IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 December 2023, accepted 4 January 2024, date of publication 26 January 2024, date of current version 5 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359028

RESEARCH ARTICLE

RETALIN: A Queue Aware Uplink Scheduling Scheme for Reducing Scheduling Signaling Overhead in 5G NR

VEERENDRA KUMAR GAUTAM, (Member, IEEE), AND BHEEMARJUNA REDDY TAMMA, (Senior Member, IEEE)
Indian Institute of Technology Hyderabad, Hyderabad 502284, India
Corresponding author: Veerendra Kumar Gautam (cs18resch01003@iith.ac.in)

ABSTRACT Flexible numerology in 5G New Radio (NR) helps to reduce End-to-End (E2E) delay by supporting different slot duration options. But, higher numerologies can increase signaling overhead (in terms of Scheduling Requests (SRs) in uplink (UL)) for applications which are UL heavy in nature; as a consequence, the E2E delay of such applications could be impacted. This sudden increase in SRs in UL happens when the timing of the Constant-Bit-Rate (CBR) UL transmission is not properly aligned with the slot duration of the numerology employed and aggressive emptying of the buffered data at the User Equipments (UEs) by the UL scheduler. In this work, to minimize SRs while meeting the latency requirements of the UL heavy applications like vehicular applications, we propose a novel UL MAC scheduling algorithm named *RETALIN*. The proposed *RETALIN* is a queue-aware radio resource scheduler that estimates the probability of SR with respect to each UE (vehicle) and anatomizes the transient queue behavior by controlling backlogs of the UEs in order to reduce SRs and thereby mitigating an adverse impact of numerology in the UL traffic while keeping a bound on E2E delay of the vehicular applications. Extensive NS-3 simulations are performed to evaluate the performance of the proposed *RETALIN* scheme with mobility traces taken from Simulation of Urban Mobility (SUMO) using OpenStreetMap. Simulation results show that the proposed *RETALIN* scheduling scheme significantly reduces link delay for different numerologies when compared with a state-of-art QoS scheduler. Further, *RETALIN* increases Packet Delivery Ratio (PDR) while reducing signaling overhead due to SRs for UL scheduling. In the case of a High Definition Map (HD Map) vehicular application, *RETALIN* assists in increasing the Offloading Success Rate (OSR) over the QoS scheduler.

INDEX TERMS 5G new radio (NR), vehicle to everything (V2X), network simulator 3 (NS-3), numerology, radio resource management, uplink scheduling.

I. INTRODUCTION

Emerging applications like Augmented Reality (AR), Virtual Reality (VR), High Definition Map (HD Map), and services with sporadic data patterns catapult the surge in uplink (UL) traffic in 5G NR [1]. As a matter of fact, UL traffic is gaining much more attention in 5G as compared to LTE [2]. UL traffic generated by these applications demands stringent Quality Of Service (QoS) guarantees in terms of End-to-End (E2E) delay and reliability. Moreover, mobility is a hot commodity in the 21^{st} century, ranging from high-speed vehicles to hyperloop, which causes an external perturbation to the UL traffic. In 5G NR Vehicle-to-Network (V2N) use cases, E2E delay is an important key performance indicator that dictates the QoS of the applications, and therefore it is given utmost importance. Here, the MAC scheduler plays a crucial role by assigning Transport Block Size (TBS) to each User Equipment (UE) in UL. TBS calculation depends on the number of radio resource

The associate editor coordinating the review of this manuscript and approving it for publication was Tiago Cruz.

blocks (RBs) assigned, Modulation and Coding Schemes (MCS), and the numerology employed.

In the realm of MAC scheduling, not much attention is given to UL scheduling in 3G/LTE as most of the traffic is in downlink (DL), which is not the case in 5G and beyond networks [3]. Hence, challenges in the UL scheduling prevail as unprecedented UL-centric traffic is expected from Vehicle to Everything (V2X) and Internet of Things (IoT) applications. Besides, scheduling parameters for UL, like Channel Quality Indicator (CQI), buffer occupancy level at UEs, and others, are locally absent at the gNodeB for efficient allocation of UL radio resources in a timely manner. This scheduling information is conveyed to gNodeB via signaling messages. As an example, a Scheduling Request (SR) is sent to gNodeB as a request for Scheduling Grant (SG) in UL; after getting an SG from gNodeB, Buffer Status Report (BSR) messages are piggybacked with UL data transmission to indicate the current buffer level of the UE,[1] and these BSR messages can be sent to gNodeB mainly to indicate UE's transmission requirements [4], [5]. The SR, BSR, and other scheduling information become the inputs for efficient scheduling of radio resources in UL by the gNodeB.

To reduce E2E delay, 3GPP introduced many technologies like flexible numerology, Massive MIMO, Bandwidth Parts (BWPs), service multiplexing, and mini-slotting in 5G NR. Specifically, numerology (μ) shortens the duration of the Orthogonal Frequency Division Multiplexing (OFDM) symbol; consequently, Transmission Time Interval (TTI) reduces, but SubCarrier Spacing (SCS) increases by a factor of 2^{μ}, $\mu \in \{0, 1, 2, 3, 4\}$, with respect to legacy 15 kHz SCS in case of numerology 0. Numerology with different SCS and cyclic prefixes can be viewed as an incontrovertible solution for reducing wireless communication latency. However, numerology creates a predicament for Constant-Bit-Rate (CBR) traffic in UL which is characterized by a fixed packet size and a fixed Inter-Packet Arrival Time (IPAT). Here, the E2E delay could increase due to increase in SR requests, which may be triggered due to varying processing times and decoding latency associated with higher numerologies. Further, if the CBR UL traffic timing is not in synchronization with the slot duration of the numerology used and the UL scheduler empties the Radio Link Control (RLC) buffer, as a consequence, SR could be generated. In addition, over-dimensional TBS (i.e., payload passed from MAC layer to PHY layer) resources cannot be used with higher numerologies due to a reduction in slot time [6]. So, we propose a queue-aware algorithm for the UL scheduling in 5G NR named *RETALIN*, which considers the probability of SR in the next TTI to decide the allocation of radio resources for the solicited BSR messages from the UEs. The algorithm dynamically controls the UEs' UL transmission rates to achieve a higher Packet Delivery Ratio (PDR) by reducing the percentage of SRs in the network, while meeting QoS requirements of the V2N applications. The key contributions of this work are as follows:

- We propose a Queue-aware radio resource scheduling scheme in UL named *RETALIN* that employs a threshold-based rule for managing backlogs in the RLC queues of the vehicles, thereby controls the UL transmissions in a way to reduce the higher signaling overhead associated with higher numerologies and high speed vehicles so that the E2E delay of V2N applications is maintained within the acceptable delay budget of the respective applications.
- We conduct extensive simulations using 5G-LENA module with mobility traces taken from SUMO using OpenStreetMap to evaluate the performance of the proposed *RETALIN* for numerologies 1 and 2. Simulation results show that *RETALIN* outperforms legacy scheduling schemes in terms of the E2E delay. *RETALIN* reduces RLC delay, consequently reducing the overall E2E delay when compared to a state-of-art QoS scheduler across both numerologies. *RETALIN* also increases PDR for numerology 2 by decreasing the percentage of SRs per packet in a vehicular environment.
- Evaluation of *RETALIN* jointly with *OCTANE* [7] (our previous work on task offloading to a MEC system) in case of HD Map application shows a further increase in the performance of HD Map application.

The rest of the paper is structured as follows: Section II briefly explains 5G NR, scheduling timings, RRC, UL grant-based handshake, and motivation for this work. Section III presents the related work in this area. Section IV presents the proposed system model, traffic model and probability of SR calculation. Section V presents the proposed *RETALIN* scheme and modified proportional metric. In Section VI, we describe the simulation setup and present performance results. Further, in Section VII, we take an HD map use case and show how *RETALIN* helps improving task offloading performance. Finally, we conclude the work in Section VIII with some future directions.

II. BACKGROUND

This section provides the necessary background on 5G NR on scheduling timings, RRC, and UL grant-based scheduling, followed by motivation which highlights the importance of reducing scheduling requests from the UEs for achieving lower E2E delay.

A. 5G NR

5G NR supports two models for the purpose of UL transmissions: OFDM (i.e., to achieve high throughput efficiency) and Direct Fourier Transform spread OFDM (DFT-s-OFDM) (i.e., to minimize the Peak-to-Average-Power-Ratio (PAPR)). 5G NR operates under multiple spectrum access paradigms. There are two Frequency Ranges (FR) that are stated: i) sub-6 GHz (FR1: 450 MHz to 6 GHz) and ii) millimeter wave (FR2: 24.25 GHz to 52.6 GHz) [8]. The maximum available

[1]Throughout this paper we use vehicles and UEs interchangeably.

bandwidths in FR1 and FR2 are significantly higher than that in LTE, that is 100 MHz and 400 MHz, respectively. As the main design principle, 5G NR redefines SubCarrier Spacing (SCS) as compared to LTE, where it was fixed to 15 KHz, giving birth to the concept of numerologies. Numerology (μ) can vary from 0 to 4, where each numerology has an SCS of $15 \times 2^{\mu}$ KHz, which shortens OFDM symbol and slot length as given in Table 1. The essence of numerology lies in the physical area occupied by RBs in a time-frequency grid which is unity despite the scaling factor of SCS *i.e.,* if it is reduced in one domain, it is compensated in another domain of the time-frequency grid. Herein, the number of OFDM symbols per slot and the number of subcarriers per RB is set to 14 and 12, respectively, unchanged across numerologies [9]. In contrast to the previous cellular technologies, different numerologies can be selected adaptively based on the carrier frequency, use case, traffic type, deployment scenario, target latency, and throughput requirements of the applications.

TABLE 1. Characteristics of 5G NR numerologies.

	$\mu = 0$	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$
SCS [kHz]	15	30	60	120	240
OFDM symbol length [us]	66.67	33.33	16.67	8.33	4.17
CP length [us]	~ 4.8	~ 2.4	~ 1.2	~ 0.6	~ 0.3
Subframes in a frame	10	10	10	10	10
Slots in a subframe	1	2	4	8	16
Slot length [us]	1000	500	250	125	62.5
OFDM symbols in a slot	14	14	14	14	14
Subcarriers in a PRB	12	12	12	12	12
PRB width [MHz]	0.18	0.36	0.72	1.44	2.88

B. SCHEDULING TIMINGS AND PROCESSING DELAYS IN 5G NR

In this section, we present different scheduling timers and processing delays incurred during scheduling procedure in 5G NR.

- K0 timer: It is used in 5G NR for scheduling of PDCCH (Physical Downlink Control Channel) and PDSCH (Physical Downlink Shared Channel) transmissions. It represents the minimum amount of time (defined in terms of *SlotOffset*) between when a UE receives a PDCCH allocation which carries Downlink Control Information (DCI) and when it should start decoding the corresponding PDSCH for the DL data. The K0 timer is started when a UE receives a DCI on the PDCCH and is used to synchronize the UE's reception of the PDSCH with the expected timing indicated by the DCI [10].
- K1 timer: It represents the amount of time between when a DL data transmission happens over PDSCH to the corresponding ACK/NACK transmission using Hybrid Automatic Repeat Request (HARQ) mechanism over Physical Uplink Control Channel (PUCCH) [11].
- K2 timer: This timer is used for scheduling PUSCH (Physical Uplink Shared Channel) transmission after receiving a UL grant over PDCCH. It is started when a UE receives an UL grant on PDCCH and starts transmitting on PUSCH. In other words, K2 is the number of slots given by the gNodeB to a UE for decoding the UL grant and transmitting UL data over PUSCH in the indicated scheduling opportunity. Here, the gNodeB conveys K2 (i.e., 0 to 32 slots [12]), mapping type, symbol start, and length of the UL scheduled transmission [10].
- N1 and N2: These represent processing delays associated during scheduling in 5G NR. N1 is the number of OFDM symbols required to process the DL data received over PDSCH and then start transmitting ACK/NACK over PUCCH. On the other hand, N2 is the number of OFDM symbols required to process the DL control info received over the PDCCH and then start transmitting the corresponding UL data over the PUSCH. N1 and N2 values change with different configurations of numerologies and UE capabilities [13], [14]. UE communicates N1 and N2 values to the associated gNodeB so that it could set the values for timers like K2 (e.g., K2 should be greater than or equal to N2).
- *L2L1Processing*: The time interval required for the gNodeB PHY/MAC layers to encode control and/or data channels is known as the encoding delay. Specifically, it represents the delay between the MAC layer's acquisition of control/data from the RLC layer and when the control/data becomes available for transmission over the air [6].
- *decodeLatency*: The time delay in which data acquisition happens from the air by the PHY layer, and the data block is available for processing at the MAC layer. In case of UL, *decodeLatency* is incurred at the gNodeB.

C. RADIO RESOURCE CONTROL

The RRC layer is responsible for connection establishment and release procedure between UE and gNodeB. Here, RRC is a piece of state machinery related to a cascade of L1/L2 control signaling messages exchanged between UEs synchronized with gNodeB using PUCCH/PDCCH. In addition, RRC signaling messages offer a solution to handle system access, interference coordination, energy saving, and mobility between a UE and gNodeB using measurement reports and resource allocation (e.g., SR and BSR control messages). Further, the RRC layer controls the scheduling of UL data of UEs by configuring the logical channel prioritization by assigning priority to each logical channel according to the mapping of service classes, thereby assisting the MAC layer in applying multiplexing and assembly procedures. In LTE, RRC has `RRC_IDLE` and `RRC_CONNECTED` states; using these two states, the UE maintains an active connection with the serving cell (i.e., eNodeB) to receive and transmit data by sending control plane signals. In the `RRC_CONNECTED` state, the context between the UE and eNodeB has been established. To do that, the UE informs eNodeB through signaling messages to initiate `RRC_CONNECTED` state, where control plane latency is incurred. Here, the UE remains in the `RRC_CONNECTED` state when the UE has data to send or

receive; otherwise, UE changes its state to RRC_IDLE. State transition also requires signaling overhead, introducing an additional delay for extra messages. In the RRC_IDLE state, upon receiving the paging message, the UE changes its state to RRC_CONNECTED for transmitting UL data to the eNodeB.

It is worth mentioning that 5G NR introduces a new independent RRC state named RRC_INACTIVE, complementing the existing states, RRC_CONNECTED and RRC_IDLE, to support diverse requirements of services in terms of power consumption, accessibility delays by means of flexible discontinuous reception configurations to assist applications like sensors, social network notification, and VoIP applications, which have small data payload size. Here RRC_INACTIVE state maintains the connection with gNodeB by storing the access stratum context when there is no UL traffic to send. UEs in the RRC_INACTIVE state can quickly switch back to RRC_CONNECTED with less control signaling overhead and avoid additional signaling delay as compared to switching back from RRC_IDLE to RRC_CONNECTED, which has high state transmission latency and more signaling overhead. RRC state diagram is shown in Fig. 1. The gNodeB configures an inactivity timer, referred to as T_{in}, for each UE in the RRC_CONNECTED state. If the gNodeB detects a UE that has not transmitted or received any packets within the duration of T_{in}, it will release the connection and transition the UE's state to the RRC_INACTIVE.

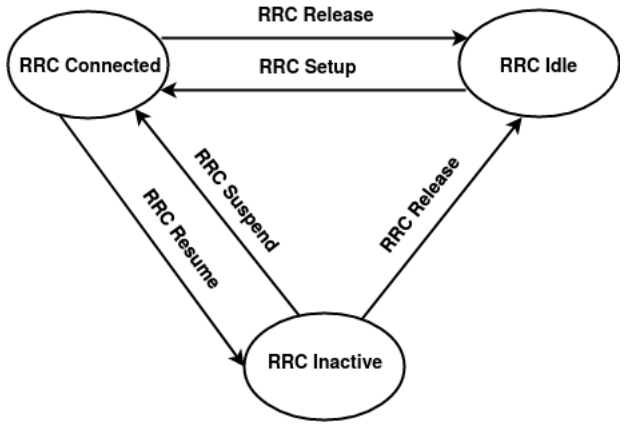


FIGURE 1. 5G NR RRC state transitions.

D. GRANT-BASED UL PROCEDURE

5G NR supports grant-based UL transmission for UEs connected to the gNodeB over the Uu interface, as illustrated in Fig. 2, which is called Mode-1 resource allocation. In this type of resource allocation, a UE seeks UL radio resources from the gNodeB by sending a SR request, which gets triggered when the UE has data in its RLC buffer and requires Scheduling Grant (SG) from the gNodeB to transmit the pending data. Here, the SR request is a flag transmitted to the gNodeB via PUCCH to perform a handshake with the gNodeB. In turn, the gNodeB sends a minimal UL grant to the UE over PDCCH, where PDCCH carries a DCI. The DCI contains MCS, RB allocation, and HARQ configuration information so that the UE uses them over PUSCH for its first UL data transmission. Along with the first UL transmission, the UE sends BSR, which contains a quantized value of the number of bytes pending in its Logical Channel Groups (LCGs). The gNodeB, in turn, sends SG messages to allocate an appropriate amount of UL radio resources (in terms of RBs) to the UE.

The E2E latency depends on packet size and TBS of the first UL scheduling assignment, which may lead to a 3-step process (SR $\rightarrow$ UL-grant $\rightarrow$ UL-data) or a 5-step process (SR $\rightarrow$ UL-grant $\rightarrow$ UL-data + BSR $\rightarrow$ UL-grant $\rightarrow$ UL-data), and the processing delays (i.e., $L2L1processing$ and $decodeLatency$) and K2 timer.

E. MOTIVATION

If the MAC scheduler assigns over-dimensional TBS, padding happens on MAC Protocol Data Unit (PDU), decreasing spectral efficiency. On the contrary, if the MAC scheduler assigns diminutive TBS, it results in segmentation of the RLC Service Data Unit (SDU), consequently increasing protocol header overhead and RLC signaling overhead [15]. The authors of [6] and [16] have shown how the E2E latency of UL traffic could be impacted negatively due to the interaction between the Inter-Packet Arrival Time (IPAT) of application traffic, scheduling timings, processing delays, and the slot length (which is numerology-dependent). This problem is partially solved by increasing the K2 timer in [16].

This predicament in UL happens due to the synergy between $decodeLatency$ of numerology, IPAT of the UL flow, and UL scheduling procedure, which may trigger extra SR messages that, in turn increase the latency as shown in Fig. 3 for higher IPAT traffic and in Fig. 4 for lower IPAT traffic in case of $\mu = 1$ with packet size $L = 1500$ bytes and $T_{in} = 6$ $slots$. Specifically, the RLC buffer becomes empty due to the high IPAT of two consecutive packets in a flow and if the allocated TBS to a UE is enough to empty the RLC buffer before the next packet arrives and T_{in} timer is expired. As a consequence, a UE requires an extra SR control message that is sent without being piggy-backed with the data. To explain why this is important, the UE must wait for a UL grant to transfer the upcoming packet, thereby increasing E2E delay of the packet. However, if IPAT of a flow is high, a UE can send data without incurring penalization of an extra SR message for a packet or using a short BSR message [4]. Moreover, the amount of RBs allocated by gNodeB MAC scheduler to the UE is discerning but oblivious concerning UE buffer level. Here, radio resources allocated to a UE is a linear function that governs the TBS for a UE. Hence, assigned RBs to a UE dictates the size of a TBS where TBS increases when the number of RBs assigned to a UE is more, and TBS also depends on numerology [10]. At the MAC layer, data sent in a given TBS depends on the packet

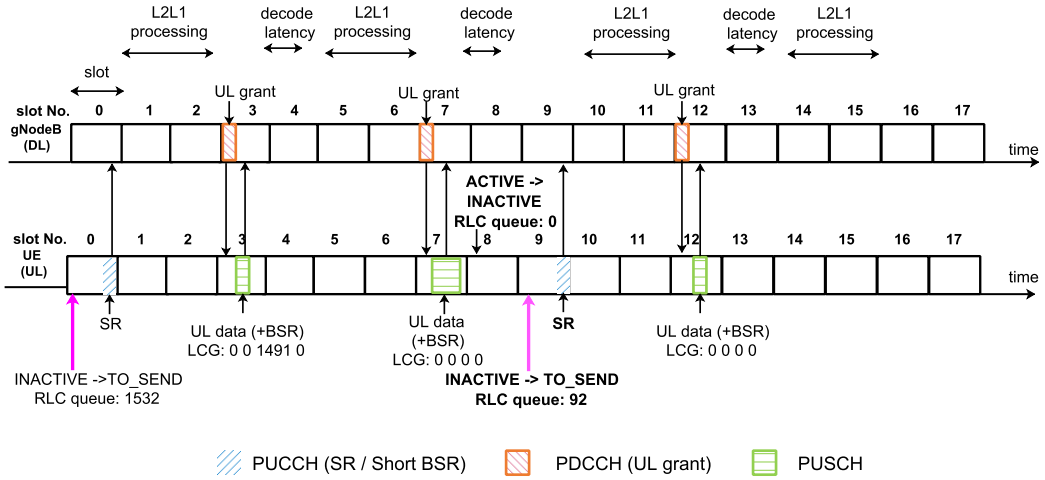


FIGURE 2. An example of grant-based UL procedure in case of $\mu = 0$ for transmitting one large packet of size $L = 1500$ bytes (with 32 bytes of header) followed by one small packet of size $L = 60$ bytes (with 32 bytes of header). The end-to-end (E2E) latency is higher in the case of 5-step process incurred for transmitting the large packet in UL.

size of a flow in a UE where the RLC queue is emptied according to RBs assigned to a UE. In this, if the IPAT of a flow is high at the RLC layer, control and signaling overhead (i.e., RLC AM Mode, periodic BSR) [16], [17], [18], [19], [20] are sent and amalgamated along with packets in the TBS. On the other hand, if the IPAT of a flow is low, then overheads may be sent alone, creating an additional SR request due to RRC timer timeout in which RRC state changes from `RRC_CONNECTED` to `RRC_INACTIVE` [12]. In addition, as slot time is reduced in higher numerology, UE may not use over-dimensional TBS due to the IPAT of the UL flow. With higher numerology, the probability of the packet arriving is low in a reduced slot time. Therefore the over-dimensional TBS left unused, unnecessarily increasing the E2E delay of the flow. These observations in UL traffic for higher numerology motivate us to design a radio resource scheduling algorithm to decrease the additional SR requests while efficiently using the over-dimensional TBS in high numerology.

III. RELATED WORK

Literature on the UL scheduling in cellular networks is scarce while a plethora of research concentrated on the downlink (DL) scheduling due to DL-heavy nature of Internet traffic [3], [21], [22], [23], [24]. In [25], the authors maximized the weighted sum rate in each UL scheduling interval but fairness among UEs was not considered. In [26], the authors proposed a low-complexity solution with fairness consideration to maximize the sum rate under individual rate and transmit power constraints for Orthogonal Frequency-Division Multiple Access (OFDMA) uplink. But these works did not consider buffer status information of the UEs, which helps to increase resource utilization efficiency while ensuring fairness among UEs [27]. The UL scheduling also needs to ensure subcarrier contiguity constraint i.e., contiguous allocation of subcarriers in the frequency domain [28], [29]. Since the contiguous subcarrier allocation is NP-Hard in nature, heuristic algorithms are proposed [30]. In this context, in [31], a queue-aware backlog-based polynomial time algorithm is proposed to maximize the average time throughput of eNodeB for UL transmission. In [32], a queue-aware uplink scheduling mechanism is employed to optimize resource utilization based on BSR messages for general traffic arrival at the UE. A state-of-art QoS-aware scheduler leveraging several parameters like the default priority level of the flow, the Proportional Fair (PF) metric, the Head-Of-Line (HOL) delay, and the Packet Delay Budget (PDB) has been proposed in [33]. Additionally, they introduced a delay budget factor (D) denoting the delay-aware weight correlated with the HOL delay and the PDB. This factor is computed as D = PDB / (PDB - HOL). Incorporating this factor allows the proposed QoS scheduler to address the low latency requirements across various traffic types effectively.

UL-centric traffic has increased in cellular networks starting from 5G NR due to introduction of diverse use cases like vehicular applications (e.g., AR, VR, HD Map). These emerging applications promulgate the need for equitable and efficient UL scheduling in 5G NR. Hence, in [34], the authors used deep learning prediction to assign radio resources for UL transmission in advance without SR and the granting process. However, the authors have not considered various traffic types and the affect of different numerologies in their study.

The introduction of different numerologies, some with extremely smaller scheduling intervals, poses challenges for radio resource scheduling as it needs to perform resource allocation with a time resolution of $\sim$ 100 μs. In [35], the authors designed a GPF+, an ultra-fast PF scheduler that decomposes the original scheduling problem into a large number of small and independent sub-problems and leverages accelerators to solve these sub-problems. In [36],

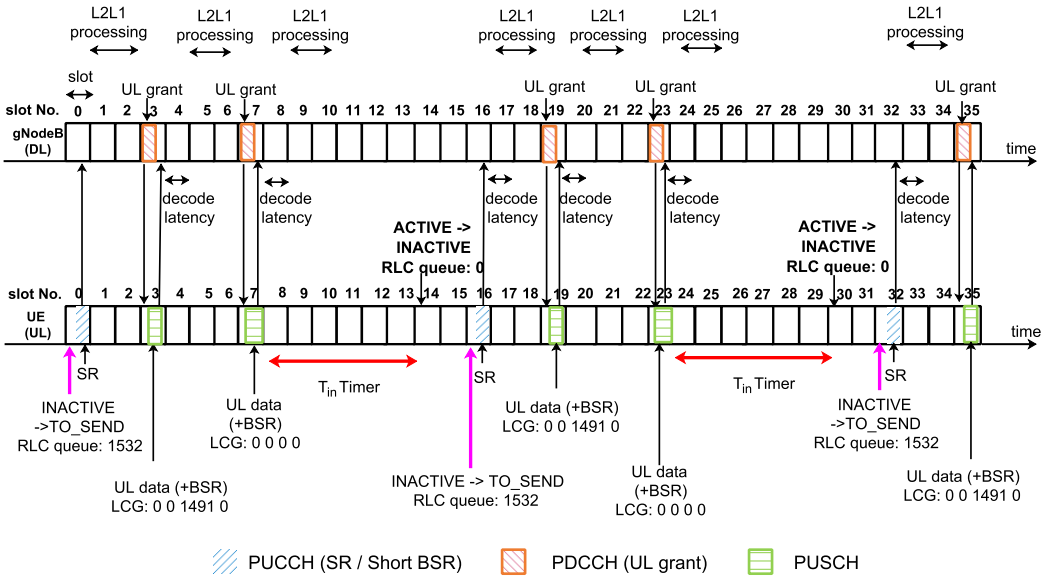


FIGURE 3. An example of grant-based UL procedure in case of $\mu = 1$ for packets of size $L = 1500$ bytes (with 32 bytes of header) with $IPAT = 16$ slots and $T_{in} = 6$ slots. UL grant-based access procedure, including RLC overhead, RLC states and processing delays for $\mu = 1$, $L = 1500$ bytes (32 bytes packet header) with K2 = 0. The E2E latency is also shown for both SR and BSR processes. Extra SR is generated with an increase in numerology from $\mu = 0$ to $\mu = 1$ because of high IPAT. The E2E delay of 1^{st} and 2^{nd} packets is the same: (8 slots + *decodeLatency* +*L2L1Processing*).

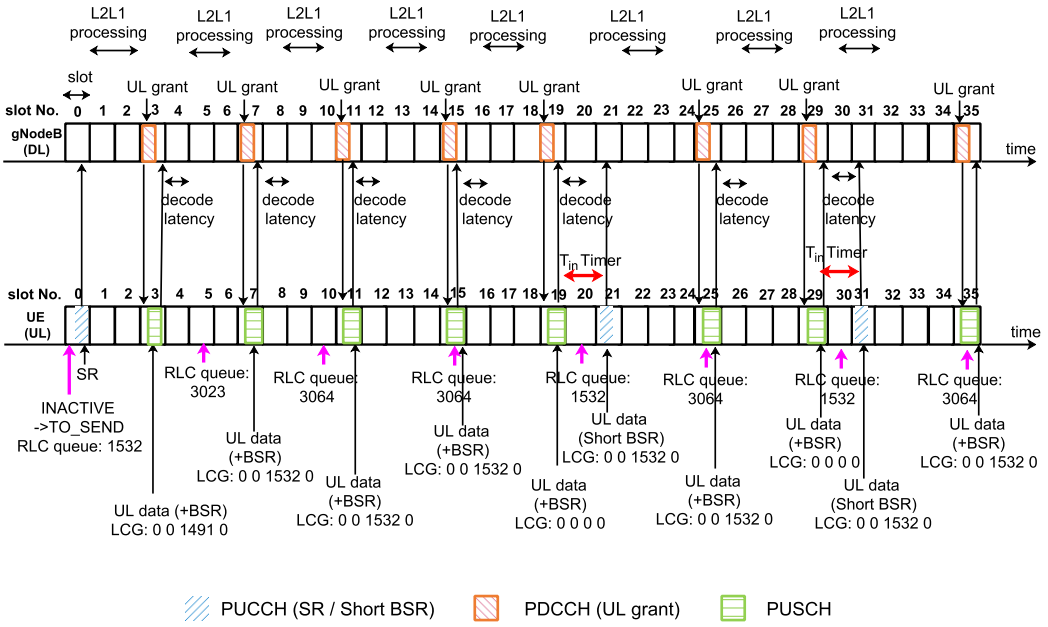


FIGURE 4. UL grant-based access procedure, including RLC overhead, RLC states and processing delays for $\mu = 1$, $L = 1500$ bytes (32 bytes packet header) with K2 = 0, E2E latency is also shown for both SR and BSR processes. Extra SR is not generated with an increase in numerology from $\mu = 0$ to $\mu = 1$ because of low IPAT. 1^{st} packet's E2E delay is (8 slots + *decodeLatency* +*L2L1Processing*). But the E2E delay of 2^{nd} packet has been reduced to (7 slots + *decodeLatency* +*L2L1Processing*).

the authors demonstrated the impact of the numerology on a sliced TDD radio access network that is multiplexed over both the sub-6 GHz and mmWave bands. It was shown that higher numerology schemes do not always translate into higher average spectral efficiency. In [6], the authors have shown a case where the E2E latency of a UL application could increase when a 5G NR network is configured with higher numerologies due to the nature of interaction between IPAT of the UL application, scheduling timings, processing delays, and the scheduling interval (which decreases for higher numerologies). It was shown that signaling overhead increased with an increase in numerology for a certain IPAT of the UL application due to rapid emptying of UE's buffer by frequent scheduling opportunities; consequently its E2E

TABLE 2. Notations.

Symbol	Description
$\mathcal{V}$	Set of UEs in a gNodeB with cardinality $\|\mathcal{V}\|= V$
T_s	Set of time slots with cardinality $\|T_s\|= T$ and duration $T_{tti}(\mu)$
$T_{tti}(\mu)$	Duration of a TTI for numerology μ
μ	5G NR numerology
$A_v[n]$	Arrival of requested bytes in n^{th} slot from v, $v \in \mathcal{V}$
$D_v[n]$	The number of bytes served in n^{th} slot at UE v, $v \in \mathcal{V}$
Λ_v	Mean arrival rate at v, $v \in \mathcal{V}$ in bytes/TTI
λ	Arrival rate for *M/M/1* queue at MAC layer of UE v, $v \in \mathcal{V}$ in bytes/TTI using BSR messages
μ_t	Mean service for *M/M/1* queue at MAC layer of UE v, $v \in \mathcal{V}$ bytes/TTI using SG messages
$P_{k,l}(T_{tti}(\mu))$	Probability of a *M/M/1* queue is in state l before been in state k for a duration $T_{tti}(\mu)$
ϵ	Violation probability threshold of a *M/M/1* queue at MAC layer of UE v, $v \in \mathcal{V}$
ϵ'	Violation probability of a *M/M/1* queue after giving RB to a UE v, $v \in \mathcal{V}$
q_{max}	Bound on the *M/M/1* queue length at the end of a TTI
$P_{v,NSR}(t)$	Probability of SR not arriving in a TTI at the MAC Layer of a UE v, $v \in \mathcal{V}$
P_{SR}	Probability of SR arriving in a TTI at the MAC Layer of a UE v, $v \in \mathcal{V}$
$P(x)$	Probability of a packet in a TTI at the MAC Layer of a UE v, $v \in \mathcal{V}$
T_{in}	RRC inactivity timer expiration of a UE v, $v \in \mathcal{V}$
T_{in}^{p}	Number of TTIs elapsed since the BSR message received at gNodeB from UE v, $v \in \mathcal{V}$
λ'	Arrival rate of user or control data at the MAC layer of a UE v, $v \in \mathcal{V}$
$PF_{v,r}(t)$	PF metric per UE v, $v \in \mathcal{V}$ per TTI for a RB r, $r \in \mathcal{R}$
α_v	Normalized backlog ratio per UE v, $v \in \mathcal{V}$ vary from 0 to 1
U_v	Set of utility metric per UE v, $v \in \mathcal{U}$
P_{NSR}^{th}	Probability threshold value of SR message
K_v	Set of RBs allocated to the UE v, $v \in \mathcal{V}$
L	Packet Size
η_v	Current MAC buffer size of UE v, $v \in \mathcal{V}$
$\mathcal{R}$	Set of RBs in a TTI
$S_{v,r}(t)$	Instantaneous data rate of a UE v, $v \in \mathcal{V}$ at RB r, $r \in \mathcal{R}$
SR_List	List of pending SRs at the gNodeB

latency increased. To tackle this problem, in [16] the K2 timer is increased from zero to keep the E2E latency under check. The authors of [37] considered an industrial scenario to show that higher numerology does not always help in reducing the E2E latency, particularly under Non-Line-of-Sight (NLOS) conditions. The authors of [38] proposed a utility-based analytical framework to choose an inactivity timer to reduce signaling overheads for mMTC traffic in 5G NR.

Although there exist numerous scheduling strategies to improve various performance metrics like throughput, fairness, latency, and resource utilization in cellular networks, to the best of our knowledge, none of them explicitly focused on controlling signaling overhead caused by higher numerologies for a certain type of UL traffic in order to keep E2E latency under check. Hence, in this work, we propose a novel radio resource allocation scheme for UL traffic in 5G NR named *RETALIN* that factors in fairness and buffer status of UEs for reducing signaling overhead in the network and thereby reducing the E2E latency of application traffic when higher numerologies are used to suit mobile use cases of 5G NR.

IV. SYSTEM MODEL

System model including traffic model is presented in this section. Afterwards, an analytical expression for the probability of a Scheduling Request (SR) not arriving within a scheduling interval (TTI) is derived. The notations used in this work are given in Table 2.

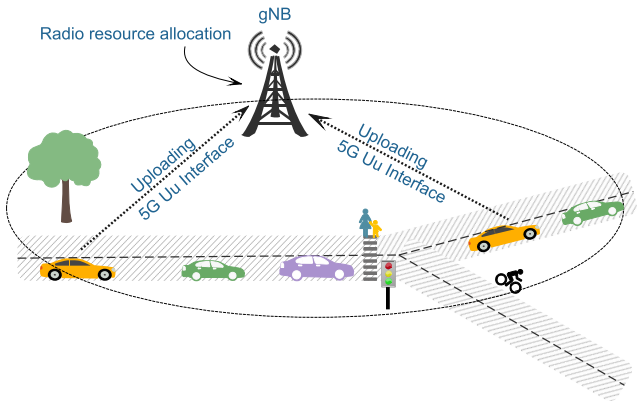


FIGURE 5. System model.

A. SYSTEM MODEL

We consider a typical scenario of V UEs under the coverage of a gNodeB in a highway environment as shown in Fig. 5. The set of UEs is denoted by $\mathcal{V} = \{1,\ldots,V\}$ and indexed by $v \in \mathcal{V}$.

Each UE generates UL-centric data periodically i.e., packets arrive according to IPAT configured for the respective application. Here, L denotes the packet size of the UE v and probability of x packets arriving in a TTI is given by $P(x)$. Since traffic at the MAC layer is a poisson process [39], [40], the arrival rate of the SR in a TTI is also a poisson process due to different slot timings for different numerologies [41], [42]. Let X be the random variable representing the number of packets at the MAC layer of UE v. X can be approximated as a poisson random variable with $\lambda' = \mathbb{E}[X]$. Here, λ' is the arrival rate of application traffic at the MAC layer of the UE and λ' is set according to traffic types [43], [44].

$$P[x=0] = \left[1 - \frac{e^{-\lambda'}(\lambda')^0}{0!}\right] \tag{1}$$

$$= \left[1 - e^{-\lambda'}\right] \tag{2}$$

In a time slot or TTI n, UL data arrived at v^{th} UE's MAC layer is denoted as $A_v[n]$, and the data served by the gNodeB is denoted as $D_v[n]$, and these two are assumed as random variables. The queue size $q_v[n]$ at v^{th} UE's MAC layer can evolve as:

$$q_v[n+1] = (q_v[n] + A_v[n] - D_v[n])^+ \tag{3}$$

where $(x)^+$ is defined as $max\{0, x\}$. The random variable $A_v[n]$ is assumed to be a poisson process with a mean value of Λ_v bytes/slot [40], [41], [45]. The departure $D_v[n]$ process can be defined as an exponential distribution which is a service given by the MAC scheduler, running at the gNodeB, to the UE by considering BSR reports, CQI, and other scheduling info from the UE. Therefore, the underlying queuing theory model for UL multiple access scheme forms a discrete time and discrete state *M/M/1* queue with the mean arrival rate of λ bytes per TTI and the mean departure rate of (i.e., service rate) of μ_t bytes per TTI.

Service rate of the UE is adjusted based on state k of the *M/M/1* UE queue such that the probability of the *M/M/1* is in a state higher than q_{max} after $T_{tti}(\mu)$ is bounded by ϵ i.e.,

$$\sum_{l=q_{max}+1}^{\infty} P_{k,l}(T_{tti}(\mu)) \leq \epsilon \tag{4}$$

The probability of the queuing system being in state l at time $T_{tti}(\mu)$, given that it was initially in state k, is denoted as $P_{k,l}(T_{tti}(\mu))$. This transient behaviour of *M/M/1* queuing system has been studied in [32], [46], and [47] which forms the following closed form expression:

$$P_{k,l}(T_{tti}(\mu)) = e^{-(\lambda+\mu)T_{tti}(\mu} \Big[\varrho^{\frac{l-k}{2}} \mathcal{I}_{l-k}(zT_{tti}(\mu)) + \varrho^{\frac{l-k-1}{2}} \mathcal{I}_{l+k+1}(zT_{tti}(\mu)) + (1-\varrho)\varrho^l \sum_{j=l+k+2}^{\infty} \varrho^{-\frac{j}{2}} \mathcal{I}_j(zT_{tti}(\mu))\Big] \tag{5}$$

By utilizing the modified Bessel function of the first kind $I_{(.)}(.)$ and $\varrho = \lambda/\mu_t$, $z = 2\mu_t\sqrt{\varrho}$, the MAC scheduler computes μ_t for the next TTI (i.e, given to the UE using SG). Here, the value of λ in bytes/TTI is determined by BSR messages, while the queue is in state k. External parameters are q_{max}, ϵ and $T_{tti}(\mu)$, which are set according to the numerology used.

B. PROBABILITY OF SR

In this section, we derive the mathematical expression for the probability of SR not arriving in a TTI, $P_{v,NSR}(t)$. So the probability of not generating a SR in a TTI is given by:

$$P_{v,NSR}(t) = [1 - P_{SR}] \tag{6}$$

where P_{SR} is the probability of generating a SR in a TTI. A UE does not generate any SR if a packet arrives in T_{in} at the MAC layer of the UE. T_{in}^p is the number of TTIs elapsed since the last BSR or SR message of the concerned UE.

$$P_{v,NSR}(t) = [P(x) + P(x)(1-P(x)) + P(x)(1-P(x))^2 + \ldots P(x)(1-P(x))^{T_{in}^p - 1}] \tag{7}$$

$$P_{v,NSR}(t) = P(x)\left[1 + (1-P(x)) + (1-P(x))^2 + \ldots (1-p(x))^{(T_{in}^p-1)}\right] \tag{8}$$

$$P_{v,NSR}(t) = P(x)\left[\frac{1(1-(1-P(x))^{T_{in}^p})}{1-1+P(x)}\right] \tag{9}$$

$$P_{v,NSR}(t) = \left[1 - [1-P(x)]^{T_{in}^p}\right] \tag{10}$$

Here, $P(x)$ is the probability of x packets arriving in a TTI. So, if SR should be generated then $x > 0$ and $P[x > 0] = [1 - P(x = 0)]$. By substituting (2) into (10), $P_{v,NSR}(t)$ is an equation where $P[x = 0]$ can be obtained.

$$P_{v,NSR}(t) = [1 - (e^{-\lambda'})^{T_{in}^p}] \tag{11}$$

V. RETALIN: A QUEUE AWARE UL SCHEDULING SCHEME FOR REDUCING SRS IN 5G NR

In this section, the proposed MAC layer scheduling algorithm named *RETALIN* is presented. The *RETALIN* uses a modified proportional fair metric that takes the probability of SR (derived in the previous section) and backlogs of UEs into account to decide the number of RBs to be assigned to different UEs for their UL transmissions in 5G NR.

A. MODIFIED PROPORTIONAL FAIR METRIC

The Proportional Fair (PF) scheduling is used to achieve a balance between maximizing the system's overall throughput and ensuring fairness among UEs by considering current channel conditions and past resource allocations to the UEs. The PF metric of UE v for a RB r in a TTI is denoted by

$PF_{v,r}(t)$ and it is given in Equation 12.

$$PF_{v,r}(t) = \left[\frac{T_v^{\alpha'}}{R_v^{\beta'}} \right] \tag{12}$$

R_v denotes the past average throughput of UE v and T_v denotes its instantaneous data rate. The parameters $0 \leq \alpha' \leq 1$ and $0 \leq \beta' \leq 1$ could be used to balance between throughput and fairness in the PF metric. The PF scheduling concerns about only throughput and fairness, hence it can result in poor E2E delay performance for latency sensitive applications like vehicular applications. To address the vehicular applications' E2E delay requirements, normalized backlog ratio can be taken into consideration, which is defined in Equation 13 [48].

$$\alpha_v = \left[\frac{\eta_v}{\sum_{v=1}^{|\mathcal{V}|} \eta_v} \right] \tag{13}$$

Normalized backlog ratio of a UE v (α_v) represents a drift in the UL buffer length of UE v as compared to the aggregated UL buffer length of all UEs in a TTI. α_v value varies from 0 to 1 for each UE. α_v will increase as the backlog of UE v increases due to bad channel conditions, thereby giving more weightage as compared to other UEs; as a result probability of getting more resources in a TTI will increase for that UE. The UEs contend for channel resources based on their normalized backlog ratios in order to minimize their respective backlogs and the total network backlog at the start of every TTI. This myopic rule maximizes network throughput and reduces the total network backlog, thereby helping latency sensitive applications meeting their strict E2E delay requirements.

The utility metric U_v is the proposed PF variant used in this work for UL scheduling at the MAC layer. The U_v has three different components as shown below.

$$U_v = \alpha_v \times P_{v,NSR}(t) \times PF_{v,r}(t) \tag{14}$$

B. RETALIN: A QUEUE AWARE UL SCHEDULING SCHEME FOR REDUCING SRS IN 5G NR

To assuage the predicament caused by numerology in the UL traffic, as aforementioned in the previous section, *RETALIN* − a queue-aware MAC scheduling algorithm − is put forward in this work. It ranks the UEs based on the proposed PF metric U_v shown in Equation (14) that assists in subsisting the generation of SRs in the system. Here, *RETALIN* policy is to facilitate latency sensitive communication services over 5G NR for opportunistic allocation of radio resources to the UEs by keeping the backlog at the UL MAC queue of the UE at q_{max}, thereby avoiding continuous connection terminations between UEs and their associated gNodeB. To maintain a balance between throughput and E2E delay, *RETALIN* clears q_{max} backlog according to the threshold value P_{NSR}^{th} of the UE afterward, updating the backlog queue information and $P_{v,NSR}(t)$ for each UE per TTI. T_{in}^{p} and $P_{v,NSR}(t)$ are reset per UE per TTI for every new arrival of BSR or SR messages.

The detailed procedure of *RETALIN* is given in Algorithm 1, and it has two phases:

Phase I (Lines 1 - 12): In the first phase, *RETALIN* clears the backlog q_{max} and the pending SRs from the *SR_List* to satisfy the QoS requirements and gives Scheduling Grants (SGs) to UEs so as to know the buffer lengths of UL MAC queues of the UEs through BSR messages. To do that, initially, the set of RBs allocated to a UE v, $v \in \mathcal{V}$ is empty, i.e., $K_v = \{\emptyset\}$, thereafter the algorithm iterates over all the UEs in $\mathcal{V}$ (Line 1), to check the condition: $P_{v,NSR}(t)$ is greater than P_{NSR}^{th} and η_v is less than q_{max} (i.e., it has checked for a new BSR message from UE v, then η_v will be greater than q_{max}) for each UE or any pending SR request is there for UE v (Line 2). Next, if the condition in Line 2 is true, then the *RETALIN* iterates over $\mathcal{R}$ RBs to clear the backlog η_v using the instantaneous service rate $S_{v,r}(t)$ on r over UE v (Line 5) and assigns the RB r to a set K_v, henceforth removing RB r from the set $\mathcal{R}$ (Line 6 - 7). Lastly, removing the UE v from the set $\mathcal{V}$ after clearing the backlog of UE v (Line 10).

Phase II (Lines 13 - 35): After the Phase I, *RETALIN* allocates remaining RBs in the set $\mathcal{R}$ to different UEs according to their U_v metric values and ensures that the violation probability ϵ' is less than a threshold value ϵ at every allocation of RB r to the UE's set, K_v. To accomplish that, *RETALIN* iterates over remaining $\mathcal{R}$ RBs until all RBs are exhausted, i.e., $\mathcal{R} = \{\emptyset\}$. First, *RETALIN* computes α_v and U_v for each UE where U_v comprises of product of three components: α_v, $P_{v,NSR}(t)$ and $PF_{v,r}(t)$ (Lines 15 - 18). Next, *RETALIN* sorts the UEs in accordance with U_v metric in descending order (Line 20). After that, *RETALIN* iterates over U_v to find current violation probability ϵ' of the UE UL MAC queue using the transient Equation 5 of the *M/M/1* queue. If the violation probability is less than the threshold violation probability ϵ, then *RETALIN* assigns the RB r to the UE v, thereafter reducing the queue level of the UE v (Lines 22 - 28). At last, *RETALIN* could not find any UEs with violation probability less than the threshold value ϵ. Thus, *RETALIN* will allocate the RB r to the UE having the highest value of U_v metric (Lines 29 - 33). For example, let us consider two UEs, A and B. UE A surpasses the 50 Bytes threshold for the queue size and has a violation probability above the specified threshold value. Conversely, UE B's queue size exceeds the 50 Bytes threshold but remains below the specified violation probability threshold. In this scenario, *RETALIN* assigns RB r to UE B since its violation probability is under the threshold value. This indicates that after a TTI, there is a good chance that the queue size will again surpass the threshold of 50 Bytes for the queue size, considering incoming UL data for UE B as its violation probability is under the threshold value.

C. TIME COMPLEXITY ANALYSIS

RETALIN scheme is executed for every TTI by the gNodeB. The UEs are assigned radio resources according to P_{NSR}^{th} in Phase I and U_v in Phase II. Hence, *RETALIN's* time complexity depends upon the number of UEs in the set $\mathcal{V}$

and the maximum number of RBs in the set $\mathcal{R}$. In the worst-case scenario, all UEs in $\mathcal{V}$ may be assigned some r, $r \in \mathcal{R}$ RBs in Phase I, thereby making Phase I's time complexity to $\mathcal{O}(|\mathcal{R}||\mathcal{V}|)$. Further, Phase II iterates over remaining RBs in the set $\mathcal{R}$ and sorting of U_v metrics has the time complexity of $\mathcal{O}(|\mathcal{V}| \times \log|\mathcal{V}|)$. The Phase II has a complexity of $\mathcal{O}(|\mathcal{R}||\mathcal{V}| \times \log|\mathcal{V}|)$. Thus, the overall time complexity of *RETALIN* consists of Phases I and II *i.e.,* $\mathcal{O}(|\mathcal{R}||\mathcal{V}| \times \log|\mathcal{V}|)$. However, with extremely short scheduling intervals at higher numerologies, radio resource scheduling faces challenges due to the need for rapid scheduling. To tackle this issue, we can leverage the capabilities of accelerators to perform scheduling decisions within the allowed time limits [35].

VI. SIMULATION SETUP AND PERFORMANCE RESULTS

In this section, we first describe the performance metrics, simulation setup and then present the performance results of the proposed *RETALIN* scheme by comparing it with a state-of-art QoS scheduler [33], baseline PF and RR scheduling algorithms.

A. PERFORMANCE METRICS

In order to evaluate the *RETALIN* scheduling strategies performance, we consider the following metrics:

- Packet Delivery Ratio (PDR): PDR is a performance metric used to evaluate the reliability of data transmission in a communication network. PDR represents the percentage of successfully delivered packets relative to the total number of packets sent.
- End-to-End delay: E2E delay refers to the total time taken for a packet to travel from the UE application to the gNodeB application in a communication network. E2E delay encompasses the time spent in all layers of the network, including the application layer, transport layer, network layer, and link layer. It also includes the time spent in various stages, such as queuing, processing, transmission, and propagation.
- RLC delay: The RLC packet delay is the duration between the moment a packet is generated at the RLC layer of a UE and when a packet is received at the RLC layer of the gNodeB.
- SR_P: SR_P is the percentage of SRs (e.g., control or signaling overhead) over the total number of packets exchanged in the network.
- BSR_{avg}: BSR_{avg} is the average number of BSR messages used to transmit a packet in a network.

Algorithm 1 RETALIN: A Queue Aware and SR Based Probabilistic Algorithm for UL Scheduling in 5G NR

input : $\mathcal{V}, \mathcal{R}, \eta, \epsilon, \mu, q_{max}, P_{NSR}(t), P^{th}_{NSR}, SR\_List$
output : K : Collection of sets of RBs assigned for various UEs.
Phase I:
forall $v \in \mathcal{V}$ **do**
1 **if** $(((P_{v,NSR}(t) \geq P^{th}_{NSR})$ && $(\eta_v \leq q_{max}))$ || $SR\_List_v! = \emptyset)$ **then**
2 **forall** $r \in \mathcal{R}$ **do**
3 **if** $(\eta_v > 0)$ **then**
// Clear backlog and process pending SR request from UE v
4 $\eta_v \leftarrow \eta_v - S_{v,r}(t)$
// Assign RB r to UE v
5 $K_v \leftarrow K_v \cup \{r\}$
6 $R \leftarrow R \setminus \{r\}$
// Remove UE v from the List $\mathcal{V}$
7 $\mathcal{V} \leftarrow \mathcal{V} \setminus \{v\}$
Phase II:
forall $r \in \mathcal{R}$ **do**
8 $\eta_v^{total} \leftarrow \sum_{v=1}^{|\mathcal{V}|} \eta_v$ **forall** $v \in \mathcal{V}$ **do**
9 $\alpha_v \leftarrow \left[\frac{\eta_v}{\eta_v^{total}}\right]$ // Normalization of UEs buffer values
10 $U_v \leftarrow \alpha_v \times P_{v,NSR}(t) \times PF_{v,r}(t)$ // Utility metric of UE v
11 $v^* \leftarrow arg\ \underset{v}{max}\ U$
// Sorting Utility values of UEs in descending order
$U \leftarrow \mathtt{Sort}\,(U, s.t. U_v > U_{v+1})$
// Number of UEs in U
$V' \leftarrow \mathtt{Cardinality}\,(U)$
12 **forall** $v \in U$ **do**
13 $l \leftarrow \eta_v - S_{v,r}(t)$
// Calculate the violation probability by using the current buffer status η_v of the UE
14 $\epsilon' \leftarrow \sum_{l=q_{max}+1}^{\infty} (P_{\eta_v})_{l(\mu)}$
// Check the violation probability threshold
15 **if** $(\epsilon' \geq \epsilon)$ **then**
16 $\eta_v \leftarrow \eta_v - S_{v,r}(t)$
// Reduce the buffer level
17 $K_v \leftarrow K_v \cup \{r\}$ // Assign RB r to UE v
18 break
// If all the UE are having low violation probability then assign RBs to the UE with the maximum U_v
19 **if** $(\mathtt{Index}(U_v) = V')$ **then**
20 $\eta_{v^*} \leftarrow \eta_{v^*} - S_{v^*,r}(t)\ K_{v^*} \leftarrow K_{v^*} \cup \{r\}$
21 break

B. SIMULATION SETUP

We consider a two-way highway scenario of Pembina Canada highway segment of 250 meters in length from the city of Winnipeg in Canada (refer Fig. 6). In this highway segment, vehicular traffic is generated by using the Rapid Cellular Network Simulation Framework (RACE) [49]. Here, the RACE framework internally uses Simulation of Urban Mobility (SUMO)[2] for customized vehicle traffic generation, and highway maps are exported using OpenStreetMap.[3] RACE uses the cellular infrastructure dataset provided by the

[2]http://www.sumo.dlr.de/userdoc/SUMO.html
[3]http://www.openstreetmap.org/

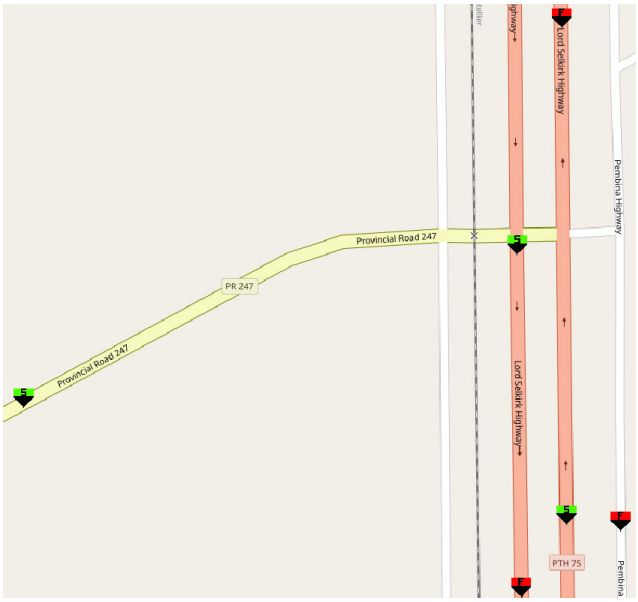


FIGURE 6. Road segment considered for simulation experiments.

Canadian organization of Innovation, Science and Economic Development (ISED),[4] which includes Canadian cellular providers like Telus, Rogers, and Bell. Here, in Fig. 6, 'S' represents the start location of vehicles and 'F' represent the finish location of vehicles. This segment is assumed to be served by a single 5G NR gNodeB. All simulation experiments are carried out over the NS-3 based 5G-LENA [50] module, which is used to realize 5G connections for the vehicles. Different simulation parameters are listed in Table 3 and set according to [6], [16], [32], [43], and [44].

In this simulated environment, each UE (vehicle) generates Constant Bit Rate (CBR) UDP traffic in the UL direction, communicating with a remote host linked to the Internet. 3GPP has established various 5QIs, encapsulating resource type, priority level, Packet Delay Budget (PDB), and more, outlined in [51]. These standardized 5QIs create a framework for services with shared characteristics, aiding in optimizing signaling based on standardized QoS attributes. All UEs are configured with the uniform 5QI value of 75 (V2X) in our experiments. However, IPAT and packet size are used to characterize UDP flows; a by-product of these two entities gives the UDP flow rate of a vehicle, which is varied across simulation experiments. Here, to show the adverse effect of 5G NR numerologies 1 and 2 on UL traffic, we measure the E2E packet delay for different packet sizes in a vehicular environment using RLC-AM (acknowledged mode). The L1L2 processing delay is numerology-dependent due to the gNodeB MAC scheduler working on a slot basis. The *decodeLatency* is set to a fixed value as decoding time is mainly related to CPU rate and available energy to perform the task. So this value is independent of the numerology used. To obtain statistical significance, the simulation experiments

[4]https://sms-sgs.ic.gc.ca/eic/site/sms-sgs-prod.nsf/eng/h_00010.html

TABLE 3. Simulation parameters.

Parameter	Value
Number of vehicles	30
Mobility model	Krauss
Average vehicle speed	20-80 $kmph$
5G NR gNodeB/UE TX power	46/23 dBm
5G NR gNodeB antenna pattern	Canadian dataset
5G NR gNodeB antena model	omni-directional
Vehicle antenna model	Isotropic
5G NR gNodeB antenna tilt	15°
5G NR gNodeB/Vehicle antenna height	25 meter/1.5 meter
Carrier frequency	5.9 GHz
Channel bandwidth	30 MHz
Channel model	UMa_LoS
5G NR numerologies (μ)	1, 2
5G NR MAC Scheduler	*RETALIN*, PF, QoS [33]
Bound on the *M/M/1* queue length (q_{max})	50 Bytes [32]
Violation probability threshold of a *M/M/1* queue (ϵ)	0.01 [32]
Arrival rate of the packet or control data (λ')	0.48 [44]
IPAT of UDP flow at UEs (v_{ipat})	2 $msec$ [16]
5G QoS Identifier (5QI)	75, GBR_V2X
P2P link Delay (S-gateway)	20 $msec$
Packet size (U) of UDP flow at vehicles	1000, 1400 Bytes [16]

are repeated 10 times and results are presented with 95% confidence intervals.

C. PERFORMANCE RESULTS

In this section, we demonstrate the effectiveness of *RETALIN* in terms of E2E delay, RLC delay (between vehicle's RLC layer to gNodeB's RLC layer), SR_p, and PDR for two numerologies (i.e., $\mu = 1$ and $\mu = 2$) in different vehicular traffic conditions. The simulation scenario is designed to answer the query like what is the impact of mobility on numerology for different packet sizes for a set of vehicles. We also compare *RETALIN* with different MAC schedulers and show how *RETALIN* reduces E2E delay and increases PDR of vehicles.

1) EFFECT OF P^{th}_{NSR} ON E2E DELAY AND PDR FOR UL TRAFFIC

RETALIN clears q_{max} of a vehicle based on the value of P^{th}_{NSR}, it implies that the vehicle's T^{p}_{in} is large. Clearing the backlog (q_{max}) means emptying the MAC buffer of the concerned UE. However, if a newly arriving UL packet (i.e., data or control packet) causes an extra SR because the T_{in} timer expires and changes the RRC state of the UE to `RRC_INACTIVE`, it can lead to potential issues. To avoid this, RETALIN prioritizes clearing q_{max} of vehicles with low values of P^{th}_{NSR}. Vehicles with low values of P^{th}_{NSR} have lower T^{p}_{in}, resulting in lower probability $P(x)$ of their MAC layer generating a new SR when T_{in} expires. By focusing on these

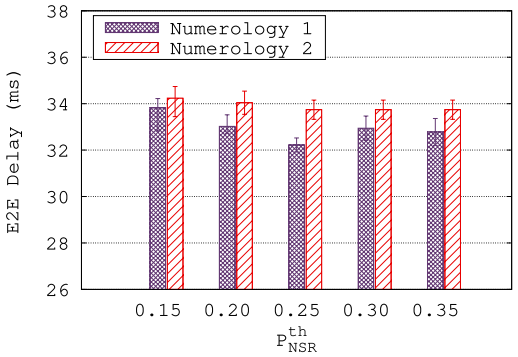


FIGURE 7. E2E delay Vs P^{th}_{NSR} for $|V| = 30$ with $\mathcal{V}_{speed} = 60kmph$ where $L = 1400$ bytes.

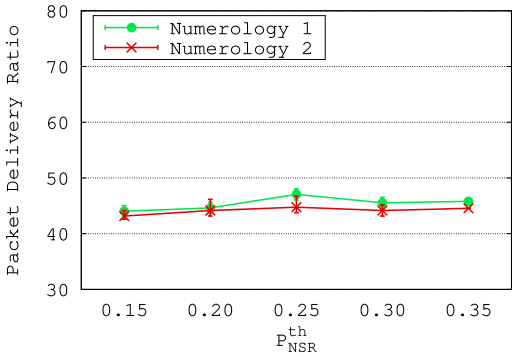


FIGURE 8. PDR Vs P^{th}_{NSR} for $|V| = 30$ with $\mathcal{V}_{speed} = 60kmph$ where $L = 1400$ bytes.

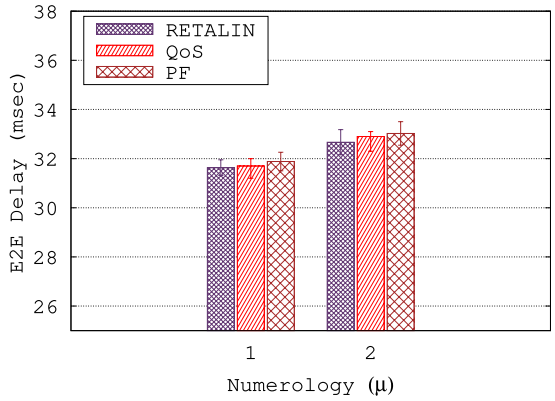
vehicles, RETALIN minimizes the likelihood of generating unnecessary SR messages. In both the cases, E2E delay and PDR are low as shown in Figs. 7 and 8. For $P_{NSR} = 0.25$, PDR is highest and E2E delay is lowest. P^{th}_{NSR} plays a dominant role in determining whether to clear q_{max} of a vehicle for every TTI or not. Tuning P^{th}_{NSR} of vehicles can help to meet their QoS requirements, and using P^{th}_{NSR}, the service rate per vehicle can be adjusted. For the rest of experiments in this work, we set $P^{th}_{NSR} = 0.25$.

2) EFFECT OF NUMEROLOGY ON UL TRAFFIC

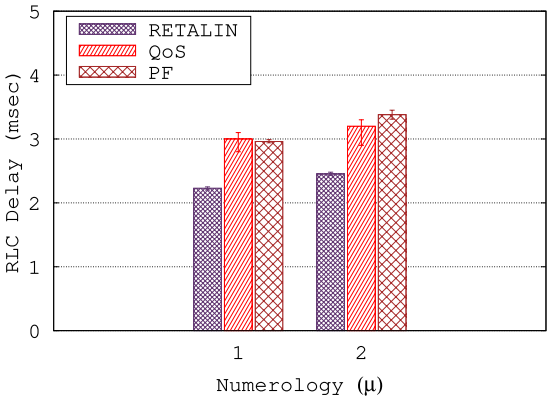
In Figs. 9a and 10a, E2E delays are plotted for PF, RETALIN and QoS scheduler [33] schemes for $\mathcal{V}_{speed} = 60kmph$ for $\mu = 1$ and $\mu = 2$ by setting packet sizes of UL traffic to $L = 1000$ bytes and $L = 1400$ bytes, respectively. Here, we can observe that with an increase in numerology from $\mu = 1$ to $\mu = 2$, the E2E delay increased from 31.882 *msec* to 33.024 *msec* in case of $L = 1000$ bytes and from 32.366 *msec* to 34.076 *msec* in case of $L = 1400$ bytes for PF. Further, we can observe that with an increase in numerology from $\mu = 1$ to $\mu = 2$, the E2E delay increased from 31.7 *msec* to 32.9 *msec* in case of $L = 1000$ bytes and from 32.3 *msec* to 33.8 *msec* in case of $L = 1400$ bytes for QoS scheme. E2E delay has seen increments of 0.42 *msec* and 1.5 *msec* for numerologies 1 and 2, respectively, when considering packet sizes of $L = 1000$ bytes. Similarly, for packet sizes of $L = 1400$ bytes, there were increments of 0.64 *msec* and 1.5 *msec* in the E2E delay for numerologies 1 and 2, respectively, in the case of PF and QoS schemes. In Figs. 9b and 10b, we have plotted the variation in RLC delay for numerologies 1 and 2 in case of $L = 1000$ bytes and $L = 1400$ bytes, respectively. The reason behind increase in RLC delay is the increased signaling overhead when we switch from $\mu = 1$ to $\mu = 2$ as shown in Figs. 9c and 10c. Here, the results indicate that increasing the numerology from $\mu = 1$ to $\mu = 2$ increases the signaling overhead for different packet sizes that in turn increases the RLC delay and thereby leads to an increase the E2E delay in case of PF and QoS schemes. On the other hand, as shown in Figs. 9d and 10d, BSR_{avg} decreases as we switch from $\mu = 1$ to $\mu = 2$ for both the packet sizes ($L = 1000$ bytes and $L = 1400$ bytes) due to transmission of more packets using SR messages than BSR messages in $\mu = 2$. This happens with an increase in numerology, and it negatively affects the E2E delay of UL traffic. The reduction in SR_p for $L = 1400$ as compared to $L = 1000$ is because of vehicles asking for more radio resources from gNodeB (through scheduling grants) in case of $L = 1400$ as compared to $L = 1000$. Due to this overall delay, the packet delay is increasing, but the IPAT remains the same in both cases. As a result, the subsequent UL packet is transmitted without an extra SR message and is transmitted with the BSR message in the case of $L = 1400$, as compared to $L = 1000$.

The E2E delay increases due to the way UL scheduling mechanism is designed in 5G NR where 3-step process (SR $\rightarrow$ UL-grant $\rightarrow$ UL-data) or a 5-step process (SR $\rightarrow$ UL-grant $\rightarrow$ UL-data + BSR $\rightarrow$ UL-grant $\rightarrow$ UL-data) is used; therefore rise in SR requests increases the E2E delay. For example, the theoretical delay of a packet will be *3 slots + 0.1 msec = 0.85 msec* for $\mu = 2$. A TBS lower than 850 bytes fits into one OFDM symbol with MCS=28, $\mu = 2$, and 100 MHz channel bandwidth. However, a packet size larger than 850 bytes requires the 5-step process and therefore its delay would be *6 slots + 0.1 msec = 1.6 msec*. On the other hand, if BSR is piggybacked with the previous UL MAC PDU, the delay comes down to 0.85 *msec* [16].

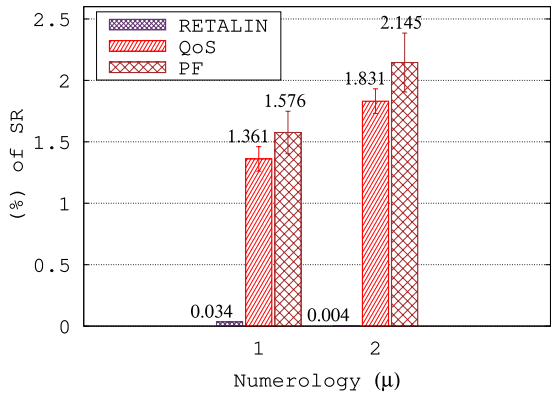
In contrast, *RETALIN* reduces SR_p as compared to PF and QoS schemes. Due to the reduction in SR_p, the average E2E delay for *RETALIN* comes down to 31.633 *msec* as compared to 100 *msec* in case of RR scheme (not plotted in the graph) and 31.882 *msec* in case of PF scheme and 31.7 *msec* in case of QoS scheme for $\mu = 1$ as shown in Fig. 9a. In case of $L = 1400$ bytes scenario, *RETALIN* has SR_p of 0.005% and 0.004% as compared to PF's 1.13% and 2.05% for $\mu = 1$ and $\mu = 2$, respectively. Due to reduction in SR_p, the RLC delay and the E2E delay decrease in *RETALIN* as compared to PF. In summary, the proposed *RETALIN* scheme is able to subsist SR_p in the network and achieve lower RLC delay and E2E delay by increasing average BSR per packet (BSR_{avg}) i.e., fragmentation of packets occurs more in *RETALIN* as compared to RR and PF schemes.



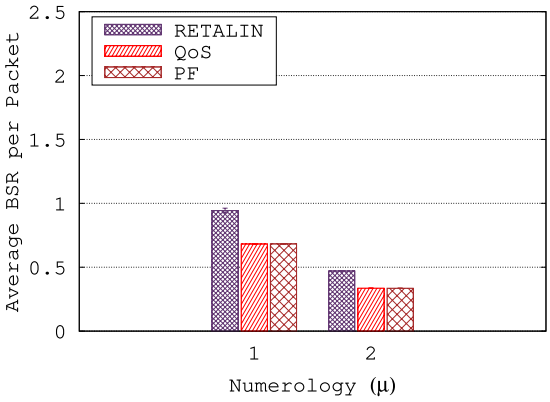
(a) E2E packet delay vs numerology.



(b) RLC delay vs numerology.



(c) (%) of SR (SR_p) vs numerology.



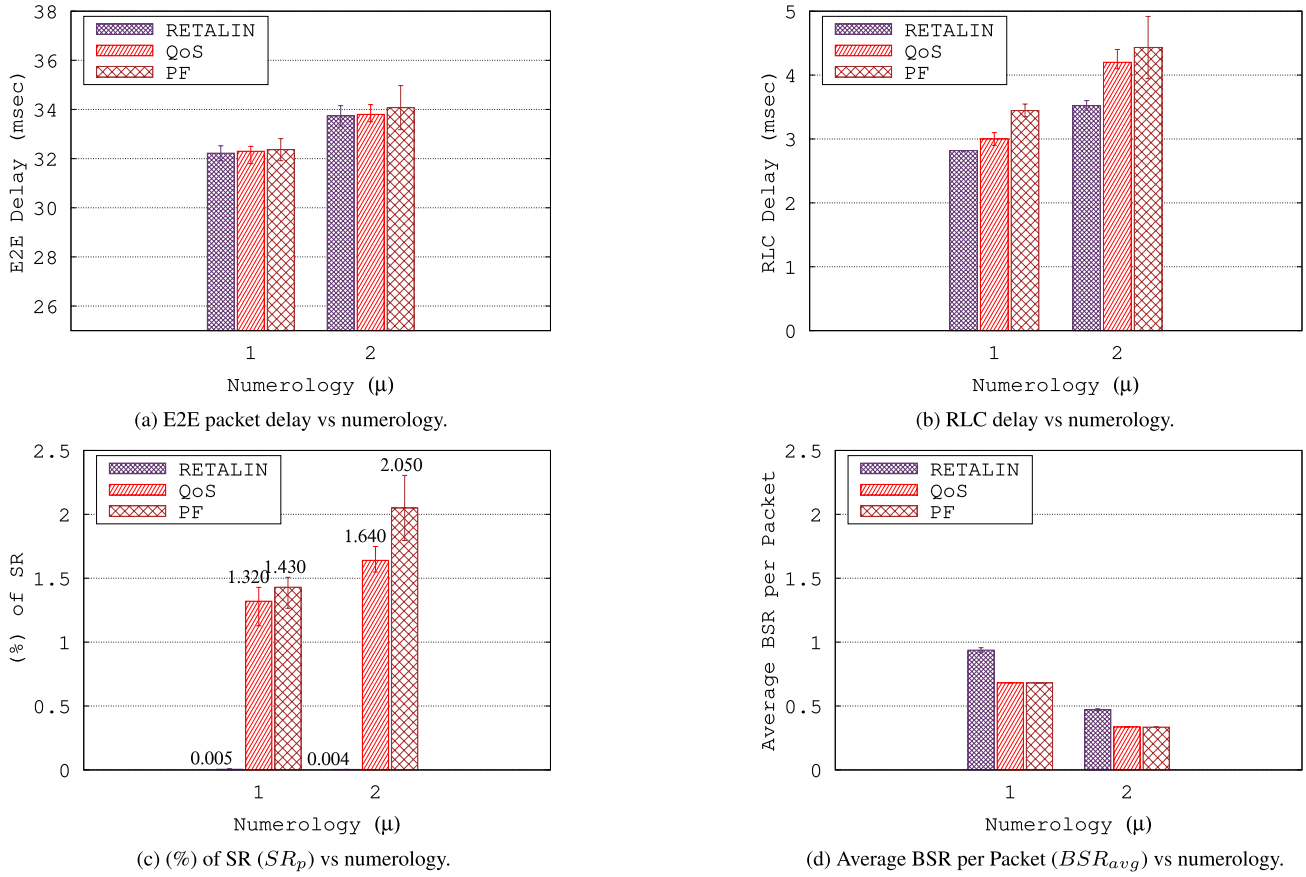
(d) Average BSR per Packet (BSR_{avg}) vs numerology.

FIGURE 9. Results observed in case of PF and RETALIN schemes by varying numerologies for UL packet size of $L = 1000$ bytes.

Since RR scheduling scheme allocates the radio resources blindly, without the knowledge of queue status and channel state information, we can observe that RR has the highest E2E delay. The PF scheduler outperforms RR as it considers the channel state information at the time of resource allocation. It guarantees proportional fairness among UEs (vehicles); without the knowledge of the backlog queue status of the vehicles. The PF scheduler may assign high-quality RBs to a vehicle that does not have much data to transmit, or it may assign surplus RBs to a vehicle that immediately empties its UL buffer and thereafter triggering an extra SR for transmission of upcoming control or data packet in the UL. Moreover, the QoS scheduler surpasses PF performance by considering Head-Of-Line (HOL) delays and Packet Delay Budget (PDB) with PF metric. This inclusion allows for more informed resource allocation decisions compared to PF, ensuring a higher level of service quality. Unlike these three schemes, *RETALIN* uses the transient equation of *M/M/1* queue to delay the transmission of packets in the UL and thereby prevents unnecessary SRs in the network. Also, a channel-aware, queue-aware, and P_{NSR} based metric is defined in Equation (14) to distribute RBs among the vehicles. By doing that, the overall E2E delay of packets in the network is reduced. In this way, *RETALIN* scheme saves resources by leaving a residual backlog (q_{max}) in the vehicle's buffer and these saved RBs are given to those vehicles with less probability of generating a SR in the TTI. Thereafter, backlogs (q_{max}) of vehicles are cleared according to P_{NSR} of vehicles violating P_{SR}^{th} at every TTI for maintaining a trade-off between throughput and delay. By doing this, we can use over-dimensional TBS and assist an upcoming control or data packet to be piggybacked with a BSR message without incurring the penalization of extra SR. In a nutshell, *RETALIN* dynamically controls the UL buffers of vehicles, overcomes the drawbacks of conventional schedulers like RR, PF, QoS scheduler and dampens the negative effect of higher numerology.

3) EFFECT OF SPEED OF VEHICLES ON UL TRAFFIC

To study the effect of speed of the vehicles on UL traffic, acceleration and speed parameters of the vehicles are varied in SUMO. The traces so obtained from SUMO are exported into NS-3 to mimic realistic movement of vehicles on the highway segment considered. In Figs. 11a and 11b, E2E delays are plotted by varying average speed of vehicles from $\mathcal{V}_{speed} = 20kmph$ to $\mathcal{V}_{speed} = 80kmph$ in case of $\mu = 1$ and $\mu = 2$ by setting the packet size of UL traffic to $L = 1000$ bytes. Here, we can observe that with an increase in numerology from $\mu = 1$ to $\mu = 2$, the E2E delays of *RETALIN* and PF increase from 31.75 *msec* to



(a) E2E packet delay vs numerology.

(b) RLC delay vs numerology.

(c) (%) of SR (SR_p) vs numerology.

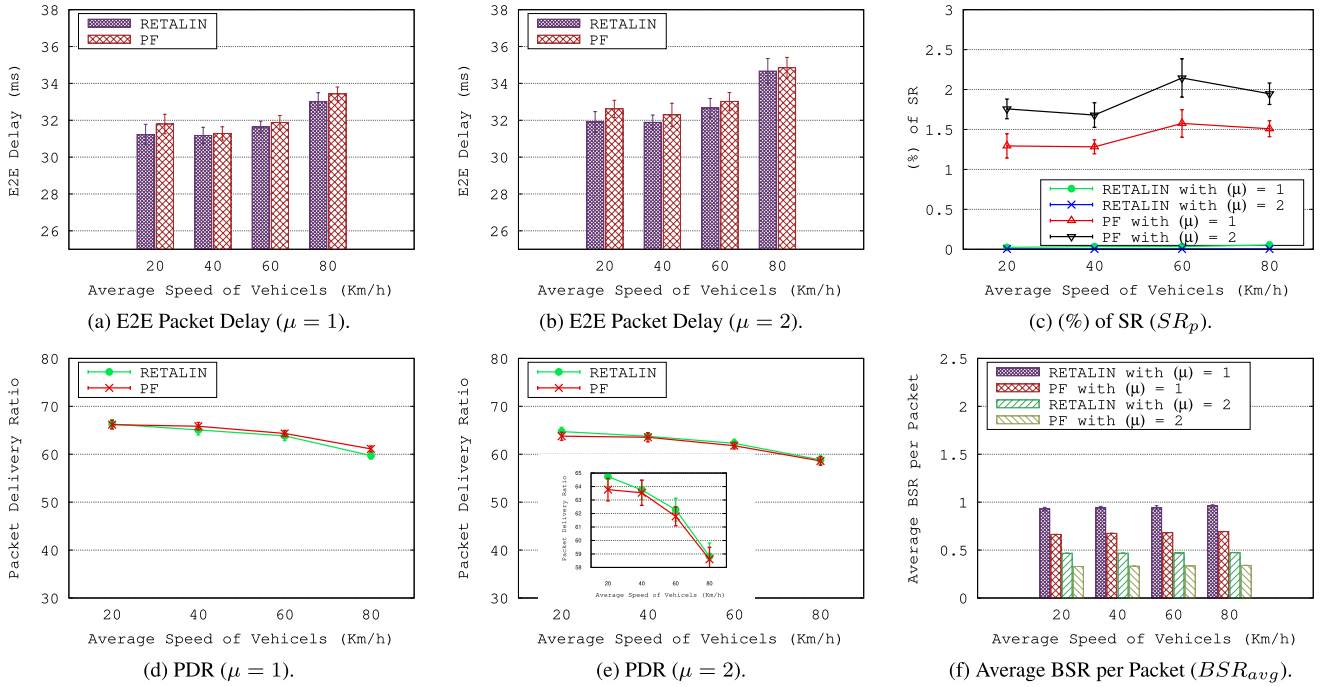
(d) Average BSR per Packet (BSR_{avg}) vs numerology.

FIGURE 10. Results observed in case of PF and RETALIN schemes by varying numerologies for UL packet size of $L = 1400$ bytes.

32.72 *msec* and from 32.13 *msec* to 33.19 *msec*, respectively. Further, as shown in Figs. 11d and 11e, PDR of vehicles for *RETALIN* and PF decreases from 63.72% to 62.38% and from 63.75% to 61.32% when we switch from $\mu = 1$ to $\mu = 2$, respectively. The reason is, as shown in Fig. 11c, itRETALIN reduces SR_p from 0.03% to 0.003% as compared to PF from 1.14% to 1.87% when we switch from $\mu = 1$ to $\mu = 2$, respectively. However, we can see that the average BSR per packet metric (BSR_{avg}) for *RETALIN* and PF reduces from 0.68 to 0.33 and from 0.68 to 0.33 when we switch from $\mu = 1$ to $\mu = 2$, respectively as shown in Fig. 11f. On the other hand, in Figs. 12a and 12b, the E2E delay is plotted by varying $\mathcal{V}_{speed} = 20kmph$ to $\mathcal{V}_{speed} = 80kmph$ in case of $\mu = 1$ and $\mu = 2$ by setting the packet size of UL traffic to $L = 1400$ bytes. Here, the E2E delay is higher than the previous scenario, while the PDR is lower.

Moreover, SR_p of *RETALIN* and PF reduces as shown in Fig. 12c. However, we can see that BSR_{avg} of *RETALIN* and PF remains almost the same for both the packet sizes as shown in Figs. 11f and 12f.

As shown in Figs. 11d, 11e, 12d, and 12e, the average speed of the vehicles negatively affects PDR of UL traffic for all the schemes under the study. With increase in the speed, channel estimation becomes difficult because the CQI reports become outdated very quickly. Impressively, policies of *RETALIN* adapt well with respect to E2E delay, as shown in Figs. 11a, and 11b, in comparison to PF and RR schemes when the speed of vehicles increases. Changing 5G NR numerology can influence the delay and throughput of the UL flows, thereby impacting PDR of the vehicles. We found that *RETALIN* in case of $\mu = 1$ performs better than $\mu = 2$ in terms of PDR. Subcarrier spacing increases with numerology; so the more distant the subcarriers are from each other, better they are protected from the varying channel conditions [52], [53]. On the contrary, increasing the subcarrier spacing slightly decreases the throughput of UL flows [54]. This happens because a different fraction of the bandwidth is utilized. For example, $\mu = 1$ uses 0.36 MHz PRB width, thereby utilizing 19.82 MHz, whereas $\mu = 2$ uses 0.72 MHz, thus utilizing 19.5 MHz bandwidth. On the other hand, for $\mu = 2$, as shown in Fig. 11e, SR_P rises with an increase in numerology with the speed of vehicles and hence PDR comes down. This happens due to a rise in SRs by which some spectrum is wasted in transmitting SRs, thereby causing deterioration in PDR achieved. The proposed *RETALIN* reduces SR_p and its effect is reflected in Fig. 11e as 1% increase in PDR as compared to PF. On the contrary, BSR_{avg} is increased for *RETALIN* as shown in Figs. 11f and 12f. Since it is a single-bit message,



(a) E2E Packet Delay ($\mu = 1$). (b) E2E Packet Delay ($\mu = 2$). (c) (%) of SR (SR_p).

(d) PDR ($\mu = 1$). (e) PDR ($\mu = 2$). (f) Average BSR per Packet (BSR_{avg}).

FIGURE 11. Result observed in case of *RETALIN* and PF by varying speed of vehicles for UL packet size of $L = 1000$ bytes.

it does not affect PDR of the vehicles. In Figs. 11c and 12c, we can see a spike in SR_p between $\mathcal{V}_{speed} = 40kmph$ to $\mathcal{V}_{speed} = 60kmph$ due to increase in packet loss with mobility. However, from $\mathcal{V}_{speed} = 60kmph$ to $\mathcal{V}_{speed} = 80kmph$, a drop can be observed in SR_p due to loss of SR messages due to low PDR of the vehicles.

VII. HD MAP: A USE CASE

The proposed *RETALIN* scheme is compared against PF and QoS by considering a HD map use case of vehicles in this section.

A. HD MAP

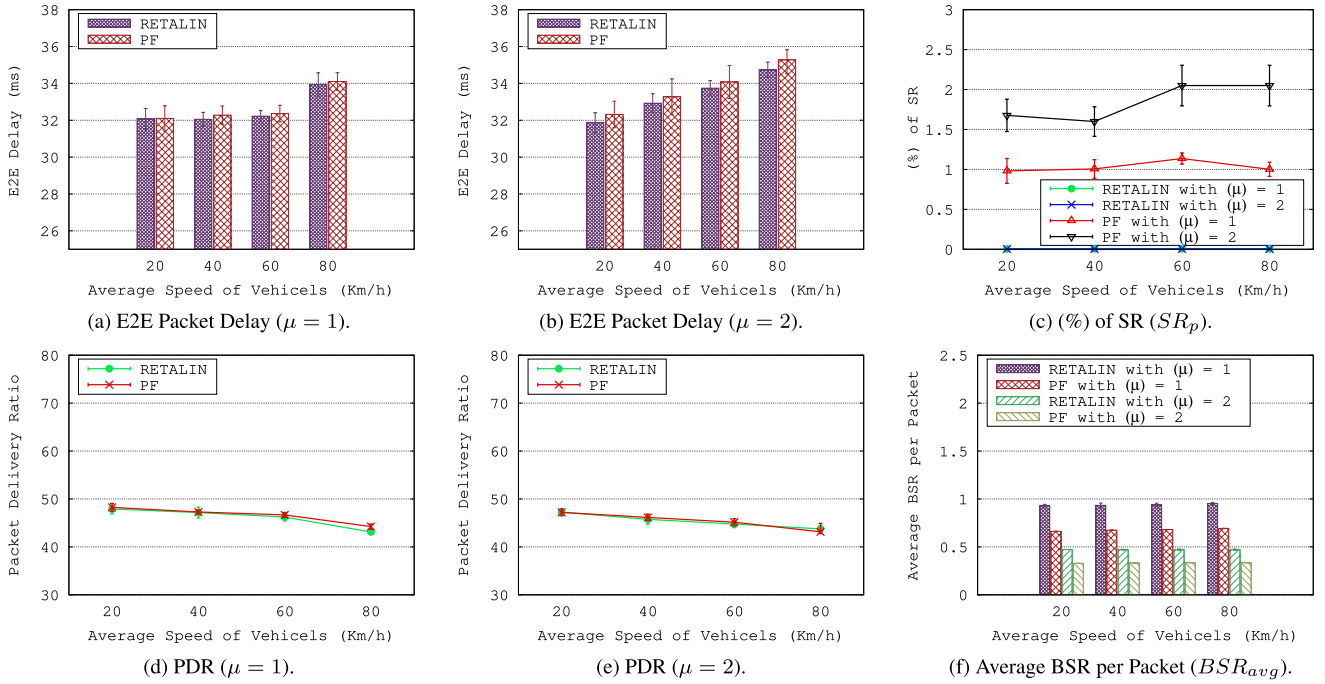
HD Map is considered as one of the key technologies for driving in future. With a centimeter-accurate visual representation of the roads and surrounding environment, it can assist (autonomous) vehicles with decision-making, perception, navigation, and localization. In fact, HD Map dissemination is a content-centric network service that puts tight constraints on latency [55]. For this reason, HD Map can be deployed at the edge of the network (MEC server) to reduce access delay. The MEC server constructs the HD Map using the on-board sensor data received from the vehicles over cellular networks (e.g., 5G NR) [56]. Vehicles do not transmit data in raw format, typically processed and aggregated data is transferred to the MEC server. The HD map so constructed is sent to the vehicles by the MEC server. In recent years, cartographers (e.g., TomTom and HERE) have released standards for high-precision maps such as OpenDrive, NDS, and vector HD Maps.

TABLE 4. Simulation parameters for HD map use case.

Parameter	Value
5G NR numerology (μ)	0, 1, 2
MEC Task scheduler	*OCTANE* [7]
5G NR MAC Scheduler	*RETALIN*, PF, QoS [33]
5G QoS Identifier (5QI)	75, GBR_V2X
CPU Clock frequency of MEC server	80 GHz [57]
Vehicle's CPU clock frequency	2 GHz [57]
Number of Jobs per vehicle	[1, 2]
Input size of the job	Mean: 6 Kb
	Variance: 10 Kb
	Bound: 5 Kb
CPU cycles required per job	[4,140] Mcycles [58]
Deadline of job	[100,150] $msec$ [59]
Resource Blocks (RBs) threshold per Job	1000 RBs
Job threshold per vehicle per second	12
Interval for maximum data transfer per vehicle	100 msec
Job generation per vehicle	0.1 sec
MCS threshold of a vehicle	5

B. SIMULATION SETUP

Using the same experimental setup as in Section VI, the vehicles in this scenario are assumed to be equipped additionally with On-Board Units (OBUs) responsible for sensing and limited processing of tasks related to the HD Map application. Additionally, an MEC server is co-located at the gNodeB, functioning as a nearby cloud infrastructure that enables vehicles to enhance their computing capabilities



(a) E2E Packet Delay ($\mu = 1$). (b) E2E Packet Delay ($\mu = 2$). (c) (%) of SR (SR_p).

(d) PDR ($\mu = 1$). (e) PDR ($\mu = 2$). (f) Average BSR per Packet (BSR_{avg}).

FIGURE 12. Result observed in case of *RETALIN* and PF by varying speed of vehicles for UL packet size of $L = 1400$ bytes.

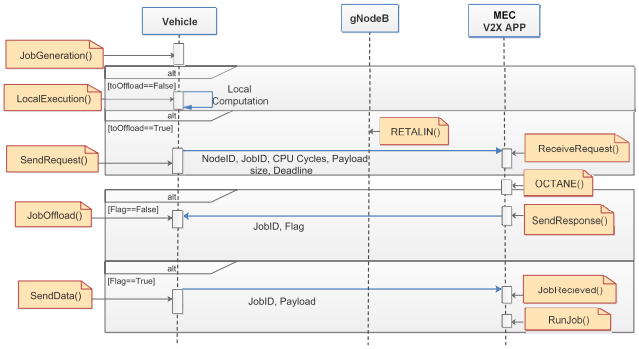


FIGURE 13. Sequence diagram of MapOffloading for HD MAP use case.

for constructing the HD Map. The system model for HD Map is shown Fig. 14, and simulation parameters are given in Table 4. Vehicles transmit various data related to the HD Map to the MEC server. Each vehicle generates various HD Map-related tasks, such as sensor data collection, sensor data analysis, and HD Map updates [60]. Notably, the analysis of sensor data is a computationally intensive task that necessitates offloading to the MEC server via 5G NR. To facilitate this process, we have developed a job (aka task) offloading application called MapOffloading, built upon the Udp-Client-Server application of NS-3. The client application is deployed on the vehicles, while the server application is hosted on the MEC server. The client application can be programmed to generate jobs with different attributes, including input sizes, deadlines, and CPU cycle requirements, at specific intervals. Vehicles retain the autonomy to determine whether to send jobs to the MEC server for execution or handle them locally. Conversely, the MEC server receives offloading requests from various vehicles and responds with offloading decisions tailored to each vehicle. Herein, the job latency adheres to the specifications outlined in the 5GCAR deliverable [58], while other parameters are set according to the guidelines provided in [57]. A sequence diagram, depicted in Fig. 13, illustrates the process of job offloading between vehicles and the MEC server over 5G NR.

The sequence diagram demonstrates the following phases and steps:

- *JobGeneration*: Vehicles generate jobs for the HD-Map application.
- *LocalExecution*: Vehicle determines whether jobs can be executed locally or need to be offloaded to a MEC server.
- *SendRequest*: If offloading is necessary, a request to offload the jobs is sent to the MEC server.
- *RequestReceived*: The MEC server collects offloading requests from all vehicles generating requests.
- *OCTANE*: At the MEC server, OCTANE [7], an iterative solution, is utilized to address the offloading decision problem, taking into account the deadline, radio, and computational resource requirements of the vehicles' jobs. OCTANE incorporates different simulation parameters, as specified in Table 4. These parameters include the job threshold, which signifies the number of jobs that can be admitted to the MEC server for a vehicle, and the MCS threshold, which indicates the minimum required MCS value for a vehicle to be considered. Furthermore, the maximum data transfer parameter per

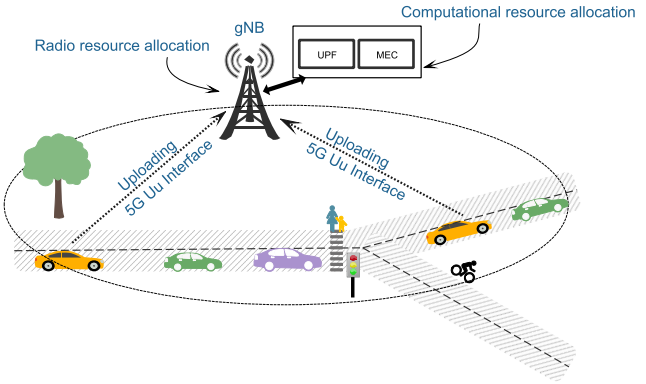


FIGURE 14. System model for HP map use case.

vehicle determines the upper limit of data that can be transmitted for a vehicle. The RBs threshold per job represents the maximum RBs that can be allocated to a job; otherwise, the job will be rejected. The primary objective of this phase is to maximize the number of successfully completed jobs for each vehicle with the assistance of the MEC server, while ensuring fairness among all the vehicles in the HD Map use case.

- *SendResponse*: The MEC server sends a response to each vehicle regarding the offloading decision.
- *JobOffload*: Vehicle receive the response from MEC server and process it.
- *SendData*: If the MEC server's response is positive, the job data is offloaded to the MEC server.
- *JobReceived*: Confirmation that the job is received at the MEC server without any packet loss.
- *RunJob*: The job is executed at the MEC server.

C. PERFORMANCE METRICS AND RELATED WORK FOR COMPARISON

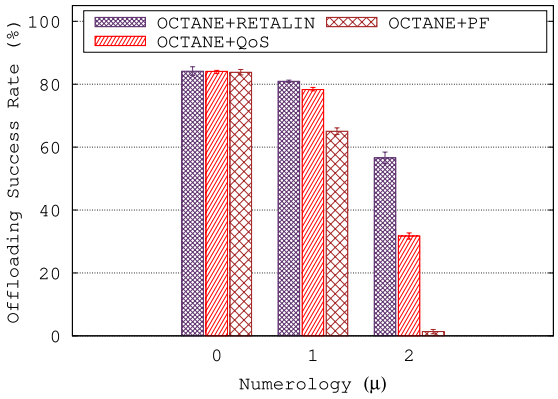
In our previous work, we introduced OCTANE [7], which operates at the application layer within the MEC server to select jobs in the HD Map use case. OCTANE aims to optimize resource utilization in the MEC server. Notably, in this use case, OCTANE runs at the application layer, while PF, QoS and RETALIN operate at the MAC layer. We conducted simulations with different combinations of algorithms, namely OCTANE+PF, OCTANE+QoS and OCTANE+RETALIN, all tested under the same scenarios and conditions.

In the HD Map use case, we consider Offloading Success Rate (OSR) as one of the key performance metrics. A job is treated as a success when it is offloaded to the MEC server and executed within its deadline. OSR is defined as the ratio of the number of successfully executed jobs by the MEC server to the total number of offloading jobs requests received by the MEC server. OSR is measured at the MEC server for different schemes in different vehicular scenarios. The Offloading Rate (OR) is another metric that we use. OR is the ratio of the number of jobs offloaded to the MEC server to the total number of requests received by the MEC server. The difference in OSR and OR is derived from the jobs which are offloaded but did not get successfully executed within their deadlines by the MEC server. Moreover, we considered Average Job Response Time (AJRT) as the delay in the time when a vehicle has requested a job to be offloaded and got the response from the MEC server. Average Job Offloading Time (AJOT) is the time taken for a vehicle to send the job to a MEC server after getting a response from MEC server. AJRT and AJOT dictate OSR and OR of a vehicle; for example, vehicles' poor channel conditions may increase AJRT and AJOT, thereby decreasing OSR and OR. If a vehicle decides to run a job locally, that job is excluded from OSR and OR calculations.

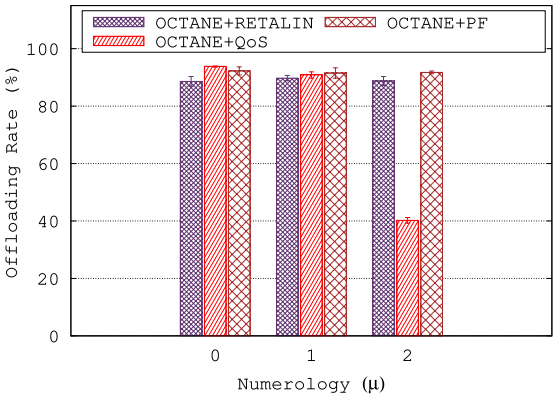
D. PERFORMANCE RESULTS

In Fig. 15a, we show the variation in OSR for *OCTANE+PF*, *OCTANE+QoS* and *OCTANE+RETALIN* for three numerologies by setting $\mathcal{V}_{speed} = 60kmph$ and $L = 1000$ bytes. We observe that by increasing the numerology from $\mu = 0$ to $\mu = 2$; OSR decreases. Here, *OCTANE+PF*, *OCTANE+QoS* and *OCTANE+RETALIN* have OSR of 83%, 64%, 2% and 84%, 78%, 31% and 84%, 81%, 56% in case of $\mu = 0$, $\mu = 1$, $\mu = 2$, respectively. On the other hand, as shown in Fig. 15b, OR remains almost same for all the numerologies. Constant OR indicates that jobs that are getting offloaded for all the numerologies considered, but OSR is degrading for higher numerologies which reflects that more and more jobs are not getting executed within the stipulated deadline of HD map application in case of higher numerologies. OSR degradation is due to increase in AJOT and AJRT values as can be seen in Figs. 15c and 15d, respectively. As compared to *OCTANE+PF* and *OCTANE+QoS*, *OCTANE+RETALIN* reduces AJOT by 16%, 6% as can be seen in Fig. 15c. As we know, AJOT and AJRT are inextricably linked together to complete the jobs within their deadlines. *OCTANE+RETALIN* decreases AJRT as compared to *OCTANE+PF* and *OCTANE+QoS*, thereby increasing OSR of vehicles for $\mu = 0$ and $\mu = 1$. But, AJRT for *OCTANE+PF* and *OCTANE+QoS* drastically increases to 135 *msec* and 125 *msec* in case of $\mu = 2$ and due to this OSR reduces to 2% and 31%, respectively. Impressively, *OCTANE+RETALIN* is able to maintain its AJRT at 89 *msec* in case of $\mu = 2$ and due to this its OSR is reasonably good (56%).

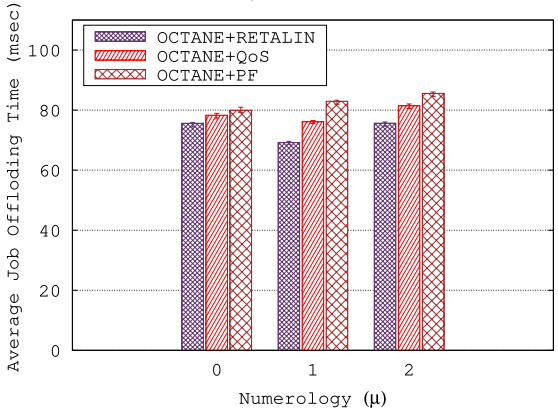
AJRT and AJOT are mainly affected due to *E2E packet delay*, as can be seen in Fig. 16a for all the numerologies. Here, we can observe that for $\mu = 2$, the E2E delay increased to 65.8 *msec* in the case of *OCTANE+PF* and 57.6 *msec* in the case of *OCTANE+QoS*. As mentioned in the previous section, this increase in delay is attributed to the interplay between factors such as *decodingLatency*, processing delays, the uplink (UL) mechanism, different IPATs of the vehicles due to job sizes and deadlines, and fragmentation of packets due to varying channel conditions. As a consequence of this interplay, SR_p and BSR_{avg} are increased as shown in



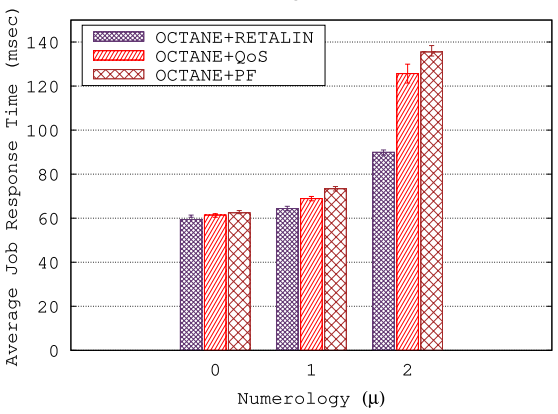
(a) Offloading Success Rate (OSR).



(b) Offloading Rate (OR).

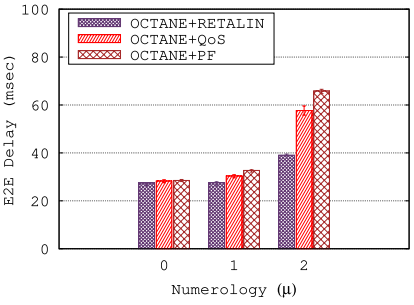


(c) Average Job Offloading Time (AJOT).

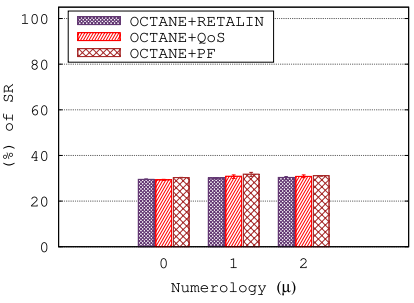


(d) Average Response Time (AJRT).

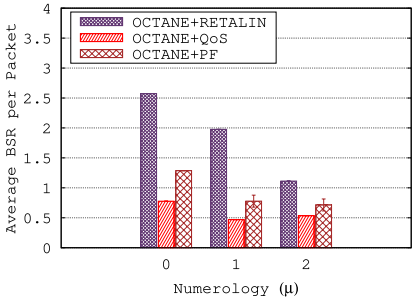
FIGURE 15. Variation in OSR, OR, AJOT, AJRT for the HD map application for different numerologies with $\mathcal{V}_{speed} = 60kmph$ and $L = 1000$ bytes.



(a) E2E Packet Delay.



(b) (%) of SR (SR_p).



(c) Average BSR per Packet (BSR_{avg}).

FIGURE 16. Performance results of the HD map application for different numerologies with $\mathcal{V}_{speed} = 60kmph$ and $L = 1000$ bytes.

Fig. 16b and Fig. 16c, which negatively affect the E2E delay. However, AJOT is reduced for *OCTANE+RETALIN* and due to this its OSR is high when compared to *OCTANE+PF*. In this way, *OCTANE+RETALIN* achieved efficient radio resource allocation, better offloading choices, and better job assignment for the HD Map use case.

VIII. CONCLUSION AND FUTURE WORK

In this work, we proposed *RETALIN*, a two-phase UL grant-based scheme for radio resource allocation in 5G NR. In the first phase, *RETALIN* clears the backlog queues of the UEs by granting resources for the pending Scheduling Requests (SRs) from the UEs based on the probability of SR in a TTI and backlog threshold of vehicles. In the second phase, *RETALIN* allocates radio resources to the UEs based on normalized buffer values and violation probability of the queues using the current RLC buffer status of UEs. By means of this two-phase radio resources allocation, *RETALIN* is capable of subsisting generation of SRs in the network, thereby increasing PDR and reducing E2E delay of the UEs. Simulation results are performed using the 5G-LENA module with mobility traces taken from SUMO

using OpenStreetMap. The outcomes demonstrated that the proposed *RETALIN* scheduling algorithm reduces link delay by 22% and 25% for numerologies 1 and 2, respectively, thereby reducing E2E delay of the applications over state-of-art schedulers QoS scheduler, as well as baseline schedulers such as Round Robin (RR) and Proportional Fair (PF). The results showed that *RETALIN* is capable of achieving a better trade-off between SR and BSR for higher numerologies with different packet sizes and traffic patterns. Further, *RETALIN* increases Packet Delivery Ratio (PDR) and reduces SRs. In the case of a High Definition Map (HD Map) application, *RETALIN* assists in increasing the Offloading Success Rate (OSR) by 17% over the QoS scheduler.

As part of the future work, one could take advantage of configured grant feature of 5G NR to reduce E2E delay further. The configured grant pre-allocates radio resources to UEs, thereby decreasing UL signaling overhead.

REFERENCES

[1] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Sahin, and A. Kousaridas, "A tutorial on 5G NR V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1972–2026, 3rd Quart., 2021.

[2] W. Fleischer, J. Dixon, A. Shapin, P. de Bruin, R. Williamson, L. Rose, G. D'Aria, A. Orlando, H. Zhang, R. Y. N. Li, and X. Han. *5GTDD Uplink v1.0, NGMN Alliance White Paper*. Jan. 2022. [Online]. Available: https://www.ngmn.org/wp-content/uploads/220117-5G-TDD-Uplink-White-Paper-v1.0.pdf

[3] Y. Özcan and C. Rosenberg, "Uplink scheduling in multi-cell OFDMA networks: A comprehensive study," *IEEE Trans. Mobile Comput.*, vol. 20, no. 10, pp. 3081–3098, Oct. 2021.

[4] *Technical Specification Group Radio Access Network; NR; Medium Access Control (MAC) Protocol Specification*, document TS 36.321, Version 15.4.0, Release 15, 3GPP, Sophia Antipolis, France, Mar. 2018. [Online]. Available: https://www.3gpp.org

[5] H.-L. Wang and S.-J. Kao, "Activity selection-based single carrier-frequency division multiple access uplink scheduling for two-tier LTE networks," *Wireless Pers. Commun.*, vol. 82, no. 1, pp. 625–642, May 2015.

[6] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "5G new radio numerologies and their impact on the end-to-end latency," in *Proc. IEEE 23rd Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, Sep. 2018, pp. 1–6.

[7] V. K. Gautam, C. Tompe, B. R. Tamma, and A. F. Antony, "OCTANE: A joint computation offloading and resource allocation scheme for MEC assisted 5G NR vehicular networks," in *Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2021, pp. 60–65.

[8] *Technical Specification Group Radio Access Network; NR; Medium Access Control (MAC) Protocol Specification*, document TR 38.912, Version 14.0.0, Release 14, 3GPP, Mar. 2017. [Online]. Available: https//www.3gpp.org

[9] *Technical Specification Group Radio Access Network; NR; Overall Description; Stage-2*, document TS 38.912, Version 15.1.0, Release 15, 3GPP, Apr. 2018. [Online]. Available: https://www.3gpp.org

[10] *NR; Physical Layer Procedures for Data; Technical Specification (TS)*, document 38.214, Version 15.7.0, 3GPP, Sep. 2018. [Online]. Available: https://www.3gpp.org

[11] *TSG RAN; NR; Physical Layer Procedures for Control*, document TS 38.213, Version 15.2.0, Release 15, Jun. 2018. [Online]. Available: https://www.3gpp.org

[12] *NR; Radio Resource Control (RRC) Protocol Specification; Technical Specification (TS)*, document 38.331, Version 15.5.1, 3GPP, Apr. 2019. [Online]. Available: https://www.3gpp.org

[13] *Summary of DL/UL Scheduling and HARQ Management*, document R1-1721515, Qualcomm, TSG RAN WG1 91 Meeting, 3GPP, Dec. 2017. [Online]. Available: https://www.3gpp.org

[14] *Remaining Issues on HARQ*, document R1-1719401, TSG RAN WG1 89 Meeting, Huawei, HiSilicon, Dec. 2017. [Online]. Available: https://www.3gpp.org

[15] H. Wu, "Efficient allocation of the amount of radio resources in 5G NR to efficient allocation of radio resources in 5G NR," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3321–3332, May 2022.

[16] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "The impact of NR scheduling timings on end-to-end delay for uplink traffic," in *Proc. IEEE Global Commun. Conf.*, Dec. 2019, pp. 1–6.

[17] *NR; Radio Link Control (RLC) Specification; Technical Specification (TS)*, document 38.322, Version 15.4.0, 3GPP, Jan. 2019. [Online]. Available: https://www.3gpp.org

[18] F. Furqan, D. B. Hoang, and I. B. Collings, "Effects of quality of service schemes on the capacity and dimensioning of LTE networks," in *Proc. IEEE 33rd Int. Perform. Comput. Commun. Conf. (IPCCC)*, Dec. 2014, pp. 1–8.

[19] L. A. N. Oliveira, M. S. Alencar, W. T. A. Lopes, and F. Madeiro, "On the performance of location management in 5G network using RRC inactive state," *IEEE Access*, vol. 10, pp. 65520–65532, 2022.

[20] Z. Zhang, S. Shi, V. Gupta, and R. Jana, "Analysis of cellular network latency for edge-based remote rendering streaming applications," in *Proc. ACM SIGCOMM Workshop Netw. Emerg. Appl. Technol.*, Aug. 2019, pp. 8–14.

[21] K. Boutiba, M. Bagaa, and A. Ksentini, "Optimal radio resource management in 5G NR featuring network slicing," *Comput. Netw.*, vol. 234, Oct. 2023, Art. no. 109937.

[22] M. K. Rana, T. Pecorella, B. Sardar, R. RaoThipparaju, and D. Saha, "A QoS improving downlink scheduling scheme for slicing in 5G radio access network (RAN)," *IEEE Trans. Veh. Technol.*, early access, 2023.

[23] M. Seguin, A. Omer, M. Koosha, F. Malandra, and N. Mastronarde, "Deep reinforcement learning for downlink scheduling in 5G and beyond networks: A review," in *Proc. IEEE 34th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2023, pp. 1–6.

[24] M. Laha, D. Roy, S. Dutta, and G. Das, "AI-assisted improved service provisioning for low-latency XR over 5G NR," *IEEE Netw. Lett.*, early access, 2023.

[25] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 2, pp. 226–234, Feb. 2009.

[26] L. Gao and S. Cui, "Efficient subcarrier, power, and rate allocation with fairness consideration for OFDMA uplink," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1507–1511, May 2008.

[27] M. Mehta, S. Khakurel, and A. Karandikar, "Buffer-based channel dependent UpLink scheduling in relay-assisted LTE networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 1777–1781.

[28] C. Wang, J.-J. Huang, and C.-Y. Su, "Buffer-aware and delay-sensitive resource allocation in the uplink of 3GPP LTE networks," *Wireless Pers. Commun.*, vol. 84, no. 3, pp. 1877–1890, Oct. 2015.

[29] O. Al-Khatib, W. Hardjawana, and B. Vucetic, "Channel- and buffer-aware scheduling and resource allocation algorithm for LTE—A uplink," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2014, pp. 1001–1006.

[30] S.-B. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, "Proportional fair frequency-domain packet scheduling for 3GPP LTE uplink," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2611–2615.

[31] X. Xiang, C. Lin, X. Chen, and X. Shen, "Toward optimal admission control and resource allocation for LTE—A femtocell uplink," *IEEE Trans. Veh. Technol.*, vol. 64, no. 7, pp. 3247–3261, Jul. 2015.

[32] A. Rizk and M. Fidler, "Queue-aware uplink scheduling with stochastic guarantees," *Comput. Commun.*, vol. 84, pp. 63–72, Jun. 2016.

[33] K. Koutlia, B. Bojovic, S. Lagén, X. Zhang, P. Wang, and J. Liu, "System analysis of QoS schedulers for XR traffic in 5G NR," *Simul. Model. Pract. Theory*, vol. 125, May 2023, Art. no. 102745.

[34] K. Kord, A. Elbery, S. Sorour, H. Hassanein, A. B. Sediq, A. Afana, and H. Abou-Zeid, "Enhanced C-V2X uplink resource allocation using vehicle maneuver prediction," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 3544–3549.

[35] Y. Huang, S. Li, Y. T. Hou, and W. Lou, "GPF+: A novel ultrafast GPU-based proportional fair scheduler for 5G NR," *IEEE/ACM Trans. Netw.*, vol. 30, no. 2, pp. 601–615, Apr. 2022.

[36] A. Hossain and N. Ansari, "5G multi-band numerology-based TDD RAN slicing for throughput and latency sensitive services," *IEEE Trans. Mobile Comput.*, vol. 22, no. 3, pp. 1263–1274, Mar. 2023.
[37] D. Segura, E. J. Khatib, J. Munilla, and R. Barco, "5G numerologies assessment for URLLC in industrial communications," *Sensors*, vol. 21, no. 7, p. 2489, Apr. 2021.
[38] W. Zhan, C. Xu, X. Sun, and J. Zou, "Toward optimal connection management for massive machine-type communications in 5G system," *IEEE Internet Things J.*, vol. 8, no. 17, pp. 13237–13250, Sep. 2021.
[39] J. A. Fernández-Segovia, S. Luna-Ramírez, M. Toril, and J. J. Sánchez-Sánchez, "A teletraffic model for the physical downlink control channel in LTE," *Telecommun. Syst.*, vol. 65, no. 3, pp. 511–523, Jul. 2017.
[40] M. Centenaro, L. Vangelista, and S. Saur, "Analysis of 5G radio access protocols for uplink URLLC in a connection-less mode," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3104–3117, May 2020.
[41] S. Dutta, M. Mezzavilla, R. Ford, M. Zhang, S. Rangan, and M. Zorzi, "Frame structure design and analysis for millimeter wave cellular systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1508–1522, Mar. 2017.
[42] H.-H. Liu and H.-Y. Wei, "5G NR multicast and broadcast QoS enhancement with flexible service continuity configuration," *IEEE Trans. Broadcast.*, vol. 68, no. 3, pp. 689–703, Sep. 2022.
[43] N. Bouchemal, N. Izri, and S. Tohme, "MAC-LTE scheduler modeling and performance evaluation in LTE network," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2014, pp. 1007–1012.
[44] N. Bouchemal, "Quality of service provisioning and performance analysis in vehicular network," Ph.D. dissertation, Dept. Comput. Sci., Université Versailles-Saint Quentin Yvelines, Versailles, France, 2015.
[45] Ö. F. Gemici, I. Hökelek, and H. A. Çirpan, "Modeling queuing delay of 5G NR with NOMA under SINR outage constraint," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2389–2403, Mar. 2021.
[46] W. Hwang, Y. Kim, and K. Lee, "A transient queueing analysis under time-varying arrival and service rates for enabling low-latency services," 2017, *arXiv:1704.07091*.
[47] J. Abate and W. Whitt, "Calculating time-dependent performance measures for the M/M/1 queue," *IEEE Trans. Commun.*, vol. 37, no. 10, pp. 1102–1104, Oct. 1989.
[48] M. Khalid, Y. Wang, X. H. Le, I.-H. Ra, and R. Sankar, "Distributed adaptive scheduling for finite horizon in wireless ad hoc networks," *J. Commun.*, vol. 7, no. 2, pp. 155–164, Feb. 2012.
[49] F. Jomrich, T. Wankhede, T. Ruckelt, D. Burgstahler, D. Bohnstedt, R. Steinmetz, and S. Knapp, "Demo: Rapid cellular network simulation framework for automotive scenarios (RACE framework)," in *Proc. Int. Conf. Networked Syst. (NetSys)*, Mar. 2017, pp. 1–2.
[50] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simul. Model. Pract. Theory*, vol. 96, Nov. 2019, Art. no. 101933.
[51] *TSG; System Architecture for the 5G System; Stage 2*, document TS 23.501, Version 15.0.0, Release 15, 3GPP, 2017. [Online]. Available: https://www.3gpp.org
[52] L. Marijanovic, S. Schwarz, and M. Rupp, "Optimal numerology in OFDM systems based on imperfect channel knowledge," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.
[53] A. A. Zaidi, R. Baldemair, V. Moles-Cases, N. He, K. Werner, and A. Cedergren, "OFDM numerology design for 5G new radio to support IoT, eMBB, and MBSFN," *IEEE Commun. Standards Mag.*, vol. 2, no. 2, pp. 78–83, Jun. 2018.
[54] C. Tang, X. Chen, Y. Chen, and Z. Li, "Dynamic resource optimization based on flexible numerology and Markov decision process for heterogeneous services," in *Proc. IEEE 25th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2019, pp. 610–617.
[55] C. W. Gran, "HD-maps in autonomous driving," M.S. thesis, NTNU, Trondheim, Norway, 2019.
[56] T. Deinlein, R. German, and A. Djanatliev, "5G-Sim-V2I/N: Towards a simulation framework for the evaluation of 5G V2I/V2N use cases," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2020, pp. 353–357.
[57] W. Zhan, C. Luo, G. Min, C. Wang, Q. Zhu, and H. Duan, "Mobility-aware multi-user offloading optimization for mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3341–3356, Mar. 2020.
[58] C. Yang, Y. Liu, X. Chen, W. Zhong, and S. Xie, "Efficient mobility-aware task offloading for vehicular edge computing networks," *IEEE Access*, vol. 7, pp. 26652–26664, 2019.
[59] *Deliverable D2.1: 5GCAR Scenarios Use Cases Requirements and KPIs*, document D2.1, Aug. 2017.
[60] *Operational Behavior of a High Definition Map Application*, Automotive Edge Computing Consortium (AECC), Wakefield, MA, USA, May 2020.

VEERENDRA KUMAR GAUTAM (Member, IEEE) received the bachelor's degree in computer applications from Symbiosis International University, Pune, India, in 2011, and the master's degree in computer application and the M.Tech. degree in computer science and engineering from the University of Hyderabad, India, in 2014 and 2017, respectively. Currently, he is pursuing the Ph.D. degree in computer science and engineering with the Indian Institute of Technology Hyderabad (IITH), India. His research interests include 5G, vehicular communication, wireless resource scheduling, and AI in mobile networks.

BHEEMARJUNA REDDY TAMMA (Senior Member, IEEE) received the Ph.D. degree from the Indian Institute of Technology (IIT) Madras, India, in 2007. Then, he was a Postdoctoral Fellow with the University of California at San Diego (UCSD), California Institute for Telecommunications and Information Technology (CALIT2), prior to taking up a faculty position with IIT Hyderabad, India, in 2010, where he is currently a Professor with the Department of Computer Science and Engineering. He has published over 100 papers in refereed international journals and conferences. His research interests include converged cloud radio access networks, SDN/NFV for 5G, network security, and green ICT. He was a recipient of Visvesvaraya Young Faculty Research Fellowship at IIT Hyderabad and iNautix Research Fellowship for his Ph.D. tenure at IIT Madras. He is a co-recipient of Top Cited Article Award from Elsevier publishers, the Best Academic Demo Award at COMSNETS 2018, the Best Poster Award at ICACCI 2018, the Second Best Paper Award at IEEE ANTS2017, and the Best Paper Award at ICACCI 2015 Conferences. He is a member of ACM. He served as the General Co-Chair for National Conference on Communications (NCC) 2018, the TCP Co-Chair for IEEE ANTS 2015, the TCP Vice Chair for IEEE ANTS 2014, and a Ph.D. student Forum Co-Chair for IEEE ANTS 2013 Conferences.