Admission Control with Latency Considerations for 5G Mobile Edge Computing

Ye Zhang, Wuyungerile Li**
Inner Mongolia Key Lab. of Wireless Networking & Mobile Computing
Engineering Research Center of Ecological Big Data, Ministry of Education
School of Computer Science, Inner Mongolia University
Hohhot, China
zhangye9852@163.com, gerile@imu.edu.cn

Winston K.G. Seah
Wireless Networks Research Group
School of Engineering and Computer Science
Victoria University of Wellington
Wellington, New Zealand
winston.seah@ecs.vuw.ac.nz

***Abstract*—The fifth generation (5G) mobile network is a new generation of broadband mobile communication technology with the potential to address the increasing demands of new user services and applications that have stringent low latency and high bandwidth requirements. Besides the enhanced mobile broadband (eMBB) which is an evolution of broadband services from previous generations, the ultra reliable low latency communication (URLLC) service comes with stringent delay requirements that are needed to support new applications like autonomous vehicles, augmented/virtual reality, etc. Mobile or multiple access edge computing (MEC) emerged to provide services and computing resources for users at the network edge to provide faster access speeds and lower end-to-end delays. To better meet user needs and maximize resource utilization, network resources need to be allocated and managed efficiently. Admission control for user requests is one of the methods used that can effectively prevent network congestion, thereby improving the overall performance of the system. In this paper, we propose a RED-based Admission Control with Latency Considerations (REDAL) algorithm for user admission control that aims to increase throughput, meet users' delay requirements and reduce packet discard rate. By explicitly accounting for user traffic delay constraints and bandwidth requirements, we are able to meet the strict delay constraints of URLLC traffic while meeting the bandwidth requirements of eMBB traffic. We validate our approach in an MEC scenario to demonstrate high resource utilization and also keeping the request discard rate below 20%.**

***Index Terms*—admission control, 5G, MEC, URLLC, eMBB**

I. Introduction

Advances in mobile communication and networking technologies continue as new and more demanding applications, such as, autonomous driving, factory automation, smart city, augmented reality (AR) and virtual reality (VR) are emerging. Driven by the demands for more diverse and stringent requirements on network capabilities, the European Union first proposed the fifth generation (5G) mobile communication technology, which is characterised high speed, low delay, and high bandwidth connectivity. One of the key technologies in the transformation from 4G network to 5G network is mobile or multiple access edge computing (MEC) which can well meet the needs of emerging services and applications.

The International Telecommunication Union (ITU) has defined three generic usage scenarios of 5G: ultra reliable low latency communication (URLLC), enhanced mobile broadband (eMBB), and massive machine type of communication (mMTC) [1]. URLLC applications, e.g., Internet of vehicles, telemedicine and industrial control, have requirements of ultra low latency and high reliability. eMBB is an evolution of the previous generation of mobile networks and supports applications centred on users' experiences. mMTC applications form the bulk of the Internet of everything comprising low-power low-cost devices that are widely distributed and numerous.

Efficient scheduling of both communication and computing resources is critical in an MEC context [2]. In a 5G network, the user equipment (UE) will first send a request for resources via the signalling channel, stating the requirements of the user traffic. The 5G base station, referred to as gNodeB or gNB, will then decide whether to accept the request and allocate the resources. When the network is busy, it is also necessary to control the admission of users' requests to avoid the degradation of the overall system performance caused by excessive resource competition.

In this paper, we show that through admission control (AC), the resource utilization rate and the number of serviceable users can be improved. Using an algorithm based on random early detection (RED), we aim to solve the problem of coexistence and resource competition between URLLC and eMBB users in a 5G MEC scenario while considering the delay/latency requirements of both traffic types. The second section describes the relevant research on resource allocation and admission control, while the third section presents the scenario and fourth section describes our algorithm. Next, we discuss the validation and performance evaluation of our algorithm before concluding the paper.

II. Related Work

The coexistence of diversified services with different requirements will become a typical scenario in 5G communication in the future. However, since the resources deployed by operators are usually limited, there will be multiple services sharing resources or even competing for resources within a

**Corresponding author

certain range, including communication resources and computing resources. By reasonably optimizing resource allocation, users' QoS and other requirements can be guaranteed. Admission control of screening the user's request can ensure the smoothness of the network and maximize the benefits of the operators.

Regarding resource allocation, Sardellitti *et al.* [3] proposed a method of jointly optimizing power, the number of bits per symbol, and the CPU cycle allocated to each application, aiming at minimizing the energy consumption of the mobile terminal. You *et al.* [4] discussed unloading the computing task to the MEC server in an efficient way, and solving the decomposed subproblem using the fast coordinate descent algorithm (BCD). To minimize resource occupation under the condition of satisfying QoS constraints, user ranking criteria to control admission has been proposed to achieve efficient computing and offloading of resources of users subject to QoS constraints [5]. However, the problem is that users may leave the system dynamically, so online resource allocation should be considered.

Admission control can improve resource utilization through screening before granting resources to more users. Admission control with the aim of maximizing the number of eMBB users admitted with guaranteed data rate while ensuring that URLLC users' requirements are always met has been formulated as an ℓ_0 minimization problem which is NP-hard [6] [7]; both studies focused on the downlink of a single-cell multiple-input single-output (MISO) system supporting eMBB and URLLC users. A suboptimal solution obtained using sequential convex programming was validated in a network scenario of 8 URLLC and 8 eMBB users and shown to achieve near-optimal performance [7]. Admission control of URLLC users was also analyzed using a realistic queueing-theoretic model, considering both homogeneous and heterogeneous users [8]. From the analysis, an admission policy was formulated and validated using simulations.

As expected, machine learning has been applied to admission control problems too. Reinforcement learning has been proposed for UE admission control in 5G networks [9], outperforming threshold-based policies under conditions of heterogeneous time-varying arrival rates and multiple UE types. Likewise, Q-learning and R-learning algorithms have also been applied to approximate the optimal admission control strategy in multi-domain 5G networks [10].

Chagdali *et al.* [11] demonstrate that the placement of the slice management function plays a crucial role in selecting the most suitable radio resource allocation scheme for URLLC slices by evaluating the impact of architecture choice on different quality of services. Similarly, the coexistence of the three major application scenarios needs to be addressed, such as, the coexistence of URLLC and eMBB [12]. Two problems are considered: one is offloading computational tasks to MEC servers in an energy-efficient manner, and the other is the coexistence of mobile users with different service requirements for eMBB and URLLC users in a cellular network, for which an energy-efficient task offloading and scheduling of eMBB and URLLC users as a mixed integer nonlinear problem is formulated. A back-propagation neural network (BPNN) based hole-punching scheduling scheme has been proposed to solve the problem of placing eMBB and URLLC traffic in hourly gaps, which effectively reduces network losses and improves network throughput [13]. Likewise, Arjun *et al.* [14] propose a joint scheduler to maximise the utility of eMBB traffic while satisfying the URLLC requirements and demonstrate that the optimal eMBB and URLLC scheduling decisions are not disconnected and require joint optimization to satisfy the dual objectives.

While formal and rigorous analytical approaches are critical for developing optimal admission control strategies for the complex scenarios that 5G networks present, these approaches are compute intensive. We opined that tested strategies adopted in the Internet can provide simple efficient solutions suitable for admission control in MEC scenarios. E.g., active queue management has been proposed for energy-efficient offloading of tasks from the edge/fog to the cloud [15] [16]. Our approach to be presented in the following sections considers the stringent latency constraints inherent in URLLC traffic while aiming to maximize resource utilization to support eMBB traffic.

III. Scenario and Problem Formulation

In this section, we first outline the MEC scenario, then provide a problem description of the admission control problem.

A. Scenario

MEC is a system that provides cloud computing capability for network edge users. Its deployment scenario is shown in Fig. 1. The mobile terminal UE, which includes all the devices that can be networked, generates computing needs to be offloaded to the nearby MEC server. The operator will deploy different amounts of resources on the MEC server, and use the gNB as a channel to connect the mobile terminal and the MEC server. By using MEC, we can better meet the relevant requirements of 5G and better serve multiple users, such as, users generating eMBB traffic from high-definition video and VR/AR, and URLLC traffic generated by the Internet of Vehicles, etc. Providing cloud-like computing resources close to the users without having to send to the cloud can significantly reduce the delay in task execution. Using MEC technology can also improve security, keeping the exchange of data between the terminal and the application without having to transmit across the network, thus ensuring data privacy.

B. Problem Description

We consider the scenario where there are URLLC and eMBB users with computation tasks to be offloaded to the MEC server. Each user packet/request u has a computation task defined by $\gamma_u = \{D_u, C_u, T_u, R_u\}$ where D_u is the total data size of user u's input data, C_u is the total number of CPU cycles required to complete a computation task, and T_u is the maximum tolerable delay the task, and R_u is the

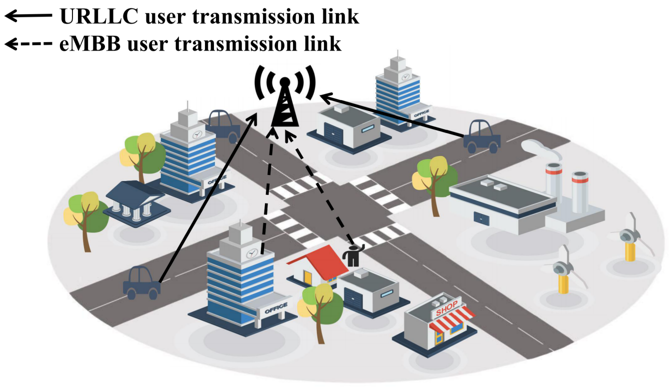

Fig. 1. Application Scenario Example

target data rate. The MEC system has communication C^U and computing C^P resources, which for simplicity, we allocate different network slices for the two service types URLLC and eMBB based on pre-defined bandwidth sharing ratios (cf: Section V-B1.) We then apply our admission control algorithm to maximize resource utilization while minimizing delay and packet discarding.

IV. RED-based Admission Control with Latency Considerations

Active queue management (AQM) is a popular admission control approach for managing congestion on routers to avoid global synchronization in the Internet. Global synchronization is the consequence of droptail queue management that affects TCP connections which are usually multiplexed over the same routing paths, resulting in multiple TCP connections going into slow-start at the same time. Global synchronisation can cause a sudden drop in network traffic, but once the network returns to normal, traffic can suddenly increase and again the network can become overloaded.

Random early detection (RED) [17] is a widely studied and applied AQM mechanism. The core idea of RED is early detection and random discarding. The goal of RED is to keep the average queue length low even in the case of packet bursts. The traditional RED algorithm handles the queueing of packets located at the network layer to perform active queue management. However, in 5G wireless communication networks when a UE needs to send data, it first requests resources from the network. The UE sends a message to the gNB, requesting uplink authorisation to send data on the uplink via the Physical Uplink Shared Channel (PUSCH); such requests are carried in Scheduling Request (SR) messages in 5G networks.

Therefore, we propose a new queue management method, which extends RED algorithm and uses the basic idea of early detection and random discarding to handle request packet queues instead of packet queues. The random discarding is not only based on the threshold, but also considers whether it can meet the specific constraints in the request packet. Admission control is based on the arrival time, delay requirements, and other information in the UE request packet. The algorithm aims to meet the QoS requirements of each UE, especially the delay constraint, and at the same time pursue higher throughput and lower packet discard rate. We call our proposed algorithm *RED-based Admission Control with Latency Considerations* (REDAL) and use it to solve the admission control problem in the coexistence of URLLC and eMBB traffic.

As shown in Fig. 2, our proposed system includes multiple URLLC and eMBB UEs, a request queue and a queue manager including resource allocation and admission control. Each UE sends small service request packets containing information flows to establish the request queue. According to the delay constraints, bandwidth requirements, types, and other data, queue manager first allocates resources and then applies the admission control algorithm to decide whether to accept the new request or not. Finally, the queue manager sends the admission decision to the UE. The UE's request that is accepted will offload its computing task to the MEC server at the assigned communication slot(s), and the UE who does not receive the admission reply can try another MEC server or upload it to the cloud for processing. Next, we introduce the details of each key component of the system.

A. Traffic Prioritization

According to the priority formula, we generate a new request queue from the user's request scheduling information. We assume data packets carrying URLLC traffic have higher priority while eMBB traffic have lower priority. Within the queues of different traffic types, it is also necessary to distinguish priorities, which are mainly based on delay constraints and bandwidth requirements. The priority computation formula of URLLC packets is as follows in Eqn. (1):

$$P_{U_n} = \alpha \frac{T_{U_n}^{max}}{T_{U_n}} + (1-\alpha) \frac{R_{U_n}}{R_{U_n}^{max}} \tag{1}$$

while the computation of eMBB priority is given in Eqn. (2):

$$P_{e_n} = \beta \frac{T_{e_n}^{max}}{T_{e_n}} + (1-\beta) \frac{R_{e_n}}{R_{e_n}^{max}} \tag{2}$$

where α $(0 \leq \alpha \leq 1)$ and β $(0 \leq \beta \leq 1)$ are the priority adjustment factors. Generally speaking, lower delays and the faster data transmission rates are better. T_v^{max} denotes the maximum delay that user v can accept; T_v denotes the actual delay incurred by user v; R_v^{max} denotes the target data rate of user v; and R_v denotes the actual data rate of user v, for $v \in \{U_n, e_n\}$. When $\frac{T_v^{max}}{T_v}$ and $\frac{R_v}{R_v^{max}}$ are less than 1, it means that the user's requirements cannot be met and the service cannot be provided. When the value approaches 1, it indicates that the resources of the network can just satisfy the requirements, and an earlier transmission time is better, implying higher priority. When this value is much greater than 1, it indicates that there are much resources left and waiting can be tolerated, implying lower priority.

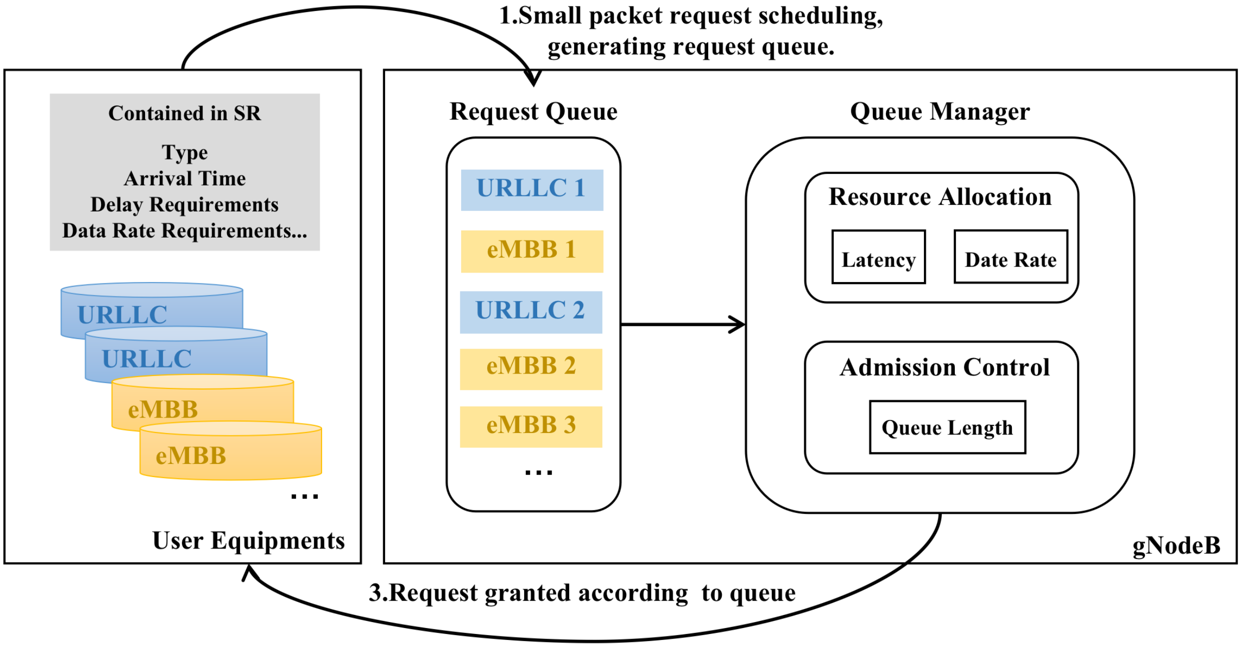

Fig. 2. 5G MEC System Architecture Diagram

B. Resource Allocation and Admission Control

For an arbitrary user u, the delay constraint includes two parts: uplink transmission delay, t_u^{up}, and server computation delay, t_u^c. It should be ensured that the sum of transmission delay and computing delay is less than the maximum tolerable delay of a user's task, that is:

$$t_u^{up} + t_u^c \leq T_u^{max} \tag{3}$$

Radio resources for the two service types are provisioned separately subject to the consition that the bandwidth resources occupied by URLLC and eMBB users should not exceed the total bandwidth resources, as shown in Eqn. (4)

$$\sum_{v \in \{U_n, e_n\}} B_v \leq B_{total} \tag{4}$$

and meet the data rate requirements of each user u as shown in Eqn. (5).

$$\frac{B_v}{D_u} \geq R_u \quad where\ v \in \{U_n, e_n\} \tag{5}$$

The basic idea of the resource allocation algorithm is to sort the packets, then filter the users meeting the QoS requirements according to the priority computation formula, while sharing the total bandwidth between the two service types.

The algorithm detects congestion by monitoring the average queue length and reduces the congestion window before the queue overflow leads to packet discard, thus alleviating network congestion. The algorithm has two key parameters: minimum threshold denoted by Th_{min} and maximum threshold denoted by Th_{max}. When the queue length L_{avg} is less than Th_{min}, an arriving packet is enqueued. When L_{avg} is between Th_{min} and Th_{max}, the packet marking probability P_a will be calculated for each arriving packet, as follows:

$$P_a = P_{max} \times \frac{L_{avg} - Th_{min}}{Th_{max} - Th_{min}} \tag{6}$$

where P_{max} is the maximum probability to drop a packet. The average queue size L_{avg} calculated as follows:

$$L_{avg} = (1 - W_p) \times L_{avg} + W_p \times q \tag{7}$$

where q is the instantaneous queue size and W_p is the time constant of the lowpass filter. If W_p is large, burst congestion will not be filtered. The probability P_a is a linear function of the average queue size and the threshold. When L_{avg} is greater than Th_{max}, $P_a = 1$. If C is the number of data packets containing the last marked data packet, then the final marking probability P is:

$$P = \frac{P_a}{1 - C \times P_a} \tag{8}$$

When a new packet arrives and the queue is full, marked packets in the queue will be discarded to accommodate the new packet. Another possible implementation of RED would be to discard packets based on the probability P_a instead of marking them.

In this system, the gNBs provide communication services for users, and each gNB allocates the necessary bandwidth resources to transmit the packets (carrying the task data) to the MEC server. Arriving packet types have different admission criteria and the gNB must ensure that the remaining resources are sufficient to serve them. URLLC traffic must meet their delay constraints while eMBB traffic have transmission rate requirements, thus users with URLLC traffic have higher

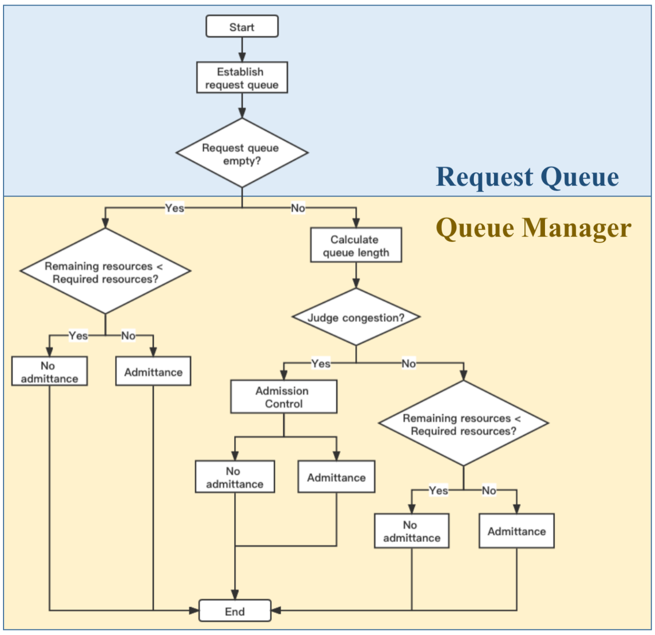

Fig. 3. REDAL Algorithm Flow Chart

priority over eMBB traffic users. URLLC packets are sporadic and small in size, while eMBB packets are often large, so the goal of REDAL is to keep the packet discard rate as low as possible and maximize the number of serviceable users while meeting the user QoS requirements. The basic idea of REDAL algorithm can be divided into three steps, and the flow chart of the algorithm is shown in Fig. 3:

1) Establish a request queue, calculate the priority of each URLLC and eMBB user request according to Eqn. (1) and Eqn. (2);
2) If the request queue is empty, use Eqn. (3) to determine whether the delay constraint is met, and use Eqn. (4) to determine whether the bandwidth constraints are met, that is, whether the remaining resources are sufficient to serve them. If both conditions are satisfied, the request is admitted; otherwise, if the remaining resources are insufficient, the request will be declined;
3) If the request queue is not empty, use Eqn. (7) to calculate the queue length to determine whether congestion is imminent. If congestion is imminent, the REDAL algorithm uses Eqn. (8) to calculate the packet discard probability. To ensure the constrains of Eqn. (3) and Eqn. (4) can be met, the request is discarded based on the computed discard probability. On the other hand, if there is no congestion and the remaining resources are sufficient, the request is admitted.

Although RED algorithm can effectively avoid congestion, it also has some limitations. It is fundamentally a congestion detection algorithm that marks packets to be discarded. If there is no mechanism to discard the marked packets, congestion and unfair resource will still occur. To avoid this, we adopt the discard approach based on the computed probability. The RED algorithm is also sensitive to parameter settings, and changing Th_{min} and Th_{max} thresholds will impact the performance.

The REDAL algorithm proposed in this paper extends the traditional RED algorithm by considering the delay requirements of users' requests, in order to meet the QoS requirements for different types of users' traffic types. This improves the throughput of the system and minimizes the delay, and most importantly, reduces the packet discard rate to achieve significantly better resource utilisation.

V. Validation and Performance

In this section, we first describe the simulation settings and scenarios used to validate our proposed algorithm, REDAL, based on throughput, delay, and packet discard rate. *Throughput* refers to the maximum of number of requests processed by the system in a unit time. Undoubtedly, throughput is a critical metric when analyzing network performance. *Delay* or *latency* refers to the time required for data packets to be transmitted from the source node to the destination node. In actual situations, users seek services, which is the sum of the time generated by the server in processing data and the time generated during transmission. In this system, the delay is an important metric since URLLC service is extremely sensitive to delay and it is necessary to analyze the delay performance. *Packet discard rate* refers to the fraction of data packets discarded by the admission control algorithm. The higher the value, the worse the user experience will be. In a 5G scenario, we assume that the underlying physical layer takes care of transmission errors while the admission control algorithm limits the amount of traffic that can be granted network access in order not to overwhelm the system resources. In this paper, one of our goals is to minimize the number of packets to be discarded while maximizing resource utilization and satisfying the delay constraints.

A. Simulation Setup

We implemented our algorithm in the ns-3 simulator with 5G-LENA module, which is a pluggable module for simulating 5G new radio (NR) cellular networks. The simulation scenario and representative traffic parameters [18] [19] are shown in Table 2, and we compare our algorithm with CoDel [20], RED [17], and Droptail which discards any new packet that arrives when the buffer is full.

To evaluate our scheme's ability to support user traffic with delay constraints, we classified user traffic into URLLC with the strictest delay constraint and eMBB with normal delay constraint. To provide a realistic scenario without incurring excessive simulation overheads, we assumed a total of 10 users, comprising 5 URLLC and 5 eMBB users, generating combined traffic data rates as shown in Table I. First, we tested the impact of different bandwidth allocation ratios on throughput, delay (or latency) and packet discard rate. After determining a suitable bandwidth allocation ratio, we evaluated the performance of our proposed algorithm and compared with other algorithms based on these metrics: throughput, delay,

TABLE I
SIMULATION PARAMETERS

	Parameter	Values
General	Simulation Time	10s
	Packet Size	8 and 220 bytes
	Traffic Data Rate	10∼100Mbps, in steps of 10Mbps
	Total Bandwidth	60Mbps (URLLC:eMBB - 1:1 & 1:2)
	Delay Constraints	1ms (URLLC) and 4 ms (eMBB)
REDAL RED	Th_{min}	50 packets
	Th_{max} (data rate)	150(10), 160(20), 170(30), 180(40), 190(50), 200(60), 210(70), 220(80), 230(90) and 240(100)
DropTail	MaxPackets	
CoDel		
	Target	5ms

and packet discard rate. We performed multiple simulation runs on each group of parameters and averaged the results to ensure that we achieved a stable representation of the performance [21].

B. Simulation Results

1) Bandwidth Allocation: Our algorithm provides differentiated services for URLLC and eMBB users. While different bandwidth allocation ratios will affect system performance, our focus is on admission control. Hence, we considered only two simple scenarios, URLLC:eMBB ratios of 1:1 and 1:2, to support multiple users seeking services.

As shown in Fig. 4, when an equal share of the bandwidth is granted to both traffic types, the throughput of eMBB is always higher than that of URLLC; furthermore, URLLC does not fully use the 30Mbps bandwidth allocated to it. Since URLLC is generally made up of small packets, the resource utilization is not high. The throughput of eMBB is higher when it is given a greater share, viz. 1:2. An interesting observation is that despite having less share of the bandwidth in the 1:2 scenario, URLLC performance was not adversely affected and more importantly, was able to stay within its 1ms delay constraint (cf: Fig. 5), even at 70Mbps data rate which exceeds 60Mbps available bandwidth. Beyond 60Mbps, the throughputs for both traffic types reach their maximum, indicating that further allocation is curbed as per the design but high resource utilization is achieved.

Fig. 5 shows the delay curve. In 1:1 allocation, the delay of URLLC is within its delay constraint of 1ms under all data rates except 100Mbps while the delay of eMBB is always within 4ms. In the 1:2 allocation, the delay of URLLC is still about 0.7ms-1.3ms, and as previously noted, keeps within 1ms until 70Mbps while eMBB delay ranges from 3.2ms to 4.3ms. It can be seen from this figure that the delay increases with the data rate, and the growth rate of eMBB is greater than that of URLLC, showing greater sensitivity to changes in bandwidth share. In our algorithm, we ensure that the priority of URLLC is always higher than that of eMBB in the same slot, because of its stringent delay constraint, which is one of the reasons for the higher delay of eMBB. In addition, the larger packet size is also another factor for the higher latency of eMBB.

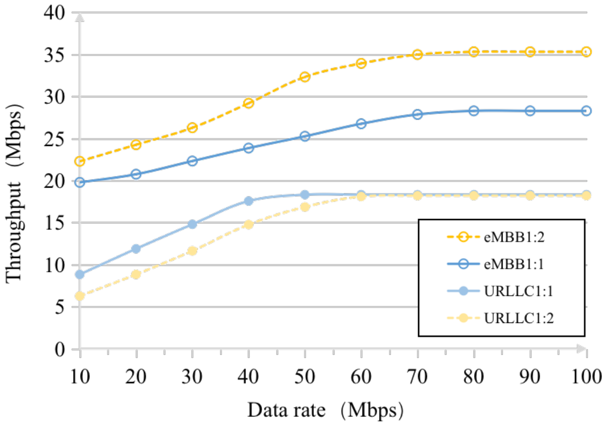

Fig. 4. Bandwidth Allocation on Throughput

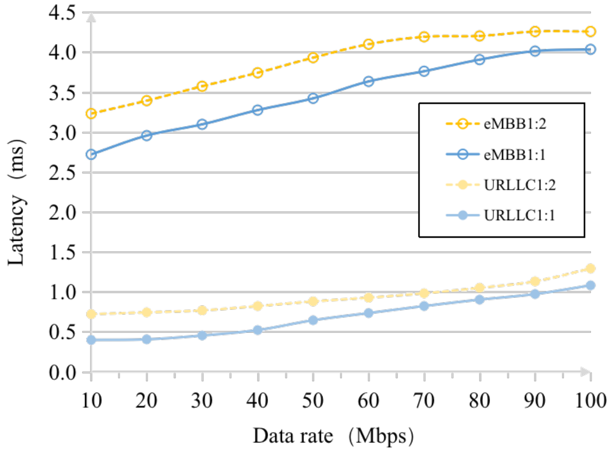

Fig. 5. Bandwidth Allocation on Latency

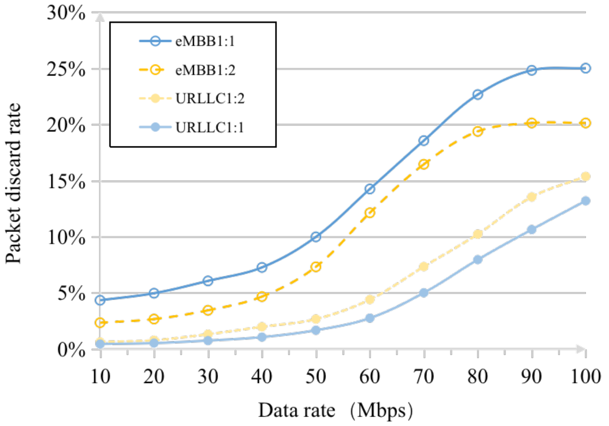

Fig. 6. Bandwidth Allocation on Discard Rate

From the packet discard rates shown in Fig. 6, it can be concluded that although the packet discard rate of URLLC users is slightly higher with 1:2 bandwidth allocation, the packet discard rate of eMBB user requests is significantly lower compared to the 1:1 allocation. Therefore, considering

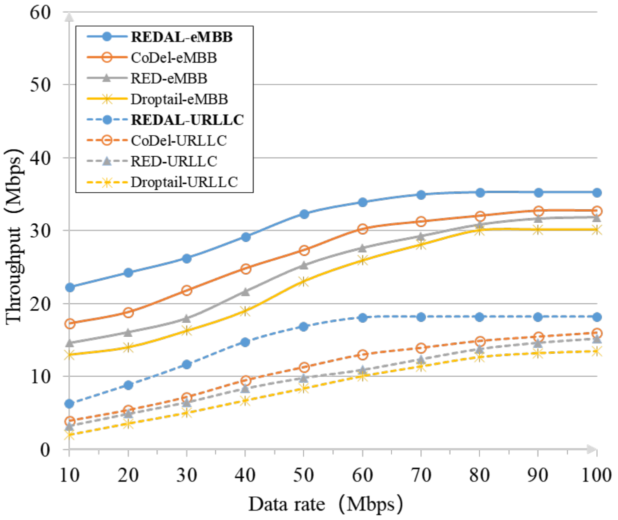

Fig. 7. Throughput Comparison

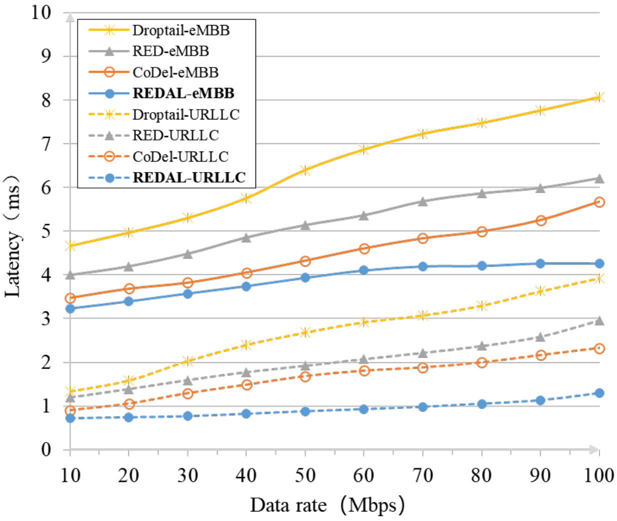

Fig. 8. Latency Comparison

all three indicators holistically, the 1:2 bandwidth allocation ratio gives better performance. As the REDAL algorithm uses delay constraints and bandwidth requirements to determine whether to admit new users' requests, we can increase system throughput by admitting users yet release more resources to subsequent requests. This reduces the delay as well as the need to discard more when congestion arises.

2) Comparison with CoDel, RED and Droptail: We also compared with other AQM algorithms, including CoDel, RED, and Droptail, using the 1:2 bandwidth allocation ratio. Fig. 7 shows the throughput comparison. Whether for URLLC or eMBB user requests, the REDAL algorithm achieved the highest throughput compared with other algorithms, and, more importantly for URLLC requests, it is able to satisfy the delay constraint up till 70Mbps.

This is very evident in Fig. 8, where REDAL delays of both URLLC and eMBB requests remain relatively low with a slow rate of increase, which indicates that the QoS requirements of user requests can be well met by the algorithm. Unlike the other algorithms, REDAL includes the delay constraints in the admission control decisions, so it can maintain the delay of URLLC and eMBB at about 1ms and 4ms respectively.

Lastly, we show in Fig. 9 relatively low packet discard rate achieved by REDAL when the network is not congested. As data rates increase, there is no sharp increase in discarding requests and, in fact, the discard rate stays within 20% which indicates that the number of serviceable user requests can be increased more using the REDAL algorithm. As the throughput is increased and the system delay is reduced, this is an indication that requests are served promptly which in turn reduces the possibility of queue congestion. With lower possibility of congestion, there is less need for requests to be discarded.

From the above results, we can see that the REDAL algorithm proposed in this paper performs better than other algorithms in terms of throughput, delay, and packet discard rate. The next best algorithm is the CoDel algorithm, and the worst performance is the Droptail algorithm. For admission control, it is critical not to aggressively discard requests to avoid congestion or violation of QoS requirements as this will lead to low throughput and resource utilization. Hence, having a low packet discard rate is an important performance criteria. On the other hand, admitting requests too liberally can easily lead to congestion and inability to satisfy the delay constraints of URLLC traffic. Achieving a fine balance between low discard rate, achieving high throughput, and satisfying the delay requirements of the supported traffic types. In this regard, the REDAL algorithm can provide differentiated services according to the requirements of different user requests, allocate resources and control admission according to different delay constraints, thus effectively avoid network congestion and reduce packet discard rate, while satisfying the delay constraint requirements.

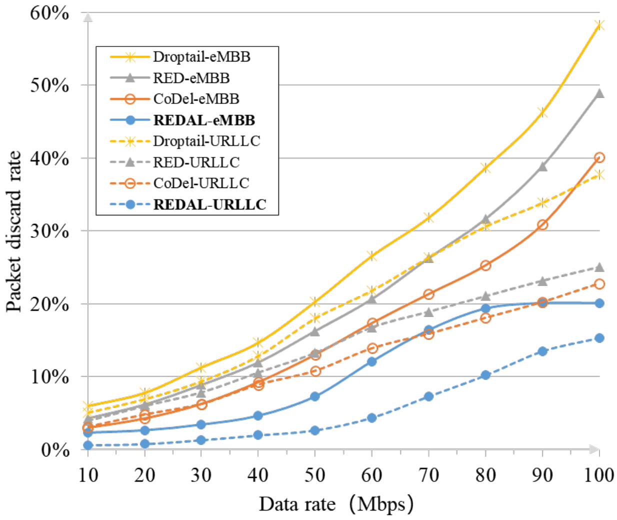

Fig. 9. Packet Discard Rate Comparison

VI. Conclusions

In this paper, the RED-based Admission Control with Latency Considerations (REDAL) algorithm is proposed which aims to solve the resource competition among different service types in 5G, and in particular, explicitly consider the latency requirements for service quality. Building on the well-tested RED algorithm for Internet routing, REDAL first classifies the users' requests, calculates whether their respective delay constraints can be met, and then come to the admission decision that can maximize the overall performance of the system. The validation based on a 5G MEC scenario shows that REDAL can more effectively reduce the probability of network congestion, reduce the end-to-end delay of users, improve the throughput of the system and reduce the packet discard rate of the system while meeting the QoS requirements of users. As our ongoing and future work, we will include the other 5G traffic scenario, mMTC, in the admission control decision as well as study more diverse bandwidth allocation scenarios.

References

[1] ITU, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," International Telecommunication Union (ITU), Report ITU-R M.2410-0, November 2017.

[2] W. K. G. Seah, C.-H. Lee, Y.-D. Lin, and Y.-C. Lai, "Combined Communication and Computing Resource Scheduling in Sliced 5G Multi-Access Edge Computing Systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 3144–3154, 2022.

[3] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89–103, 2015.

[4] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, 2017.

[5] X. Chen, W. Li, S. Lu, Z. Zhou, and X. Fu, "Efficient Resource Allocation for On-Demand Mobile-Edge Cloud Computing," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8769–8780, 2018.

[6] A. Slalmi, H. Chaibi, R. Saadane, and A. Chehri, "Call Admission Control Optimization in 5G in Downlink Single-Cell MISO System," in *Proc. of the 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)*, Szczecin, Poland, 8-10 Sep 2021, pp. 2502–2511.

[7] N. U. Ginige, K. B. Shashika Manosha, N. Rajatheva, and M. Latva-aho, "Admission Control in 5G Networks for the Coexistence of eMBB-URLLC Users," in *Proc. of the IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, Antwerp, Belgium, 25-28 May 2020, pp. 1–6.

[8] F. Mehmeti and T. F. La Porta, "Admission Control for URLLC Users in 5G Networks," in *Proc. of the 24th International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, Alicante, Spain, 22-26 Nov 2021, pp. 199–206.

[9] Y. Raaijmakers, S. Mandelli, and M. Doll, "Reinforcement learning for Admission Control in 5G Wireless Networks," in *Proc. of the IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, 7-11 December 2021, pp. 1–6.

[10] B. Bakhshi, J. Mangues-Bafalluy, and J. Baranda, "R-Learning-Based Admission Control for Service Federation in Multi-domain 5G Networks," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, 7-11 December 2021, pp. 1–6.

[11] A. Chagdali, S. E. Elayoubi, and A. M. Masucci, "Impact of Slice Function Placement on the Performance of URLLC with Redundant Coverage," in *Proc. of the 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Thessaloniki, Greece, 12-14 October 2020, pp. 1–6.

[12] Y. K. Tun, M. Alsenwi, N. H. Tran, Z. Han, C. S. Hong *et al.*, "Energy efficient communication and computation resource slicing for eMBB and URLLC coexistence in 5G and beyond," *IEEE Access*, vol. 8, pp. 136 024–136 035, 2020.

[13] Q. Shang, F. Liu, C. Feng, R. Zhang, and S. Zhao, "A BP Neural Network Based Punctured Scheduling Scheme Within Mini-slots for Joint URLLC and eMBB Traffic," in *Proc. of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Ottawa, ON, Canada, 11-14 November 2019, pp. 1–5.

[14] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 477–490, 2020.

[15] R. Sosa, C. Kiraly, and J. D. Parra Rodriguez, "Offloading Execution from Edge to Cloud: A Dynamic Node-RED Based Approach," in *Proc. of the IEEE International Conf. on Cloud Computing Technology and Science (CloudCom)*, Nicosia, Cyprus, 10-13 Dec 2018, pp. 149–152.

[16] A. H. Ismail, N. A. El-Bahnasawy, and H. F. A. Hamed, "AGCM: Active Queue Management-Based Green Cloud Model for Mobile Edge Computing," *Wireless Personal Communications*, vol. 105, no. 3, pp. 765–785, 2019.

[17] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, 1993.

[18] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S. A. Ashraf, J. Huschke, I. Aktas, and J. Ansari, "Wireless Communication for Factory Automation: an opportunity for LTE and 5G systems," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 36–43, 2016.

[19] P. Popovski, Č. Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[20] K. Nichols and V. Jacobson, "Controlling Queue Delay: A Modern AQM is Just One Piece of the Solution to Bufferbloat," *Queue*, vol. 10, no. 5, pp. 20–34, may 2012.

[21] F. E. Ritter, M. J. Schoelles, K. S. Quigley, and L. C. Klein, "Determining the number of model runs: Treating cognitive models as theories by not sampling their behavior," in *Human-in-the-loop simulations: Methods and practice*. Springer-Verlag, 2011, pp. 97–116.