MSTS Methods in Science and Technology Studies

futuristech PT. Teknologi Futuristik Indonesia

Received: May 20, 2026; Revised: June 29, 2026; Accepted: July 12, 2026; Published: July 13, 2026; Current version: July 13, 2026. e-ISSN: 3123-4232
Digital Object Identifier 10.64539/msts.v2i2.2026.513

Article

MLHP: A Multi-Level Handoff Prioritization Framework for Service Differentiation in Buffered Systems

Oluwasanya Oyedele[1,*], Akinniyi Stellamaris Omowunmi[2], Akinola Kayode Emmanuel[3]

[1] Department of Software Engineering, McPherson University, Seriki Sotayo, Ogun State, Nigeria; e-mail: oyedeleo@mcu.edu.ng O. Oyedele).
[2] Department of Information Technology and Information Communication Technology, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria; e-mail: akinniyiso@funaab.edu.ng (A. S. Omowunmi).
[3] Department of Information Technology, McPherson University, Seriki Sotayo, Ogun State, Nigeria; e-mail: akinolaoe@mcu.edu.ng (A. K. Emmanuel).
* Correspondence Author

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Abstract: In modern wireless networks, efficient handoff strategies are crucial for several services with various Quality of Service (QoS) requirements. However, a significant research gap exists as most current handoff techniques treat all internet traffic uniformly, leading to performance degradation, latency, and glitches for time-sensitive applications like Ultra-Reliable Low Latency Communication (URLLC) during network transitions. To address this, the main objective of this study was to develop the Multi-Level Handoff Prioritization (MLHP) framework specifically for buffered handoff setups. The MLHP system integrates three core components: a multi-level service classifier, a dedicated buffering architecture with dynamic thresholds, and a hybrid scheduling mechanism combining Strict Priority and Weighted Fair Queuing. Simulation results reveal that MLHP significantly outperforms both traditional Non-Prioritized Buffered Handoff (NPBH) and Dynamic Queue Management (DQM) schemes. Key findings indicate that MLHP maintains a low dropping probability of approximately 6% under high handoff frequencies and achieves an aggregate throughput exceeding 44 Mbps during high mobility scenarios, while successfully maintaining sub-10 ms delays specifically for mission-critical URLLC traffic. The broader implications of this study suggest that MLHP provides a scalable and flexible solution for handoff management, effectively meeting the stringent requirements of 5G-and-beyond networks. By ensuring granular service differentiation, the framework enhances overall network reliability and user experience in increasingly heterogeneous mobile environments.

Keywords: Multi-Level QoS; Buffered Handoff System; Service Differentiation; Prioritization; Service Classification.

Copyright: © 2026 by the authors. This is an open-access article under the CC-BY-SA license.

1. Introduction

The ever-growing use of mobile devices in diverse services, from real-time multimedia streaming to mission-critical applications, has necessitated the need for an efficient handoff strategy to be deployed in wireless communication systems [1]. Handoff is a fundamental operation in both cellular and other wireless networks. It is used to transfer an active connection from one base station (BS) to another as a mobile user (MU) moves [2]. If not well managed, handoff can lead to service interruption, increased latency, and connection drops which greatly affect user experience negatively.

Existing handoff methods usually treat all connections the same, forgetting the variations in QoS requirements of different applications. For instance, delay and jitter are more sensitive in video calls than a simple email transfer. Since the emergence of 5G and beyond networks, attention has been drawn to different services like enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC). For effective use of this services, better QoS-aware solutions must be birth [3]-[5]. With the introduction of buffered handoff systems, a promising approach seems to be offered. In this system,

data packets are stored temporarily during transition to avoid packet loss. However, with heavy network load or frequent handoffs, these buffers can become bottleneck without a proper prioritization mechanism resulting in increased delays and potential buffer overflows [6]-[8].

This paper introduces a Multi-level Handoff Prioritization (MLHP) framework designed specifically for buffered handoff systems. MLHP is birth to categorized traffic into different priority levels based on their sensitivity to delay, packet loss, and throughput. After which it manages the handoff buffer and scheduling accordingly. The aim of framework is to provide superior QoS to high-priority tasks while ensuring lower-priority tasks runs smoothly. This approach is necessary in heterogeneous demands of modern wireless networks, where seamless and differentiated connectivity is of essence.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of existing literature concerning handoff management and QoS differentiation in wireless networks. Section 3 details the methodology of the proposed MLHP framework, including its architectural components, mathematical modeling, and the discrete-event simulation environment setup. Section 4 presents the simulation results and a comparative performance analysis against established schemes. Section 5 discusses the limitations of the current study and outlines potential future research directions, while Section 6 concludes the paper with a summary of the key findings and contributions.

2. Literature Review

A lot of research has been carried out in handoff management and QoS provision in wireless network, with meaningful advancement in recent years. Early research work on handoff mainly considered reduction of handoff dropping probability and increasing call admission rate [9]-[11]. But with the emergence of more intelligent services, the idea of QoS has appeared in handoff decision-making and resource allocation.

In improving QoS during handoffs, several approaches have been proposed in research. A popular approach is the resource pre-reservation, where resources are reserved at the target base station (TBS) long before handoff operation takes place [12]-[14]. This strategy, though effective in the reduction of handoff dropping, can lead to inefficient resource utilization with a failed handoff process. Another approach is the predictive handoff techniques. This technique utilizes mobility patterns and signal strength measurements to predict handoffs. This allows for proactive resource allocation and buffer management [15]-[17]. Recent works have explored machine learning techniques in handoff prediction to enhance QoS in dynamic environments [18]-[20].

Towards improvement, buffered handoff systems have gained attention as a means to experiencing an uninterrupted handoff process by temporarily storing data packets on transit to avoid loss. [20], [21] examined buffer management strategies in software-defined networking - enabled handoff architectures. The benefits of centralized control for QoS was highlighted, however, performance degradation of the system is inevitable with the introduction of buffer without intelligent prioritization.

In networking, QoS differentiation has been a major concern. Both Differentiated Services (DiffServ) and Integrated Services (IntServ) architectures have been deployed to provide QoS at the network layer [22]-[24]. Researchers have explored various prioritization schemes in the context of wireless handoff, among which are: priority queueing mechanisms that give precedence to high-priority calls on requests [19], [25], a fuzzy logic-based QoS-aware handoff decision scheme for vehicular ad-hoc networks [26], [27], a priority-aware adaptive buffering scheme for 5G industrial IoT applications [22], a deep reinforcement learning approach for dynamic QoS provisioning during handoff [28], [29].

While existing literature offers valuable insights into various aspects of handoff and QoS, there is a need to develop a system that will explore the advantage of both concepts. Hence, this work that integrates buffer management with service differentiation at multiple levels. a comprehensive multi-level QoS prioritization framework specifically designed for buffered handoff systems, that rigorously integrates buffer management with service differentiation at multiple levels, remains an area for further exploration. Many current schemes focus on a binary high/low priority or a limited set of service classes. By examining the specific QoS requirements of different applications in modern wireless networks, this framework provides a more detailed and flexible solution.

3. Methodology

The proposed MLHP framework designed for buffered handoff systems comprises of three key components, which are: service classification, handoff buffer management, and prioritized packet scheduling.

3.1. System Architecture

As illustrated in Figure 1, the framework architecture consists of three primary components: Mobile Users, the Source Base Station, and the Target Base Station.

3.1.1. Mobile Users (MUS)

These entities serve as the originators of network traffic, generating a variety of different service types that must be processed by the system.

3.1.2. Source Base Station (SBS)

Functioning as the central hub of the framework's operation, the Source Base Station is responsible for managing core logic. It is equipped with several critical

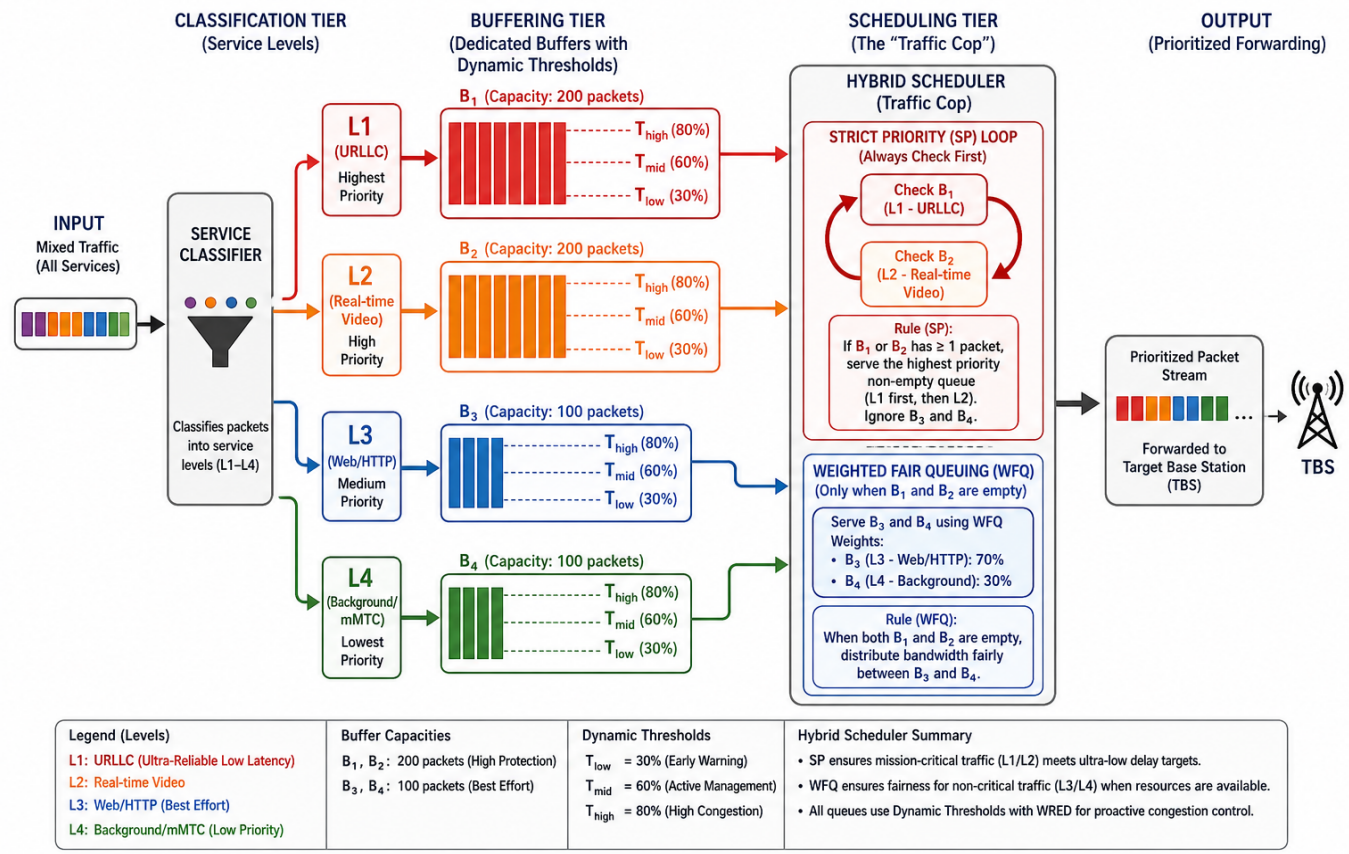

Figure 1. System architecture.

components, including a Service Classifier for identifying traffic, a QoS Policy Database for maintaining service standards, and a dedicated handoff buffer management module.

3.1.3. Target Base Station (TBS)

The Target Base Station facilitates the transition of services and is equipped with its own QoS Policy Database. Additionally, it possesses packet forwarding capabilities to ensure the continuity of data transmission during handoffs.

The service classification process at the Source Base Station (SBS) serves as the foundational step for traffic management, where incoming data is differentiated according to specific Quality of Service (QoS) requirements.

3.2. Service Classification

3.2.1. Priority Hierarchy

The framework organizes services into a hierarchical structure consisting of P priority levels, ranging from L1 to LP. Within this system, L1 is designated as the highest priority tier, while LP represents the lowest.

3.2.2. Classification Criteria

To determine the appropriate priority level, the system inspects packets based on three critical performance metrics: the maximum tolerable delay, the required bandwidth, and the acceptable packet loss rate.

3.2.3. Priority Mapping and Application

This classification system maps specific technologies to appropriate tiers based on their urgency. Level 1 is reserved for mission-critical Ultra-Reliable Low-Latency Communications (URLLC), such as autonomous driving or remote surgery. Intermediate tiers, specifically Levels 2 and 3, handle real-time applications including video conferencing, VoLTE, and online gaming. The lowest tier, Levels 4 is utilized for non-urgent traffic such as general web browsing, email, and background massive machine-type communications (mMTC).

3.2.4. Operational Mechanism

The service classifier at the SBS implements this logic by consulting the QoS Policy Database. By referencing this database, the classifier tags each individual packet and routes it to the appropriate queue to ensure it receives the necessary network resources.

3.3. Handoff Buffer Management

Following the classification phase, services are directed to a sophisticated buffering system at the Source Base Station (SBS) designed to mitigate data loss during the transition between base stations.

3.3.1. Dedicated Buffer Architecture

Rather than utilizing a single, uniform storage pool, the system implements dedicated buffers (B_k) for each

Table 1. Traffic classes.

Priority Level	Service Type	Traffic Model	Packet Size	QoS Requirement
Level 1	URLLC (Mission Critical)	Constant Bit Rate (CBR	64 – 128 bytes	Delay < 10 ms
Level 2	Real-time video	Variable Bit Rate (VBR)	1500 bytes	Low jitter
Level 3	Web browsing (HTTP)	Poisson distribution	Variable	Best effort
Level 4	Background (mMTC)	Bulk data transfer	1024 bytes	High loss tolerance

distinct priority level. This architecture ensures that traffic from different service classes remains isolated and manageable according to its specific requirements.

3.3.2. Dynamic Resource Allocation

The system employs dynamic rather than static buffer capacities. To minimize the probability of data drops for critical services, higher-priority levels such as L_1 and L_2 are allocated larger storage capacities, ensuring greater resilience during peak traffic periods.

3.3.3. Tiered Management via Multi-Level Thresholds

To maintain stability and manage congestion before it reaches a critical state, each buffer is governed by dynamic thresholds (T_k). These thresholds act as specific trigger points for different management actions:

- **Low Threshold:** This level serves as an early warning indicator, signaling the onset of congestion.
- **Mid Threshold:** At this stage, the system activates Weighted Random Early Detection (WRED). This mechanism specifically targets lower-priority packets for early management to prevent the buffer from reaching capacity.
- **High Threshold:** Upon reaching this final limit, the system initiates active packet dropping. This process specifically targets the lowest-priority data, effectively clearing space to ensure that high-priority packets are preserved and successfully transmitted.

While the MLHP framework is designed to support adaptive thresholding, for the purposes of this simulation, these thresholds (T_{low}, T_{mid}, T_{high}) were implemented as fixed percentages (30%, 60%, and 90%) of the allocated buffer capacity. This approach was taken to establish a consistent baseline for comparative analysis against existing schemes.

3.4. Prioritized Packet Scheduling

This is the "traffic cop" mechanism that decides the order in which buffered packets are forwarded to the Target Base Station (TBS) during a handoff. It uses a hybrid approach:

- Strict Priority (SP): This is applied to the highest levels (L_1 and L_2). The scheduler will always send L_1 and L_2 packets first. As long as there is data in the L_1 or L_2 buffers, no packet from lower levels are transmitted. This ensures near-zero latency for mission-critical apps.
- Weighted Fair Queuing (WFQ): This is applied to the lower levels (L_3 to L_P). Once the high-priority buffers are empty, the remaining bandwidth is shared among the lower-priority queues. Each queue is assigned a weight (w_i): This prevents "starvation," a common problem where low-priority data (like an email) never gets sent because high-priority data (like a video) is constantly arriving.

The scheduler maximizes the delivery of critical data while ensuring that background tasks still receive a "fair share" of the remaining network capacity.

3.5. Simulation Setup

Several discrete-event simulations were conducted to evaluate the effectiveness of the MLHP framework. The test environment was designed to match a 5G-and-beyond cellular network with many users and heterogeneous traffic during frequent handoff events.

3.5.1. Simulation Environment and Tool

The proposed framework was implemented and evaluated using NS-3 (version 3.35), utilizing the LTE and 5G-NR (LENA) modules to model the radio access network. To ensure the reproducibility of the results, all simulations were executed on a Linux-based environment (Ubuntu 20.04 LTS). Each simulation scenario was performed for 50 independent runs using different random seeds provided by the NS-3 GlobalValue RngRun generator. This approach ensures that the observed performance variations are a result of the mobility and traffic models rather than a specific random sequence.

3.5.2. Traffic Modelling

To validate the multi-level architecture, we mapped four distinct traffic models to the $L1$–$L4$ priority tiers. This selection provides a representative cross-section of modern network demands, allowing us to evaluate the framework's ability to handle the specific QoS requirements (delay, jitter, and loss tolerance) assigned to each tier in Table 1.

3.5.3. Mobility and Handoff Scenario

Through Gauss-Markov mobility model, the MU is able to change its speed and direction. To simulate the framework under different stress levels, the handoff

Table 2. Simulation Parameters.

Parameter	Value
Simulation duration	500 seconds
System bandwidth	20MHz
Carrier frequency	3.5GHz (C-band)
Total link capacity (C)	50Mbps
Propagation model	Log-distance path loss model
Handoff threshold (H_{th})	-100dBm
Hysteresis margin (HYS)	3dB
Scheduling algorithms	SP (L1,L2), WFQ (L3, L4)
Comparative schemes	NPBH (FIFO), DQM (Dynamic sizing)

frequency was changed between 2 to 10 handoffs per minute. This was done by increasing the MU speed from 10km/h to 120km/h and reducing the cell overlap margin.

3.5.4. Framework Parameter Configuration

The MLHP specific parameters were configured as follows:

i) Buffer Capacity (K_i): Levels 1 and 2 were given 200 packets each, while 100 packets were given to Levels 3 and 4 to focus on high-priority storage.

ii) Thresholds for WRED:
- T_{low}: 30% of buffer capacity (Trigger warning)
- T_{mid}: 60% of buffer capacity (Start WRED for L3/L4)
- T_{high}: 90% of buffer capacity (Strict dropping of L4)

iii) Scheduling Weights (w_i): For the WFQ component, Level 3 was assigned a weight of 0.7 and Level 4 a weight of 0.3.

3.5.5. Summary of Simulation Parameters

Table 2 shows the summary of the main parameters used in the simulation.

3.5.6. Comparative Analysis Setup

The MLHP framework was compared against two other systems:

1) Non-Prioritized Buffered Handoff (NPBH): A standard FIFO buffer where all packets are treated the same, regardless of service type.
2) Dynamic Queue Management (DQM): An adaptive scheme that changes buffer size based on traffic load but lacks multi-level strict prioritization.

3.5.7. Statistical Validation and Reliability

To establish the statistical reliability of the performance metrics, the results reported in Section 4 represent the arithmetic mean of all 50 simulation runs. Statistical confidence was measured using a 95% confidence interval (CI). In the subsequent figures, the magnitude of the confidence intervals remained within ± 3–5% of the mean values, confirming the stability of the MLHP framework across diverse stochastic iterations.

3.6. Performance Metrics

To rigorously evaluate the efficacy of the MLHP framework, the following performance metrics are defined and utilized:

- **Handoff Dropping Probability (P_{drop}):** This metric represents the likelihood that a packet arriving at the handoff buffer will be discarded rather than successfully forwarded. In this framework, dropping occurs fundamentally due to buffer overflows when a priority queue reaches its maximum capacity (B_k) or when the high threshold (T_{high}) triggers active packet discarding to protect higher-priority traffic. It is calculated as the ratio of the total number of dropped packets to the total number of packets generated during the handoff window. A lower P_{drop} is critical for maintaining service continuity, particularly for URLLC and real-time applications.
- **Average Queueing Delay (D_{avg}):** This measures the mean duration a data packet resides in the handoff buffer at the Source Base Station (SBS) before being scheduled and transmitted to the Target Base Station (TBS). While buffering prevents packet loss during the "silent period" of a handoff, it inherently introduces latency. D_{avg} accounts for the time elapsed from the moment a packet enters the queue until the hybrid SP/WFQ scheduler initiates its transmission. Minimizing this delay is the primary objective for Level 1 and Level 2 services to prevent glitches and lagging.
- **Throughput (T):** Throughput is defined as the rate of successful data delivery over the communication channel, measured in Megabits per second (Mbps). It represents the effective utilization of the available 50 Mbps link capacity during the handoff process. This metric indicates the system's ability to maintain high data transfer rates despite the overhead of classification, buffering, and prioritization. Aggregate throughput reflects the total volume of data (across all four priority

levels) successfully received at the TBS per unit of time.

3.7. Mathematical Model

3.7.1. Buffer Occupancy Dynamics

The rate of change of the queue length $q_i(t)$ for a specific priority level L_i is defined by the difference between arrival and service rates:

$$\frac{dq_i(t)}{dt} = \lambda_i - \mu_i(t) \quad (1)$$

where λ_i is the packet arrival rate and $\mu_i(t)$ is the service rate at time t.

3.7.2. Prioritized Service Rate (μ_i)

The service rate depends on the scheduling policy. For Strict Priority (L_1, L_2):

$$\mu_i(t) = \frac{C}{S_1} if\ q_i(t) > 0 \quad (2)$$

$$\mu_2(t) = \frac{C}{S_2} if\ q_i(t) = 0\ and\ q_2(t) > 0 \quad (3)$$

where C is link capacity and S_i is the average packet size).

For Weighted Fair Queuing (L_i where $i \geq 3$):

If $q_i(t) = 0$ and $q_2(t) = 0$, the available capacity is distributed as:

$$\mu_i(t) = \frac{w_i}{\sum_{j=3}^{P} w_j \cdot \coprod(q_j(t) > 0)} \cdot \frac{C}{S_i} \quad (4)$$

where $\coprod(.)$ is an indicator function that is 1 if the queue is non-empty.

3.7.3. Packet Dropping Probability ($P_{drop,i}$)

Using an M/M/1/K queue approximation, the probability that a packet i is dropped due to buffer overflow is:

$$P_{drop,i} = \frac{(1-\rho_i)\,\rho_i{}^{B_i}}{1-\rho_i{}^{B_i+1}} \quad (5)$$

where $\rho_i = \lambda_i / \mu_i{}^{eff}$ is the utilization factor, and $\mu_i{}^{eff}$ is the effective service rate for that priority level.

3.7.4. Average Queuing Delay (W_1)

Applying Little's Law, the average time a packet spends in the handoff buffer is:

$$W_i = \frac{n_1}{\lambda_i(1-P_{drop,i})} \quad (6)$$

where n_1 is the average number of packets in buffer B_i.

3.7.5. Optimization Objective

The framework aims to minimize total QoS violations by adjusting buffer sizes B_i and weights w_i:

$$\min \sum_{i=1}^{P} (\alpha_i, \max(0, W_i - D_i^{max}) + \beta_i \cdot P_{drop,i}) \quad (7)$$

Subject to:

$$\sum_{i=1}^{P} B_i \leq B_{total} \quad (Physical\ memory\ constant) \quad (8)$$

$$\sum w_i = 1 \quad (Weight\ constraint) \quad (9)$$

$$0 \leq q_i(t) \leq B_i \quad (Occupancy\ constraint) \quad (10)$$

3.8. Multi-Level Handoff Prioritization (MLHP) Algorithm

MLHP Algorithm shown in Algorithm 1. The algorithm is executed at the Source Base Station (SBS) to manage packets as a Mobile User (MU) transition to Target Base Station (TBS).

4. Result and Discussion

The simulation results presented below are derived from averaged data across 50 independent trials, with error bars (where applicable) representing 95% confidence intervals to ensure the credibility and consistency of the comparative analysis.

To evaluate the performance of the proposed MLHP framework, we conducted simulations comparing it against Non-Prioritized Buffered Handoff (NPBH) and Dynamic Queue Management (DQM) schemes. The NPBH scheme uses a single, shared buffer for all traffic and employs a First-In, First-Out (FIFO) scheduling policy, similar to many traditional buffered handoff systems while the DQM for 5G handoff [30] incorporates some level of differentiation but primarily focuses on dynamic buffer sizing based on overall traffic load rather than granular multi-level QoS. DQM changes buffer sizes at both the SBS and TBS, following the predicted handoff events and traffic levels. This dynamic adjustment helps to reduce packet loss. However, this technique failed to use multi-level prioritization within the buffer. Instead, it usually treats all data as one group based on aggregate traffic features.

It should be noted that while the simulation results presented here focus on four distinct service classes (L1–L4), this configuration was specifically selected to represent the most critical 5G traffic profiles: URLLC, real-time multimedia, standard broadband, and background mMTC. The MLHP framework is termed 'multi-level'

Algorithm 1. MLHP Algorithm.

Step 1: Initialization

1. Define P Priority Levels: $L_1, L_2, \dots, L_P$ (e.g., L_1= URLLC, L_P= Background).

2. Define QoS Targets: For each level i, set maximum tolerable delay (D_i^{max}), minimum bandwidth and maximum loss rate.

3. Allocate Resources:

Initialize P dedicated buffers ($B_1, B_2, \dots, B_P$)

Set buffer capacities $Cap_1, Cap_2, \dots, Cap_P$

Define thresholds $T_{i,1}$ (Warning), $T_{i,2}$ (WRED), and $T_{i,3}$ (Drop) for each buffer.

Assign WFQ weights w_i for lower priorities (i > 2)

Step 2: Service Classification and Buffer Management

For every incoming packet (PKT):

1. Analyze PKT: Inspect header/service type and assign to level L_i

2. Handoff check:

If Handoff State is INACTIVE: Forward PKT through the normal data plane.

If Handoff State is ACTIVE:

Attempt to enqueuer PKT in B_i

Congestion control:

If Occupancy(B_i)> $T_{i,2}$: Trigger Weighted Random Early Detection (WRED)

If Occupancy(B_i)> Cap_i: Drop PKT (starting with the lowest priority packet currently in B_i)

Step 3: Handoff Detection:

1. Continuously monitor Signal Strength (RSSI).

2. If $RSSI_{SCS} < Threshold_1$ and $RSSI_{TBS} > Threshold_2$:

Set Handoff State = ACTIVE

Begin buffering incoming data.

Notify TBS to reserve resources for the service profile.

Step 4: Prioritized Packet Scheduling

Execute periodically during the handoff window:

1. Strict Priority (SP): Check B_1 and B_2. If packets exist, dequeue and transmit to TBS immediately.

2. Weighted Fair Queuing (WFQ): If B_1 and B_2 are empty:

Calculate available capacity C.

Allocate transmission slots to $B_3 \dots B_P$ proportional to weights w_i.

Dequeue and transmit selected packets to TBS.

3. Completion: If all buffers are empty and signal at TBS is stable:

Set Handoff State = COMPLETE.

Clear buffer allocations.

because its mathematical foundation and algorithmic logic, specifically the recursive application of Strict Priority (SP) and Weighted Fair Queuing (WFQ) are designed to support an arbitrary number of levels (*P*). The four-tier evaluation serves as a baseline proof-of-concept for 5G environments, demonstrating that the system successfully isolates mission-critical traffic regardless of the congestion state in lower-priority tiers.

Figure 2 shows the superior performance of the MLHP framework, especially for high-priority traffic. With a handoff frequency of 6 handoffs/min, NPBH experiences a dropping probability of 10%, indicating significant service disruption. DQM shows an improvement, keeping dropping probability below 2% at the same frequency due to its dynamic buffer resizing.

However, MLHP consistently keeps the lowest dropping probability for its highest priority services. It also performs better than DQM regarding the aggregate dropping probability due to its smart packet dropping policies for lower priority traffic. At 10 handoffs/min, MLHP's total dropping probability is about 6%. This is much better than NPBH and only slightly higher than DQM, whose focus is the overall throughput. It is observed that the drops in MLHP system are largely from lower priority services.

It should also be noted that a comparative analysis of Figure 2 reveals an apparent trade-off between the MLHP and DQM schemes. At a handoff frequency of 6 handoffs/min, the DQM scheme achieves a lower aggregate dropping probability (below 2%) compared to the aggregate dropping rate of MLHP. However, the claim of MLHP's superiority is rooted in its service differentiation capability. While DQM resizes buffers to reduce overall packet loss across the aggregate traffic, it does not prevent loss within mission-critical streams. In contrast, MLHP's 6% total dropping probability at high frequencies (10 handoffs/min) is strategically composed almost entirely of Level 4 (Background) and Level 3 (Web) packets. By proactively dropping non-urgent data to maintain a near-zero dropping probability for Level 1 (URLLC) and Level 2 (Real-time Video), MLHP ensures that the most sensitive applications meet their strict QoS targets. Consequently, MLHP is superior not in raw aggregate loss reduction, but in the reliability of high-priority transmission, which is the primary requirement for 5G-and-beyond networks where a 1% loss in URLLC is more detrimental than a 20% loss in background data.

Figure 3 shows the happenings of the three strategies when a handoff frequency of 8 handoffs/min is considered. The results highlight the main benefit of MLHP, which is service differentiation (the ability to treat different types of data differently). For Level 1 which is the URLLC, MLHP keeps the average queuing delay very low, staying under 10 ms. Level 2 which is the video, also has much lower delays compared to NPBH and DQM. While Level 3 (web) and Level 4 (background) though possess higher delays, they remain within the limits required for their QoS. DQM performs better than NPBH by deploying dynamic buffer management to improve delay performance. However, the absence of fine-grained control and strict prioritization as found in MLHP, results in higher delays for critical services compared to MLHP. For example, Level 2 (video) in DQM experiences a delay of approximately 210 ms. This is better than NPBH but much higher than the delay seen in MLHP. The non-prioritized scheme shows extremely high delays for all priority levels, especially for Level 1 (URLLC), exceeding 4000 ms (4 seconds). This

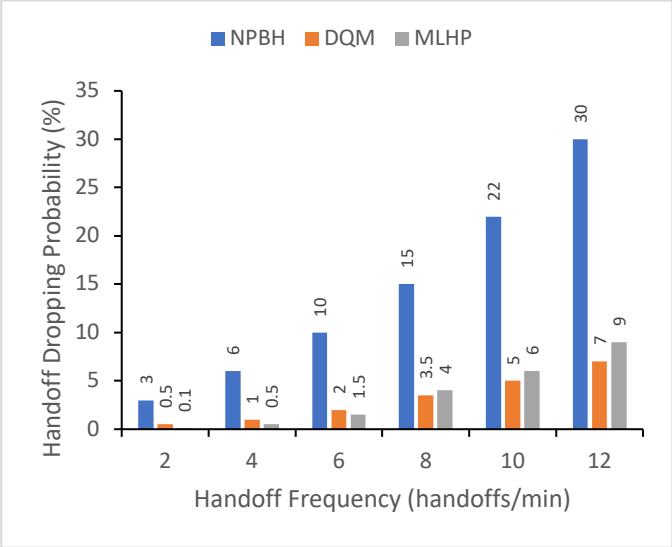

Figure 2. Handoff Dropping Probability against Handoff Frequency.

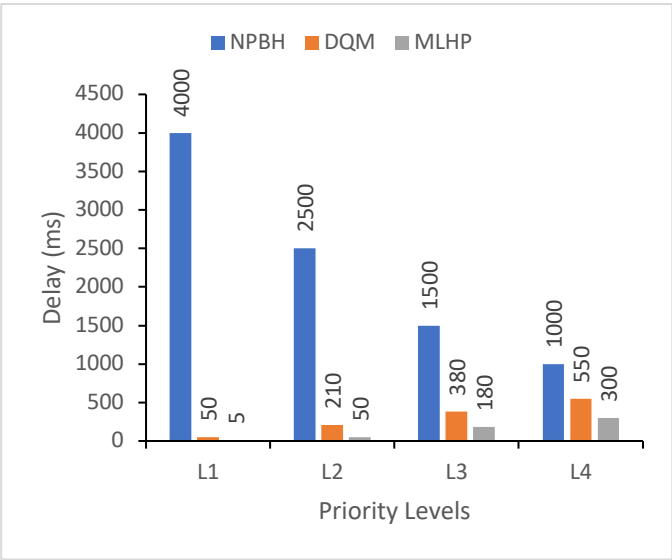

Figure 3. Delay against Priority Levels.

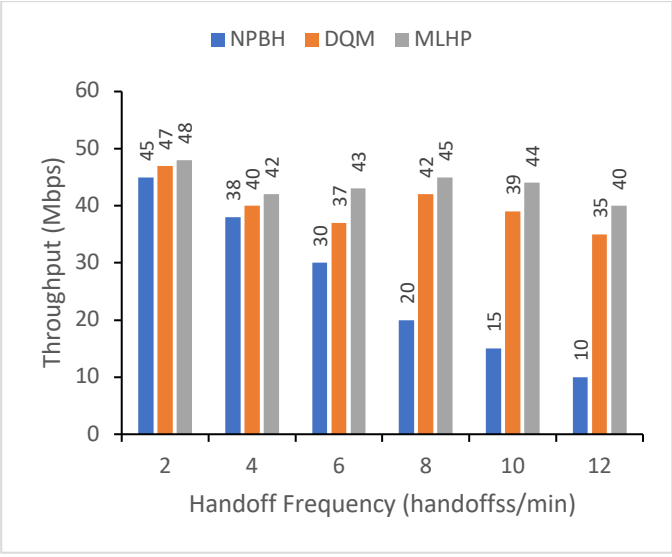

Figure 4. Throughput against Handoff Frequency.

demonstrates the inadequacy of traditional FIFO buffering for time-sensitive applications during handoff.

Figure 4 shows the aggregate data delivered by each scheme under varying handoff frequencies. MLHP consistently achieves higher aggregate throughput, especially at increased handoff frequencies. This is because by efficiently prioritizing critical traffic and judiciously dropping lower-priority packets when congestion is unavoidable, MLHP ensures that the network resources are utilized optimally to deliver the most important data. At 10 handoffs/min, MLHP achieves over 44 Mbps, while DQM achieves 39.5 Mbps, demonstrating its efficiency. DQM also shows a good throughput, adapting to higher handoff frequencies, reaching around 42 Mbps at 8 handoffs/min. Its dynamic buffer sizing helps to maintain a good flow of data. The NPBH scheme struggles with throughput as handoff frequency increases. The non-differentiated queuing leads to buffer overflows and retransmissions, resulting in lower effective throughput. At 4 handoffs/min. it delivers around 38 Mbps, which degrades significantly under higher load.

An analytical examination of the relationship between Figure 2 (Dropping Probability) and Figure 4 (Throughput) is necessary to explain why aggregate throughput remains robust despite a rise in packet dropping at higher handoff frequencies. This phenomenon is a direct result of MLHP's Selective Bit-Efficiency policy.

In traditional NPBH systems, packet dropping is 'blind'; the system is equally likely to drop a large, high-throughput video packet (Level 2) as it is a small background packet (Level 4). When a high-bitrate packet is dropped, the throughput 'cliffs' due to the collapse of the TCP congestion window or the loss of large UDP frames. In contrast, MLHP's hybrid scheduler and WRED thresholds ensure that at 10 handoffs/min, the 6% dropping probability is almost exclusively concentrated in the lowest-priority tiers (L3 and L4).

Analytical justification for this behavior rests on three factors:

- **Payload Weighting**: By sacrificing multiple small L4 packets (e.g., 1024 bytes) to ensure the delivery of a single large L2 video frame (1500 bytes) or a sequence of L1 URLLC commands, the system maintains a high *bit-per-second* delivery rate even if the *packet-loss count* increases
- **Congestion Window Preservation:** By protecting L1 and L2 traffic from loss, the framework prevents the catastrophic retransmission timeouts and 'slow-start' phases that typically degrade throughput in non-prioritized buffered systems.
- **Link Saturation:** MLHP ensures that during the brief windows of connectivity between handoffs, the 50 Mbps link is saturated with 'high-value' data (L1/L2) rather than being contested by background traffic.

Consequently, the high aggregate throughput of 44.2 Mbps at high mobility rates is not a contradiction of the 6% dropping rate, but rather a validation that the framework successfully prioritizes successful bit delivery over raw packet count preservation.

While the current study evaluates the MLHP framework as an integrated system, it is important to note the functional interdependency of its three core components: classification, buffering, and scheduling. In the context of MLHP, an ablation of service classification would result in a single-queue system (effectively the NPBH baseline), while removing prioritized scheduling would render the dedicated buffers (B_k) non-functional, as no logic would exist to differentiate their egress. The comparison against the Non-Prioritized Buffered Handoff (NPBH) baseline serves as a macro-level ablation study, illustrating the performance gap when classification and prioritized scheduling are absent. These results confirm that the framework's gains are not derived from a single component, but from the synergy between them: classification identifies the requirements, the buffering tier provides the necessary isolation, and the hybrid scheduler ensures the execution of QoS guarantees.

5. Limitations and Future Research Directions

Despite the performance gains demonstrated by the MLHP framework, several limitations offer opportunities for further enhancement. Currently, the system utilizes static buffer thresholds and fixed capacities which may not fully adapt to the extreme traffic volatility or the ultra-dense deployments expected in real-world 6G environments. Furthermore, the reliance on granular packet inspection for classification introduces potential computational overhead at the Source Base Station, while the memory requirements for maintaining multiple dedicated buffers may face scalability challenges in massive machine-type communication (mMTC) scenarios. To address these constraints, future research will focus on integrating reinforcement learning to enable the autonomous, real-time tuning of WRED thresholds and buffer sizes based on predictive traffic modeling. Additionally, exploring cross-layer optimization that incorporates instantaneous Channel State Information (CSI) and aligning the framework with end-to-end network slicing will be crucial for 5G-and-beyond deployments. Future efforts will also prioritize energy-aware prioritization mechanisms to improve network sustainability and involve the validation of MLHP on Software-Defined Radio (SDR) testbeds to assess its performance under live hardware and signal constraints.

6. Conclusion

In this paper, we introduced a multi-level QoS prioritization framework (MLHP) designed to manage the complexities of buffered handoffs in modern wireless networks. By integrating service-aware classification with a hybrid SP/WFQ scheduling architecture, our system provides a granular method for protecting mission-critical traffic during base station transitions. Our simulation results confirm that MLHP effectively isolates URLLC and real-time video services from the latency and packet loss typically associated with high mobility. Specifically, the framework achieves sub-10 ms delays for Level 1 URLLC traffic, fulfilling its stringent requirements, while simultaneously ensuring that Level 2 real-time video services maintain a significantly lower latency profile compared to current dynamic queue management or non-prioritized schemes. Future research will explore the use of machine learning to predict traffic bursts and dynamically tune the WRED thresholds in real-time.

7. Declarations

7.1. Author Contributions

Oluwasanya Oyedele: Conceptualization, Methodology & Mathematical Modeling, Software & Simulation and Writing - Original Draft; **Akinniyi Stellamaris Omowunmi:** Project Administration, Writing, Writing – Review and Editing and Validation; **Akinola Kayode Emmanuel:** Literature Review, System Architecture and Formal Analysis.

7.2. Institutional Review Board Statement

Not applicable.

7.3. Informed Consent Statement

Not applicable.

7.4. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.5. Acknowledgment

Not applicable.

7.6. Conflicts of Interest

The authors declare no conflicts of interest.

8. References

[1] I. Shayea, M. Ergen, M. H. Azmi, S. A. Çolak, R. Nordin, Y. I. Daradkeh, "Key Challenges, Drivers and Solutions for Mobility Management in 5G Networks: A Survey," *IEEE Access*, vol. 8, pp. 172534 – 172552, 2020. https://doi.org/10.1109/ACCESS.2020.3023802.

[2] Y. Ullah, M. B. Roslee, S. M. Mitani, S. A. Khan, and M. H. Jusoh, "A survey on handover and mobility management in 5G HetNets: Current state, challenges, and future directions," *Sensors*, vol. 23, no. 11, p. 5081, Jun. 2023. https://doi.org/10.3390/s23115081.

[3] G. Kumar, M. Varshney, and P. K. Srivastava, "Handoff management advances in enhancing mobility management in 5G networks," *African J. Biomedical Res.*, vol. 27, no. 4S, pp. 5590–5604, 2024. https://doi.org/10.53555/AJBR.v27i4S.4649.

[4] O. O. Erunkulu, A. M. Zungeru, C. K. Lebekwe, M. Mosalaosi, and J. M. Chuma, "5G mobile communication applications: A survey and comparison of use cases," *IEEE Access*, vol. 9, pp. 97251–97295, 2021. https://doi.org/10.1109/ACCESS.2021.3093213.

[5] B. Adhikari, M. Jaseemuddin, and A. Anpalagan, "Resource allocation for co-existence of eMBB and URLLC services in 6G wireless networks: A survey," *IEEE Access*, vol. 12, pp. 552–581, 2024. https://doi.org/10.1109/ACCESS.2023.3343250.

[6] V. Kovtun, M. Yukhimchuk, J. A. J. Alsayaydeh, and D. Berdysheva, "Stochastic modelling of delays and buffering in 5G-IoT ecosystems with programmable P4 switches based on BMAP," *PLOS ONE*, vol. 20, no. 8, Art. no. e0330526, 2025. https://doi.org/10.1371/journal.pone.0330526.

[7] H. Haile, K.-J. Grinnemo, S. Ferlin, P. Hurtig, and A. Brunstrom, "End-to-end congestion control approaches for high throughput and low delay in 4G/5G cellular networks," *Computer Networks*, vol. 186, Art. no. 107692, 2021. https://doi.org/10.1016/j.comnet.2020.107692.

[8] J. Lorincz, Z. Klarin, and J. Ožegović, "A comprehensive overview of TCP congestion control in 5G networks: Research challenges and future perspectives," *Sensors*, vol. 21, no. 13, Art. no. 4510, 2021. https://doi.org/10.3390/s21134510.

[9] E. E. Ekechukwu, O. C. Nosiri, and M. I. Ajumuka, "Efficient call admission control algorithm for mobility management in LTE networks," *Int. J. Netw. Commun.*, vol. 12, no. 1, pp. 28–38, 2022. http://article.sapub.org/10.5923.j.ijnc.20221201.02.html.

[10] H. Shih-Cheng and L. Shieh-Shing, "Fuzzy call admission control combined with distributed dynamic channel assignment and reassignment for cellular mobile systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, pp. 1–12, 2015. https://doi.org/10.1186/s13638-015-0330-5.

[11] E. U. Udo, U. Q. Oleh, and G. O. Odo, "Analysis and computer simulation of handoff decision in 4G networks," *NIPES Journal of Science and Technology Research*, vol. 3, no. 4, pp. 97–108, 2021. https://journals.nipes.org/index.php/njstr/article/download/272/287.

[12] M. Abdulwaheed and S. Adeniran, "Performance analysis of cellular mobile networks for QOS optimization," *Technosci. J. Commun. Dev. Africa*, vol. 3, pp. 131–139, 2024. https://journals.kwasu.edu.ng/index.php/technoscience/article/view/257.

[13] B. Saoud ,I. Shayea, M. A. Alnakhli, and H. Mohamad, "Mobility and handover management in 5G/6G networks: Challenges, innovations, and sustainable solutions," *Technologies*, vol. 13, no. 8, pp. 1–38, 2025. https://doi.org/10.3390/technologies13080352.

[14] E. Otsetova-Dudin., "Handoff in various mobile network technologies," *Transport and Communications*, vol. 2, pp. 28–32, 2019. https://doi.org/10.26552/tac.C.2019.2.6.

[15] A. Hamarsheh, H. R. K. El-Taj, A. Al-Qerem, and M. Alauthman, "Optimizing handover performance in 5G networks using dual connectivity," *Wireless Pers. Commun.*, 2026. https://doi.org/10.1007/s11277-026-11944-2.

[16] T. Ankome and E. Hanada, "Reactive to predictive mobility management: A systematic review of ML-driven handover optimization in 5G and beyond," *Machine Learning and Knowledge Extraction*, vol. 8, no. 5, pp. 1–29, 2026. https://doi.org/10.3390/make8050133.

[17] M. Mukhtar, F. Yunus, A. S. Mohd Noor, Z. Ali, M. Junaid, and M. Ahmed, "Handoff decision-making in 5G cellular networks using deep learning," *Computers, Materials & Continua*, vol. 87, no. 3, 2026. https://doi.org/10.32604/cmc.2026.076246.

[18] Y. Wei, C.-H. Lung, S. A. Ajila, and R. Paredes Cabrera, "Pre-connect handover management for 5G networks using multi-agent deep Q-networks," *IEEE Access*, vol. 13, pp. 176176–176197, 2025. https://doi.org/10.1109/ACCESS.2025.3618587.

[19] J. R. Behera, A. L. Imoize, S. S. Singh, S. S. Tripathy, and S. Bebortta, "Optimizing priority queuing systems with server reservation and temporal blocking for cognitive radio networks," *Telecom*, vol. 5, no. 2, pp. 416–432, 2024. https://doi.org/10.3390/telecom5020021.

[20] A. Nadeem, A. Ali, N. Ali, and S. Manzoor, "QoS-aware handoff framework for performance optimization in software-defined Wi-Fi networks," *Telecommunication Systems*, vol. 89, no. 32, pp. 4–12, 2026. https://doi.org/10.1007/s11235-026-01410-6.

[21] D. Battulga, "Handover with buffering for distributed mobility management in software defined mobile networks," *J. Telecommun. Digit. Econ.*, vol. 6, no. 1, pp. 26–40, 2018. https://doi.org/10.18080/ajtde.v6n1.137.

[22] S. Nazmus and D. Rui, "A survey of quality-of-service and quality-of-experience provisioning in information-centric networks," *Network*, vol. 5, no. 2, pp. 1–29, 2025. https://doi.org/10.3390/network5020010.

[23] R. Dobrescu and D. Cârciumărescu, "Interoperability of integrated services and differentiated services architectures," *U.P.B. Scientific Bulletin*, vol. 71, no. 1, pp. 21–32, 2009. https://www.scientificbulletin.upb.ro/static/pdfs/full9873.pdf.

[24] D. Strzęciwilk, "Differentiated service quality analysis based on QoS traffic prioritisation," *International Journal of Electronics and Telecommunications*, vol. 71, no. 1, pp. 285–292, 2025. https://doi.org/10.24425/ijet.2025.153572.

[25] P. E. Ugege, T. E. Akhigbe-Mudu, B. A. Saka, and S. A. Akee, "A flexible handoff prioritization scheme for improved quality of service in mobile networks," *Journal of Information Engineering and Applications*, vol. 11, no. 1, pp. 18–28, 2021. https://doi.org/10.7176/JIEA/11-1-03.

[26] B. Malakreddy, I. S. Rajesh, H. G. Mohan, U. N. Ranjitha, M. S. Krishnamurthy, and C. Maithri, "FLQL-VANET: A hybrid of fuzzy logic and Q-learning schemes for QoS-aware routing in VANET," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 6, pp. 1268–1283, 2024. https://doi.org/10.22266/ijies2024.1231.92.

[27] R. Men, X. Fan, A. Shan, and G. Yuan, "Fuzzy logic based binary computation offloading scheme in V2X communication networks," *IEEE Access*, vol. 12, pp. 45507–45518, 2024. https://doi.org/10.1109/ACCESS.2024.3376607.

[28] M. Abbasi, A. Shahraki, M. J. Piran, and A. Taherkordi, "Deep reinforcement learning for QoS provisioning at the MAC layer: A survey," *Eng. Appl. Artif. Intell.*, vol. 102, Art. no. 104234, 2021. https://doi.org/10.1016/j.engappai.2021.104234.

[29] T. S. Kumar, R. Mardeni, J. Jayapradha, U. Yasir, S. Chilakala, M. I. M. Sufian, F. O. Anwar, and Z. A. Fatimah, "Advanced handover optimization (AHO) using deep reinforcement learning in 5G networks," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 115, pp. 1–14, 2025. https://doi.org/10.1007/s44443-025-00124-0.

[30] A. M. Anwar, M. Shehata, S. M. Gasser, and H. El Badawy, "Handoff scheme for 5G mobile networks based on Markovian queuing model," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 30, no. 3, pp. 348–361, 2023. https://doi.org/10.37934/araset.30.3.348361.