Received: 04 Mar 2026 | Revised: 18 May 2026 | Accepted: 01 Jun 2026 | Published: 30 Jun 2026
Volume 4(1) Pages 106-127 DOI: https://doi.org/10.56279/jicts.v4i1.447

JICTS
Journal of ICT Systems

A Queue-Aware MEC-Assisted Radio Resource Management Framework for Latency and Congestion Optimization in 5G Networks: Analytical and Simulation-Based Evaluation

Godfrey William Chugulu[1] Hadija Ramadhani Mbembati

Department of Electronics and Telecommunications Engineering, University of Dar es Salaam, Dar es Salaam, Tanzania

[1]**Corresponding author**
Email:
chugulu.godfrey@udsm.ac.tz

Funding information

Not Applicable.

Keywords

5G
Latency optimization
Multi-access edge computing
Queue-aware scheduling
Radio resource management
Resource allocation

Abstract

The rapid growth of latency-sensitive and data-intensive applications in fifth-generation (5G) networks has exposed limitations of centralized radio resource management (RRM), including backhaul dependency, core-network congestion, and limited responsiveness to dynamic traffic conditions. This paper proposes a queue-aware Multi-Access Edge Computing (MEC)-assisted RRM framework for latency and congestion optimization in heterogeneous 5G networks supporting enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC) services. The framework integrates latency-aware traffic offloading, queue-aware scheduling, and adaptive edge-assisted resource allocation. Delay-sensitive traffic is processed at MEC-enabled gNodeBs, while radio resources are allocated based on queue state, traffic priority, and channel quality. Performance is evaluated using M/M/1 queueing analysis and OMNeT++/Simu5G simulations under realistic 5G conditions. Results show that, under high traffic loads, the proposed framework reduces queueing delay by approximately 74.5% and end-to-end latency by 78.4% compared with centralized RRM, while maintaining packet loss below 2%. The framework also improves resource utilization and scalability, demonstrating its effectiveness for congestion control and latency optimization in heterogeneous 5G networks.

Contribution to Sustainable Development Goals (SDGs)

SDG 9 – Industry, Innovation and Infrastructure

This research contributes to SDG 9 by improving the efficiency, reliability, scalability, and responsiveness of 5G communication infrastructure through a queue-aware MEC-assisted radio resource management framework.

1. Introduction

The rapid increase in mobile data traffic caused by Internet of Things (IoT) applications, multimedia services, autonomous systems, and industrial automation continues to place significant demands on fifth-generation (5G) wireless networks [1]. These emerging applications require low latency, high reliability, and efficient utilization of available radio resources. As a result, Radio Resource Management (RRM) has become increasingly important in ensuring efficient scheduling, interference management, power control, and spectrum allocation within modern wireless systems [2]. However, conventional centralized RRM architectures, where most control and processing tasks are handled in the core network, often experience increased transmission delay, backhaul congestion, and slow adaptation to changing traffic conditions, especially in dense and heterogeneous 5G environments [3].

To improve network performance, several studies have investigated dynamic RRM techniques based on artificial intelligence, adaptive spectrum allocation, and network slicing [4, - 6]. Machine-learning-based scheduling approaches have shown the ability to improve load balancing and resource allocation efficiency in highly dynamic environments, while network slicing enables differentiated Quality of Service (QoS) support for enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC) services [7, 8]. Despite these improvements, centralized architectures suffer from additional latency due to the long communication path between user equipment (UE), base stations, and centralized cloud servers [9].

Multi-Access Edge Computing (MEC) has recently emerged as a practical solution for reducing this dependency on centralized processing by moving computation and control functions closer to the radio access network [10]. With MEC, latency-sensitive traffic can be processed locally at edge servers located near gNodeBs, thereby reducing end-to-end delay and minimizing backhaul congestion. Previous studies have shown that MEC-assisted scheduling and caching techniques can improve service responsiveness, throughput stability, and resource utilization in delay-sensitive applications [11, 12] . In addition, combining MEC with intelligent RRM enables localized traffic management, adaptive power control, and faster scheduling decisions based on real-time network conditions [13].

Although MEC-assisted resource management has attracted significant research interest, there are limited studies that combine both analytical modelling and simulation-based evaluation under realistic network conditions [14]. Most existing works focus on either MEC optimization or standalone RRM enhancement without fully examining their combined operational impact [15]. Therefore, there is a need for an integrated evaluation framework capable of analysing how MEC-assisted RRM improves latency, congestion control, and overall network performance in practical 5G environments [16]. Table 1 presents a comparative analysis of existing MEC-assisted and RRM improvement approaches, highlighting the contribution of the proposed framework.

This paper presents a queue-aware MEC-assisted RRM framework for latency and congestion optimization in 5G networks. The proposed framework integrates edge-assisted traffic processing, adaptive load balancing, queue-aware scheduling, and channel-aware resource allocation to improve network performance under varying traffic conditions. A hybrid analytical and simulation-based evaluation was conducted to investigate the effectiveness of the framework under realistic 5G traffic scenarios.

Table 1. Related works on MEC-assisted and RRM improvement approaches.

Study	Focus	Limitation	This work's contribution
Fontes Pupo et al., 2024 [17]	Low-complexity AI scheduling for 5G MBS across eMBB, URLLC, and mMTC	No MEC integration; centralized scheduling; no queue-state awareness or edge congestion control	Proposes distributed MEC-assisted scheduling incorporating real-time queue state to reduce edge congestion beyond AI-only centralized approaches
Marwani & Kaddoum, 2026 [18]	Event-based temporal graph neural network for dynamic radio resource management in wireless networks	No MEC offloading; no queueing delay model; scheduling driven by graph topology rather than queue state; no heterogeneous 5G simulation	Combines queue-state-driven scheduling with MEC offloading under heterogeneous 5G traffic, providing analytical and simulation validation absent in GNN-only approaches
Kumar & Ahmad, 2022 [7]	ML-based QoS prediction and dynamic network slicing for heterogeneous eMBB, URLLC, and mMTC services	Centralized slicing-based QoS; no edge offloading or queue-state-driven resource allocation; no MEC processing at gNodeB level	Complements slicing-based QoS differentiation with edge-native queue-aware PRB allocation, directly reducing per-user queueing delay at the radio access layer
Iradier et al., 2022 [19]	P-NOMA-based RRM analysis for URLLC latency and throughput capacity in 5G NR	No MEC layer; no queue-aware scheduling; no packet-level simulation; limited to URLLC traffic class only	Integrates a queue-aware scheduling metric over URLLC, eMBB, and mMTC jointly, validated through analytical modelling and OMNeT++/Simu5G simulation
Sarah et al., 2023 [8]	Resource allocation frameworks for MEC-enabled 5G-and-beyond networks supporting heterogeneous services	No queue-aware scheduling algorithm; no OMNeT++/Simu5G simulation; congestion and queueing dynamics not modelled	Adds queue-aware scheduling and discrete-event simulation to MEC resource allocation, providing the congestion analysis missing in prior MEC works
Ebrahimi & Afghah, 2025 [20]	Deep reinforcement learning for intelligent task offloading and resource management in MEC-enabled 5G	No analytical queueing model; no queue-aware PRB allocation; no heterogeneous traffic; no packet-level simulation	Provides an M/M/1 analytical queueing model and queue-aware PRB allocation under heterogeneous traffic, adding a tractable theoretical foundation to MEC offloading

Study	Focus	Limitation	This work's contribution
Gautam & Tamma, 2024 [21]	Queue-aware uplink MAC scheduler for 5G NR controlling UE backlogs to reduce signaling overhead in V2N scenarios	No MEC; limited to uplink V2N; no heterogeneous eMBB/URLLC/mMTC traffic; no queueing delay decomposition model	Extends queue-aware scheduling into a MEC-assisted RRM framework with heterogeneous traffic and end-to-end latency decomposition absent in RETALIN
Chrysikos, 2023 [22]	Two-stage analytical queue model with priority-based RRM for 5G/6G cloud-based star-topology networks	Centralized core/cloud architecture only; no MEC offloading or edge scheduling; no packet-level simulation; limited heterogeneous traffic	Relocates queue-aware RRM to the network edge via MEC, enabling faster scheduling decisions validated through discrete-event simulation

The results demonstrate significant reductions in queueing delay, end-to-end latency, and packet loss, together with improved throughput stability compared with conventional centralized RRM approaches.

2. Methods and Materials

2.1 General 5G Network Architecture

The considered network architecture represents a typical 5G edge-enabled heterogeneous cellular deployment consisting of four major components: User Equipment (UE), 5G gNodeBs, MEC hosts, and the 5G core network. These components collectively support radio access, computation offloading, and centralized control functions required for advanced radio resource management.

UE represents mobile devices generating heterogeneous traffic types, such as enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC). Each UE periodically transmits uplink data packets containing service requests, channel state information, and mobility updates that are used by the network for scheduling and resource allocation decisions [23].

5G gNodeBs (Base Stations) provide radio access functionality, including packet scheduling, link adaptation, handover management, and interference coordination across neighboring cells. In the proposed architecture, gNodeBs also act as the primary aggregation points for edge computing services, forwarding selected traffic flows either to the local MEC platform or to the centralized core network depending on latency and processing requirements.

Each gNodeB is equipped with an associated MEC server responsible for executing latency-sensitive applications and localized network control functions. The MEC host performs edge-level traffic processing, predictive resource allocation, localized scheduling optimization, and context-aware service provisioning using real-time network information [24]. By processing selected user traffic locally, the MEC platform significantly reduces backhaul signaling load and enables faster response to dynamic network conditions.

The 5G core network performs centralized control functions, such as subscriber management, mobility management, authentication, and large-scale traffic routing. Delay-tolerant or computation-intensive services that do not require real-time processing are forwarded to centralized cloud servers through the core network infrastructure [25].

In the proposed system model, traffic flows are dynamically classified according to latency sensitivity and computational requirements. Time-critical services are processed at the MEC layer, while non-critical traffic is routed to the centralized core. This hierarchical processing approach enables efficient resource utilization, reduces transmission latency, and improves overall Quality of Service (QoS) performance. Furthermore, localized decision-making at the edge allows rapid adaptation to user mobility, traffic fluctuations, and interference conditions, making the architecture highly suitable for dense 5G deployments and emerging real-time applications.

Figure 1 illustrates the proposed queue-aware MEC-assisted radio resource management architecture for 5G networks. The architecture consists of UEs, including eMBB users, URLLC users, and mMTC/IoT devices, which communicate with MEC-enabled gNodeBs through the Radio Access Network (RAN). Each gNodeB is connected to a nearby MEC server responsible for localized traffic processing and edge-assisted resource management. Latency-sensitive traffic is processed at the edge to reduce backhaul congestion and end-to-end delay, while non-critical traffic is forwarded to the centralized core network for further processing and external network access. This distributed architecture enables faster decision-making, improved traffic handling, and more efficient utilization of network resources in dense 5G environments.

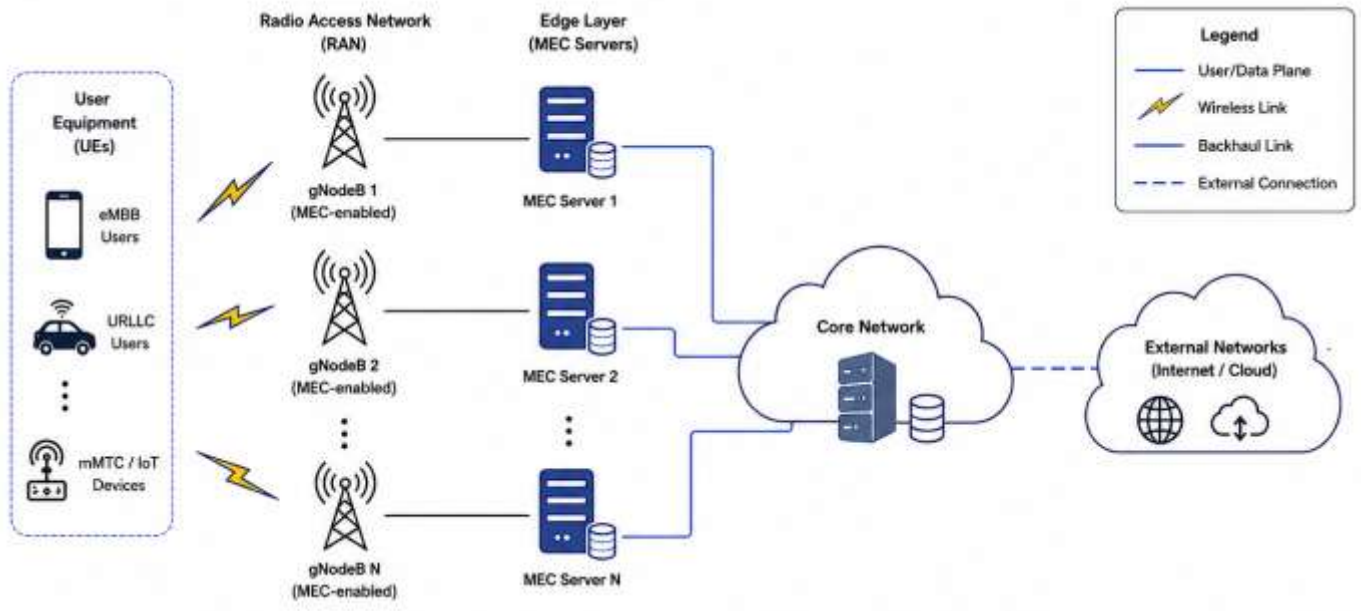


Figure 1. Queue-aware MEC-assisted RRM architecture for 5G networks.

2.2 Analytical Performance Modeling

2.2.1 Latency Model

The end-to-end communication delay experienced by a user packet is modelled as the sum of transmission, queueing, processing, and core network delays:

$$D_{total} = D_{tx} + D_{queue} + D_{proc} + D_{core} \quad (1)$$

The variables from (1) are defined as follows: D_{tx} represents the radio transmission delay between UE and serving gNodeB, D_{queue} denotes the waiting time in transmission buffers at intermediate network nodes, D_{proc} is the computation and scheduling processing delay, and D_{core} corresponds to the additional latency introduced when packets traverse the centralized core network.

In conventional centralized architectures, packets must travel to the core network for processing, leading to increased latency [26]. When Multi-Access Edge Computing (MEC) is deployed, latency-sensitive processing tasks are executed locally at the edge server attached to the gNodeB, making the core network traversal delay negligible:

$$D_{core} \approx 0 \quad (2)$$

Thus, the effective latency in MEC-assisted systems becomes

$$D_{total}^{MEC} \approx D_{tx} + D_{queue} + D_{proc}, \quad (3)$$

which significantly reduces overall service delay, particularly for time-critical applications.

2.2.2 Queueing Model

Packet arrivals at the network scheduler are assumed to follow a Poisson process, while service times are exponentially distributed, allowing the system to be modeled as standard single-server queueing model M/M/1. The average queueing delay is therefore given by

$$W_q = \frac{\rho}{\mu(1-\rho)}, \quad (4)$$

where $\rho = \frac{\lambda}{\mu}$

represents the system utilization, λ is the packet arrival rate, and μ is the service rate. High utilization levels increase the queueing delay exponentially, leading to congestion and packet loss in centralized architectures [27].

With MEC deployment, a portion of the traffic is processed locally at the edge, effectively reducing the arrival rate directed to the core network [28]. The new utilization becomes

$$\rho_{new} = \frac{\lambda_{edge}}{\mu}, \quad (5)$$

where $\lambda_{edge} < \lambda_{core}$.

This reduction in traffic load decreases queue occupancy and improves service stability, resulting in lower average waiting time [29].

2.2.3 Interference Model

Wireless channel performance is influenced by co-channel interference from neighboring cells [30]. The received signal quality is characterized by the Signal-to-Interference-plus-Noise Ratio (SINR):

$$\text{SINR} = \frac{P_S}{(I+N)} \quad (6)$$

where P_S is the received signal power, I represents aggregate interference from adjacent transmitters, and N denotes thermal noise power. In the proposed framework, adaptive power control mechanisms dynamically adjust transmission power to maintain

$$SINR \geq SINR_{target}, \quad (7)$$

ensuring reliable communication, stable throughput, and improved spectral efficiency even in dense network deployments. The integration of edge-based processing further enables faster interference-aware scheduling decisions,

enhancing overall radio resource utilization efficiency [31].

3. Proposed MEC-Assisted RRM Framework and Evaluation Methodology

3.1 Design Rationale

Conventional centralized RRM in 5G networks relies on core network processing, which introduces additional latency due to backhaul traversal and centralized queueing. Under high traffic load, this leads to increased queueing delay, packet loss, and degraded QoS, particularly for latency-sensitive applications, such as real-time monitoring and control.

To address these limitations, this work proposes a MEC-assisted RRM framework that enables distributed decision-making at the network edge. By relocating control and scheduling functions closer to UE, the proposed approach reduces end-to-end latency, alleviates congestion in the core network, and improves responsiveness to dynamic traffic conditions. The framework is designed to support heterogeneous traffic classes and to adaptively balance load between edge and core processing domains.

3.2 System Model and Operation

The considered system follows a hierarchical 5G architecture consisting of UE, gNodeBs, MEC servers deployed at the network edge, and a centralized core network. In the conventional centralized mode, most processing, scheduling, and control decisions are performed at the core network, resulting in increased backhaul utilization and additional latency due to centralized queueing. In contrast, the proposed MEC-assisted framework distributes processing tasks between the edge and the core, enabling localized decision-making, reduced communication overhead, and faster response to dynamic network conditions. This architectural shift allows latency-sensitive traffic to be processed closer to the source, thereby improving overall system efficiency and scalability.

The proposed framework operates in a hierarchical 5G architecture (Figure 1) consisting of the following:

i. UE: Generates heterogeneous traffic, including URLLC, eMBB, and mMTC flows.
ii. gNodeB (gNB): Acts as the radio access point and performs real-time scheduling decisions.
iii. MEC Server: Deployed at or near the gNodeB to enable localized processing and edge-based optimization.
iv. Core Network: Handles centralized processing for delay-tolerant and high-computation tasks.

3.2.1 System Inputs

At each scheduling interval, the gNodeB collects the following information:

i. Queue length, $Q_i(t)$, for each UE
ii. Channel Quality Indicator (CQI)
iii. Traffic classification (URLLC, eMBB, mMTC)
iv. Network congestion indicators:
    a. Backhaul delay, $D_{bh}(t)$
    b. Core network load
    c. MEC utilization, $U_{mec}(t)$

3.2.2 Decision Process

Based on the collected inputs, the system performs:

1. Traffic classification to assign priority levels
2. Offloading decision to determine whether traffic is processed at MEC or the core
3. Queue-aware resource allocation to ensure efficient and fair distribution of radio resources

This adaptive process ensures that latency-sensitive and high-priority traffic is processed at the edge, while delay-tolerant traffic is routed to the core network.

3.3 MEC-Assisted RRM Algorithm

The proposed framework performs queue-aware scheduling and allocates radio resources dynamically among active users to ensure efficient utilization of network capacity. Physical Resource Blocks (PRBs), which are the smallest allocatable units of radio resources in 5G systems comprising a set of subcarriers over a defined time interval, are assigned by the scheduler to users for data transmission.

The scheduling process leverages channel quality indicators (CQI) as practical approximations of channel state information (CSI), enabling link-adaptive resource allocation. By incorporating both queue state and channel conditions into the scheduling metric, the proposed algorithm achieves a balance between delay minimization and spectral efficiency. This ensures that users experiencing favorable channel conditions are efficiently served while maintaining fairness and prioritization for latency-sensitive traffic.

To operationalize the proposed MEC-assisted queue-aware RRM framework, Algorithm 1 summarizes the MA-QA-RRM decision-making and resource allocation procedure. The algorithm first acquires the queue, channel, traffic, backhaul, and MEC states of active UEs. It then determines traffic priorities and makes an offloading decision based on latency and congestion conditions. A queue-aware scheduling metric is subsequently computed by jointly considering queue length, achievable data rate, channel quality, and traffic priority. PRBs are allocated according to the resulting metric, while queue control is applied when the queue length exceeds a predefined threshold. The procedure is repeated as the network state changes, enabling the framework to adapt to dynamic traffic and radio conditions. The complete MA-QA-RRM procedure is summarized in Algorithm 1.

Algorithm 1: MEC-Assisted Queue-Aware RRM (MA-QA-RRM)

Inputs:

- Queue length $Q_i(t)$
- Channel Quality Indicator $CQI_i(t)$
- Traffic class c_i
- Backhaul delay $D_bh(t)$
- MEC utilization $U_MEC(t)$

Outputs:

- Offloading decision $o_i(t)$
- Resource allocation $r_i(t)$

Procedure:

1. Acquire network state from all active UEs.
2. Classify traffic and assign priority weights.
3. Perform offloading decision based on latency and congestion.
4. Compute scheduling metric:

$M_i(t) = w_i [Q_i(t) / (\bar{R}_i(t) + \varepsilon)] f(CQI_i(t))$ where $\bar{R}_i(t)$ is the average achievable data rate of UE i, ε is a small positive constant introduced to avoid division by zero, and $f(\cdot)$ is the CQI-based channel-quality function.

5. Allocate PRBs in descending order of $M_i(t)$.
6. Apply queue control if $Q_i(t)$ exceeds threshold.
7. Update system metrics and repeat.

3.4 Simulation Scenarios and Experiment Design

To quantify the benefits of edge-assisted resource management, two simulation scenarios were implemented under identical traffic, mobility, and channel conditions.

3.4.1 Scenario 1: Centralized RRM

In this baseline scenario, all traffic processing, scheduling optimization, and application-level computations are performed in centralized cloud/core network servers. Packets transmitted by UEs traverse the radio access network and backhaul links before reaching the centralized processing nodes, introducing additional processing and transmission delays. Resource allocation decisions are also executed centrally, resulting in slower adaptation to local network dynamics.

3.4.2 Scenario 2: MEC-Assisted RRM

In the proposed scenario, a MEC server is deployed at each gNodeB, enabling localized processing of latency-sensitive traffic flows and edge-based scheduling optimization. Real-time channel state information and local traffic statistics are used by the MEC platform to perform adaptive resource allocation, interference-aware scheduling, and traffic prioritization. Only delay-tolerant or computation-intensive traffic is forwarded to the centralized core network, significantly reducing backhaul utilization and processing latency.

3.4.3 Load Conditions

Both scenarios were simulated across multiple traffic load levels and user densities to evaluate system behavior under light, moderate, and heavy network congestion conditions. The comparative performance results obtained from these simulations provide quantitative insights into the effectiveness of MEC-assisted RRM in improving latency, throughput stability, and overall QoS.

3.4.4 Simulation Environment and Performance Metrics

The simulation study was conducted using the OMNeT++ discrete-event network simulator integrated with the Simu5G framework, which provides detailed modelling of 5G New Radio (NR) protocol stacks, radio propagation, scheduling mechanisms, and mobility management. The simulation environment represents a multi-cell heterogeneous 5G deployment in which UE communicates with gNodeBs equipped with optional MEC servers.

User mobility is modeled using the Gauss–Markov mobility model, which captures realistic movement patterns with controllable randomness and temporal correlation in velocity [32]. This model enables evaluation of the proposed framework under dynamic network conditions, including varying user density and mobility scenarios.

To emulate heterogeneous service environments, multiple traffic categories were generated:

- VoIP traffic, representing ultra-low-latency conversational services [33].
- Video streaming traffic, representing high-throughput enhanced mobile broadband (eMBB) services [34].
- IoT data traffic, representing periodic low-rate massive machine-type communication (mMTC) transmissions [35].

Network performance was evaluated using the following key metrics:

a. Throughput: Average successful data delivery rate per user and per cell;
b. Latency: End-to-end packet delivery delay from transmission to reception;
c. Queue length: Average buffer occupancy at base stations and processing nodes;
d. Packet loss ratio: Percentage of packets dropped due to congestion or transmission errors;
e. Spectral efficiency: Achieved throughput normalized by allocated bandwidth.

These metrics provide a comprehensive evaluation of both radio-layer efficiency and

system-level performance under centralized and MEC-assisted RRM architectures.

The following section presents the analytical performance evaluation, followed by simulation-based results and discussion.

3.4.5 Detailed Simulation Assumptions and System Parameters

To improve transparency and reproducibility, this subsection provides additional details on the network configuration, simulation assumptions, and algorithm parameters used in the study. The same parameter settings were applied to both centralized RRM and MEC-assisted RRM scenarios to ensure fair comparison.

The simulated environment consists of a seven-cell 5G New Radio (NR) macro-cell deployment, where each gNodeB serves 20–60 active users depending on traffic load conditions. MEC servers were attached to selected gNodeBs to support localized processing and edge-assisted scheduling. The network operates at a carrier frequency of 3.5 GHz with a bandwidth of 100 MHz.

Three traffic classes were considered: VoIP traffic for URLLC services, video streaming for eMBB traffic, and periodic sensor traffic for mMTC applications. Packet sizes were configured as 64 bytes for VoIP, 1,024 bytes for video traffic, and 128 bytes for IoT traffic. Packet arrivals follow a Poisson process, while service times are exponentially distributed in accordance with the adopted M/M/1 queueing model.

The gNodeB transmission power was set to 40 W and UE transmission power to 23 dBm. MEC processing delay ranged between 1–3 ms, while centralized core-network processing delay varied between 10–20 ms. User mobility was modeled using the Gauss–Markov model with speeds between 3–30 km/h, and each simulation was executed for 1,000 seconds using multiple random seeds to improve statistical reliability. Average results were obtained over multiple simulation runs and confidence variations remained within acceptable limits.

At each scheduling interval, the proposed algorithm collects queue length, CQI, traffic class, MEC utilization, and backhaul delay information to determine traffic offloading and PRB allocation decisions. Latency-sensitive traffic with high queue occupancy is prioritized for edge processing, while a queue-control mechanism is applied whenever queue length exceeds a predefined threshold, Q_{th}, to maintain system stability and reduce congestion.

Performance evaluation was based on queueing delay, end-to-end latency, packet loss ratio, throughput stability, queue occupancy, and spectral efficiency. Appendix A presents simulation parameters used in this work.

3.5 Hybrid Analytical and Simulation-Based Performance Evaluation

To evaluate the performance of the proposed MEC-assisted RRM framework, a hybrid analytical and simulation-based evaluation approach was adopted. Analytical modelling was used to investigate system-level characteristics, including queueing behavior, traffic intensity, latency trends, congestion dynamics, and parameter sensitivity, while OMNeT++/Simu5G simulations were used to validate the framework under realistic 5G network conditions. The simulation environment captured packet-level interactions, such as heterogeneous traffic generation, user mobility, radio resource scheduling, and queue dynamics under both centralized and MEC-assisted architectures. This combined approach enables comprehensive assessment of both theoretical performance behavior and practical network operation under varying traffic and deployment conditions.

To complement the simulation analysis, analytical models derived from queueing theory and latency decomposition were used to evaluate delay performance and congestion behavior. The end-to-end delay for centralized and MEC-assisted architectures is expressed as:

$$D_{cloud} = D_{tx} + D_{queue}^{core} + D_{proc}^{core} + D_{bh} \quad (8)$$

$$D_{MEC} = D_{tx} + D_{queue}^{edge} + D_{proc}^{edge} \quad (9)$$

Since the MEC-assisted framework reduces backhaul dependency and minimizes core-network queueing, the resulting delay satisfies:

$$D_{MEC} < D_{cloud} \quad (10)$$

Queueing behavior was modeled using an M/M/1 queueing system, where edge-assisted traffic offloading reduces the effective arrival rate directed to the centralized core network:

$$\lambda_{core}^{MEC} = \lambda_{total} - \lambda_{offloaded} \quad (11)$$

This leads to lower traffic intensity and reduced queueing delay:

$$\rho_{MEC} < \rho_{cloud} \Rightarrow W_q^{MEC} < W_q^{cloud} \quad (12)$$

Packet loss was further modeled as a function of buffer overflow probability:

$$P_{loss} \propto P(Q > Q_{\max}) \quad (13)$$

Since MEC-assisted processing reduces queue occupancy and congestion levels, the probability of buffer overflow is significantly reduced, resulting in improved packet delivery reliability.

The analytical evaluation was further used to generate performance curves illustrating the effects of user density, traffic load, and MEC offloading efficiency on latency, scalability, and congestion performance. The close agreement between the analytical results and OMNeT++/Simu5G simulations validates the consistency, reliability, and practical applicability of the proposed MEC-assisted RRM framework under varying network conditions.

4. Results and Discussion

The evaluation combines analytical modeling and OMNeT++/Simu5G simulations to investigate system behavior under realistic 5G traffic and mobility scenarios. The proposed framework is compared with conventional centralized RRM using identical network settings and simulation parameters to ensure fair comparison. Key performance metrics considered include queueing delay, end-to-end latency, packet loss, throughput stability, and resource utilization. The obtained results demonstrate the effectiveness of MEC-assisted resource management in reducing congestion, improving latency performance, and enhancing overall network scalability and stability.

4.1 Queueing Delay Performance

The average queueing delay observed in the centralized framework is significantly higher compared with the MEC-assisted framework, particularly under moderate and heavy traffic conditions. In the centralized case, traffic from multiple UEs accumulates at the core network, leading to increased queue lengths and higher waiting times.

In contrast, the MEC-assisted architecture maintains consistently low queueing delay due to localized processing and traffic offloading. By reducing the effective arrival rate at the core, the system operates at a lower traffic intensity (see Section 3.5). Consequently, queue buildup is minimized and delay variability is significantly reduced. Figure 2 shows that queueing delay increases with the number of users in both architectures; however, the growth rate is significantly higher in the centralized RRM case.

This is due to traffic aggregation at the core network, which increases queue occupancy and waiting time. In contrast, the MEC-assisted framework maintains lower and more stable queueing delay by offloading traffic to edge servers and reducing the effective arrival rate at the core.

This behavior is consistent with the analytical model in Section 3.5, where reduced traffic intensity leads to lower queueing delay under MEC operation. Figure 3 shows queueing delay performance in OMNeT++/Simu5G simulations.

The OMNeT++ results in Figure 3 confirm that MEC-assisted RRM significantly reduces queue buildup compared with centralized RRM, consistent with the analytical reduction in traffic intensity.

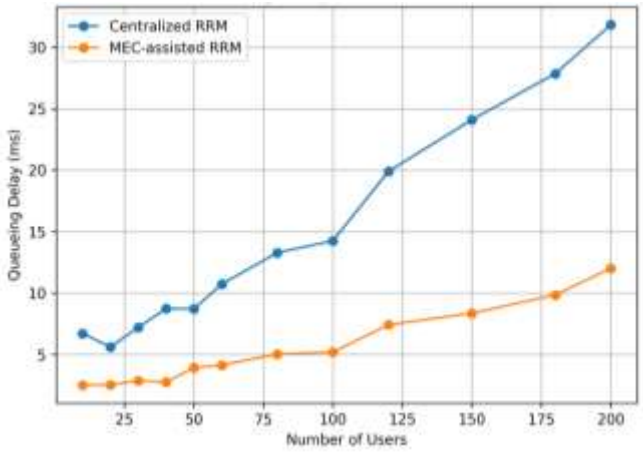


Figure 2. Queueing delay versus number of users for centralized and MEC-assisted RRM.

4.2 End-to-End Latency

The end-to-end latency results show a substantial reduction when using MEC-assisted RRM compared with centralized processing. In the centralized architecture, packets experience additional delay due to backhaul transmission and core-network processing. Under high traffic conditions, this leads to latency spikes and increased jitter.

The MEC-assisted framework eliminates backhaul delay for latency-sensitive traffic and reduces queueing delay through edge-based scheduling. As a result, latency remains low and stable across different load conditions. This behavior is consistent with the analytical latency model in equations (8), (9), and (10).

Figure 4 illustrates that end-to-end latency increases with user density for both architectures, but remains significantly lower under MEC-assisted RRM. The centralized approach experiences additional delay due to backhaul transmission and core-network processing, which becomes more pronounced at higher loads. In contrast, MEC eliminates the backhaul component

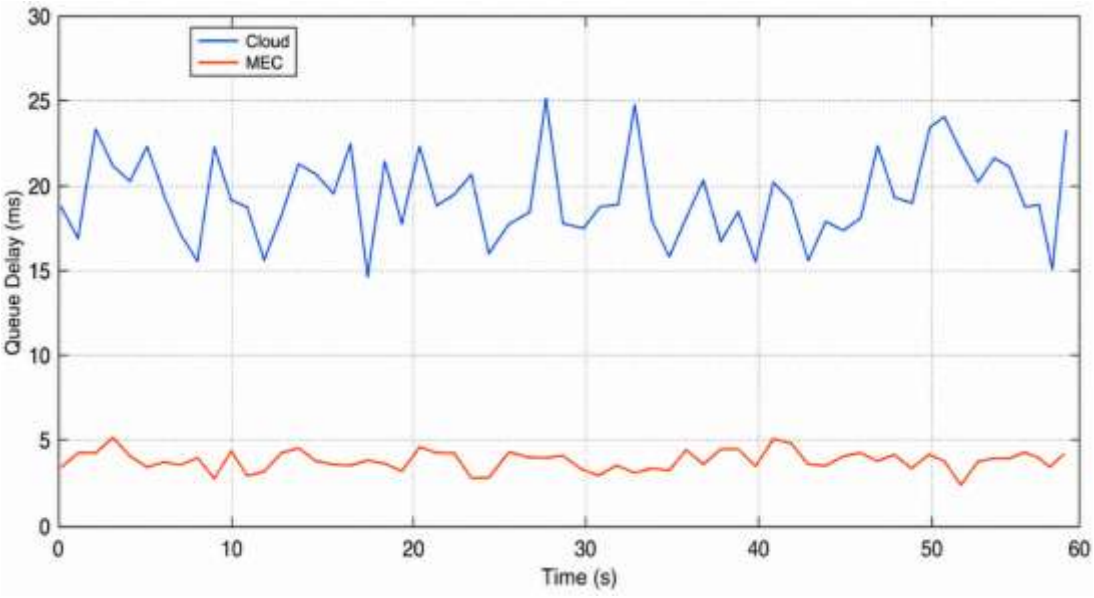


Figure 3. Queue length comparison between centralized and MEC-assisted RRM obtained from OMNeT++ simulations

for latency-sensitive traffic and reduces queueing delay through localized processing. This validates the analytical result that in (10), demonstrating the effectiveness of edge-based scheduling in minimizing latency.

Figure 5 illustrates end to end delay (latency) as simulated in OMNeT++. The reduction in delay also validates the theoretical model in (10). MEC-assisted RRM reduced average end-to-end delay to approximately 5 ms, compared with over 20 ms under centralized RRM.

4.3 Packet Loss Performance

Packet loss is significantly reduced in the MEC-assisted framework compared with the centralized approach. In the centralized system, high queue occupancy and congestion at the core network lead to increased packet drops due to buffer overflow. In contrast, the proposed MEC-assisted RRM framework distributes traffic processing across edge nodes, thereby reducing congestion and preventing queue overflow. This leads to improved reliability and higher packet delivery ratios. This result is consistent with the analytical relationship (13). Figure 6 indicates that packet loss increases with network load in both architectures, with the centralized system exhibiting a significantly higher loss rate.

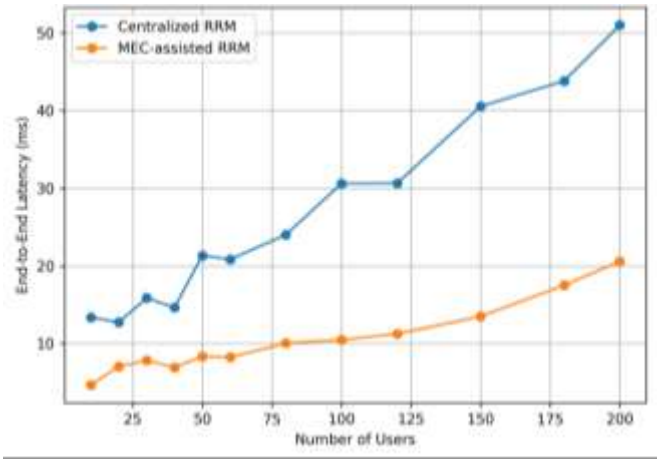


Figure 4. End-to-end latency versus number of users under centralized and MEC-assisted RRM.

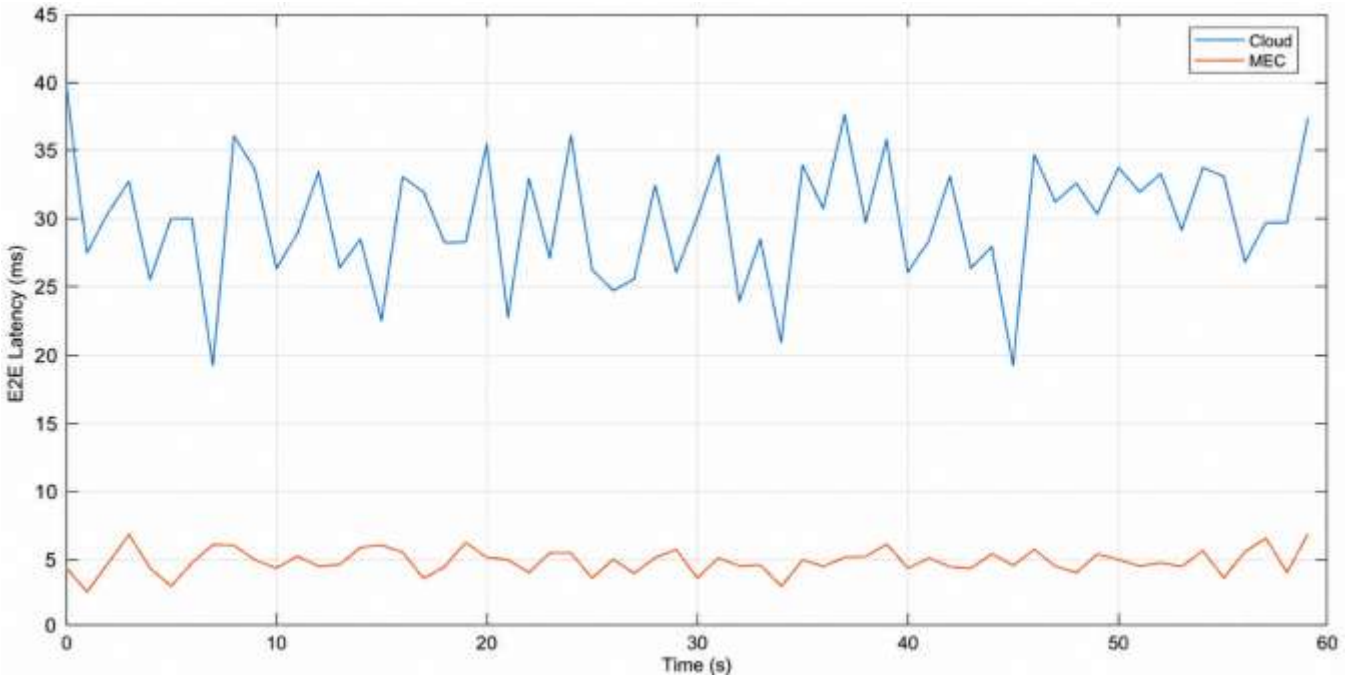


Figure 5. End-to-end delay comparison under centralized and MEC-assisted RRM (OMNeT++ results).

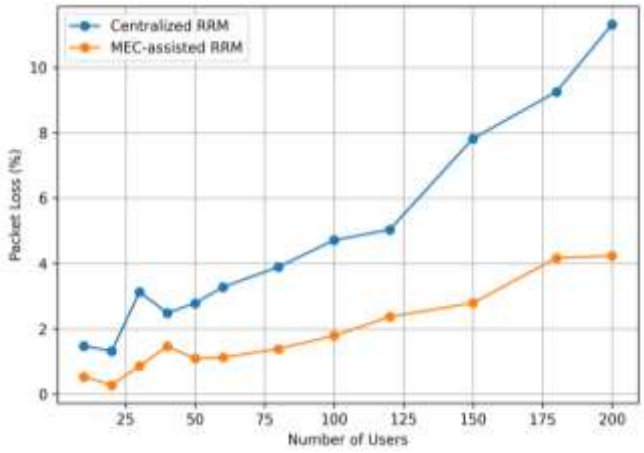


Figure 6. Packet loss versus number of users for centralized and MEC-assisted RRM architectures.

This is primarily due to buffer overflow caused by congestion at the core network. The MEC-assisted framework mitigates this effect by distributing traffic processing and reducing queue buildup, thereby lowering the probability of buffer overflow. This observation aligns with the analytical relationship between queue length and packet loss presented in Section 3.5. When users are less than 100, the packet loss was below 2%.

Figure 7 presents the comparison of packet loss performance between the centralized and MEC-assisted RRM architectures under varying traffic conditions. Results show that the proposed MEC-assisted framework achieves significantly lower packet loss compared with the centralized approach. In particular, the packet loss ratio under MEC-assisted RRM remains below 2% across most operating conditions, while the centralized architecture experiences packet loss exceeding 6% under high traffic load due to increased congestion and buffer overflow at the core network. Lower packet loss shown in Figure 6 is attributed to reduced congestion and improved buffer management at the edge (key features of MEC-assisted RRM).

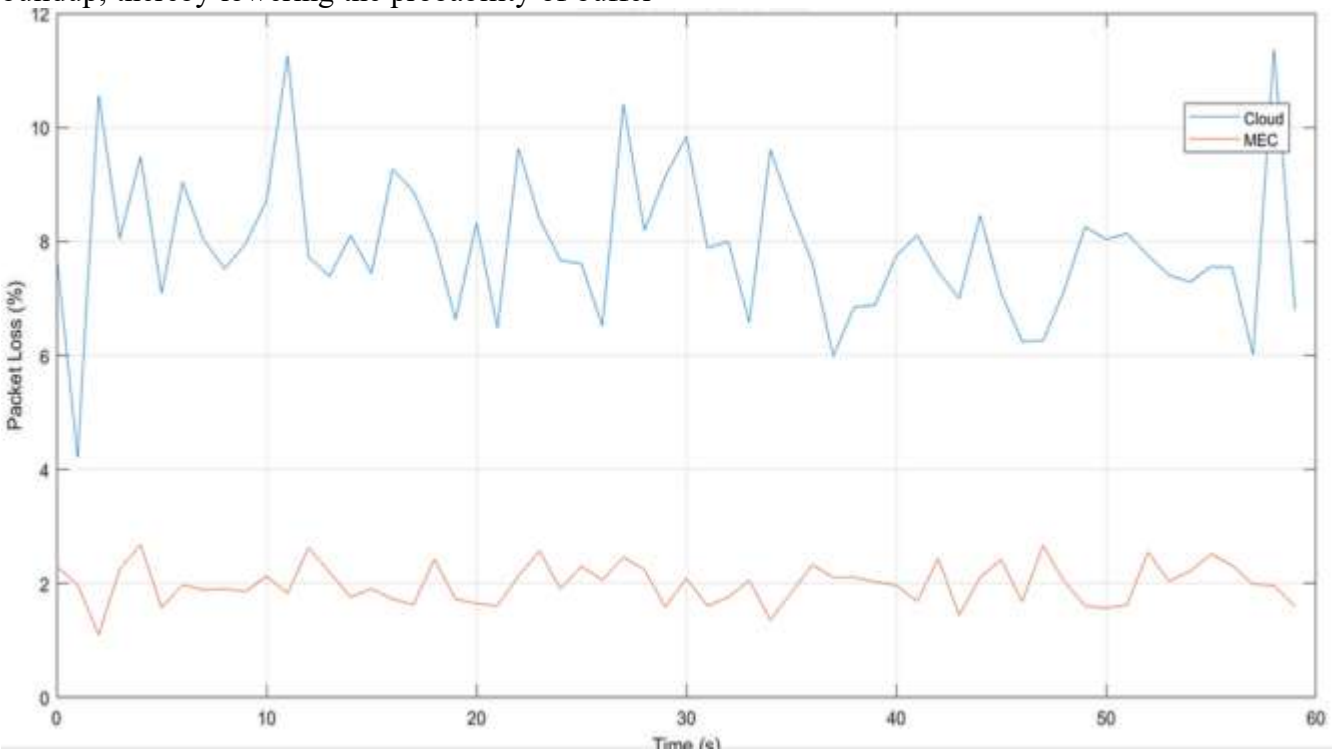


Figure 7. Packet loss comparison under varying traffic conditions for centralized and MEC-assisted architectures.

4.4 Throughput and Resource Utilization

Figure 8 presents the throughput performance of the proposed queue-aware MEC-assisted RRM framework compared with the conventional centralized RRM architecture under varying network load conditions. Results show that the proposed framework maintains more stable throughput performance as traffic intensity and user density increase. In contrast, the centralized RRM architecture experiences noticeable throughput degradation under heavy traffic conditions due to increased queue buildup, backhaul congestion, and centralized processing delays.

The proposed MEC-assisted framework demonstrates noticeable improvement in throughput stability under high traffic load conditions compared with the centralized RRM approach. This improvement is primarily attributed to localized edge processing, adaptive traffic steering, and queue-aware PRB allocation, which collectively reduce congestion at centralized processing nodes and improve continuity of packet delivery.

The results further indicate that throughput performance is strongly influenced by both radio-layer conditions and packet-level congestion dynamics. Under increased user density, the centralized RRM architecture exhibits larger throughput fluctuations caused by resource contention, queue accumulation, and higher processing delay. In contrast, the MEC-assisted framework distributes processing tasks across MEC-enabled nodes, thereby reducing bottlenecks associated with centralized scheduling and improving traffic-handling efficiency under dense network conditions.

Furthermore, reduced queue occupancy and lower processing delay decrease retransmission probability and improve overall network responsiveness, which contributes to more stable throughput behavior across heterogeneous traffic classes, including eMBB, URLLC, and mMTC services. The close agreement between the analytical evaluation and OMNeT++/Simu5G simulation results further validates the consistency and reliability of the proposed framework under varying traffic and deployment scenarios.

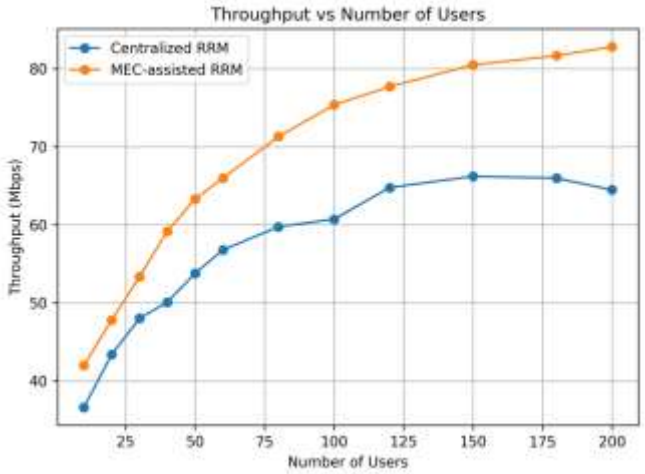


Figure 8. Throughput versus number of users showing improved stability under MEC-assisted RRM.

4.5 Comparative Performance Analysis

To provide a consolidated assessment of the proposed MEC-assisted RRM framework, a comparative performance analysis was conducted using both analytical evaluation and OMNeT++/Simu5G simulation results. The proposed framework was compared with the conventional centralized RRM architecture under identical network conditions and simulation parameters. The analysis focused on key performance metrics, including queueing delay, end-to-end latency, packet loss, throughput stability, resource utilization, and scalability under varying traffic conditions.

Table 2 presents a qualitative comparison between the centralized and MEC-assisted RRM architectures, highlighting the general performance behavior of each approach under increased traffic load. Table 3 provides corresponding numerical estimates and percentage improvements derived from the analytical and simulation results. Together, the two tables demonstrate that the proposed MEC-assisted framework achieves lower congestion levels, improved latency performance,

better resource utilization, and enhanced scalability primarily due to localized edge processing, adaptive traffic steering, and queue-aware scheduling.

Table 2. Performance Comparison between Centralized and MEC-Assisted RRM.

Performance Metric	Centralized RRM	MEC-Assisted RRM	Improvement Trend
Queueing Delay	High; increases rapidly with load due to core congestion	Low; grows slowly due to distributed processing	Significant reduction
End-to-End Latency	High; includes backhaul and core delays	Low; edge processing eliminates backhaul delay	Strong reduction
Packet Loss Ratio	High under heavy load due to buffer overflow	Low; reduced congestion at edge	Noticeable reduction
Throughput Stability	Fluctuates under high load	Stable; efficient resource allocation	Improved stability
Resource Utilization	Inefficient under congestion	Efficient via dynamic PRB allocation	Improved efficiency
Scalability	Limited; degrades at high density	High; maintains performance under load	Enhanced scalability

Table 3. Revised numerical performance comparison with justifications.

Metric	Centralized RRM	MEC-RRM	Improvement (%)	Technical Justification
Queueing Delay (ms)	21.08	5.37	≈74.5%	MEC reduces queue buildup by processing latency-sensitive traffic locally at edge servers, thereby lowering traffic intensity at the centralized core network.
End-to-End Latency (ms)	37.12	7.02	≈78.4%	Edge-assisted processing minimizes backhaul transmission delay and reduces centralized processing overhead for delay-sensitive traffic.
Packet Loss Ratio (%)	8–15%	<=2% for less than 100 users and 2-5% for over 100 users.	≈65–75%	Reduced congestion and lower queue occupancy at MEC nodes minimize buffer overflow and packet dropping under high traffic conditions.
Average Throughput (Mbps)	62–68	78–85	≈20–25%	Adaptive PRB allocation and reduced retransmissions improve continuity of packet delivery and overall throughput stability.

Metric	Centralized RRM	MEC-RRM	Improvement (%)	Technical Justification
Throughput Stability	High fluctuation	Low fluctuation	≈50–60% reduction	Distributed edge processing reduces congestion spikes and maintains more stable traffic handling under varying load conditions.
Resource Utilization Efficiency (%)	60–70%	80–90%	≈25–30%	Queue-aware scheduling and channel-aware PRB allocation improve utilization of available radio resources.
Scalability Capacity (Users)	≤100 users	≥200 users	≈100%	Distributed MEC processing reduces centralized bottlenecks, enabling the network to support significantly higher user density.
Average Queue Occupancy (%)	70–85%	25–40%	≈45–60%	Traffic offloading to edge nodes reduces accumulation of packets at centralized queues and improves queue stability.
Average Processing Delay (ms)	10–20	1–3	≈70–85%	MEC-enabled local processing reduces dependence on distant core-network computation resources.
Spectral Efficiency (bps/Hz)	3.5–4.2	5.0–6.1	≈30–40%	Channel-aware scheduling and interference-aware PRB allocation improve radio resource utilization efficiency.

4.6 Discussion Summary

The results presented in this section demonstrate the effectiveness of the proposed queue-aware MEC-assisted RRM framework in improving network performance compared with conventional centralized radio resource management approaches. Both analytical modeling and OMNeT++/Simu5G simulations consistently showed significant reductions in queueing delay, end-to-end latency, and packet loss, together with improved throughput stability under varying network load conditions. In particular, the proposed framework achieved approximately 74.5% reduction in queueing delay and 78.4% reduction in end-to-end latency, mainly due to localized edge processing, reduced backhaul dependency, and adaptive queue-aware scheduling.

The observed performance improvements are primarily attributed to the integration of distributed MEC processing, latency-aware traffic steering, and dynamic PRB allocation based on queue state and channel conditions. The results further indicate that network performance is influenced not only by radio-layer characteristics but also by packet-level dynamics, such as queue occupancy, buffering behavior, and processing delay. In addition, the close agreement between the analytical models and OMNeT++/Simu5G simulation results validates the consistency and reliability of the proposed framework.

Nevertheless, the results also indicate that the achievable performance gains may depend on deployment conditions such as MEC penetration rate, traffic intensity, signaling overhead, and mobility characteristics. Under scenarios involving MEC server overload, partial edge deployment, or highly dynamic mobility conditions, the performance improvement may decrease due to increased processing contention and handover-related signaling overhead. Despite these practical limitations, the findings demonstrate that MEC-assisted queue-aware RRM remains a scalable and

effective approach for supporting latency-sensitive and high-density services in 5G and next-generation wireless networks.

5. Conclusion

This paper presents a queue-aware MEC-assisted RRM framework designed to address latency and congestion challenges in heterogeneous 5G networks. By integrating latency-aware traffic offloading, queue-state-driven physical resource block scheduling, and adaptive load balancing between edge and core processing domains, the framework significantly reduces backhaul bottlenecks and slow adaptation cycles associated with conventional centralized RRM architectures.

Evaluation using a hybrid analytical and discrete-event simulation approach demonstrated substantial and consistent performance gains over the centralized baseline. The proposed framework reduced queueing delay by approximately 74.5% and end-to-end latency by 78.4% under high traffic load conditions, while constraining packet loss below 2%, improving resource utilization by 25–30%, and effectively doubling usable network capacity. These gains were observed across eMBB, URLLC, and mMTC service classes, confirming the suitability of the framework for heterogeneous 5G deployments with diverse QoS requirements.

Certain simplifying assumptions, including the M/M/1 queueing model and moderate user mobility, bound the scope of these results. Researchers may further extend the framework to partial MEC deployment scenarios, high-mobility environments, and AI-assisted scheduling integration, with direct benchmarking against state-of-the-art RRM approaches. Nevertheless, the results establish that queue-aware, edge-native resource management is not merely a theoretical improvement but a practically viable and scalable approach for meeting the stringent latency and reliability demands of next-generation wireless networks.

ACKNOWLEDGEMENT

The authors extend their special thanks to the research assistant, Mr. Denis Mbuya, for his significant contributions and dedicated assistance that greatly facilitated the successful completion of this study.

CONTRIBUTIONS OF CO-AUTHORS

Godfrey W. Chugulu	[ORCID: 0000-0001-7660-8643]	Conceived the idea and wrote the paper
Hadija R. Mbembati	[ORCID: 0000-0001-8748-4674]	Conducted experiments and analyzed results

APPENDIX

APPENDIX A: Simulation Parameters Used in This Study

Parameter	Centralized RRM	MEC-Assisted RRM
Network Architecture	Centralized core-network processing	Edge-assisted processing with MEC-enabled gNodeBs
Number of Cells	7 macro cells	7 macro cells
Users per Cell	20–60 UEs	20–60 UEs
Carrier Frequency	3.5 GHz	3.5 GHz
System Bandwidth	100 MHz	100 MHz
Traffic Types	VoIP, Video Streaming, IoT	VoIP, Video Streaming, IoT
VoIP Packet Size	64 bytes	64 bytes
Video Packet Size	1024 bytes	1024 bytes
IoT Packet Size	128 bytes	128 bytes
Packet Arrival Model	Poisson process	Poisson process
Queueing Model	M/M/1	M/M/1
gNodeB Transmission Power	40 W	40 W
UE Transmission Power	23 dBm	23 dBm
Mobility Model	Gauss–Markov	Gauss–Markov
UE Speed Range	3–30 km/h	3–30 km/h
Simulation Duration	1000 s	1000 s
Scheduler Type	Centralized scheduling	Queue-aware MEC-assisted scheduling
Traffic Processing Location	Core network	MEC server / Edge
MEC Processing Delay	Not Applicable	1–3 ms
Core Network Processing Delay	10–20 ms	Reduced due to edge offloading
Queue Monitoring	Centralized	Real-time edge-assisted queue monitoring
CQI-Based PRB Allocation	Limited	Enabled
Traffic Offloading	Not Supported	Supported
Congestion Control	Centralized	Queue-aware adaptive control
Evaluation Metrics	Delay, Throughput, Packet Loss, Queue Length	Delay, Throughput, Packet Loss, Queue Length

REFERENCES

[1] U. J. Ibe, Emekaoranebo, and S. Adaobi, "Performance metrics of 5G mobile communication network and its impact on communication systems," *Int. J. Sci. Eng. (IJSE)*, vol. **1**, no. 1, pp. 90–100, Nov. 2025.

[2] E. Iradier, A. Abuin, L. Fanari, J. Montalban, and P. Angueira, "Throughput, capacity and latency analysis of P-NOMA RRM schemes in 5G URLLC," *Multimed. Tools Appl.*, vol. **81**, no. 9, pp. 12251–12273, 2022, doi: 10.1007/s11042-021-11086-6.

[3] G. Subrahmanyam, M. Puttaramaiah, G. Kalnoor, and P. B. Metre, "Joint radio resource management soft computing technique for 5G and beyond 5G wireless networks," *Procedia Comput. Sci.*, vol. **265**, pp. 17–24, 2025, doi: 10.1016/j.procs.2025.07.151.

[4] E. Fontes Pupo, C. Carballo González, J. Montalban, P. Angueira, M. Murroni, and E. Iradier, "Artificial intelligence aided low complexity RRM algorithms for 5G-MBS," *IEEE Trans. Broadcast.*, vol. **70**, no. 1, pp. 110–122, 2024, doi: 10.1109/TBC.2023.3311337.

[5] N. H. Nwobodo-Nzerebe and E. U. Chidi, "Optimizing network radio resource management using machine learning technique for traffic improvement," vol. **4**, no. 64, pp. 814–826, 2025.

[6] M. Marwani and G. Kaddoum, "Event-based temporal graph neural network for radio resource management," *IEEE Trans. Wireless Commun.*, vol. **25**, pp. 10855–10868, 2026, doi: 10.1109/TWC.2026.3654880.

[7] N. Kumar and A. Ahmad, "Machine learning-based QoS and traffic-aware prediction-assisted dynamic network slicing," *Int. J. Commun. Netw. Distrib. Syst.*, vol. **28**, no. 1, pp. 27–42, 2022, doi: 10.1504/IJCNDS.2022.120298.

[8] A. Sarah, G. Nencioni, and M. M. I. Khan, "Resource allocation in multi-access edge computing for 5G-and-beyond networks," *Comput. Netw.*, vol. **227**, Art. no. 109720, 2023, doi: 10.1016/j.comnet.2023.109720.

[9] A. Ebrahimi and F. Afghah, "Intelligent task offloading: Advanced MEC task offloading and resource management in 5G networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2025, pp. 1–6, doi: 10.1109/WCNC61545.2025.10978559.

[10] R. Xavier, R. S. Silva, M. Ribeiro, W. Moreira, L. Freitas, and A. Oliveira-Jr, "Integrating multi-access edge computing (MEC) into open 5G core," *Telecom*, vol. **5**, no. 2, pp. 433–450, 2024, doi: 10.3390/telecom5020022.

[11] Q.-V. Pham *et al.*, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. **8**, pp. 116974–117017, 2020, doi: 10.1109/ACCESS.2020.3001277.

[12] X. Ma, A. Zhou, S. Zhang, and S. Wang, "Cooperative service caching and workload scheduling in mobile edge computing," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2020, pp. 2076–2085, doi: 10.1109/INFOCOM41043.2020.9155455.

[13] A. G. Burr, A. G. Armada, C. Smeenk *et al.*, "COST INTERACT whitepaper on signal processing for communications, localization, and integrated sensing and communication," 2024.

[14] G. B. Shettar and A. D. Devangavi, "A comprehensive study and future prospects on multi-access techniques for efficient resource allocation in MEC," in *Proc. 6th Int. Conf. Control, Commun. Comput. (ICCC)*, 2025, pp. 1–6, doi: 10.1109/ICCC64910.2025.11077352.

[15] N. Rasouli, C. Klein, and E. Elmroth, "Resource management for mission-critical applications in edge computing: Systematic review on recent research and open issues," *ACM Comput. Surv.*, vol. **58**, no. 3, 2025, doi: 10.1145/3762181.

[16] A. Virdis, G. Nardini, G. Stea, and D. Sabella, "End-to-end performance evaluation of MEC deployments in 5G scenarios," *J. Sens. Actuator Netw.*, vol. **9**, no. 4, Art. no. 57, 2020, doi: 10.3390/jsan9040057.

[17] E. Fontes Pupo, C. Carballo González, J. Montalban, P. Angueira, M. Murroni, and E. Iradier, "Artificial intelligence aided low complexity RRM algorithms for 5G-MBS," *IEEE Trans. Broadcast.*, vol. **70**, no. 1, pp. 110–122, 2024, doi: 10.1109/TBC.2023.3311337.

[18] M. Marwani and G. Kaddoum, "Event-based temporal graph neural network for radio resource management," *IEEE Trans. Wireless Commun.*, vol. **25**, pp. 10855–10868, 2026, doi: 10.1109/TWC.2026.3654880.

[19] E. Iradier, A. Abuin, L. Fanari, J. Montalban, and P. Angueira, "Throughput, capacity and latency analysis of P-NOMA RRM schemes in 5G URLLC," *Multimed. Tools Appl.*, vol. **81**, no. 9, pp. 12251–12273, 2022, doi: 10.1007/s11042-021-11086-6.

[20] A. Ebrahimi and F. Afghah, "Intelligent task offloading: Advanced MEC task offloading and resource management in 5G networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2025, pp. 1–6, doi: 10.1109/WCNC61545.2025.10978559.

[21] V. K. Gautam and B. R. Tamma, "RETALIN: A queue aware uplink scheduling scheme for reducing scheduling signaling overhead in 5G NR," *IEEE Access*, vol. **12**, pp. 18900–18916, 2024, doi: 10.1109/ACCESS.2024.3359028.

[22] T. Chrysikos, "QoS-aware resource management in 5G and 6G cloud-based architectures with priorities," *Information*, vol. **14**, no. 3, Art. no. 175, Mar. 2023, doi: 10.3390/info14030175.

[23] S. Mostafa, M. P. Mota, Á. Valcarce, I. Labrador, and S. Barrachina, "Intent-aware DRL-based uplink dynamic scheduler for 5G-NR," 2024.

[24] Q.-V. Pham, F. Fang, V. N. Ha, D. Kim, and H. Arslan, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," 2022.

[25] H. Zhou and M. Erol-Kantarci, "Knowledge transfer based radio and computation resource allocation for 5G RAN slicing," 2022.

[26] B. Coll-Perales, M. C. Lucas-Estañ, T. Shimizu, and J. F. Monserrat, "End-to-end V2X latency modeling and analysis in 5G networks," 2022.

[27] X. Wu, D. Wu, and E. Modiano, "Queueing delay minimization in overloaded networks," 2023.

[28] R. Solozabal, A. Sanchoyerto, M. Cava, J. L. Pipa, D. López-Pérez, and M. Luise, "Providing mission-critical services over 5G radio access network," 2018.

[29] X. Xu, Y. Xia, F. Zeng, and Q. Jiang, "A novel vehicular task deployment method in hybrid MEC," 2022.

[30] Y. Pan, A. Nix, and M. A. Beach, in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Singapore, 2008.

[31] Z. Chen, S. Zou, Y. Tang, B. Fu, J. Li, and J. Cao, "Radio resource coordination and scheduling scheme in ultra-dense cloud-based small cell networks," 2018.

[32] H. Zaaraoui, "End-to-end resource management and service quality control with mobility in LTE/LTE-Advanced networks," 2017.

[33] M. Di Mauro and A. Liotta, "An experimental evaluation and characterization of VoIP over an LTE-A network," 2020.

[34] A. Baratè *et al.*, "5G technology for music education: A feasibility study," 2019.

[35] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," 2019.