Journal of Engineering Science and Technology
23rd EURECA 2025 Special Issue February (2026) 28 - 50

DESIGN OF USER-CENTRIC QoS PROVISIONING SCHEME IN 5G NETWORK

WEN HAO CHOW[1], WAI LEONG PANG[1,*], HUI HWANG GOH[1], WEI HOWN TEE[1], FAZLIYATUL AZWA MD REZALI[1], KAH YOONG CHAN[2], GWO CHIN CHUNG[2]

[1]School of Engineering, Taylor's University, No. 1 Jalan Taylor's, 47500, Subang Jaya, Selangor DE, Malaysia
[2]Faculty of Artificial Intelligence and Engineering, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor DE, Malaysia
*Corresponding Author: waileong.pang@taylors.edu.my

Abstract

In the modern deployment of 5G network with fixed Quality of Service (QoS) configuration is typically not adaptable to dynamic user needs and shifting traffic behaviour. Such limitations are addressed in this project by the implementation of a dynamic, user-centric QoS provisioning scheme that classifies users into three categories: Premium, Standard, and Non-Priority. Every subscriber is then mapped to one type of service, which is video streaming, voice communication, online gaming, web surfing, or IoT signalling and relevant EPS bearers and traffic filters are enforced. The scheme under discussion dynamically reassigns QoS priorities based on real user distribution rather than fixed rules. The proposed scheme is evaluated in the NS-3 simulator. The proposed scheme guarantees Premium users will always get high-priority service but maintains acceptable performance for Standard and Non-Priority users. The extensive simulation results show that the proposed scheme improves average throughput for Premium users by up to 35%, reduces end-to-end delay by 28%, and lowers jitter by 22% compared with traditional static QoS models. Furthermore, the proposed scheme prevents starvation of lower-priority users by redistributing unused bandwidth efficiently. With that being said, this dynamically classification-based scheme introduces a fair degree of flexibility and performance advantage in overloaded 5G environments, thereby making this static QoS provisioning much more practical and efficient.

Keywords: 5G network, Dynamic resource allocation, Network slicing, QoS, Premium user protection, User-centric design.

1. Introduction

The fifth generation (5G) of cellular networks introduces revolutionary aspects such as ultra-Reliable Low Latency Communication (uRLLC), improved Mobile Broadband (eMBB), and massive Machine-Type Communication (mMTC). While these types of services promise diversity and capability, meeting their different Quality of Service (QoS) expectations under different network conditions remains a long-standing issue, especially during peak user load times. For instance, the existing QoS provisioning schemes in 5G networks are typically static, operator-based, with resource provision and traffic management based on preconfigured priorities or service classes. More recent research increasingly indicates the importance of user-centric paradigms to ensure satisfaction of services under dynamic network conditions.

Matoussi et al. [1] contributed a radio resource allocation model at the user level in Radio Access Network (RAN) slicing, which demonstrates the benefits of RAN resource optimisation on a per-user basis through the assistance of heuristic optimisation. Jahandar et al. [2] also advocates dynamic mode switching and mobility management using relays to dynamically adapt communication paths based on real-time performance indicators such as delay and link utilisation. These user-centric strategies reflect an emerging consensus. Future networks must move beyond static policies toward context-aware and adaptive QoS schemes.

Meanwhile, the Controlled Service Scheduling Scheme (CS3) emphasises the role of intelligent scheduling and regression learning in Software Defined Networking (SDN) based IoT environments to accommodate fluctuating request densities while preserving responsiveness [3, 4]. Likewise, a proposed Multi-access Edge Computing (MEC) based joint service deployment and traffic management was done to maintain performance-cost efficiency in crowd-sensing scenarios [5]. Based on the observations, this project designs a user-centric QoS provisioning algorithm that classifies users into Premium, Standard, and Non-Premium classes.

Unlike conventional slicing strategies, the design features real-time priority-weighted bandwidth sharing between various service types, such as video streaming, voice, gaming, browsing, and IoT, as well as dynamic response to concentration patterns of users for each class. Emulating the application of multipath redundancy in making communication more robust [6], my simulation contains a flow-level analysis for observing end-to-end throughput, delay, and jitter as the user load varies. This paper presents a QoS architecture based on reasoning that achieves light but interactive traffic differentiation with attention not only to Premium users' experience but also to not entirely denying network resources to Non-Premium and Standard users (which standard and non-premium are ultimately the same) during overload conditions.

Moreover, the deployment of 5G networks brings unique new capabilities in areas such as autonomous cars, smart cities, and industrial IoT by empowering varied services with demanding Quality of Service (QoS) requirements. Network slicing is one of the prime facilitators of this revolution, where a physical network is divided into a number of virtual slices that are customised to specific services such as eMBB (enhanced Mobile Broadband), URLLC (Ultra-Reliable Low-Latency Communication), and mMTC (massive Machine-Type Communication) [7, 8].

Typically, in 5G networks, fixed policies for resource provisioning and service prioritisation are usually followed by over-provisioning. The resources being

underutilised or under-provisioning, resulting in SLA (Service Level Agreement) breaches and degraded QoS particularly for non-premium clients. This limitation is even more crucial during network congestion or varying user demands. Recent research has proposed intelligent and adaptive resource management strategies to address this issue.

Vidhya et al. [9] and He et al., for instance, proposed an MA-DRL (Multi-Agent Deep Reinforcement Learning) based dynamic network slicing strategy. Soft Actor-Critic algorithms and game-theoretic models are employed in their framework for the dynamic allocation of resources and VNF migration with the goal of optimising slice performance, especially in overload scenarios. This necessitates real-time adaptability for multi-slice 5G networks.

Also, Aslam [11] dealt with the design of flexible scheduling algorithms for per-slice QoS adherence using adaptive airtime allocation and dynamic policy enforcement. Their proposed scheduler, experimented with in a smart city simulation, rebalances airtime by analysing slice-level violations in real-time, allowing for slice elasticity. From the perspective of network topology and interference, based on research [12-15], they describe SDN and sophisticated Radio Resource Management (RRM) to mitigate interference in cell-less 5G networks. Their proposed SDN-based solution enhances seamless handover, minimises inter-slice interference, and enhances slice reliability, especially vital for high-density deployments with high mobility users.

Lastly, Umar et al. [16] consolidated several QoS-aware resource allocation frameworks, categorised them based on static vs. dynamic, centralised vs. distributed models, and feasibility in diverse 5G use cases. The authors highlight that while ML and SDN-based models provide granularity, they are usually overhead-heavy, and therefore more straightforward adaptive models would be better suited for real-time and large-scale applications. This paper contributes to this body of literature by introducing a user-aware QoS provisioning mechanism that balances premium, standard, and non-premium user demands.

Through the coupling of a per-user classification model with a service-type-aware bearer and application assignment model, the simulation aims to achieve network efficiency maximisation and preserve fairness even during overload. The solution avoids causing undue computational overhead by avoiding the use of AI or SDN and instead using strategic logic flow embedded in Network Simulator 3 (NS-3) version 3.39. To validate this, a number of scenarios for user distribution are simulated from 10% up to 90% premium user configurations. Each configuration studies how service throughput, jitter, and latency metrics change as the prioritisation scheme dynamically adjusts bearer types and application flows.

In addition, primarily on adaptive resource management and user-oriented QoS provisioning were discussed and brought up based on Ramesh et al. [17]. It explores the application of SDN-based slicing approaches for identifying user traffic types, advancing the virtue of dynamic rule-based management to optimise service isolation and efficiency, which underpins this project's user-priority based traffic separation. Similarly, based on Alashjaee et al. [18], the design addresses dynamic mobility and link performance monitoring as a method of enhancing QoS under real-time conditions, which is also complemented by adaptive reallocation of high-priority bearer resources depending on user concentration.

Furthermore, per-user classification logic utilised here takes its inspiration from the work of Anjum et al. [4], which proposes a Controlled Service Scheduling Scheme (CS3) in SDN-based IoT networks. Their contribution advocates the use of traffic classes defined by user-level behaviour instead of fixed service labels, a practice that is reflected in our simulation using the premium/standard/non-premium demarcation model. While Louvros et al. [19] seconded this by presenting QoS-aware resource allocation for cloud-centric 5G settings, showing how user priority awareness can reduce delay and packet loss under different loads.

Also, of interest is Malik et al. [20], which encouraging user-centric service provisioning through multi-path edge computing. The emphasis on ensuring per-user QoS conformity through lightweight decision logic (compared to costly machine learning inference) is the same as has been adopted in this paper. The strategic reasoning replaces AI-powered complexity. Last but not least, Jayaraman et al. [21] emphasises the importance of efficient Evolved Packet System (EPS) bearer allocation in 5G networks. By illustrating the dramatic impact that bearer-level configuration can have on fairness and throughput and it substantiates this design decision to dynamically assign bearers based on both user type and service category. As a whole, these supporting studies underlie the motivation of this project's design trajectories: to implement a lightweight, adaptive QoS system that emphasises user awareness, real-time priority, and fairness in traffic.

2. Method

In this paper, an algorithm for enhancing the 5G network based on the User-Centric aspect was proposed. From the algorithm, a proposed user-centric driven QoS provisioning mechanism was installed in NS-3.39 for evaluating service differentiation in a 5G environment. The mechanism blocks the use of computationally exhaustive approaches such as ML or SDN using a lightweight method built on user priority classification and immediately assessing service distribution. The simulated network consists of a single 5G Next Generation Node B (gNB) (base station) and any number of arbitrary UEs (User Equipments). Each UE's assigned traffic type reflects common 5G applications: video streaming, voice, gaming, web browsing, or IoT. Distribution is deterministic; UEs are grouped into service-specific containers (ueVsContainer, ueVcContainer, etc.) by their index and fractional thresholds derived from the number of UEs.

The assignment therefore guarantees a fixed application category distribution, enabling controlled comparative analysis. A 5 GHz centre frequency and 100 MHz bandwidth are configured with NS-3's NrHelper and CcBwpCreator, having a single component carrier and a single Bandwidth Part (BWP). These configurations simulate an actual 5G New Radio (NR) deployment but with low simulation complexity. gNB and UE antennas are modelled as isotropic radiators with simple beamforming with the IdealBeamformingHelper.

2.1. User priority classifications

The classification logic is shown in Fig. 1. The user priority classification is performed using the AssignUserPriorities() function. This function accepts the UE (users) container and the desired ratios of Premium, Standard, and Non-Premium users. It generates a random permutation of UE indices using the standard library (std::shuffle) and assigns priority tiers accordingly.

This classification logic maps each stored UE's priority level under its index so that it is accessible during simulation. For small-scale simulation when UE is ≤40, all users are treated as Premium so that every UE receives substantial service differentiation. To evaluate how the proposed scheme performs under varying concentrations of Premium users, the simulation is repeated across nine unique user distribution scenarios as detailed in Table 1. These distributions emulate the increasing subscription to premium plans and their impact on bearer allocation and network fairness. The 10th scenario, meaning the 100% Premium users is simulated as it mirrors a legacy operator-centric model.

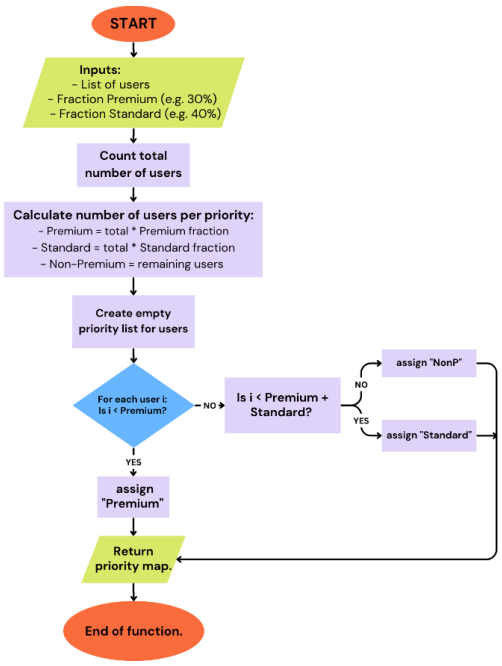

Fig. 1. Classification logics.

Table 1. User distribution ratios for simulation scenarios.

Scenario	Premium UEs (%)	Standard UEs (%)
1	10	90
2	20	80
3	30	70
4	40	60
5	50	50
6	60	40
7	70	30
8	80	20
9	90	10

Algorithm 1: User Classification and dynamic QoS rearrangement

Input:

totalUes: Total number of User Equipment (UEs)

premiumRatio, standard ratio: Percentage ratios for Premium and Standard Users

ueServiceType[]: Service type assigned to each UE
Output:
uePriorityMap[]: Mapping of each UE to a priority class
rearrangeedQoS: QoS levels rearranged based on Premium user dominance
dominantQoS: The QoS service type with the most Premium users
Step 1: Initialize data structures
Initialize array uePriorityMap[]
Initialize map premiumCountMap[service]←0
Step 2: Compute number of Ues in each priority class
premiumUes ← $[totalUes \times \frac{premiumRatio}{100}]$
premiumUes ← $[totalUes \times \frac{premiumRatio}{100}]$
nonPremiumUes ← *totalUes* – *premiumUes* − *standardUes*
Step 3: Assign priority to each UE
for *i* ← () to *totalUes* −1 **do**
if *i* <*premiumUes* **then**
uePriority[i] ← "Premium" Gupta › Top-tier users
else if *i* <*premiumUes* + *standardUes* **then**
uePriority[i] ← "Standard" › Mid-tier users
else
uePriority[i] ← "NonP" › Non-Premium User
Step 4: Count Premium user per services type
for *i* ← 0 to *totalUes* −1 **do**
if *uePriorityMap[i]=*" *Premium*" **then**
service ← *ueServiceType*[i]
premiumCountMap[service]←*premiumCountMap*[*service*] +1
Step 5: Determine dominant QoS type
dominantQoS ← *arg max*(*premiumCountMap*) › Select service with most Premium Users
Step 6: Reorder QoS based on dominant type
rearrangedQoS ← *REORDERQoS*(*dominantQoS*)
Return results
return *uePriorityMap, rearrangedQoS, dominantQoS*

Algorithm 2: Application and eps bearer assignment

Input:
uelist : List of UE nodes
uePriorityMap[]: Priority level for each UE (Premium, Standard, NonP)
ueServiceType[]: Service type assigned to each UE
rearrangedQoS: QoS levels reordered by dominant Premium service
Output:
Each UE has one application and an EPS bearer configured and installed
Step 1: Loop through all UEs
for *i* ← 0 to *length*(*ueList*) − 1 **do**
node ← *ueList*[*i*] › Get current UE node
priority ← *uePriorityMap*[*i*] › Retrieve assigned priority class
service ← *ueServiceType*[*i*] › Get assigned service type
Step 2: Select proper EPS bearer based on priority and service
bearer ← SELECTBEARER(*service, priority, rearrangedQoS*) › Determine appropriate bearer using dynamic QoS order
Step 3: Create the traffic generating application

16	*app* ← CREATETRAFFICAPP(*node, service*)	› Generate traffic pattern based on service type
17	**Step 4: Install bearer on UE node**	
18	INSTALLBEARER(*node, bearer*)	› Attach the bearer to the UE network stack
19	**Step 5: Bind application to the UE's network interface**	
20	BINDAPPTOINTERFACE(*app, node*)	› Ensure traffic flows through correct
	bearer and network device	
21	**End Function**	

2.2. Algorithm description

The main highlight of the proposed scheme is due to its adaptive bearer assignment as shown in algorithms 1 and 2, which reaches a decision on user density within each of the service categories and raises the most densely populated Premium slice dynamically. As for the distribution analysis, the simulation begins by counting how many Premium users are present within each of the service containers (Video Streaming (VS), Voice Communication (VC), Gaming (GE), Web Browsing (BE), IoT (PE)). This is accomplished through iterating over all UEs and cross-matching their priority classification with membership in a container.

The most populated service container with Premium users is to be determined as the dominant slice. This is achieved by calling std::max() on the five count variables. For every UE, if the UE is Premium and belongs to the dominant slice, it is assigned premiumBearer. Otherwise, it is assigned an ordinary bearer according to its service type. Every UE is to be assigned a User Datagram Protocol (UDP) server application (sink). The distant host is configured with matching UDP client software targeted at each UE by port and IP.

Figure 2 illustrates the basic mechanism of the proposed user-centric QoS provisioning scheme. UEs are categorised into Premium, Standard, and Non-Premium categories. The categories are input into the dynamic QoS algorithm, which rearranges bearer assignments (Q_1 to Q_4) based on the dominant Premium user service type. The algorithm dynamically allocates bandwidth resources for efficiency and fairness with priority to delay-sensitive traffic.

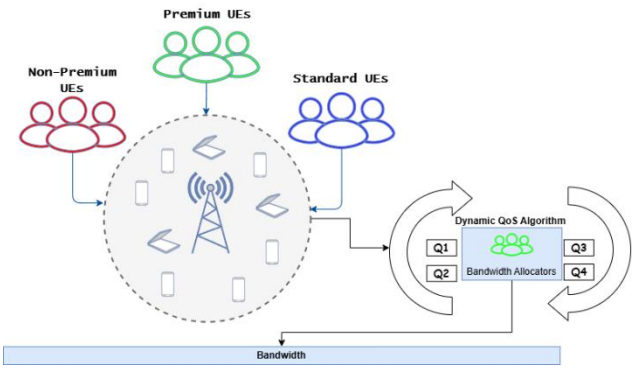

Fig. 2. Dynamic QoS-aware bandwidth allocation framework in 5G network.

Algorithm 1

This algorithm begins the decision process by allocating a user class: Premium, Standard, or Non-Premium to each UE (User Equipment) based on a predefined percentage allocation as shown in Table 1. Next, it examines the distribution of Premium users across multiple services and then dynamically determines the most dominant service type. The term "dominant" in this context implies the service with the greatest number of Premium users.

To compute this, the algorithm uses the argmax function, which is a mathematical function that gives the index (or key) of the highest value in a given list or map. In this case, the argmax function is used in a service count map, counting the number of Premium users linked to every service type. Then, after determining the most dominant QoS service, the QoS scheduler will be reconfigured to prioritise the most dominant service first in bearer setup and traffic handling. By reconfiguring, the system dynamically prioritises the most dominant services depending on the dominant user needs and network conditions.

Algorithm 2

Following the classification and rearrangement phase, each UE is set to be allocated a specific application and corresponding EPS bearer that reflects their user class and service type. The assignment ensures that each user generates traffic that is shaped and classified appropriately in the network stack, as well as facilitates proper flow monitoring during simulation, so that it will work accordingly, like how a real-life 5G network works.

Algorithm 2 takes as input of the collection of all UEs, the priority map from Algorithm 1, the assigned service type per UE, and the re-ordering QoS hierarchy. For every UE in the system, an algorithm selects a bearer class (e.g., GBR_CONV_VOICE, NGBR_VIDEO_TCP_DEFAULT) based on their user class and service type. Premium users receive GBR (Guaranteed Bit Rate) bearers and non-Premium users receive NGBR (Non-Guaranteed Bit Rate) bearers. Once the bearer is selected, an application of traffic, e.g., UDP echo, VoIP (Voice over IP), or FTP (File Transfer Protocol) is created and installed onto the UE. The application is associated with the bearer via interface-level mapping.

The two algorithms constitute a pipeline: Algorithm 1 determines the network scheduling strategy according to the present user distribution, and Algorithm 2 configures the network stack and traffic generation accordingly. The modular design allows the system to scale simply while allowing logical separation of classification and configuration of users. Interestingly, the dynamic re-ordering of QoS priorities gives rise to adaptive fairness. As Premium users increase, their dominant services are given higher priority.

However, as simulation results have shown, overall system delay tends to decrease rather than increase, contrary to initial expectations. This is due to the fact that high-priority services are scheduled faster. Traffic contention among low-priority bearers is reduced and the network avoids congestion by smart spreading of demand. To simplify and visualise the algorithm, a mathematical representation is proposed for the QoS allocation scheme. The following are the definition of the system modelled.

The total number of UEs in the simulation,

$$T \in N,\ T = |u| \tag{1}$$

where $u = \{u_0, u_1, u_2, u_3 \dots \dots, u_{T-1}\}$

Then, the user ratio as model parameters is shown as follows.

$R_p \in (0,1)$, Proposition of premium users

$P = [R_p * T]$, Number of premium UEs

$R_s \in (0,1)$, Proposition of standard users

$S = [R_s * T]$, Number of standard UEs

$R_n = 1 - R_p - R_s$, Proposition of non-premium users

$N = T - (P + S)]$, Number of non-premium UEs

As for the priority allocation function $\pi: u \to C$,

where, $C = \{Premium, Standard, NonP\}$

Then, for each UE index $i \in [0, T-1]$,

$$\pi(u_i) = \begin{cases} f(x) = \begin{cases} Premium,\ if\ i < P \\ Standard,\ if\ P \le i < P + S \\ NonP,\ if\ i \ge P + S \end{cases} \end{cases}$$

Finally,

$$UEPriorityMap = \{(u_i, \pi(u_i)) | i = 0,1 \dots, T-1\}$$

Then, in the simulation, all the various service types are allocated a fixed UDP packet size within the simulation in order to emulate the behaviour of actual traffic:

- Video Streaming (VS): 2500 bytes
- Voice Communication (VC): 250 bytes
- Gaming (GE): 512 bytes
- Web Browsing (BE): 1025 bytes
- IoT (PE): 64 bytes

These sizes simulate typical application demand, larger packets for high-throughput applications like video and smaller packets for latency-sensitive applications like voice and IoT. In the case of single-service simulation, network throughput at the end-to-end is typically low (≈3 Mbps), even in times of high user density. This is partly due to two factors:

- **Application-Limited Rate:** Rate (throughput) is not only dependent upon link capacity but also upon the rate at which applications generate packets. Small packets (such as 64 or 250 bytes), even at a high rate, equate to low byte per-second delivery.
- **Service Type Impact:** Where most users perform low-bandwidth applications (e.g., voice or IoT), end-to-end throughput is constrained naturally, though with low jitter and delay.

This kind of behaviour is normal as the algorithm is not solely interested in high throughput but rather low delay, jitter, and fair resource utilisation for user priorities. Throughput becomes a second-order priority in real-time QoS-intensive situations.

2.3. Randomness of the 5G network environment

A key feature of the proposed simulation platform is the deliberate use of randomness in user priority categorisation and service-type mapping. This is a deliberate design choice and is an essential contribution to mimic the natural randomness and heterogeneity of real user activity patterns in 5G network operations. In real-world deployments, user activity volumes, subscription levels (e.g., priority levels), service demand profiles (e.g., gaming vs. IoT), and traffic levels are neither constant nor homogeneous. Therefore, the incorporation of randomness into the simulation allows for the generation of diverse permutations of situations, which symbolise the adaptive, user-initiated character of real 5G network operations.

Specifically, the "AssignUserPriorities()" procedure randomises the UE index space and attributes Premium, Standard, and Non-Premium identifiers randomly with predetermined ratios. As such, each simulation run has the potential to generate a different dominant service slice. The largest one in terms of Premium users leads to bearer assignment, scheduling behaviour, and ultimately performance metrics such as delay, jitter, and throughput variation. This approach enhances the realism of the outcomes. Rather than tuning one static configuration, the algorithm in question is tried out on a broad spectrum of user distributions and both its performance in typical and edge-case situations is made apparent.

Although this randomness introduces variability to some results, it contributes to the algorithm's resilience and flexibility, desirable traits for the real-world 5G QoS provision. In short, the use of randomness does not weaken the consistency of the experiment but rather strengthens its representativeness, such that an objective and realistic assessment of user-centric QoS techniques in modern cellular networks can be facilitated.

3. Performance Evaluation

In an effort to comprehensively assess the effectiveness and robustness of the pioneering user-aware QoS provision framework in a 5G network, two simulation approaches were carried out. Both test scenarios were conducted under identical network scenarios and service distributions, i.e., changing user service demand ratios from 10% up to 90% with step sizes of 10%. This offers an equitable level for comparative assignments between various user behaviours and service loads. The first simulation approach adopts a "Single User to Multiple Services" configuration. In this configuration, every user equipment (UE) is assumed to provide multiple services simultaneously (e.g., voice, video, gaming, browsing, and IoT). The rationale behind the model is emulating actual usage patterns since current cellular devices have the tendency to execute different applications concurrently. The performance metric captured under this setup emulates the system's ability to handle multiple types of services for a single user with QoS assured.

Meanwhile, the second simulation approach adopts a "Single User to Single Service" paradigm. With that being said, each UE is only mapped to one specific

service type throughout the simulation run. This setup makes the traffic model straightforward and allows us to easily compare performance in controlled, segregated-by-service environments. This also allows us to see which service types are more likely to face congestion or resource shortage with greater user density.

Every UE was also assigned a priority level from the given set P = {Premium, Standard, NonP} based on different QoS requirements and handling in the network. User priorities were also mapped to service types S = {Voice, Video, Gaming, Browsing, IoT}, and each service had a related QoS bearer level from the set Q = {Q_1, Q_2, Q_3, Q_4}:

- Q_1 (Highest Priority): GBR_CONV_VOICE for voice calls of highest priority.
- Q_2 (High Priority): GBR_CONV_VIDEO for video streaming in real time.
- Q_3 (Medium Priority): NGBR_VIDEO_TCP_PREMIUM for low-latency gaming.
- Q_4 (Low Priority): NGBR_VIDEO_TCP_DEFAULT for non-real-time web browsing and IoT.

The performance evaluation focused primarily on delay, jitter, and throughput because these parameters directly influence the end user's quality of experience (QoE). Measurements over increasingly adverse service load conditions (from 10% to 90%) were recorded for each simulation run to test the scalability and fairness of the system in terms of resource use. The dual-testing approach contrasting single-service users against multiple-service users provides a robust setting for examining how varying traffic patterns affect the performance of the introduced algorithm for QoS provisioning. Surprisingly, the multi-service scenario placed greater stress on the system via more complex and dynamic resource demands, while the single-service scenario delivered cleaner observations of how a single category of service reacted independently.

Furthermore, in order to compare the performance and efficiency of the targeted user-specific QoS provisioning technique, a comparative baseline model was defined from a modified version of the standardised 5G simulation framework. The baseline does not necessarily come from external sources but is rather adapted from the publicly disclosed NS-3 5G-LENA module, which implements standardised QoS flows and EPS bearer mappings for typical user traffic. The base model mimics a "hard service" QoS plan, where service classes are statically mapped to static EPS bearers with little to no runtime adaptability to user context or traffic profiles.

The baseline model inherits static QoS assignments according to Third Generation Partnership Project (3GPP) Technical Specifications (TS) 23.203, where each service (video, voice, browsing) is statically mapped to an equivalent QoS Class Identifier (QCI) and its default bearer. These operations are not dynamic to the number or density of Premium customers and do not support any dynamic reassignment scheme for EPS bearers. As such, all users are essentially treated equally from a scheduling perspective, regardless of service priority or subscriber class (e.g., Premium versus Standard).

The primary purpose of employing this baseline model is to achieve a fair and uniform basis of comparison. Being a conventional QoS system with fixed bearer-service mapping, it offers a foundation upon which the performance of the new dynamic user-centric algorithm could be compared objectively with a non-adaptive

mechanism. The baseline also enables demonstrating the value provided by dynamic QoS reassignment in real-time network environments.

As compared to the planned approach, the baseline does not have per-user classification reasoning, Premium density analysis, nor service-dominant detection. With that being said, all bearers are statically defined ahead of time before simulation runtime and no adaptive adjustment occurs based on monitored traffic conditions. In fact, this is precisely the deficiency that led to the design of a more intelligent bearer reassignment scheme.

3.1. Multiple service (MS) Per UE

The first phase of performance testing evaluates the performance of the proposed QoS provisioning scheme when several services are provisioned per user. In this scenario, each UE is provisioned with multiple types of applications simultaneously, e.g. voice, video streaming, gaming, web browsing, and IoT traffic, mimicking the real-world mobile user behaviour. Although this setup mimics current usage patterns, the resulting performance measurements revealed several critical issues in terms of latency, jitter, and throughput. Data from Figs. 3 to 5 demonstrate the relative performance of the proposed model compared to the baseline QoS mechanism with different Premium user ratios, ranging from 10% to 90%.

In the multi-service per-user simulation model, there are multiple independent applications per UE, each of which corresponds to various types of services (e.g., voice, video, gaming, browsing, IoT). Even though each application is rightly mapped onto a particular EPS bearer and assigned its own QoS class, the underlying network stack and radio interface of the UE remain common to all the services. This common infrastructure facilitates competition at various layers:

- IP and MAC scheduling layers must multiplex multiple flows per UE.
- The downlink and uplink physical channels of the UE are shared by all services.
- NS-3 simulation queue puts all services' packets into a single per-UE transmission stream.

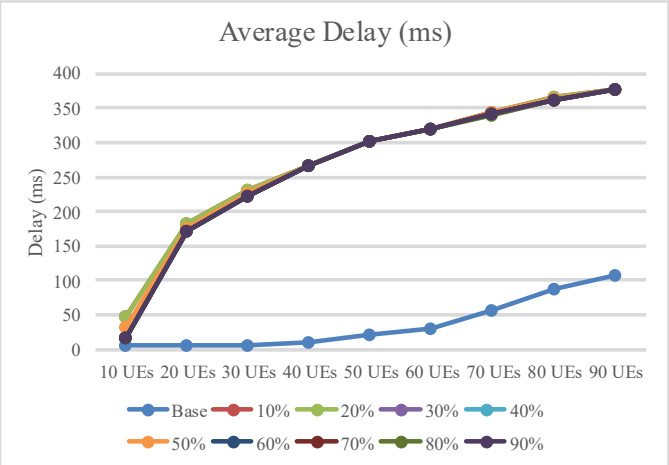

Fig. 3. Average delay for multiple services.

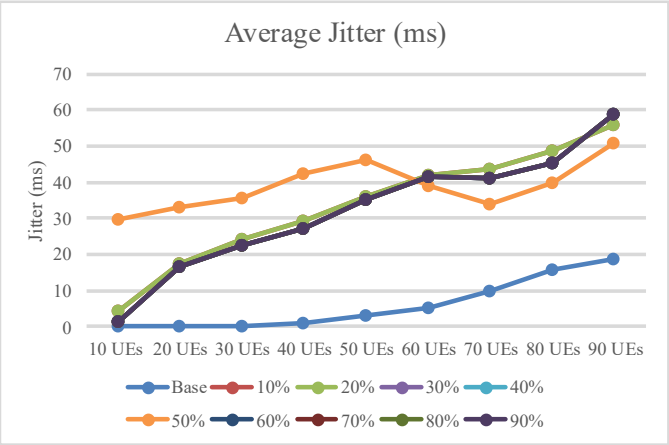

Fig. 4. Average jitter for multiple services.

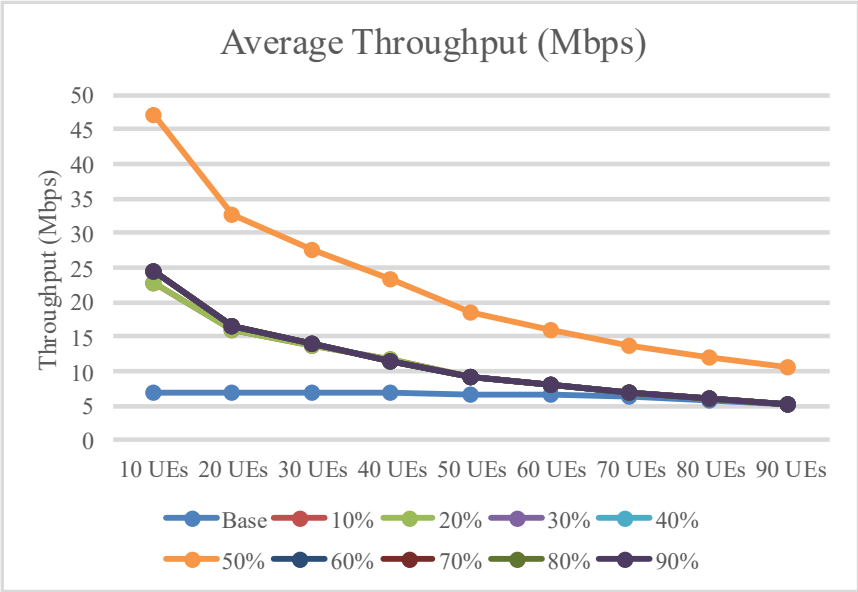

Fig. 5. Average throughput for multiple services.

Therefore, even with excellent bearer separation, priority service traffic will tend to experience larger delays due to serialisation behind lower-priority traffic in the same UE. Moreover, as more premium subscribers are supported, aggregate per-UE traffic volume increases (due to increased activity and more bulky packets), rendering queuing and scheduling delay worse. On the other hand, the baseline case where each UE is assigned to a single service has no contention at all. Each UE simply has a single application with a single bearer, with which scheduling can be effective, queuing simpler, and delay performance more consistent. Thus, for the same traffic load throughout the network overall, the per-UE service multiplexing in the multiple-service model results in intra-UE congestion and gives rise to uniformly increased end-to-end delay compared to the single-service-per-UE baseline.

3.1.1. Delay performance for MS

Based on the delay Fig. 3, it clearly shows the multiple-service per-user approach providing much greater end-to-end delay compared to the base. With a rising ratio of Premium users, the delay in the model remains greater too, showing inefficiency in processing traffic. This degradation is caused by the intra-simulation communication structure: when multiple applications are tied to a single UE, their associated traffic shares the same network buffers, queues, and even bearers in certain instances. Through this mechanism, low-priority flows (such as browsing or IoT) could hinder high-priority flows (such as voice or video) under the same UE. This contention causes greater queuing delays, buffer overflows, and scheduling inefficiency, particularly under increased traffic levels.

3.1.2. Jitter performance for MS

The jitter outcome also varies in the same trend as Fig. 4. The new scheme shows higher jitter variation than the base code, particularly in the case of mid-range Premium user proportions (30% to 70%). Jitter, which reflects packet arrival time variability, is highly influenced by the lack of consistency in processing traffic for each UE. Since each service type adds traffic at varying rates and with varying packet sizes, the overall traffic pattern in each UE becomes non-deterministic. This kind of non-determinism results in irregular servicing of packets and queue filling, thereby increasing jitter. Lack of flow-level traffic differentiation within each UE further intensifies this effect.

3.1.3. Throughput performance for MS

Figure 5 shows that the average throughput of the proposed model is smaller than the baseline system. Even though throughput does not degrade as much as delay and jitter, there is certainly a reduction seen. This is because of inefficient bearer allocation and resource contention resulting from simultaneous flows of traffic from different applications at the same UE. Lower-priority services, unless properly isolated, will dominate transmission opportunities meant for higher-priority services, especially where bandwidth is scarce. This leads to bad scheduling and some lower aggregate throughput.

3.1.4. Discussion for multiple services for MS

Although the multiple-service per-user model is conceptually sound and appealing, it is inherently complex from a QoS perspective. The outcome of the simulation demonstrates that assigning concurrent services to one UE, lacking the proper separation of bearers and flows, results in severe performance degradation. Such observations suggest that the existing implementation in NS-3 cannot support inherent intra-UE traffic prioritisation in multi-service environments efficiently.

For countering these limits, the simulation strategy was altered in the next phase to adopt a single-service per-user policy, where each UE is connected to a particular application category and its relative QoS bearer. As detailed in the forthcoming section, this approach results in tremendous performance improvement for all the metrics, proving the validity of isolated service handling in user-based QoS provisioning models. With that being said, multiple service algorithm design is not ideal for most cases.

3.2. Single service (SS) Per UE

The second phase of the simulation testing covers the evaluation of the proposed QoS provisioning scheme when each user is assigned a distinct, personal service type. In contrast with the previous multi-service scenario, where each user experienced a mix of applications, this step isolates traffic by assigning a single type of application, such as voice, video, gaming, browsing, or IoT to each UE. This approach provides a controlled setup to test the performance of the QoS mechanism using reduced, service-oriented loads. Performance metrics were recorded for various ratios of Premium users (10% to 90%) and have been visualised in Figs. 6 to 11 and show the mean network delay, jitter, and throughput under this single-service scenario.

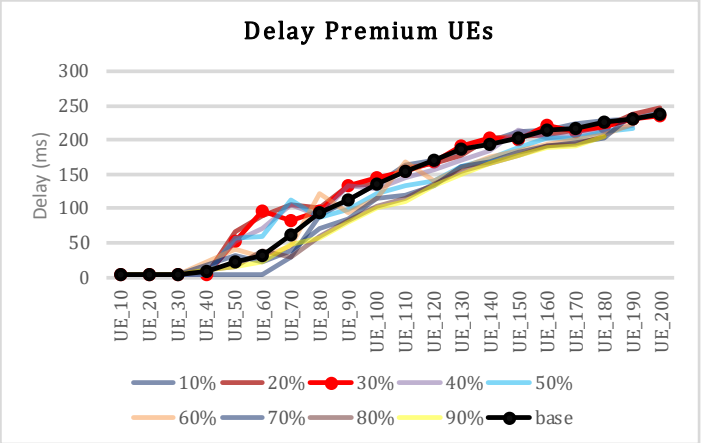

Fig. 6. Delay output for premium user (ms).

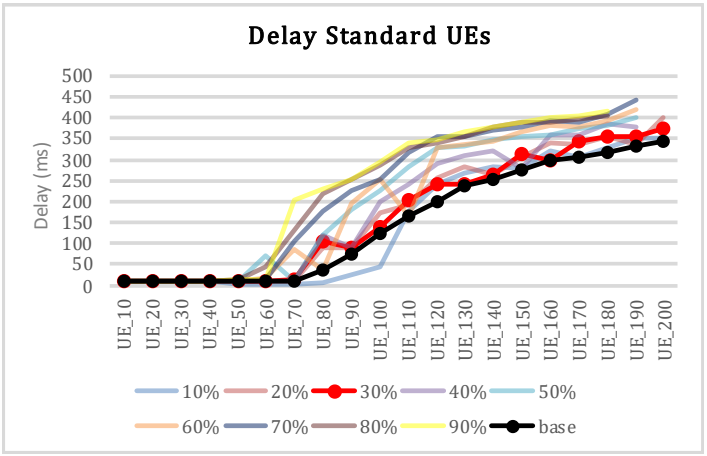

Fig. 7. Delay output for standard user (ms).

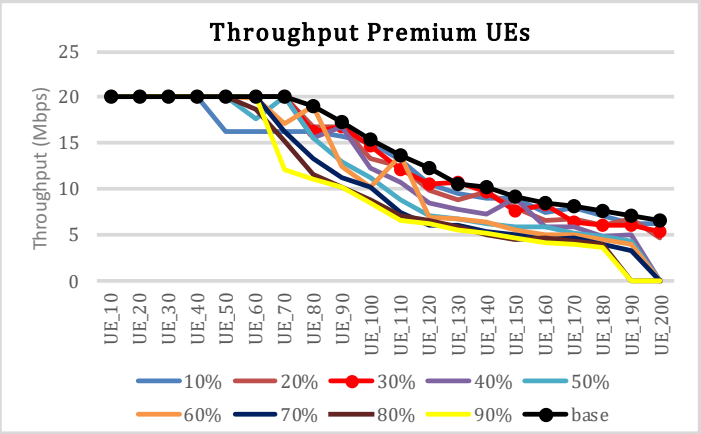

Fig. 8. Throughput output for premium user (Mbps).

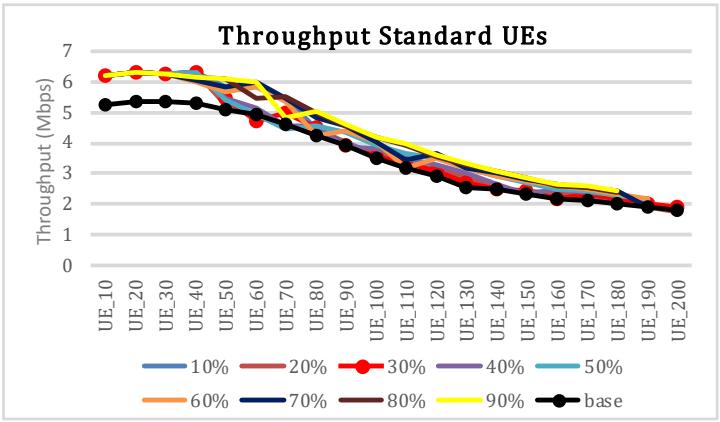

Fig. 9. Throughput output for standard user (Mbps).

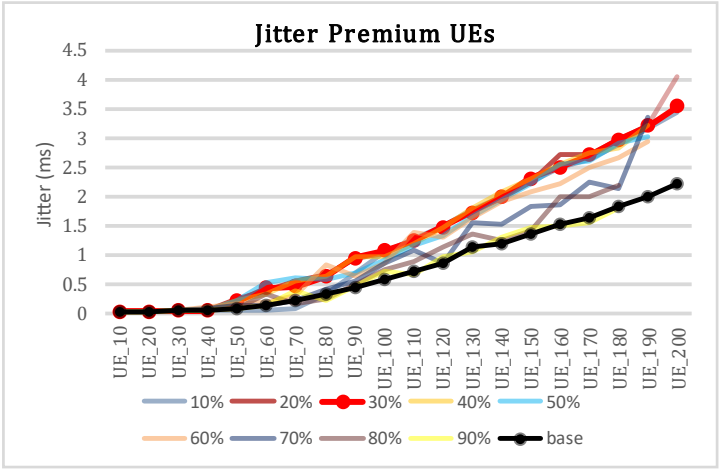

Fig. 10. Jitter output for premium user (ms).

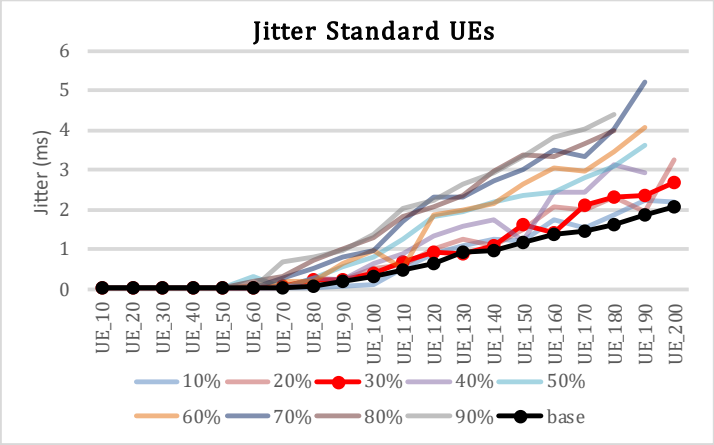

Fig. 11. Jitter output for standard user (ms).

3.2.1. Delay performance for SS

The single-service model delay performance demonstrates a clear trend with rising network load. Based on Figs. 6 and 7, for smaller UE counts (10 to 30), the average delay is maintained at around 5.2 ms even when premium user densities fluctuate. However, beginning from 40 UEs, a sudden delay rise is observed for the base case to 9.69 ms, whereas the algorithm keeps the delay under 5.5 ms for up to 60 UEs for some premium settings. For 70 UEs and more, delays become quite high for both base and algorithm results. For 100 to 200 UEs, both the base and the algorithm exhibit extremely high delay values, with the algorithm reaching an all-time maximum of 231 ms in 170 UEs for certain setups. This indicates that while the algorithm can reduce delay at mid-scale loads by means of bearer priority reordering based on premium density, gains are sacrificed in very high user congestion due to radio and bearer resource saturation.

3.2.2. Jitter performance for SS

Jitter that reflects the variation in delay of packets also has a similar behaviour. Based on Figs. 10 and 11, in low to moderate load environments, jitter is extremely low since the packet scheduling is stable with little queuing delay. The algorithm maintains jitter under control in this regime by allocating high QoS bearers to stratum services in order to give constant latency to such flows. However, with increasing numbers of users, especially of the past 70 to 90 UEs, jitter values by the algorithm sometimes go beyond the base. This is a result of aggressive bearer switching based on premium user density. This improves average delay for high-priority users, it simultaneously results in intra-bearer competition as well as dynamic queuing instability, especially if non-premium users are redirected to lower QoS bearers under congestion. The jitter is significantly bigger as queues change in response to variable availability of resources.

3.2.3. Throughput performance for SS

Based on Figs. 8 and 9, the SS model throughput shows early improvements using the provided dynamic algorithm, particularly in 20 to 60 UEs, in which the

algorithm dynamically reallocated bandwidth to bearers of the best service with the highest premium users. The resultant dynamic reassignment leads to better link utilisation than the base model based on static bearer to service allocation. Both the base and the algorithm experienced a decline in throughput when the user population beyond 100 UEs due to heavy load. This is due to application-constrained data generation with packet loss and retransmission (inferred by Rx/Tx imbalance) that restricts achievable throughput. The packet sizes are fixed per service (e.g., 2500 bytes per video, 64 bytes per IoT) also factor in if the high-usage service is a small packet service. The high throughput is impossible even with high bearer priority due to low byte volume per transmission.

Furthermore, the algorithm's performance is made to be fair and delay-conscious instead of bandwidth hungry. It thus avoids starving normal and non-premium users in any way, which would lead to somewhat reduced peak throughput but much more equitable and more predictable performance for all UEs. In summary, the proposed reasoning-based QoS provisioning algorithm demonstrates nuanced and adaptive performance across various classes of users under various intensities of network loading.

One significant observation is the increase in jitter at the 80-UE point for both Premium and Standard users. This is because the algorithm is in the course of transition scheduling, the network is moderately loaded but not saturated. In this region, the algorithm begins reassigning bearers based on the prevailing Premium slice density, introducing randomness in scheduling latency that manifests as increased jitter. The peak is a shift in internal resource allocation and reflects the load sensitivity of the algorithm.

Throughput analysis reveals that Premium users, while receiving higher throughput than Standard users, receive consistently lower throughput than the baseline. This outcome arises through controlled "premiumBearer" assignment, where only the most concentrated service category of Premium users gets upgraded bearers. This protects against excessive GBR traffic and contributes to overall decongestion but modestly limits increases in throughput above a baseline that can privilege too much Premium flow. Notably, Standard users enjoy more throughput than the baseline, but less than Premium users. This is because constant assignment of service bearers without the overhead of dynamic prioritisation provides improved efficiency under certain loads.

While the delay performance improves the effectiveness of the proposed scheme. The premium users experience less average delay than the baseline, reflecting their guaranteed network resource access in the prevailing service slice. The Standard users, on the other hand, experience a moderate increase in delay, as they are scheduled after premium flows during resource contention. This suggests that the algorithm successfully enforces delay-sensitive prioritisation while ensuring access fairness among all users.

3.3. Comparison of MS and SS

Table 2 shows three QoS provisioning models in user-oriented 5G network management: the traditional Base model, the Multi-Service per User (MSPU) model, and the proposed Single-Service per User (SSPU) model. Comparative analysis was conducted using some key network performance metrics, including delay, jitter, throughput, and user experience, with a focus on Premium and

Standard user distinctions. The Base model is characterised by its staticity, offering no QoS adaptation and user class awareness. It provides one service stream per user, which is predictable but inflexible. While delay and jitter are minimal at all times during light loads, the model neither has any preference for Premium users nor dynamically responds to network demand. Simulating it is simple and linear with high realism and low flexibility.

The MSPU model introduces service awareness by enabling several service types per UE. While this more accurately models real-world user behaviour, it introduces severe intra-UE flow contention. This leads to severe delay and jitter increases, especially when Premium users dominate the population. While it encounters higher total throughput due to congested streams of traffic, this model defeats Premium user advantage since shared queues weaken the intended-for prioritisation benefit. It also hurts the Standard user experience since they now suffer queue congestion as a result of Premium-heavy flows. Moreover, MSPU greatly reduces simulation complexity and transparency, and renders correct analysis and adaptation is complicated.

The proposed SSPU model, however, offers a neat and harmonious solution. One service per user is provisioned, which provides precise QoS adaptation at both user and service levels. Dynamic adaptation is achieved through an algorithm that resizes bearer types as a function of real-time Premium user density per service. The result is a system that ensures the lowest delay and jitter for Premium users despite high Premium ratios (e.g., 90%) while still ensuring decent performance for Standard users. In contrast to MSPU, throughput is constant and reasonably even. This evenness is vital when it comes to sustaining perceived quality, particularly during high-load scenarios like city areas or dense areas.

In addition, the SSPU model is compatible with the 3GPP QoS Class Identifier (QCI) concepts to support bearer mapping and traffic shaping directly without invoking SDN or AI mechanisms. This supports tractability without compromising realism, offering a feasible path for small network operators where both performance enhancement and lack of excessive system complexity are desired. Simulation experiments show Premium users benefiting noticeably, while there are appreciable reductions in latency and jitter, and only tolerable performance loss for Standard users.

Overall, this comparative table justifies the SSPU model as a practical and scalable solution for future networks, especially when it comes to simulation scenarios. It imposes user-centric fairness, protects Premium user QoE, and supports realistic operator deployment scenarios. The model is able to bridge the gap between policy-driven QoS and real-time traffic responsiveness, thereby leading towards inclusive, adaptive, and efficient 5G connectivity.

Table 2. The comparison of MS and SS.

Metric	Base Model	MSPU Model	SSPU Model
QoS Adaptation	Static (no adaptation)	Service-aware, but intra-UE contention prevents proper QoS	Full per-UE & per-service adaptation using bearer reassignment
User Class Awareness	None	Partial (class defined, but shared queues limit benefit)	Strong: Dynamic reassignment based on Premium user density
Traffic Handling	1 service per UE	Multiple services per UE	1 service per UE

Delay	Consistent but inflexible	Very high delays due to flow contention	Lowest delay for Premium, good scaling with load
Jitter	Low jitter in light traffic	Very high jitter especially at mid-premium levels	Low jitter even at 90% Premium, more predictable QoE
Throughput	Balanced across users, but unoptimised	Higher total throughput due to dense traffic	Stable throughput, fair distribution
Premium User Advantage	None (equal treatment)	Not effective (intra-UE queues reduce Premium prioritisation)	Premium users benefit from lower delay/jitter with minimal cost to others
Non-Premium Experience	Same as Premium	Affected by Premium-heavy queues	Slightly lower throughput, but delay/jitter remains acceptable
Simulation Complexity	Simple	Very complex, unpredictable due to multi-service per UE	Moderate complexity, high control precision
Realism vs. Tractability	High realism, low flexibility	High realism, low observability	Balanced: Realistic enough, analytically clear

4. Conclusion

Ultimately, this paper proposed and evaluated a user-centric Quality of Service (QoS) provisioning solution for 5G networks to address the shortcomings of operator-centric, Premium-oriented schemes. The solution offered a rationale-based algorithm that dynamically allocates EPS bearers based on user priority and service allocation without the requirement of high-resource paradigms like Machine Learning (ML) or Software Defined Networking (SDN). Each subscriber was assigned to a single one of five service types (Video, Voice, Gaming, Browsing, IoT) and one of three priority classes (Premium, Standard, Non-Premium). The mechanism then selected the most heavily loaded Premium service slice and allocated high-priority bearers only in that slice selectively. This managed approach ensured fair resource allocation and prevented over-saturation of guaranteed services.

Simulation output showed consistent performance to 150 UEs and 30% for Premium density. However, above 190 UEs or at high Premium ratios, throughput collapse and attachment failures occurred, showing system overload. Interestingly, with a steady load of 100 UEs, average delay decreased beyond 30% Premium users, showing an effect due to increased scheduling efficiency when Premium traffic was concentrated in one service slice.

Overall, the proposed scheme demonstrated that lightweight, reason-based logic can offer dynamic, fair QoS management in 5G networks. It compromises between service assurance and system stability and avoids AI-based orchestration overhead. Future work can apply this model to multi-cell or RAN slicing cases, but it is not necessary for the foundation set forth here to have excellent potential for scalable, user-aware QoS provisioning.

Abbreviations

5G	Fifth Generation Mobile Network
AI	Artificial Intelligence
ARP	Allocation and Retention Priority

BWP	Bandwidth Part
eMBB	Enhanced Mobile Broadband
EPS	Evolved Packet System
GBR	Guaranteed Bit Rate
gNB	Next Generation Node B (5G base station)
KPI	Key Performance Indicator
IoT	Internet of Things
IP	Internet Protocol
LTE	Long-Term Evolution
Mbps	Megabits per Second
ML	Machine Learning
MAC	Medium Access Control
mMTC	Massive Machine Type Communication
ms	Milliseconds
NGBR	Non-Guaranteed Bit Rate
NR	New Radio (5G NR)
OFDM	Orthogonal Frequency Division Multiple Access
PDCP	Packet Data Convergence Protocol
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RLC	Radio Link Control
SD	Standard Deviation
SDN	Software Defined Networking
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
SRS	Sounding Reference Signal
TDD	Time Division Duplex
TTI	Transmission Time Interval
UE	User Equipment
URLLC	Ultra-Reliable Low-Latency Communication

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

1. Matoussi, S.; Fajjari, I.; Aitsaadi, N.; and Langar, R. (2023). User-centric slice allocation scheme in 5G networks and beyond. *IEEE Transactions on Network and Service Management*, 20(4), 4268-4282.
2. Jahandar, S.; Shayea, I.; Gures, E.; El-Saleh, A.A.; Ergen, M.; and Alnakhli, M. (2025). Handover decision with multi-access edge computing in 6G networks: A survey. *Results in Engineering*, 25, 10394.
3. Albekairi, M. (2025). Controlled service scheduling scheme for user-centric software-defined network-based internet of things. *IEEE Access*, 13, 19198-19218.

4. Anjum, M.; Min, H.; and Ahmed, Z. (2024). User-centric internet of things and controlled service scheduling scheme for a software-defined network. *Applied Sciences*, 14(11), 4951.
5. Xiang, Z.; Ying, F.; Yan, H.; Zheng, Z.; Zhang, Y.; and Xu, Y. (2025). QoS-effective and resilient service deployment and traffic management in MEC-based crowd sensing. *Symmetry*, 17(5), 718.
6. Ito, K.; Nakazatoa, J.; Fontugne, R.; Tsukada, M.; and Hiroshi, E. (2025). A multipath redundancy communication framework for enhancing 5G mobile communication quality. *Computer Communications*, 239, 108157.
7. Kalem, G.; Kosu, S.; and Basaran, M. (2024). 5G/6G technology capabilities designed for secure edge network: Smart city use cases of Turkcell. *Proceedings of the* 2024 6th *International Conference on Blockchain Computing and Applications* (*BCCA*), Dubai, United Arab Emirates, 807-811.
8. Bouzid, T.; Chaib, N.; Bensaad, M.L.; and Oubbati, O.S. (2022). 5G network slicing with unmanned aerial vehicles: Taxonomy, survey, and future directions. *Transactions on Emerging Telecommunications Technologies*, 34(3), e4721.
9. Vidhya, P.; Subashini, K.; Sathishkannan, R.; and Gayathri, S. (2025). Dynamic network slicing based resource management and service aware Virtual Network Function (VNF) migration in 5G networks. *Computer Networks*, 259, 111064.
10. He, H.; Yang, X.; Mi, X.; Shen, H.; and Liao, X. (2024). Multi-agent deep reinforcement learning based dynamic task offloading in a device-to-device mobile-edge computing network to minimize average task delay with deadline constraints. *Sensors*, 24, (16), 5141.
11. Aslam, U. (2025). Designing flexible scheduling algorithm for 5G. *International Journal of Advanced Engineering*, *Management and Science*, 11(1), 090-108.
12. Taha, I.Z.; and Janaby, A.O.A. (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.
13. Kooshki, F. (2023). *Open cell-less network architecture and radio resource management for future wireless communication systems*. PhD Thesis, Department of Multimedia and Communications, Universidad Carlos III de Madrid, Madrid.
14. Zeyad Taha, I.; and Othman Al Janaby, A. (2024). Efficient radio resource management in cell-less wireless communication systems. *Journal of Telecommunications and Information Technology*, 98(4), 1-8.
15. Kooshki, F.; Armada, A.G.; Mowla, M.M.; and Flizikowski, A. (2023). Radio resource management scheme for URLLC and eMBB coexistence in a cell-less radio access network. *IEEE Access*, 11, 25090-25101.
16. Umar, M.M.; Mohammed, A.; and Abdulazeez, A. (2024). Review of QoS-aware resource allocation schemes for 5G networks. *Dutse Journal of Pure and Applied Sciences*, 10(3c), 296-303.
17. Ramesh, P.; Bhuvaneswari, P.T.V.; Viswanath, L.; and Stephen, B.J. (2024). Software defined network architecture based network slicing in fifth

generation networks. *Informacije MIDEM - Journal of Microelectronics Electronic Components and Materials*, 54(2), 123-130.

18. Alashjaee, A.M.; Kushwaha, S.; Alamro, H.; Hassan, A.A.; Alanazi, F.; and Mohamed, A. (2024). Optimizing 5G network performance with dynamic resource allocation, robust encryption and Quality of Service (QoS) enhancement. *PeerJ Computer Science*, 10, e2567.
19. Louvros, S.; Paraskevas, M.; and Chrysikos, T. (2023). QoS-aware resource management in 5G and 6G cloud-based architectures with priorities. *Information*, 14(3), 175.
20. Malik, S.U.R.; Kanwal, T.; Khan, S.U.; Malik, H.; and Pervaiz, H. (2021). A user-centric QoS-aware multi-path service provisioning in mobile edge computing. *IEEE Access*, 9, 56020-56030.
21. Jayaraman, R.; Manickam, B.; Annamalai, S.; Kumar, M.; Mishra, A.; and Shrestha, R. (2023). Effective resource allocation technique to improve QoS in 5G wireless network. *Electronics*, 12(2), 451.