the necessity of combining multiple QoS mechanisms in practical deployments.

**Adaptive Traffic Engineering and QoS or Quality of Experience (QoE) Management**

Effective QoS management in 5G increasingly depends on adaptive traffic engineering, particularly during congestion events where traffic steering and prioritisation decisions directly influence service stability (Beshley et al., 2024). The service-oriented Software-Defined Networking approaches have been introduced to strengthen QoS and QoE. The Service-Oriented Software-Defined Networks (SOSDN) with centralised programmability and supports runtime control of traffic flows and resource policies. This allows prioritisation to be adjusted automatically according to the Service-Level Agreement (SLA) requirements. This capability preserves low-latency paths and critical bandwidth during network congestion. SOSDN employs real-time optimisation instead of static rules to refine prioritisation decisions based on ongoing network measurements. The architecture can accommodate diverse service demands, ranging from latency-sensitive applications such as video streaming to less time-critical traffic. The flow priorities are tuned dynamically according to QoS indicators i.e. delay, throughput, and packet loss. The QoE-oriented approaches have gained attention, particularly for delay-sensitive services where packet loss and jitter have a pronounced impact on perceived user experience (González et al., 2026). For example, a multipath redundancy framework has been introduced to minimise packet loss and latency through redundant paths and intelligent packet reordering, and it was validated through real-vehicle testing and Web Real-Time Communication (WebRTC) streaming experiments (Ito et al., 2025). In addition, SDN-based Radio Resource Management (RRM) has been explored

in a cell-less architecture to enable dynamic rerouting and interference mitigation (Zeyad & Al Janaby, 2025). This is particularly relevant for delay-sensitive mobile users, where mobility and interference significantly influence QoS continuity.

From a cross-study perspective, SOSDN-oriented prioritisation mechanisms are most effective for SLA-driven congestion handling, whereas multipath redundancy approaches show stronger reliability benefits for mobility-dominated streaming scenarios. SDN-based RRM models, in contrast, are better aligned with maintaining QoS continuity in cell-less architectures. Despite these strengths, practical deployment remains constrained by scalability limits and added control-plane overhead in real-time operation.

This paper separates the existing works that were carried out through simulations under controlled environments or experimental tests under the real network limitations. The simulations are carried out in ideal conditions with stable channels, simplified network topology and low signalling overhead. The performance of the proposed solutions may be different in the real 5G deployment. These review strategies to strengthen the practicality of the existing QoS solutions proposed in a real 5G with various network conditions, such as unstable handover, high signalling load and complex user mobility trends.

**QoS-driven Network Slicing Management**

QoS-aware network slicing has gained substantial research interest as a possible solution to tackle different service requirements in 5G networks, particularly in multi-service environments where resource contention is extremely high. On the other hand, in vehicular communication environments, flexible slice-based resource allocation models have been proposed to ensure that traffic like Vehicle-to-Infrastructure (V2I), which can be referred to as a critical service, is given high priority compared to other non-critical traffic. Less critical services, such as infotainment, are allocated proportionally fewer resources (Tam et al., 2024). Despite the research mentioned above being vehicular service-oriented, it makes clear the general idea concerning real-time QoS-aware resource allocation, where bandwidth allocation happens on a dynamic basis according to service needs and demand. In this regard, slice-aware resource allocation aligns with user-centric QoS requirements, whereby resource allocation policies can adapt continuously according to evolving network conditions and different service requirements.

In general, network slicing enables the logical partitioning of network resources into multiple slices to support the QoS requirements of various traffic. Dynamic slice management is central to sustaining differentiated services (Debbabi et al., 2021). The service-aware orchestration has combined VNF migration with slice-level resource allocation to limit service interruptions and maintain QoS continuity (Vidhya et al., 2025). Despite the isolation and differentiation that slicing provides, real-world deployments still confront coordination and monitoring overheads as well as scalability challenges for real-time control in dense 5G traffic.

QoS-driven slicing frameworks perform well in scenarios that demand strong service isolation and real-time differentiation, such as vehicular and mission-critical applications. The large-scale 5G deployments expose practical limitations related to monitoring overhead, orchestration latency, and the complexity of cross-domain coordination. These continue to affect scalability and operational feasibility.

The coexistence of the URLLC, eMBB, and mMTC is realistic, and the adaptive QoS-based network slicing solutions are proposed to support various types of traffic. Network slicing isolates each service class by allocating dedicated resources. It guarantees the latency of the URLLC, provides stable throughput to eMBB and high-density support for mMTC devices. However, the spectrum efficiency drops if isolation is deployed alone. Adaptive network slicing reallocates the underutilised resources in real time and maintains the QoS of each service. This solution offers a balance between isolation and efficiency for real 5G network deployment.

## QoS Provisioning for 5g Network

QoS provisioning in 5G requires coordinated management across multiple layers, spanning radio resource scheduling, access procedures, core-network resource control, and traffic management. 5G deployments highly depend on adaptive mechanisms to preserve QoS stability, particularly when multiple services compete for limited spectrum resources, rather than relying on the static priority-based differentiation.

Although the application of the MDMA techniques improves the capacity of the access network when dealing with competing traffic, the application of the techniques is hindered by the coordination of interferences and the lack of scalability of the system. Beyond physical-layer access mechanisms, the QoS provisioning has been extensively investigated in cloud-based 5G and emerging 6G architectures, where resource allocation increasingly depends on flexible control and multi-metric optimisation strategies. In this context, the importance of QoS-aware resource management capable of supporting service heterogeneity and rising traffic demands has been highlighted in cloud-native network environments (Louvros et al., 2023). A multi-criteria decision-making approach has been proposed to allocate resources based on service priority across multiple QoS dimensions, including latency, jitter, and throughput, supporting a more balanced allocation under high-load conditions. Despite all these efforts, there is still a constraint towards the real-world implementation due to the complexity of orchestration and control overhead in real-time monitoring and frequent policy updates.

Although better performance has been shown in such techniques, challenges persist in terms of theoretical and practical differences in QoS provisioning frameworks. Therefore, even though MDMA might provide better bandwidth utilisation along with supporting high-prioritised traffic, high complexities along with computation might affect its effectiveness

in a larger environment (Rathod & Saxena, 2024). Similarly, in terms of QoS decision-making in a cloud environment, better resource balancing might be achieved, but in such a case, staging delays might also occur. The problem in RAN environments where strict latency requirements need to be met (Louvros et al., 2023). Therefore, any form of QoS provisioned mechanisms requires more than just simulated results but also a practical approach in terms of traffic realities, along with high-density users.

In this context, QoS-based approaches to resource management have reinforced the importance of developing resource management models depending on actual deployment scenarios by identifying the existence of a gap between assumptions and actual scenarios (Umar et al., 2024). Moreover, flexible scheduling approaches have been proposed to facilitate the adaptation of scheduling operations depending on the urgency of traffic and services, improving the alignment between resource allocation and QoS requirements (Aslam, 2025).

Comparatively, though QoS provisioning in MDMA improves access layer access, QoS in native clouds allows the decision control for multiple QoS metrics at the infrastructure level. Nevertheless, both approaches face challenges in dealing with the complexities and scalability issues in achieving timely feedback for QoS.

**Functional Split Optimisation in Cloud-RAN**

The implementation of functional split strategies demands an appropriate balance of resources and fronthaul bandwidth, especially to meet the QoS requirements while ensuring efficient management of energy consumption. In this context, dynamic and adaptive strategies have also been put forward to efficiently manage the application of resources. This includes the optimisation-based split selection using the Split Hybrid Particle Swarm Optimisation (SPLIT-HPSO) algorithm (Matoussi, 2020a). This strategy is expected to improve the efficiency of decision-making with respect to functional split computation-intensive functions while reducing the cost of deployment. In addition to this, learning-oriented split selection strategies have also been put forward with the application of deep learning models to learn appropriate split selection based on real-time operational parameters. This is expected to improve latency and computation efficiency based on training data and measurements (Ly & Yao, 2021). However, optimisation-driven and learning-based approaches both face practical limitations. The former is often constrained by computational burden, while the latter depends heavily on data quality and training scale, which may restrict real-time use in dense deployments.

In addition, Cloud-RAN introduces the challenge of selecting appropriate split points between centralised and edge units to balance latency, load distribution, and resource utilisation. Medium access strategies that account for functional split constraints and multi-service coexistence have also been discussed (Khodakhah et al., 2025). Scheduling

techniques have been proposed to adapt to fluctuating service demands while optimising control signalling overhead across distributed network nodes. This supports more effective C-RAN operation in dense urban 5G scenarios, although frequent split adaptation may increase coordination overhead across distributed nodes.

### User-centric RAN Slice Allocation

User-centric RAN slice resource allocation has emerged as an important research direction for improving QoS flexibility in 5G systems. The End-to-End User Slice Allocation (E2E-USA) approach has been investigated using hybrid optimisation methods to address the constraints across radio, transport, and computing resources. This strategy iteratively determines slice placement while maintaining QoS targets under varying traffic conditions (Matoussi, 2020a). Vertical slicing has been incorporated into 5G QoS architectures, where traffic classes such as URLLC and eMBB are mapped to dedicated slices according to their performance requirements (Xie et al., 2022). Such designs enable more efficient resource utilisation by redistributing bandwidth when slices operate below capacity or experience reduced demand. Despite their emphasis on isolation and adaptive control, both approaches continue to face challenges in achieving scalable and consistently responsive operation, particularly during abrupt traffic changes and in dense deployment environments.

Studies on user-centric slice allocation have explored models that dynamically adjust slice properties according to user profiles, location, and mobility patterns. These models optimise the usage of the spectrum resources while also providing the service level assurances in the presence of varying user conditions, which underlines the importance of personalisation at the user level in the RAN to manage slices (Matoussi et al., 2023b).

### Deep Learning for QoS Management

Deep learning has been drawing a lot of attention as an emerging solution for efficient management of QoS in 5G communications, owing to its capability of providing real-time adaptability and data-driven decision support. Besides that, learning-based user slice assignment has been researched using deep learning-based neural networks, looking at achieving fast user RAN prediction or slicing with efficient utilisation of network resources (Ly & Yao, 2021). Meanwhile, context-driven network selection solutions have also been researched for enhancing network access decisions, considering various real-time parameters like user requirement, received signal strength, as well as the network congestion status (Honarvar et al., 2022). These approaches highlight the advantage of predictive intelligence in adapting QoS decisions to evolving network conditions, especially in environments with fluctuating network demand and mobility-driven variation. However, there are also significant operational complexities incorporated in the QoS frameworks with the help of deep learning models. In most cases, large-scale training data is used

with these models, which are then required to be periodically retrained to be able to cope with the non-stationary network conditions. In addition to this, the inference latency and operational feasibility are also the major concerns when deep learning models are incorporated into the QoS control loops. Thus, when we talk about beyond deep learning, there are also some light learning-based strategies considered. As discussed above, these strategies have also been incorporated into QoS optimisation problems, wherein the major concern is to maintain responsiveness with a reduction in computational complexity. In this regard, link quality modelling with a Hidden Markov Model (HMM) has also been used in order to enable real-time adaptation of network coding redundancy, which enhances bandwidth efficiency with a reduction in unnecessary retransmissions under varying loss conditions (Yin, et al., 2020).

Comparatively, the deep learning-based QoS management solution has stronger adaptability based on prediction in terms of slice allocation and access, while the lightweight ML solution, such as HMM-based adaptation, provides lower-complexity adaptation mechanisms for bandwidth optimisation. However, one of the most important research gaps is to achieve scalable and deployable learning-based QoS solutions for dense 5G networks.

The ML-based QoS schemes support adaptability, but they come with fundamental performance degradation. Extra resources are required by the high-performance deep learning models (LSTM, CNN, etc.) that require higher energy consumption to support the larger amount of computation with longer computational time, which is not available at the edge nodes. It may hurt the latency-sensitive URLLC. The higher accuracy provided by the SDN scheme comes at the trade-off of the heavy computational load (Jiang et al., 2023). Lightweight models are proposed for edge nodes with faster judgment, but the accuracy is lower in active scenarios (Yin et al., 2020). A balance between the accuracy, latency and edge cost for the ML-driven provisioning schemes. The edge nodes support the lightweight models for the latency-sensitive deployment. The deep learning models are preferable for high-precision prediction applications.

The additional energy consumption by the complex ML-based QoS control, especially at the edge, remains a challenge. Deep learning models require longer processing cycles and larger memory access, which significantly increases the power consumption. The proposed solution should maintain a balance between maintaining energy efficiency and adaptive QoS provisioning. The lightweight ML models with lower energy consumption are preferable for edge deployment.

### Traffic Management through Network Slicing

Traffic management using network slicing is extensively researched as an important QoS support feature in the context of 5G networks, particularly during congestion states where multiple services are competing to access shared infrastructure resources.

Vertical slicing enables the segmentation of service traffic according to the requirements of the application, thereby providing more predictable QoS performance through service isolation. Moreover, the use of slice-aware bandwidth re-allocation has been recognised as one of the key utilisation methods, whereby idle slices' underutilised resources may be used to support active slices to avoid underperformance for high-priority traffic (Xie et al., 2022). Apart from traffic segregation, another approach that has been mentioned as a secondary option that could potentially reduce fronthaul bandwidth requirements, without compromising QoS for high-priority services, is the functional split approach, particularly in computation-intensive RAN scenarios (Matoussi, 2020a). These techniques improve flexibility in slice-level traffic handling, but they introduce additional management complexity, particularly during abrupt traffic surges that require frequent reconfiguration. Slice-aware traffic control mechanisms have been investigated across both edge and core network domains. For instance, the robust distribution of traffic in Multi-access Edge Computing (MEC) based crowd sensing has been researched to minimise the deterioration of services in case of an overload situation (Xiang et al., 2025). SDN-based slicing architectures have also been investigated to enhance the centralised control of traffic in slices to improve isolation while providing effective orchestration (Ramesh et al., 2024). A cross-comparison of these methods indicates that vertical slicing, together with dynamic redistribution of unused resources, can improve utilisation efficiency and service differentiation. The function splitting mechanisms are helpful in optimising fronthaul for latency-critical services, SDN and MEC-based control that can improve slice isolation in overload situations. Nevertheless, the scalability of 5G networks is a challenge in the implementation of real-time slices due to orchestration, monitoring, and scalability issues.

Though the simulation-based investigations are beneficial from a theoretical point of view, in practical application scenarios, issues regarding scalability as well as latency are more challenging to overcome. For instance, the SDN infrastructure has demonstrated its efficiency across a range of testbed environments, but poses inherent issues regarding its control plane overhead when applied to 5G scenarios (Rathod & Saxena, 2024). ML models have demonstrated a superior capacity for increased efficiency under testbed environments but remain more challenging to retain when dealing with dynamic environments that are perpetually changing (Louvros et al., 2023). Field tests performed by (Ito et al., 2025) have effectively demonstrated the QoS improvement for video streaming by leveraging the benefits of multipath redundancy.

Table 2 presents a comparison of leading user-centric QoS application methods in terms of their design focus, limitations, and applicability. One study investigates the application of AI-based assistance for a 6th-generation network that focuses on a novel shift from a network function-based approach to a user-centric approach. Despite its exploratory value, inadequacies regarding its scalability and QoS metrics affect its application and applicability, respectively (Gkatzios et al., 2024). In another study, real-time user feedback

for IoT systems was investigated to promote a shift in its resource allocation approach, changing it to a form of a closed-loop system to allow for increased adaptability. Despite its potential and value, a limitation regarding its ability to perform real-time changes to the QoS (Chow et al., 2026).

Table 2
*Comparative analysis of research on QoS enhancement and resource allocation*

| Title | Project statement | Objective | Limitation |
|---|---|---|---|
| A Proof-of-Concept Implementation of an AI-assisted User-Centric 6G Network (Gkatzios et al., 2024) | Explores the shift from network function-centric to user-centric designs in 6G slicing techniques. This change aims to address the illusion of user-centricity in 5G caused by shared network functions serving multiple users. | Proposes a self-organising, AI-assisted network that adapts to individual user requirements, enhancing scalability and user-centric service delivery. | Limited scalability due to reliance on narrow QoS metrics, difficulties in managing Message Queuing Telemetry Transport (MQTT) feedback frequency, and risks of network overhead. |
| Design-oriented user-centric QoS scheme for 5G (Chow et al., 2026) | Proposes a user-centric QoS provisioning scheme for 5G that prioritises the user requirements. | Designed an SDN-based control framework integrated with ML prediction and network slicing to guarantee the users under various loads. | Handover variability and multi-vendor orchestration challenges were not evaluated. |
| QoS-driven Selection of Web Services Considering Group Preference (Wang et al., 2015) | Examines the group preferences and QoS attributes influence web service selection in composite architectures. | Balances individual and group preferences in selecting web services, ensuring alignment with QoS criteria and stakeholder expectations. | Computational complexity escalates with the number of services and attributes, lacking real-world dynamic validations. |
| Edge User Allocation with Dynamic Quality of Service (Lai et al., 2019) | Proposes a dynamic allocation mechanism that prioritises QoS metrics for edge servers in fluctuating network conditions. | Adapt allocation dynamically to real-time changes in network conditions, improving resource utilisation and user satisfaction. | Susceptible to instability in resource-constrained or wide-scale scenarios, with insufficient focus on long-term signalling overhead. |
| Self-adaptive Resource Allocation for Cloud-based Software Services Based on Iterative QoS Prediction Model (Chen et al., 2020) | Suggests a predictive resource allocation framework for cloud services using iterative QoS forecasting. | Optimises resource use and ensures SLA compliance by predicting and adapting to changing workloads and conditions in real time. | Faces challenges in volatile environments due to predictive inaccuracies and significant computational overhead, limiting real-time applicability. |

Additional slicing-related SDN architectures are also discussed beyond Table 2. While several works explore the use of 5G network clustering techniques in predicting users' behaviours for better QoS provisioning. These papers also have some challenges in terms of the poor quality of data and scalability when used in larger systems (Ramesh et al., 2024). Another paper has been written on quality-of-service-driven web service selection based on individual and group preferences in such a manner that this selection is often restricted due to challenges in terms of being computationally complex and not being well validated in real scenarios (Wang et al., 2015). Similarly, dynamic quality-of-service-aware users' allocation mechanisms in edge servers in scenarios where network conditions may become unstable have been explored to a greater degree (Lai et al., 2019). This review differs from prior studies by others due to its emphasis on not only algorithmic interpretations but also using human-centred approaches to this topic in a better manner. It critiques existing static frameworks and advocates for dynamic solutions using SDN and ML integration (Chen et al., 2020). Furthermore, an iterative QoS prediction model for cloud-based services enhances resource utilisation. However, it also poses challenges in terms of predictive inaccuracies and computational overhead in volatile environments (Vidhya et al., 2025).

As such, it reveals that there is an inevitable need for an adaptable solution which can satisfy the continuously expanding needs of 5G as well as future technologies, while at the same time showing evident gaps in the subject matter which remain unaddressed by researchers. SDN-based QoS management has an advantage in real-time traffic management, but there are challenges with scalability, especially for high-density urban areas. On the other hand, the ML-based methods have the advantage of high flexibility and adaptability to dynamic traffic conditions, but are challenged by high computation needs, which makes them suitable for IoT and edge computing. Network slicing, particularly for URLLC services, has the advantage of high reliability but is challenged by interference between slices.

Every approach has its own set of benefits targeting specific user scenarios. For instance, SDN-based approaches support centralised control and reconfiguration of QoS but may suffer from single points of failure and scalability problems. ML-based approaches support prediction-based flexibility but pose overhead concerns that must be optimised for effective deployment. Slice-based allocation of resources supports robust QoS guarantees but requires a tighter orchestration to prevent resource conflict. At the same time, network coding and redundancy schemes enhance reliability and performance in lossy or wireless environments, but at the cost of added complexity and bandwidth usage. As shown in Table 3, the comparative analysis demonstrates that no single approach is universally superior. Instead, future research should explore hybrid architectures, e.g., combining SDN with lightweight ML agents at the edge or integrating slicing with redundancy management to optimise user-centric QoS for diverse 5G and B5G application domains.

To further clarify the difference in techniques shown in Figure 3, the taxonomy chart categorises user-centric QoS approaches on four dimensions, i.e. architectural model, intelligence layer, goal of performance, and validation method. Architecture-based QoS solutions are proposed in different layers, such as RAN, the core network, or the edge/fog layer, often by using SDN, Network Function Virtualisation (NFV), or container services. Intellectually, the approaches range from traditional rule-based configurations (e.g. 5QI, Access Class Barring (ACB), and QCI) all the way to advanced solutions powered by ML (e.g., Hidden Markov Models (HMM), Deep Reinforcement Learning (DRL), and Federated Learning (FL)). The objectives span multiple target domains, including URLLC, eMBB, mMTC, and maintaining QoE consistency under mobility. The validation methods include theoretical modelling, NS-3 and MATLAB simulations, experimental tests, and real-world deployment experiments with different levels of realism.

Table 4 shows that the selected case studies demonstrate the usability and efficiency of user-centric QoS approaches through real-world tests and simulations. Techniques like IP-layer path redundancy, adaptive network coding based on machine learning, and dynamic user profile-based slice allocation have exhibited measurable QoS improvement in metrics like packet loss, throughput, and fairness.

Table 3
*Comparative summary of user-centric QoS techniques*

| Technique Category | Strengths | Limitations | Use Case Focus | Validation Method | Research Gap |
|---|---|---|---|---|---|
| SDN-based QoS | Programmable control, dynamic policy updates | Centralised bottlenecks, latency in large-scale networks | eMBB, general mobile broadband | Simulation, real-world prototypes | Need for distributed SDN controllers |
| ML-based prediction | Real-time adaptability, behaviour-aware decisions | High computational demand at edge/cloud | High mobility, IoT, VR/AR | Simulation and testbeds | Lightweight and energy-efficient models |
| RAN slice allocation | Fine-grained isolation, customisable service levels | Inter-slice interference, complex orchestration | URLLC, mission-critical services | Conceptual and simulation | QoS assurance across slices |
| Multipath redundancy | Enhanced reliability and robustness to handovers | Buffering overhead, complexity in path selection | Mobile video streaming, autonomous systems | Real-world field testing | Smart path switching algorithms |
| Adaptive network coding | Improved throughput, error recovery | High redundancy cost, coding complexity | Vehicular communications, low-latency uplinks | Simulation and small-scale testing | Scalable coding under mixed traffic |

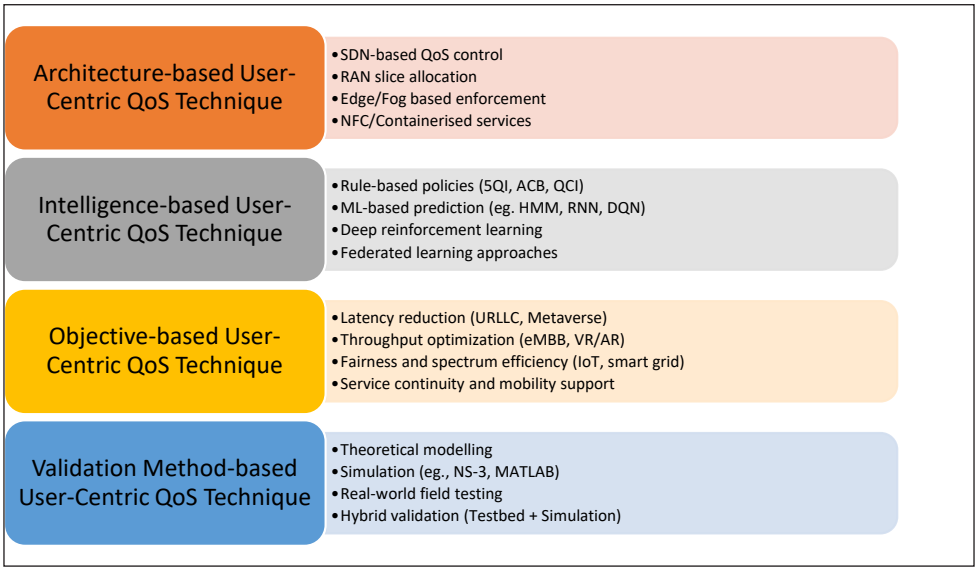


*Figure 3.* 5G QoS user-centric hierarchical structure

The results confirm the theoretical explanations summarised in this paper and reveal promising directions for scalable, smart QoS management in future 5G deployment. Furthermore, the experimental results from the selected case studies corroborate the applicability of user-centric QoS solutions for 5G networks. For the Multipath Redundancy for Media Streaming (Ito et al., 2025), a real-time vehicular deployment with GENEVE tunnelling achieved a packet loss ratio as low as 0.002%, stable delivery of bitrate and only a slight rise in latency (~38ms), verifying the stability of the system at high-mobility situations. The Adaptive Network Coding case (Yin et al., 2020) with HMM-based prediction on NS-2 achieved 18-25% throughput gain and up to 40% PLR reduction over static methods. Similarly, the User-Centric RAN Slice Allocation study from (Matoussi et al., 2023b) shows improved spectrum utilisation by up to 15%, then fairness index by 12%, and inter-slice contention reduction.

These results highlight that adaptive, ML-founded, and user-aware strategies provide tangible gains in essential QoS metrics, namely in dynamic and heterogeneous 5G environments. The case studies also reveal gaps and avenues for enhancement. For example, most of the implementations were tested under control or single-scenario conditions. Future work should thus target:

- Scale testing under gigantic user equipment (UE) density and mixed traffic types.
- Trade-offs in energy efficiency during ML inference at the edge.
- Long-term service continuity via handovers or network slicing reconfigurations.
- Deployment of real-time orchestration tools for runtime automation of QoS enforcement.

Table 4
*Comparison of the case studies*

| Case Study Title | Scenario Type | Technique | Evaluation Platform | Key QoS Metrics Improved |
|---|---|---|---|---|
| Multipath Redundancy for Mobile Communication (Ito et al., 2025) | Real-World Field Test | IP-layer multipath & GENEVE tunnelling & packet reordering | Live vehicular test with WebRTC | ↓ Packet loss (to 0.002%), ↑ Bitrate stability, ↔ Latency (minor increase) |
| Adaptive Network Coding for Vehicular Network (Yin et al., 2020) | Simulation & Experimental | HMM-based packet loss prediction & adaptive coding | NS-2 + real vehicular setup | ↓ Packet Loss Ratio (PLR), ↑ Throughput efficiency |
| User-Centric RAN Slice Allocation (Matoussi et al., 2023b) | Simulation | Mobility & profile-based dynamic slice assignment | Custom 5G network simulator | ↑ Spectrum utilisation, ↑ Fairness index, ↓ Inter-slice contention |

Many works are carried out through simulations or small testbeds. A practical evaluation framework is used to judge the practicality of real 5G deployment for each work. The framework analyses the similarity of the network topology used in the simulation with the real 5G network. For example, the conditions such as traffic loads, users' mobility, interference and handover are modelled according to the real 5G network. The ability of the framework to support a scalable number of users is analysed. The feasibility and complexity of deploying the proposed framework into the existing 5G network are studied. The robustness of the proposed framework under the overloaded conditions is evaluated. The review evaluates the performance in the controlled testbeds and the practicality of each QoS solution in the real 5G network.

Collectively, these studies emphasise the urgency of coming up with cross-layer, modular, and policy-aware QoS solutions that can scale up readily to user behaviour, network conditions, and deployment constraints in real 5G and beyond-5G environments. Furthermore, despite the long upgrades and changes in the 5G network, there are still some outstanding research gaps in terms of the User-Centric QoS aspect. Despite the advancements, several research challenges remain unresolved in the current literature as follows:

- Lack of real-world testing for ML-driven QoS models under dynamic traffic conditions.
- Limited support for seamless interworking of QoS across heterogeneous networks (e.g., 5G, Wi-Fi, IoT).
- Minimal integration of regulatory policies (e.g. General Data Protection Regulation (GDPR), SLA enforcement) into adaptive QoS mechanisms.
- Inadequate exploration of lightweight, privacy-preserving ML models suitable for mMTC.
- Underdeveloped feedback loops between QoE metrics and network-level QoS adjustments.

The user-centric QoS schemes support the 5G network and do not fully evaluate the interoperability of heterogeneous networks with the coexistence of WiFi, fiber optic networks and 4G. It remains a challenge for the real deployment of the QoS schemes across the heterogeneous network with various performance constraints and requirements. The future lightweight user-centric QoS schemes should support cross-layer operations and guarantee consistent QoS across the heterogeneous network. Future research must also protect the privacy issues of processing and collecting the user data, especially in the user-centric framework. The mobility patterns, service usage and preference are sensitive and need protection. The proposed solutions must deploy privacy protection schemes to protect all users' information. The adaptive decisions made by the solutions protect the users' privacy.

To consolidate the comparative insights derived from the reviewed QoS frameworks and selected case studies, Table 5 summarises the most suitable user-centric QoS mechanisms across representative 5G deployment scenarios, while accounting for key practical deployment constraints. The comparative insights from the reviewed SDN, ML, and slicing-driven QoS frameworks were consolidated by mapping each major solution class to the 5G scenario where it is most technically effective. It provides a scenario-oriented interpretation of the user-centric QoS approaches in 5G. The same mechanism may yield significant gains in practical deployments for one condition (e.g., mobility-driven streaming), but it may be impractical for another condition (e.g., ultra-dense mMTC). The potential reasons that caused the performance degradation are overhead, scalability, or latency constraints. It states the advantages and limitations of suitable QoS approaches to be deployed in the mobile network.

From the comparative mapping, SDN-based mechanisms are most suitable in scenarios requiring rapid and policy-driven traffic control, such as congestion-aware traffic engineering for eMBB services. Their key advantage is centralised programmability, which allows real-time enforcement of prioritisation and routing policies across flows. However, the performance of SDN degraded when the network is overloaded.

These may be caused by the frequent flow updates, control-plane overhead, and scalability constraints due to dynamic policy reconfiguration activities. In contrast, ML-based QoS frameworks are more efficient in highly variable environments where predictive traffic awareness enables more proactive allocation decisions, such as adaptive bandwidth allocation and mobility-aware optimisation. Nevertheless, the effectiveness of ML-based models may be dependent on the quality of the training data, the latency of the models, and the associated cost of the models, which may limit the applicability of the models to meet the URLLC deadlines.

Table 5 also highlights that network slicing provides the strongest performance isolation for scenarios requiring deterministic QoS guarantees, such as URLLC and mission-critical applications. Slice-aware mechanisms improve reliability by reserving and isolating resources for critical services, but they remain constrained by orchestration complexity, monitoring overhead, and cross-domain coordination issues when deployed end-to-end across RAN and core networks. Furthermore, mobility-centric strategies, such as redundancy techniques and adaptive coding schemes, have been proven to be very effective in vehicular streaming and handover-aware scenarios, where packet loss is minimised, and throughput is maximised. Nevertheless, such techniques are accompanied by additional bandwidth, diversity, and coordination overhead, which might be more pronounced in heavily loaded networks.

There is no universally optimal user-centric QoS mechanism for all 5G scenarios. Instead, a more viable approach to QoS provisioning would be facilitated by a hybrid approach that leverages the capabilities of SDN for policy enforcement, ML for traffic intelligence, and slicing for isolation. This, in turn, points to the primary research gap within the existing literature, where most of the solutions are still validated in simulations or small-scale testbeds, with the feasibility of deployment at scale, as well as the constraints of real-time operation under mixed traffic conditions, being inadequately investigated.

A practical user-centric QoS framework that consists of SDN for fast deployment, network slicing for isolation and lightweight ML for prediction should have the following properties. A simple operation cycle for the edge deployment. The real-time monitoring of the key performance metrics. SDN and network slicing enforced the policy translated by the orchestration layer for QoS provisioning in real time. The user-level data is protected by the framework.

## Challenges in 5G QoS Security

As shown in Figure 4, the deployment of user-centric QoS models in 5G networks is hindered by a set of interdependent challenges at computational, security, and regulatory levels. Computational complexity remains a key barrier, where high model latency and constrained edge-device resources weaken the real-time practicality of ML-assisted QoS enforcement.

Table 5
*Scenario-based comparative summary of user-centric QoS approaches in 5G*

| 5G Scenario / Service Context | Most Suitable QoS Approach | Why Is It the Best Fit (critical reasoning) | Main Limitation/ Trade-off |
|---|---|---|---|
| URLLC (mission-critical/low latency) | QoS-driven slicing + strict priority scheduling | Ensures isolation and latency guarantees under heavy load | Orchestration delay and control overhead |
| eMBB congestion (high throughput demand) | SDN-based traffic engineering + adaptive prioritisation | Supports centralised control and dynamic policy enforcement | Scalability bottleneck in dense deployments |
| High mobility (vehicular streaming) | Multipath redundancy (path diversity + reordering) | Reduces loss during handovers and stabilises QoE | Extra bandwidth and coordination overhead |
| Loss-sensitive channels (vehicular uplink) | Adaptive Network Coding (HMM/ML-based prediction) | Improves throughput by reducing retransmissions under burst loss | Coding overhead and computation cost |
| Dense multi-user fairness scenario | Fairness-based allocation (PF / WFQ / Max-Min) | Prevents starvation and stabilises user experience | Slower reaction to sudden traffic bursts |
| mMTC/massive IoT | Lightweight ML + rule-based prioritisation | Scales better with device density and avoids heavy inference cost | Limited accuracy vs deep learning |
| B5G/edge-cloud service placement | Edge-cloud orchestration + DRL-based optimisation | Enhances continuity across user mobility regions | Retraining needs, deployment complexity |

This implies that while user-centric QoS approaches can provide performance benefits in a controlled environment, their practical applicability depends on whether such approaches can function within a tight delay constraint without overburdening the edge resources.

From a security point of view, the nature and types of threats emerging in the context of QoS management in a 5G environment appear to be getting increasingly varied and sophisticated. This includes authentication threats, data privacy issues, network slicing issues, edge computing risks, architectural risks in SDN, and those associated with mMTC. Man-in-the-middle attacks have also continued to pose a significant threat, especially in situations where traffic management and reconfiguration occur frequently, thus increasing the chances for such attacks to occur. The effectiveness and suitability of QoS-optimisation approaches should be evaluated through a security lens that evaluates their ability to preserve trust and confidence in operation.

The large-scale deployment of QoS adaptation mechanisms may also face constraints related to regulatory and policy requirements. Especially, when the adaptation of QoS is related to the enforcement of SLA and data protection between different countries.

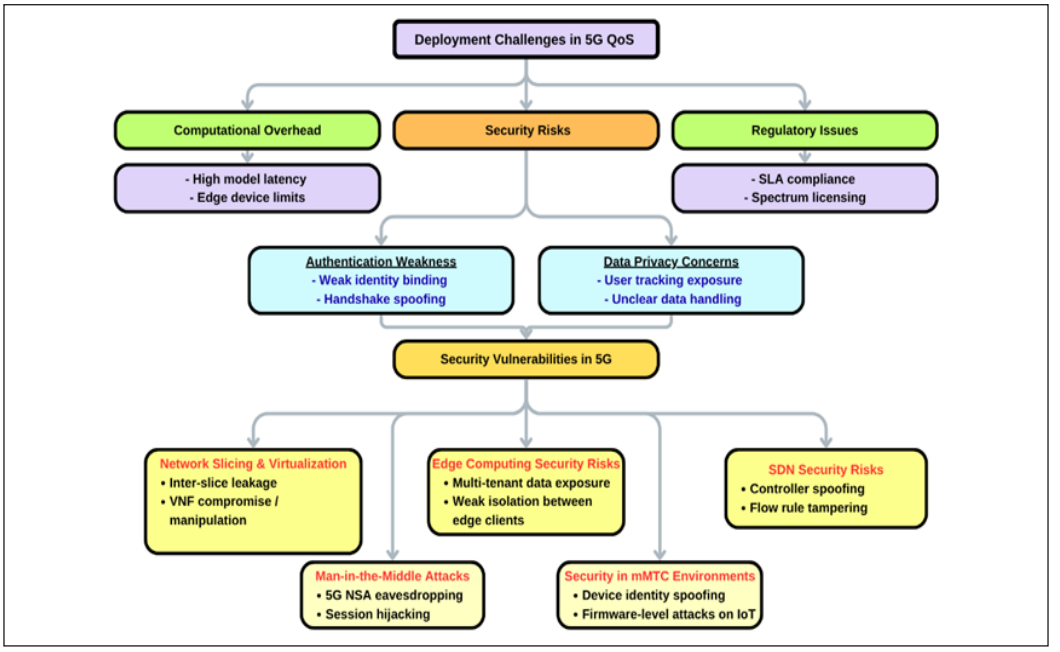


*Figure 4.* 5G QoS Deployment challenges, security and operational risks flowchart (Fakhouri et al., 2023)

This becomes more apparent when the system needs to cope with multiple operators and multiple vendor environments, especially when there are requirements related to the balancing of policy requirements with the need for high QoS responsiveness. Thus, since security and regulatory issues are anticipated in this manner, it is important to ensure that QoS techniques in 5G networks always remain reliable and safe.

## Deployment Challenges and Mitigation Strategies

As discussed earlier, the deployment of user-centric QoS frameworks in real-world 5G environments remains beset with a variety of challenges, ranging from the ability to cope with different network environments, which require a non-trivial integration of AI SDN-QoS with conventional 5G RAN QoS, as well as a variety of scaling challenges, particularly in ultra-dense network environments as well as IoT-centric use cases. It should also be pointed out that the inherent latency overhead of real-time traffic classification and actual decision-making also contributes to the challenges in terms of responsiveness.

To reduce the challenges associated with the above issues, certain strategies have already been experimented with, as described below. One such strategy for the deployment of fog computing within the network includes the “OpticalFog Node Architecture.” In an operational surveillance deployment within a university environment, this architecture reported improvements in latency performance, energy efficiency, and bandwidth utilisation,

while applying SDN-based shortest-path routing to enhance traffic steering efficiency (Singh & Sood, 2020). The emerging B5G has the potential to support diverse mobile user requirements by combining edge-cloud orchestration with deep reinforcement learning. A Double Deep Q-Learning Recurrent Neural Network (DDQL-RNN) framework has been applied to predict user behaviour and support service placement through a water-filling allocation strategy. It considers both the latency and cost. Simulation-based evaluation suggests that this approach can help maintain service availability across the user journey (Farhoudi et al., 2023).

These architectures reviewed report notable QoS improvements. They also reveal a recurring limitation. The validations are still largely confined to simulation-based or tightly controlled test environments. The complementary strategies, such as federated or lightweight ML, intent-driven SDN, and containerised VNFs, improve the deploy ability and the implementation cost. These can help to narrow the gap between simulation studies and operational deployment for the success of deploying the user-centric QoS models in the real 5G networks.

## Open Challenges and Future Research Areas

The user-centric QoS frameworks still face open challenges that affect scalability, interoperability, and practical deployment. The development of low-latency, privacy-aware ML models that can run efficiently at the network edge is highly demanded. These models enable real-time decision-making without imposing excessive computational overhead on end devices. Secondly, QoS interoperability over heterogeneous networks, including Wi-Fi 7, satellite, and legacy LTE, is still unexplored, especially for handover scenarios or multi-access convergent scenarios. Security and privacy concerns continue to exist as well, especially in AI/ML-based systems where inference models become vulnerable to adversarial attack or data compromise. On the other hand, another important challenge to be addressed is the integration of regulatory compliance, such as SLA compliance and GDPR, with dynamic QoS systems. From an operational point of view, there is no common orchestration framework that can enforce QoS policies across different network domains, vendors, and services.

The above deficiencies must be remedied by subsequent research efforts in developing as follows:

- Edge-compatible, resource-aware ML models for QoS adaptation.
- Secure, regulation-aware QoS management frameworks.
- Federated learning-based architectures for privacy-aware traffic classification.
- Cross-layer QoS methods involving application, transport, and network policies.
- Unified orchestration platforms across cloud, edge, and RAN domains.

## CONCLUSION

Overall, SDN and ML-based mechanisms remain central enablers for adaptive QoS provisioning in 5G. SDN supports real-time, agile control of network resources while ML enhances the ability of the network in terms of predicting the pattern of traffic and further optimises bandwidth allocation and dynamically readjusts QoS priorities. When combined with techniques such as network slicing, these mechanisms can improve service differentiation across high and low-priority traffic classes. Despite such significant progress, the current implementation of most of these works is usually not satisfactory to meet diversified user demands, especially under high-traffic scenarios. This review identifies a persistent research gap in achieving seamless and scalable integration between SDN control logic and ML-driven decision modules for real-time QoS adaptation. Although SDN and ML demonstrate strong potential individually, their integrated deployment has not yet been fully exploited for practical QoS optimisation in 5G networks. The approaches discussed emphasise the need to ensure that premium users get the bandwidth guaranteed to them, while at the same time using the free resources intelligently to satisfy other non-premium users. The reallocation of resources has the effect of not just improving network performance but also of providing a better user experience by balancing efficiency with satisfaction. In contrast, several areas will need further research in the future. First, the scalability of the solutions of SDN-ML must be ensured for large and densely populated networks. The existing systems cannot bear such a high demand and variability typical of 5G environments. Moreover, the improvement of the accuracy of ML models used in traffic prediction and energy-efficient models will be of prime importance to enhance the performance of the 5G network while being more sustainable. In addition, the issue of interoperability with other complementary access technologies, such as Wi-Fi and IoT, should be further investigated to enable the development of hybrid QoS models that are capable of seamless operation across different heterogeneous environments. Finally, it is recommended that federated learning be further investigated as a technique that can be used to enhance the decision-making of adaptive QoS models while maintaining the privacy of users and minimising the dependency of these models on centralised data. In order to make these models more practical, it is necessary that extensive experiments are conducted while keeping the 5G network evolving so that the full potential of 5G is achieved by integrating SDN and ML.

In summary, this critical review establishes a framework of reference for future research on scalable and user-centric QoS management by critically assessing the positive aspects as well as the limitations of SDN, ML and slicing-based approaches. Further research is needed to enhance the feasibility, scalability, and real-time efficiency of these approaches, especially in view of the increasing demands of network applications for higher reliability, lower latency and more efficient resource use.

These areas will be critical for achieving more adaptive, scalable, and energy-efficient QoS management in future 5G and beyond 5G networks.

## ACKNOWLEDGEMENT

This research was supported by the Ministry of Higher Education (MOHE), Malaysia, through the Fundamental Research Grant Scheme (FRGS/1/2024/ICT06/TAYLOR/02/1).

## REFERENCES

Abuajwa, O., Roslee, M., Yusoff, Z. B., Chuan, L. L., & Leong, P. W. (2022). Resource allocation for throughput versus fairness trade-offs under user data rate fairness in NOMA systems in 5G networks. *Applied Sciences, 12*(7), 3226. https://doi.org/10.3390/app12073226

Akinola, O. I., Olaniyi, O. O., Ogungbemi, O. S., Oladoyinbo, O. B., & Olisa, A. O. (2024). Resilience and recovery mechanisms for software-defined networking (SDN) and cloud networks. *Journal of Engineering Research and Reports, 26*(8), 112-134. https://doi.org/10.9734/jerr/2024/v26i81234

Alabarce, M. G., Bravalheri, A., & Marino, P. P. (2020). INSPIRING-SNI: Investigating SDN programmability improving optical south-and north-bound interfaces. In *Proceedings of the 2020 22nd International Conference on Transparent Optical Networks (ICTON)* (pp. 1-4). https://doi.org/10.1109/ICTON51198.2020.9203409

Albekairi, M. (2025). Controlled service scheduling scheme for user-centric software-defined network-based Internet of Things. *IEEE Access, 13*, 19198-19218. https://doi.org/10.1109/ACCESS.2025.3533310

Al-Shammari, B. K. J., Al-Aboody, N., & Al-Raweshidy, H. S. (2018). IoT traffic management and integration in the QoS-supported network. *IEEE Internet of Things Journal, 5*(1), 352-370. https://doi.org/10.1109/JIOT.2017.2785219

Amin, R., Rojas, E., Aqdus, A., Ramzan, S., Casillas-Perez, D., & Arco, J. M. (2021). A survey on machine learning techniques for routing optimisation in SDN. *IEEE Access, 9*, 104582-104611.https://doi.org/10.1109/ACCESS.2021.3099092

Anjum, M., Min, H., & Ahmed, Z. (2024). User-centric Internet of Things and controlled service scheduling scheme for a software-defined network. *Applied Sciences, 14*(11), 4951. https://doi.org/10.3390/app14114951

Aslam, U. (2025). Designing flexible scheduling algorithm for 5G. *International Journal of Advanced Engineering Management and Science, 11*(1), 90-108. https://doi.org/10.22161/ijaems.111.7

Baz, A., Logeshwaran, J., Natarajan, Y., & Patel, S. K. (2024). Enhancing mobility management in 5G networks using deep residual LSTM model. *Applied Soft Computing, 165*, 112103.https://doi.org/10.1016/j.asoc.2024.112103

Beshley, M., Kryvinska, N., Beshley, H., Panchenko, O., & Medvetskyi, M. (2024). Traffic engineering and QoS/QoE supporting techniques for emerging service-oriented software-defined network. *Journal of Communications and Networks, 26*(1), 99-114. https://doi.org/10.23919/JCN.2023.000065

Chen, X., Wang, H., Ma, Y., Zheng, X., & Guo, L. (2020). Self-adaptive resource allocation for cloud-based software services based on iterative QoS prediction model. *Future Generation Computer Systems, 105*, 287-296.https://doi.org/10.1016/j.future.2019.12.005

Chia, R., Pang, W. L., Phang, S. K., Goh, H. H., & Chan, K. Y. (2025). Machine learning-driven analysis of user bandwidth allocation and performance in 5G network. *IEEE Access, 13*, 173081-173095. https://doi.org/10.1109/ACCESS.2025.3615398

Chow, W. H., Pang, W. L., Goh, H. H., Tee, W. H., Md Rezali, F. A., Chan, K. Y., & Chung, G. C. (2026). Design of user-centric QoS provisioning scheme in 5G network. *Journal of Engineering Science and Technology*, 28-50.

Ciceri, O. J., Astudillo, C. A., Zhu, Z., & da Fonseca, N. L. S. (2024). Bandwidth allocation for multiple functional splitting options over TWDM-EPON networks with multi-ONU customers. In *ICC 2024-IEEE International Conference on Communications* (pp. 468-473). https://doi.org/10.1109/ICC51166.2024.10622322

Debbabi, F., Jmal, R., & Chaari Fourati, L. (2021). 5G network slicing: Fundamental concepts, architectures, algorithmics, projects, practices, and open issues. *Concurrency and Computation: Practice and Experience, 33*(20). https://doi.org/10.1002/cpe.6352

Debnath, R., Akinci, M. S., Ajith, D., & Steinhorst, S. (2023). 5GTQ: QoS-aware 5G-TSN simulation framework. In *2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)* (pp. 1-7). https://doi.org/10.1109/VTC2023-Fall60731.2023.10333533

Dias, I., Ruan, L., Ranaweera, C., & Wong, E. (2023). From 5G to beyond: Passive optical network and multi-access edge computing integration for latency-sensitive applications. *Optical Fiber Technology, 75*, 103191. https://doi.org/10.1016/j.yofte.2022.103191

Etxezarreta, X., Garitano, I., Iturbe, M., & Zurutuza, U. (2023). Software-defined networking approaches for intrusion response in industrial control systems: A survey. *International Journal of Critical Infrastructure Protection, 42*, 100615. https://doi.org/10.1016/j.ijcip.2023.100615

Fakhouri, H. N., Alawadi, S., Awaysheh, F. M., Hani, I. B., Alkhalaileh, M., & Hamad, F. (2023). A comprehensive study on the role of machine learning in 5G security: Challenges, technologies, and solutions. *Electronics, 12*(22), 4604. https://doi.org/10.3390/electronics12224604

Farhoudi, M., Shokrnezhad, M., & Taleb, T. (2023). QoS-aware service prediction and orchestration in cloud-network integrated beyond 5G. In *GLOBECOM 2022 -2022 IEEE Global Communications Conference*. https://doi.org/10.1109/GLOBECOM54140.2023.10436905

Gkatzios, N., Koumaras, H., Fragkos, D., & Koumaras, V. (2024). A proof-of-concept implementation of an AI-assisted user-centric 6G network. In *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)* (pp. 907-912). https://doi.org/10.1109/EuCNC/6GSummit60053.2024.10597020

González, C. C., Pupo, E. F., Floris, A., Porcu, S., Murroni, M., & Atzori, L. (2026). OTT–MNO collaboration for a network-layer ML-based QoE prediction for video streaming over 5G O-RAN. *Computer Networks*, 112152. https://doi.org/10.1016/j.comnet.2026.112152

Haji, S. H., Zeebaree, S. R. M., Saeed, R. H., Ameen, S. Y., Shukur, H. M., Omar, N., Sadeeq, M. A. M., Ageed, Z. S., Ibrahim, I. M., & Yasin, H. M. (2021). Comparison of software defined networking with traditional networking. *Asian Journal of Research in Computer Science*, 1-18. https://doi.org/10.9734/ajrcos/2021/v9i230216

Honarvar, R., Zolghadrasli, A., & Monemi, M. (2022). Context-oriented performance evaluation of network selection algorithms in 5G heterogeneous networks. *Journal of Network and Computer Applications, 202*, 103358. https://doi.org/10.1016/j.jnca.2022.103358

Hoyhtya, M., Lahetkangas, K., Suomalainen, J., Hoppari, M., Kujanpaa, K., Trung Ngo, K., Kippola, T., Heikkila, M., Posti, H., Maki, J., Savunen, T., Hulkkonen, A., & Kokkinen, H. (2018). Critical communications over mobile operators' networks: 5G use cases enabled by licensed spectrum sharing, network slicing and QoS control. *IEEE Access, 6*, 73572-73582. https://doi.org/10.1109/ACCESS.2018.2883787

Hussain, S. M. S., Aftab, M. A., & Ustun, T. S. (2020). Performance analysis of IEC 61850 messages in LTE communication for reactive power management in microgrids. *Energies, 13*(22), 6011.https://doi.org/10.3390/en13226011

Hyder, H. K., & Lung, C.-H. (2018). Closed-loop DDoS mitigation system in software defined networks. In *2018 IEEE Conference on Dependable and Secure Computing (DSC)* (pp. 1-6). https://doi.org/10.1109/DESEC.2018.8625125

Ito, K., Nakazato, J., Fontugne, R., Tsukada, M., & Hiroshi, E. (2025). A multipath redundancy communication framework for enhancing 5G mobile communication quality. *Computer Communications*, 108157. https://doi.org/10.1016/j.comcom.2025.108157

Jiang, W., Han, H., He, M., & Gu, W. (2023). ML-based pre-deployment SDN performance prediction with neural network boosting regression. *Expert Systems with Applications, 241*, 122774. https://doi.org/10.1016/j.eswa.2023.122774

Khairi, M. H. H., Ariffin, S. H. S., Latiff, N. M. A., Yusof, K. M., Hassan, M. K., Al-Dhief, F. T., Hamdan, M., Khan, S., & Hamzah, M. (2021). Detection and classification of conflict flows in SDN using machine learning algorithms. *IEEE Access, 9*, 76024-76037.https://doi.org/10.1109/ACCESS.2021.3081629

Khan, N. A. (2022). 5G network: Techniques to increase quality of service and quality of experience. *International Journal of Computer Networks and Applications, 9*(4), 476. https://doi.org/10.22247/ijcna/2022/214508

Khodakhah, F., Mahmood, A., Österberg, P., & Gidlund, M. (2025). Adaptive user pairing with non-orthogonal medium access choices for balanced coexistence of mission-critical and eMBB services in cellular IoT. *IEEE Open Journal of the Communications Society, 6*, 5414-5433. https://doi.org/10.1109/OJCOMS.2025.3578727

Kunasegran, M. P., Pang, W. L., & Phang, S. K. (2025a). Optimising resource allocation in 5G networks: Balancing URLLC and eMBB traffic under gNB congestion. In *Proceedings of the 2025 Multimedia University Engineering Conference (MECON 2025)* (pp. 1-6). IEEE. https://doi.org/10.1109/MECON67253.2025.11277096

Kunasegran, M. P., Pang, W. L., & Phang, S. K. (2025b). SLA-aware DRL for joint slice admission control and load balancing in multi-service 5G RAN. In *Proceedings of the 9th International Conference on Recent*

*Advances and Innovations in Engineering (ICRAIE 2025)* (pp. 207-212). IEEE. https://doi.org/10.1109/ICRAIE65839.2025.11239148

Lai, P., He, Q., Cui, G., Xia, X., Abdelrazek, M., Chen, F., Hosking, J., Grundy, J., & Yang, Y. (2019). Edge user allocation with dynamic quality of service. In *Proceedings of the conference* (pp. 86-101). https://doi.org/10.1007/978-3-030-33702-5_8

Louvros, S., Paraskevas, M., & Chrysikos, T. (2023). QoS-aware resource management in 5G and 6G cloud-based architectures with priorities. *Information, 14*(3), 175. https://doi.org/10.3390/info14030175

Ly, A., & Yao, Y. D. (2021). A review of deep learning in 5G research: Channel coding, massive MIMO, multiple access, resource allocation, and network security. *IEEE Open Journal of the Communications Society, 2*, 396-408. https://doi.org/10.1109/OJCOMS.2021.3058353

Mahmood, A., Abedin, S. F., Sauter, T., Gidlund, M., & Landernäs, K. (2021). Factory 5G: A review of industrial-centric features and deployment options. *IEEE Industrial Electronics Magazine, 16*(2), 24-34. https://doi.org/10.36227/techrxiv.17089265.v1

Matoussi, S., Fajjari, I., Aitsaadi, N., & Langar, R. (2020a). User slicing scheme with functional split selection in 5G cloud-RAN. In *2020 IEEE Wireless Communications and Networking Conference (WCNC)* (pp. 1-8). https://doi.org/10.1109/WCNC45663.2020.9120828

Matoussi, S., Fajjari, I., Aitsaadi, N., & Langar, R. (2023b). User-centric slice allocation scheme in 5G networks and beyond. *IEEE Transactions on Network and Service Management, 20*(4), 4268-4282.https://doi.org/10.1109/TNSM.2023.3284206

Nikolaidis, P., Zoulkarni, A., & Baras, J. (2023). Bandwidth provisioning for network slices with per user QoS guarantees. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium* (pp. 1-9). https://doi.org/10.1109/NOMS56928.2023.10154366

Prodromos, N., Diasakos, D., Kokkinos, V., Gkamas, A., Bouras, C., & Pouyioutas, P. (2024). Dynamic bandwidth allocation in MIMO 5G networks. In *2024 International Wireless Communications and Mobile Computing (IWCMC)* (pp. 97-102). https://doi.org/10.1109/IWCMC61514.2024.10592577

Raeisi, M., & Sesay, A. B. (2023). Handover reduction in 5G high-speed network using ML-assisted user-centric channel allocation. *IEEE Access, 11*, 84113-84133. https://doi.org/10.1109/ACCESS.2023.3297982

Ramesh, P., Mohan, B., Viswanath, L., & Stephen, B. J. (2024). Software defined network architecture based network slicing in fifth generation networks. *Informacije MIDEM- Journal of Microelectronics, Electronic Components and Materials, 51*(3). https://doi.org/10.33180/InfMIDEM2024.205

Rathod, I., & Saxena, S. (2024). Design, simulation and analysis of multi-dimensional multiple access (MDMA) schemes using MATLAB for quality of service (QoS) enhancement. *Journal of Intelligent Systems and Internet of Things, 11*(2), 111-128. https://doi.org/10.54216/JISIoT.110210

Rawshan, F., Hossen, M., & Islam, Md. R. (2024). Dynamic wavelength and bandwidth allocation using service class prioritisation for upstream in 100 Gb/s NG-EPON. *Results in Engineering, 24*, 103151.https://doi.org/10.1016/j.rineng.2024.103151

Shameli, R., & Rajkumar, S. (2026). Design of an AI-driven secure 5G-SDN framework with federated reinforcement learning for anomaly detection, mitigation, and attack forensics. *Frontiers in Artificial Intelligence, 9*, 1701944. https://doi.org/10.3389/frai.2026.1701944

Shrivastava, V. K., Baek, S., & Baek, Y. (2022). 5G evolution for multicast and broadcast services in 3GPP Release 17. *IEEE Communications Standards Magazine, 6*(3), 70-76. https://doi.org/10.1109/MCOMSTD.0001.2100068

Singh, K. D., & Sood, S. K. (2020). QoS-aware optical FoG-assisted cyber-physical system in the 5G-ready heterogeneous network. *Wireless Personal Communications, 116*(4), 3331-3350.https://doi.org/10.1007/s11277-020-07855-5

Sufyan, A., Khan, K. B., Khashan, O. A., Mir, T., & Mir, U. (2023). From 5G to beyond 5G: A comprehensive survey of wireless network evolution, challenges, and promising technologies. *Electronics, 12*(10), 2200. https://doi.org/10.3390/electronics12102200

Tam, P., Ros, S., Song, I., & Kim, S. (2024). QoS-driven slicing management for vehicular communications. *Electronics, 13*(2), 314. https://doi.org/10.3390/electronics13020314

Tanuja, K. S., Shanmukaswamy, C. V., Gurushankar, H. B., & Dinesh, H. A. (2023). Dynamic bandwidth allocation scheme for enhanced performance in 5G point-to-point networks. *ICTACT Journal on Communication Technology, 14*(2), 2945-2951. https://doi.org/10.21917/ijct.2023.0438

Toor, W. T., Basit, A., Maroof, N., Khan, S. A., & Saadi, M. (2019). Evolution of random access process: From legacy networks to 5G and beyond. *Transactions on Emerging Telecommunications Technologies, 33*(6). https://doi.org/10.1002/ett.3776

Trifan, R.-F., Lerbour, R., & Le Helloco, Y. (2015). Mirroring LTE scheduler performance with an adaptive simulation model. In *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)* (pp. 1-5).https://doi.org/10.1109/VTCSpring.2015.7145931

Umar, M. M., Mohammed, A., & Abdulazeez, A. (2024). Review of QoS-aware resource allocation schemes for 5G networks. *Dutse Journal of Pure and Applied Sciences, 10*(3c), 296-303.https://doi.org/10.4314/dujopas.v10i3c.28

Vidhya, P., K, S., R, S., & S, G. (2025). Dynamic network slicing-based resource management and service-aware virtual network function (VNF) migration in 5G networks. *Computer Networks*, 111064.https://doi.org/10.1016/j.comnet.2025.111064

Wang, H.-C., Chiu, W.-P., & Wu, S.-C. (2015). QoS-driven selection of web service considering group preference. *Computer Networks, 93*, 111-124. https://doi.org/10.1016/j.comnet.2015.10.014

Wu, C., Lu, H., Chen, Y., & Qin, L. (2024). Cross-layer optimisation for statistical QoS provision in C-RAN with finite-length coding. *IEEE Transactions on Communications, 72*(6), 3393-3407.https://doi.org/10.1109/TCOMM.2024.3370817

Xiang, Z., Ying, F., Yan, H., Zheng, Z., Zhang, Y., & Xu, Y. (2025). QoS-effective and resilient service deployment and traffic management in MEC-based crowdsensing. *Symmetry, 17*(5), 718. https://doi.org/10.3390/sym17050718

Xie, M., Gonzalez, A. J., Gronsund, P., Lonsethagen, H., Waldemar, P., Tranoris, C., Denazis, S., & Elmokashfi, A. (2022). Practically deploying multiple vertical services into 5G networks with network slicing. *IEEE Network, 36*(1), 32-39. https://doi.org/10.1109/MNET.001.2100361

Yang, D., & Tsai, W.-T. (2024). SDN-based congestion control and bandwidth allocation scheme in 5G networks. *Sensors, 24*(3), 749. https://doi.org/10.3390/s24030749

Yaqoob, J. I., Pang, W. L., Wong, S. K., & Chan, K. Y. (2020). Enhanced exponential rule scheduling algorithm for real-time traffic in LTE network. *International Journal of Electrical and Computer Engineering (IJECE), 10*(2), 1993-2002. https://doi.org/10.11591/ijece.v10i2.pp1993-2002

Yaqoob, Y. J. I. A., Pang, W. L., Wong, S. K., & Chan, K. Y. (2019). Performance evaluation of video streaming on LTE with coexistence of Wi-Fi signal. *Bulletin of Electrical Engineering and Informatics, 8*(3), 890-897. https://doi.org/10.11591/eei.v8i3.1580

Yin, C., Dong, P., Du, X., Zheng, T., Zhang, H., & Guizani, M. (2020). An adaptive network coding scheme for multipath transmission in cellular-based vehicular networks. *Sensors, 20*(20), 5902. https://doi.org/10.3390/s20205902

Zeyad, I., & Al Janaby, A. O. (2025). Interference mitigation for dynamic user connectivity using SDN and radio resource management in cell-less networks. *Journal of Engineering Science and Technology, 20*(2), 520-535.