IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 20, 2020, accepted August 26, 2020, date of publication September 18, 2020, date of current version October 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024790

Session-Level Reliability Analysis for Multi-Service Communication in Autonomous Vehicular Fleets

VITALY PETROV[1], (Student Member, IEEE), NATALIA YARKINA[2], DMITRI MOLTCHANOV[1], SERGEY ANDREEV[1], (Senior Member, IEEE), AND KONSTANTIN SAMOUYLOV[2,3]

[1]Unit of Electrical Engineering, Tampere University, 33720 Tampere, Finland
[2]Department of Applied Mathematics and Probability Theory, Peoples' Friendship University of Russia (RUDN University), 117198 Moscow, Russia
[3]Federal Research Center "Computer Science and Control," Institute of Informatics Problems, Russian Academy of Sciences, 119333 Moscow, Russia

Corresponding author: Vitaly Petrov (vitaly.petrov@tuni.fi)

The work was supported by the Project RADIANT of the Academy of Finland, in part by the 5G-FORCE Project, and in part by the RAAS Connectivity RTF Framework. The work of Natalia Yarkina was supported by the RUDN University Program 5-100. The work of Konstantin Samouylov was supported by the RFBR under Project 18-00-01555 (18-00-01685) and Project 19-07-00933.

ABSTRACT The support for reliable communication in the increasingly large fleets of autonomous vehicles is one of the important challenges for emerging 5G systems. The presence of heterogeneous data streams of different priority between the connected vehicles (coordinated autonomous driving, platooning, passenger entertainment services, etc.) calls for new methods to estimate the performance characteristics of these systems. In this article, a novel mathematical framework is proposed to model the process of dynamic radio resource (re-)allocation across multiple competing data streams in autonomous vehicular fleets equipped with 5G cellular capabilities. The developed framework is subsequently applied to: (i) study the coexistence of multiple traffic types having different service requirements and (ii) quantify the impact of session prioritization schemes. Our study reveals that the prioritization scheme initially offloading high-rate entertainment sessions, named *ESpreempt*, in most of the setups achieves a 5-30% performance gain in comparison with the scheme initially offloading low-rate platooning sessions, named *PSpreempt*. It is also shown that higher variations in the traffic load of autonomous driving sessions have a distinctly negative impact on system-level performance. The outlined framework can be applied in a wide range of 5G vehicular scenarios, as well as extended to capture other categories of data streams in future wireless networks.

INDEX TERMS Resource management, computer network reliability, vehicular and wireless technologies.

I. INTRODUCTION

The envisioned emergence of connected autonomous vehicles is one of the major disruptions introduced on the way from 4G and 4G+ to the 5G-grade communication systems [1]–[3]. From the networking perspective, these autonomous vehicles – and soon their large-scale connected fleets – represent a whole new class of intelligent mobile users that combine a number of heterogeneous data services [4], [5]. First and foremost, these include *mission-critical vehicular transmissions* to support collective driving operation [6]–[8]. This category of traffic patterns is characterized by relatively small but delay-critical messages [9]. The volumes of critical vehicular traffic in a certain area of interest may vary significantly: from regular levels in normal operating conditions [10] to extreme levels in case of a sudden disruption (emergency brake, car accident, etc.) [11].

Autonomous vehicular fleets are also envisaged to participate in *distributed sensing and platooning* functions as part of the Internet of Things (IoT) [12]–[14]. Hence, mission-critical data communication is complemented by delay-tolerant message transmissions from onboard vehicular sensors as well as from those deployed around the connected car (e.g., on the surrounding buildings and roadside infrastructure). Intelligent autonomous vehicles will also facilitate *rich office work and entertainment experience* of their passengers, who will have more time for such activities on the move with increased levels of driving autonomy.

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

This category may support a wide range of services: from high-quality video streaming to remote desktop applications [15]–[17]. Other categories of services may emerge soon due to the rapid developments in the field.

Due to stringent capacity, space, cost, and communication range limitations, high-rate multi-service communications in connected autonomous fleets require intelligent radio resource provisioning for improved system robustness and reliability at vehicular speeds [3], [18]. This is especially true during the early stages of deploying the high-rate 5G systems, where substantial overprovisioning everywhere can hardly be expected. Similar concerns apply to the subsequent stages of the 5G network deployment, specifically during the ''busy hours''.

To support the operational flexibility of heterogeneous data streams within a single highly-mobile connectivity platform [19], especially in the presence of intermittent connectivity and unexpected link failures, the design of dynamic survivability strategies becomes central [20]. One of the key challenges in achieving sustainable network performance is to effectively share the resources among multi-service data streams with intelligent admission control and session management procedures [21]. Such sharing needs to provide guaranteed reliability levels for mission-critical vehicular communication, while at the same time not hindering the operation of other network services [22]. High service availability has to also be maintained together with the efficient utilization of radio resources, which is crucial for prospective system operators due to demanding quality of service requirements [23]–[25].

In this article, we study the coexistence of traffic types having different service requirements at the 5G cellular interface. We particularly analyze intelligent admission control and session management procedures that aim to accept the new session for service only when there are sufficient radio resources available to serve this session. For this purpose, we develop a mathematical framework that is capable of evaluating a wide range of performance characteristics, with a particular emphasis on serving multi-service data streams in large fleets of connected autonomous vehicles. The contributed model is later applied to assess and compare different service schemes prioritizing certain categories of traffic over the others. Particularly, two candidate schemes are evaluated for dynamic preemption of the sessions in the case where a session with higher priority arrives.

The rest of this article is organized as follows. We review the state of the art and highlight the key novelty of our framework in Section II. We then detail the considered scenario and the system model employed by this study in Section III. We elaborate our mathematical framework in Section IV. In Section V, we present and discuss the key numerical results and also analyze the effect of different stream prioritization schemes that the network can employ. The paper concludes with important general remarks collected in Section VI.

II. BACKGROUND AND RATIONALE

A. RELATED WORK SUMMARY

The aspects of resource allocation for mission-critical communication have been studied for years now with a number of sound solutions proposed to date [26]. The research community concentrated in the past on investigating static resource sharing between dissimilar traffic categories, so that each of them operates on its own virtual subband in isolation from other streams [27], [28]. Such solutions are featured by lower implementation complexity [29], but are not fully applicable in the 5G-grade vehicular context, since the rate of mission-critical communication can vary significantly over the relatively short periods of time, as outlined in [4], [6].

Further, advanced wireless network design options emerged to offer a dynamic adaptation of the resource shares in response to the immediate changes in the traffic demand. These assume the utilization of a certain fraction of the radio resources on the first-come-first-served basis [30]. As shown in [31], the performance of such approaches is better than that for static resource allocation; however, the corresponding architectures still do not reach the desired levels of flexibility, since the instantaneous intensities of the counterpart streams may vary faster than the system adaptation rate [32].

Hence, more elaborate solutions are envisioned for the next-generation 5G systems [33], which can associate each of the traffic categories with its own priority class to then conduct dynamic (re-)allocation of radio resources directly in the packet scheduler. Consequently, no radio resources remain idle and lower priority streams are always served unless there is a competing data transmission with a higher priority. In the latter case, lower priority transmissions may be dropped or offloaded to release the system resources demanded by higher priority traffic. These contemporary approaches enable guaranteed reliability of mission-critical traffic, which is instrumental in delivering information both timely and successfully, while at the same time ensuring sufficient degrees of network availability for fault-tolerant data streams. As a result, system radio resources are utilized more efficiently [34].

B. PROBLEM FORMULATION AND OUR CONTRIBUTION

Together with notable performance- and reliability-specific benefits, the introduction of intelligent admission control procedures and dynamic radio resource provisioning mechanisms leads to the increased degrees of network complexity and may thus compromise its robustness. Hence, there is a prompt demand in novel powerful methods to analyze the respective system-level performance, which can carefully model the complex process of dynamic resource (re-)allocation and capture the intricate dynamics in the underlying multi-service data streams.

While the real-world 5G networks are inherently packet-switched, there are stringent quality of service demands

defined at higher layers for different categories of *data sessions* [35]. Hence, analysis of the network performance at the session level in the presence of multiple heterogeneous data streams, on one side, and of the intelligent mechanisms for dynamic (re-)allocation of radio resources for these sessions, on the other, becomes of interest.

In this article, we respond to the said need and develop a comprehensive mathematical framework capable of analyzing such flexible and dynamic systems. The contributions of this work are thus summarized as follows:

- *Enabling mathematical framework:* The key novelties of the developed framework are: (i) accounting for time-varying behavior in the offered load of vehicular communication flows; (ii) evaluating a more practical setup where the network (re-)allocates in real-time the resources individually for each of the user sessions following a given prioritization algorithm that takes into account the current set of active sessions. The latter allows us to mathematically model more advanced and flexible resource management policies envisioned for 5G/5G+ networks. The proposed framework can be adapted to a given number of traffic streams with their dissimilar properties, and also to preferred priority levels and corresponding strategies of radio resource (re-)allocation.
- *In-depth assessment of traffic prioritization schemes:* A thorough investigation of the multi-stream system operation is also contributed. Our study considers three heterogeneous data streams coming from connected vehicles: (i) mission-critical traffic related to collective autonomous driving; (ii) platooning traffic between the vehicles and the road-side infrastructure; and (iii) multimedia data streams for in-vehicle entertainment systems. For the considered setup, a wide range of user- and network-centric performance indicators are assessed. We particularly emphasize the effect of the employed preemption schemes, when a new high-priority session can acquire the radio resources originally provisioned for another session with a lower priority.

The following section details an illustrative scenario and a system model employed by our study.

III. CONSIDERED SYSTEM MODEL

In this section, we introduce our system model. We begin by describing the considered scenario and the network deployment in subsection III-A. Traffic models are then introduced in subsection III-B. We clarify the service model and the employed preemption schemes in subsection III-C. Finally, the selected performance indicators are enumerated in subsection III-D.

A. SCENARIO DESCRIPTION

We consider a single (tagged) network cell in an urban environment with a number of autonomous vehicles equipped with millimeter-wave (mmWave) cellular radio capabilities that reside within the cell coverage (see Fig. 1).

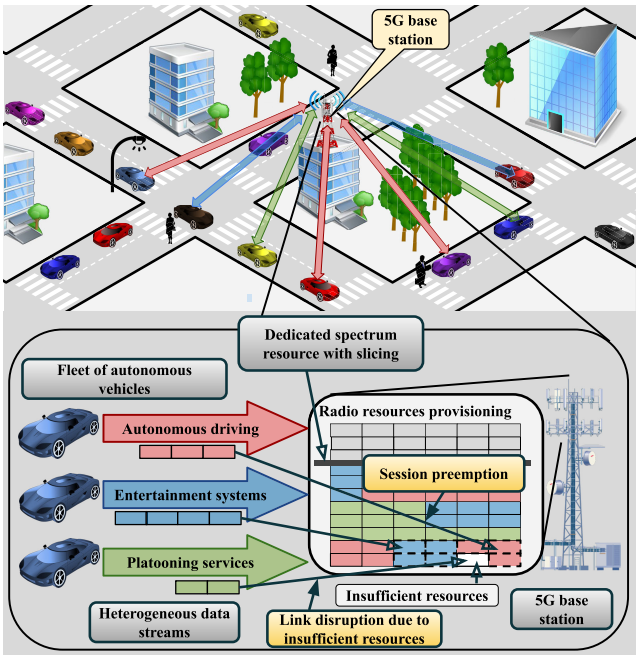


FIGURE 1. Considered autonomous vehicular fleet communications scenario with dynamic reallocation of radio resources.

Due to potentially high volumes of traffic generated by autonomous vehicle fleets, we assume that all the communication between them and the network infrastructure is conducted primarily over such mmWave radio links. The system operator is assumed to utilize the state-of-the-art network slicing mechanisms [36] to reserve a certain fraction of the mmWave radio resources exclusively for autonomous vehicles.

The data traffic is modeled at the session level. Each of the user sessions is characterized by its category as well as, depending on the category, session duration, or the number of radio resources required to successfully serve it. Admission control and service processes are modeled individually for every session, where we assume that uplink and downlink traffic flows share the same set of radio resources. In the case where the network does not currently have sufficient resources available in the pool reserved for vehicular communications, the session is assumed to e.g., be offloaded onto another radio access technology.[1] While the emerging 5G-grade networks are expected to be packet-switched, the end-to-end quality of service (QoS) is envisioned to still be maintained at the session level [37], [38]. Therefore, we argue that modeling the data traffic at the session level is appropriate for the first-order performance analysis attempted in this work.

[1] In real-world 5G networks, the session that currently cannot be accommodated by the considered RAT can be either offloaded onto another RAT (if available in proximity) or dropped/postponed. Using our proposed framework, the network operators may estimate the numbers and characteristics of the traffic flows that do not fit into the high-rate 5G component. Hence, it becomes possible to make conclusions on the density of base stations of other RATs to be deployed to prevent demand losses with a desired margin.

B. TRAFFIC CATEGORIES

Each of the connected autonomous vehicles is a complex and intelligent service platform, which is capable of producing different categories of data sessions. As an illustrative example, in this work we consider the following three categories:

1) AUTONOMOUS DRIVING (AD)

This category of traffic represents mission-critical communications related to autonomous driving. To capture occasional realistic spikes in the offered load due to this traffic pattern, we represent the process of session arrivals using the *Markovian arrival process (MAP)* [39], [40]. The number of states and the characteristics of the arrival process for each of the states may correspond, e.g., to *regular-load regime* and *extreme-load conditions* as well as be extended to support finer granularity of intensity levels. The number of radio resources required to serve a new AD session is exponentially distributed with the mean β_{AD}^{-1}.

MAP for the AD traffic is defined using two $K \times K$ matrices $\mathbf{Q_0}$ and $\mathbf{Q_1}$. Transitions between the states can occur both with ($\mathbf{Q_1}$) and without ($\mathbf{Q_0}$) producing a new session arrival. The matrix $\mathbf{Q} = \mathbf{Q_0} + \mathbf{Q_1}$ is the infinitesimal generator of the Markov chain corresponding to transitions between K states of the source. The elements of matrix $\mathbf{Q_1}$ are non-negative and $\mathbf{Q_1} \neq \mathbf{0}$; matrix $\mathbf{Q}$ is irreducible. Let $\mathbf{q}$ denote the row vector of the stationary state probabilities and let $\mathbf{1}$ be the column vector of ones. The average autonomous driving session arrival rate is thus given by $\alpha_{AD} = \mathbf{qQ_1 1}$.

2) PLATOONING SERVICES (PS)

This category of traffic is inspired by recent research, which envisions that connected vehicles may also act as so-called *data mules* [6], [12], [28]. Accordingly, an autonomous vehicle is assumed to continuously receive updates from various sensors and meters in its proximity, aggregate thus received data, and send this traffic to the remote application servers as a dedicated stream. Aiming to capture the random behavior of aggregate demands, this class of traffic is modeled as a Poisson process with rate α_{PS}, and the mean resource request of β_{PS}^{-1}.

3) ENTERTAINMENT SYSTEMS (ES)

This category of traffic is related to bandwidth-hungry entertainment or work applications. Since the driving process is becoming increasingly autonomous, the passengers are assumed to be more active by engaging in remote desktop, online gaming, and high-definition video services. Entertainment data sessions are assumed to arrive randomly according to a Poisson process with rate λ, and their duration is also random, distributed exponentially with parameter μ. We particularly assume a constant-bitrate video codec and consider the minimum rate requirements that are to be satisfied at all times. In the case if the amount of radio resources allocated to a session drops below its minimal requirement, the session may not be served successfully.

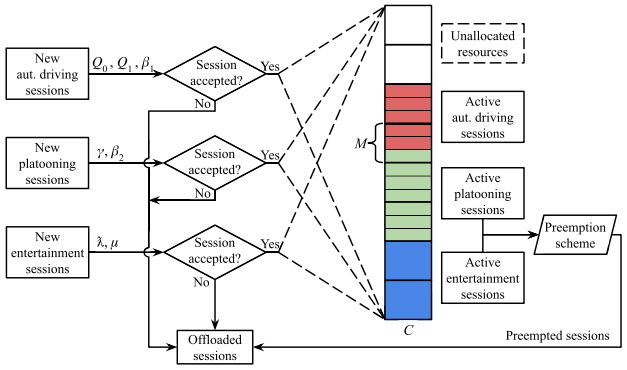


FIGURE 2. Queuing system with preemptive priority service.

The listed set of traffic categories is not exhaustive. The contributed model can be further adapted to other categories of traffic, such as high-priority data coming from emergency or law enforcement vehicles [18], voice and video calls performed by the drivers [16], online/social gaming [41], augmented reality [42], and many others.

C. SERVICE MODEL

A high-level illustration of the considered service model is offered in Fig. 2. We assume that the pool of radio resources dedicated to vehicular communications is fixed and equals to $C > 0$ resource units (RUs). The RU is defined as the mean rate attained at a single resource block and is thus measured in bits per second. Each of the entertainment sessions is assumed to occupy a single resource unit. An arriving session of this type is accepted if there is at least one free RU. Platooning and autonomous driving sessions are expected to demand lower rates as compared to the bandwidth-oriented entertainment streams and thus occupy fractions of the RU.

Particularly, platooning and autonomous driving sessions are served according to the Egalitarian Processor Sharing (EPS) service discipline where up to M such sessions share a single RU. If there are together s active AD and PS sessions, they jointly occupy and equally share $\lceil s/M \rceil = \min\{y \in \mathbb{N} : y \geq s/M\}$ RUs. Hence, $1, 2, \ldots, M$ sessions are served by one RU, $M+1, M+2, \ldots, M+M$ sessions require two RUs, $2M+1, 2M+2, \ldots, 2M+M$ need three RUs, and so forth. Correspondingly, the service rate can vary between a maximum of 1 and a minimum of $1/M$ share of a single resource unit.

In the case if there is currently a lack of available resources to serve a new session, the network follows a preemptive priority admission process as illustrated in Fig. 3; it either offloads the incoming session or accepts it by withdrawing the resources from one of the active sessions with a lower priority. The mathematical framework developed in this article can be adapted to any specified priority order between the considered traffic classes. For the sake of exposition, we particularly consider two preemptive priority schemes, namely, *PSpreempt* and *ESpreempt*.

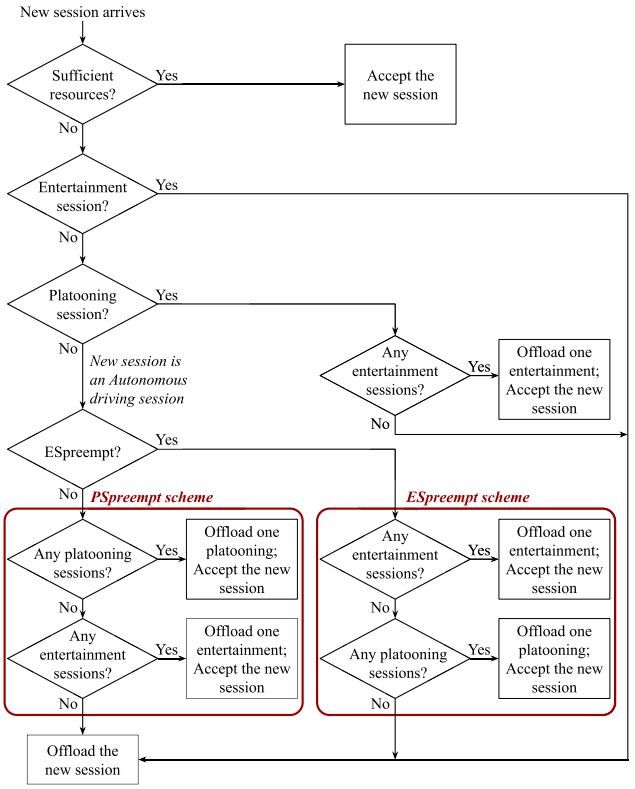


FIGURE 3. Modeled cascade preemptive priority service schemes.

In the considered model, ES sessions have the lowest priority and do not preempt other sessions. Incoming PS sessions preempt ES sessions if all RUs are busy. AD sessions have the highest priority and can preempt either ES or PS sessions. If both ES and PS sessions are currently present in the system, and all the RUs are busy, an arriving AD session will preempt a randomly-selected ES session in the case of *ESpreempt* scheme and a randomly-selected PS session in the case of *PSpreempt* scheme (all the preempted sessions are offloaded, see Fig. 2). We compare the network characteristics with these two schemes later in this article.

D. METRICS OF INTEREST

The following set of essential metrics of interest is captured by our mathematical framework:

- *Offloading probability upon arrival.* The probability that a new session will become offloaded upon arrival.
- *Session preemption probability.* The probability that a session will be preempted by another session having a higher priority (see Fig. 3).
- *Session offloading rate.* The probability that a session will become offloaded either upon arrival or when preempted by another session during service.
- *Average number of active sessions.* The average number of sessions for a given traffic category, which are served simultaneously.

This set of parameters allows assessing both user-centric (the first three) and network-centric (the last one) performance indicators for the considered vehicular network. In the next section, we detail our mathematical framework that is developed to model the described operation of a multi-service autonomous vehicular fleet.

IV. PROPOSED ANALYTICAL FRAMEWORK

In this section, we formalize the service process of multi-service traffic streams with priorities by leveraging an appropriate queuing model with a preemptive service process. To this aim, we first formalize the admission process and the considered priority schemes. We then employ a matrix-analytic formulation to describe and solve the resulting queuing model. Ultimately, performance metrics of interest are derived. The essential notation is summarized in Table 1.

A. ADMISSION AND SERVICE PROCESS WITH PREEMPTIVE PRIORITIES

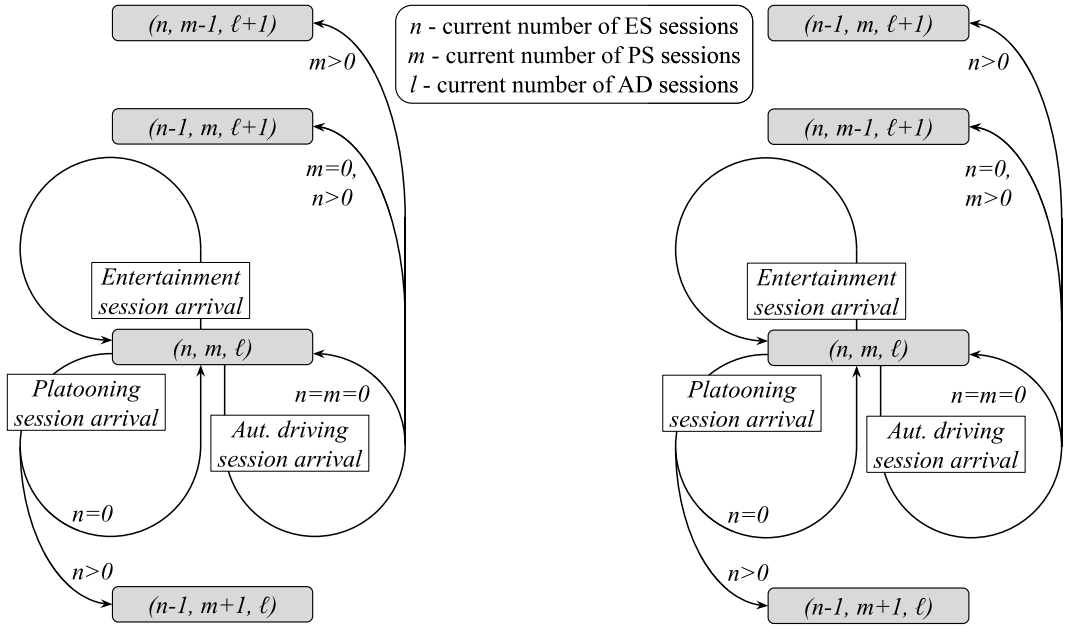
Our preemptive priority assumptions are illustrated in Fig. 4, where l, m, and n denote the numbers of active autonomous driving, platooning, and entertainment sessions, respectively. First, we formalize the *PSpreempt* priority scheme. Assume that there are l AD and m PS sessions in the considered setup, $s = l + m$. Let $c(s) = \lceil s/M \rceil$ denote the number of RUs occupied by these sessions.

Hence, upon arrival of $(m+1)$-th platooning session, one of the following outcomes occurs:

1) If $c(s+1) = c(s)$ then the session is accepted for service, but no additional RUs are allocated. The capacity of the RUs already occupied by the platooning and autonomous driving sessions is then equally redistributed among $s+1$ sessions, so that each session occupies $c(s+1)/(s+1) = c(s)/(s+1)$ RUs instead of $c(s)/s$, and its service rate thus decreases.
2) If $c(s+1) > c(s)$ and there is at least one unoccupied resource unit in the system, then the session is accepted and one additional RU is allocated to the platooning and autonomous driving sessions. In this case, the RU capacity is redistributed and the service rate of all the platooning and autonomous driving sessions increases.
3) If $c(s+1) > c(s)$ but $C - c(s)$ RUs are occupied by the entertainment sessions then a randomly chosen entertainment session is preempted and becomes offloaded while the arriving platooning session is accepted to thus vacated RU. The corresponding entertainment session is thus preempted and becomes offloaded.
4) If $c(s+1) > C$ then the platooning session is offloaded.

Now recall that the system has l AD sessions and m PS sessions, $s = l + m$. Hence, upon arrival of $(l+1)$-th AD session, one of the following outcomes occurs:

1) If $c(s+1) = c(s)$ then the arriving session is accepted for service without allocating additional RU and the capacity of the RUs allocated to the autonomous driving and platooning sessions is equally redistributed among $s+1$ sessions.



(a) *PSpreempt* scheme (b) *ESpreempt* scheme

FIGURE 4. Illustration of considered preemptive priority schemes running where there are insufficient resources to handle an arriving session: $c(l+m)+n=C$.

2) If $c(s+1) > c(s)$ and there is at least one free RU then the session is accepted and an additional RU is allocated.
3) If $c(s+1) > c(s)$ but $C - c(s) \geq 0$ RUs are busy serving the entertainment sessions and $m > 0$ then the arriving autonomous driving session preempts one platooning session and occupies its radio resources. The preempted platooning session becomes offloaded.
4) If $c(s+1) > c(s)$ but $C - c(s) > 0$ RUs are busy serving the entertainment sessions and $m = 0$ then the arriving autonomous driving session preempts an entertainment session and becomes accepted with the allocation of thus vacated RU. The preempted platooning session becomes offloaded.
5) If $c(l+1) > C$ then the arriving session is offloaded.

Note that the *PSpreempt* preemption scheme introduced above implies that the autonomous driving sessions arriving in the system will first preempt the platooning sessions when all of the RUs are occupied. If no platooning sessions are present in the system, entertainment sessions are preempted.

We are also interested in *ESpreempt* scheme that can be defined by modifying the autonomous driving sessions acceptance rules 3) and 4) from the previous list as follows:

3) If $c(s+1) > c(s)$ and $C - c(s) > 0$ RUs are busy serving entertainment sessions then the arriving autonomous driving session preempts an entertainment session and is accepted for service with the allocation of thus vacated RU. The preempted entertainment session becomes offloaded. The service rates of the autonomous driving and platooning sessions increase.
4) If $c(s+1) > c(s) = C$ and $m > 0$ then the arriving session preempts one platooning session and occupies its resources. The preempted platooning session becomes offloaded.

Assume that there are altogether s platooning and autonomous driving sessions. Then, upon a departure of one platooning or autonomous driving session, the RU capacity is redistributed among the remaining platooning and autonomous driving sessions as follows:

1) If $c(s-1) = c(s)$ but the remaining platooning and autonomous driving sessions cannot be served by a smaller number of RUs then no RU is vacated and the capacity of $c(s)$ RUs is redistributed equally among $s-1$ sessions, which then have their service rate increased.
2) On the other hand, if $c(s-1) < c(s)$ then one RU is vacated and the remaining autonomous driving and platooning sessions occupy $c(s-1)$ RUs, which results in a decrease of their service rate.

B. MATRIX ANALYTIC FORMULATION AND SOLUTION

We now proceed with studying the system by using the matrix analytic methods [43]. The analyzed system is classified as a loss system with cascaded preemptive priorities [44]. Let $n(t)$, $m(t)$, and $l(t)$ denote the numbers of ES, PS, and AD sessions in the system, respectively, and let $k(t)$ be the state of the AD sessions MAP model at time $t \geq 0$.

We represent the system at hand by a multidimensional Markov chain $\{X(t) = (n(t), m(t), l(t), k(t)), t \geq 0\}$ defined over the state space

$$\mathcal{X} = \{(n, m, l, k) : n \geq 0, m \geq 0, l \geq 0, 1 \leq k \leq K\}, \quad (1)$$

with the additional constraint of $c(m+l)+n \leq C$.

TABLE 1. Notation used by our mathematical framework.

Parameter	Definition
	System-wide parameters
C	Total number of resource units
M	Max. number of autonomous/platooning sessions in a RU
	Communication traffic patterns for autonomous vehicles
K	Number of states in autonomous driving traffic pattern
$\mathbf{Q}_0$,$\mathbf{Q}_1$,$\mathbf{Q}$	Matrices of order K defining the MAP model
$\mathbf{q}$	Stationary state vector of MAP model
α_{AD}	Average arrival rate of autonomous driving sessions
β_{AD}^{-1}	Mean autonomous driving resource request
α_{PS}	Arrival rate of platooning sessions
β_{PS}^{-1}	Mean platooning resource request
λ	Arrival rate of entertainment sessions
μ^{-1}	Mean duration of an entertainment session
	Performance measures
$\bar{c}$	Average number of occupied resource units
U	Resource utilization of the system
N_{ES}	Mean number of active entertainment sessions
N_{AD}	Mean number of active autonomous driving sessions
N_{PS}	Mean number of active platooning sessions
c_{AD}	Mean number of RUs occupied by autonomous driving
c_{PS}	Mean number of RUs occupied by platooning
B_{AD}	Autonomous driving session offloading probability
B_{PS}^{arr}	Loss probability of platooning sessions upon arrival
B_{ES}^{arr}	Loss probability of entertainment sessions upon arrival
B_{PS}^{pr}	Preemption probability of platooning sessions
B_{ES}^{pr}	Preemption probability of entertainment sessions
T_{AD}	Mean duration of autonomous driving sessions
	Auxiliary notation
k	State of autonomous driving traffic pattern
l	Number of active autonomous driving sessions
m	Number of active platooning sessions
n	Number of active entertainment sessions
$p_{n,m,l,k}$	Stationary probability of system state (n, m, l, k)
$c(l+m)$	RUs occupied by l autonomous and m platooning sessions
$M(n)$	Max. number of autonomous driving and platooning sessions in the system with n active entertainment sessions
$L(n,m)$	Max. number of autonomous driving sessions in the system with n entertainment and m platooning sessions
$\mathbf{1}$	Column vector of ones of size K
$\mathbf{I}$	Identity matrix of order K

The infinitesimal generator $\mathbf{A}$ of the Markov chain $\{X(t), t \geq 0\}$ can be written in the block-tridiagonal form

$$\mathbf{A} = \begin{pmatrix} \mathbf{D}_0 & \mathbf{G}_0 & & & \\ \mathbf{H}_1 & \mathbf{D}_1 & \mathbf{G}_1 & & \\ & \mathbf{H}_2 & \ddots & \ddots & \\ & & \ddots & \mathbf{D}_{C-1} & \mathbf{G}_{C-1} \\ & & & \mathbf{H}_C & \mathbf{D}_C \end{pmatrix}, \tag{2}$$

or $\mathbf{A} = \text{tridiag}(\mathbf{H}_n, \mathbf{D}_n, \mathbf{G}_n)$, $n = 0, \ldots, C$.

The diagonal blocks of $\mathbf{A}$ correspond to the transitions that are not related to entertainment sessions. They are also block-tridiagonal for $n = 0, \ldots, C$ and defined as

$$\mathbf{D}_n = \text{tridiag}(\mathsf{B}_{n,m}, \Delta_{n,m}, \Gamma_{n,m}), \ m = 0, \ldots, M(n), \tag{3}$$

where $M(n) = M \times (C - n)$ is the maximum number of autonomous driving and platooning sessions in the system with n active entertainment sessions.

Matrices $\mathbf{D}_n$ are square of size $(M(C-n)+1)(M(C-n)+2)/2K$, and each of their blocks at the intersection of the block row i and the block column j is a block matrix comprising square matrices of order K and having the block size of $(L(n,i)+1) \times (L(n,j)+1)$, where $L(n,m) = (M(n) - m)$ is the maximum number of autonomous driving sessions in the system with m platooning and n active entertainment sessions.

Let $\mathbf{I}$ denote the identity matrix of order K. The diagonal blocks of $\mathbf{D}_n$ representing transitions related to autonomous driving sessions are block-tridiagonal matrices of the form

$$\Delta_{n,m} = \text{tridiag}(\varphi_{l,m}\beta_{AD}\mathbf{I}, \mathbf{F}_{n,m,l}, \mathbf{Q}_1), \ l = 0, \ldots, L(n,m), \tag{4}$$

where $\mathbf{F}_{n,m,l}$ is given by

$$\begin{cases} \mathbf{Q} - \varphi_{l,m}\beta_{AD}\mathbf{I}, & (n,m,l,k) \in \mathcal{B}_{AD}, \\ \mathbf{Q}_0 - (\varphi_{l,m}\beta_{AD} + \varphi_{m,l}\beta_{PS} + n\mu)\mathbf{I}, & (n,m,l,k) \in \mathcal{B}_{PS} \setminus \mathcal{B}_{AD}, \\ \mathbf{Q}_0 - (\varphi_{l,m}\beta_{AD} + \varphi_{m,l}\beta_{PS} + n\mu + \alpha_{PS})\mathbf{I}, & (n,m,l,k) \in \mathcal{B}_{ES} \setminus \mathcal{B}_{PS}, \\ \mathbf{Q}_0 - (\varphi_{l,m}\beta_{AD} + \varphi_{m,l}\beta_{PS} + n\mu + \alpha_{PS} + \lambda)\mathbf{I} & \text{otherwise}, \end{cases} \tag{5}$$

and $\varphi_{i,j}$ are defined as

$$\varphi_{i,j} = \begin{cases} 0, & i = 0, j = 0, \\ ic(i+j)/(i+j), & i \geq 0, j \geq 0. \end{cases} \tag{6}$$

Note that in (5) we define the subset of states, where autonomous driving, platooning, and entertainment sessions are accepted to the system, i.e.,

$$\begin{aligned} \mathcal{X}_{AD} &= \{(n,m,l,k) \in \mathcal{X} | l < M(0)\}, \\ \mathcal{X}_{PS} &= \{(n,m,l,k) \in \mathcal{X} | l + m < M(n)\} \cup \\ &\quad \{(n,m,l,k) \in \mathcal{X} | n > 0\}, \\ \mathcal{X}_{ES} &= \{(n,m,l,k) \in \mathcal{X} | c(l+m) + n < C\}, \end{aligned} \tag{7}$$

which implies that $\mathcal{B}_{AD}$, $\mathcal{B}_{PS}$, and $\mathcal{B}_{ES}$ in (5) are

$$\mathcal{B}_{AD} = \mathcal{X} \setminus \mathcal{X}_{AD}, \ \mathcal{B}_{PS} = \mathcal{X} \setminus \mathcal{X}_{PS}, \ \mathcal{B}_{ES} = \mathcal{X} \setminus \mathcal{X}_{ES}, \tag{8}$$

such that $\mathcal{B}_{AD} \subset \mathcal{B}_{PS} \subset \mathcal{B}_{ES}$.

The superdiagonal blocks of $\mathbf{D}_n$ correspond to transitions related to platooning session arrivals. These are non-square block matrices with dimensions $(L(n,m)+1) \times (L(n,m+1)+1)$ defined as follows

$$\Gamma_{n,m} = \begin{pmatrix} \text{diag}(\alpha_{PS}\mathbf{I}) \\ \mathbf{0} \end{pmatrix}. \tag{9}$$

The subdiagonal blocks of $\mathbf{D}_n$ represent transitions related to platooning session departures and, consequently, differ depending on the preemption scheme. Define $\mathsf{B}_{n,m} = \mathsf{B}_{n,m}^{PS}$

for *PSpreempt* scheme and $\mathsf{B}_{n,m} = \mathsf{B}^{ES}_{n,m}$ for *ESpreempt* scheme. Let us also denote

$$\Phi_{n,m} = \begin{pmatrix} \varphi_{m,0}\beta_{PS}\mathbf{I} & & & \\ & \varphi_{m,1}\beta_{PS}\mathbf{I} & & \\ & & \ddots & \\ & & & \varphi_{m,L(n,m)-1}\beta_{PS}\mathbf{I} \end{pmatrix}. \quad (10)$$

Now, for all $m = 1, \ldots, M(n)$ we have for *PSpreempt* scheme

$$\mathsf{B}^{PS}_{n,m} = \begin{pmatrix} \Phi_{n,m} & \mathbf{0} \\ & \mathbf{Q}_1 \end{pmatrix}, n = 0, \ldots, C, \quad (11)$$

whereas in the case of *ESpreempt* scheme

$$\mathsf{B}^{ES}_{n,m} = \begin{cases} \mathsf{B}^{PS}_{0,m}, & n = 0, \\ \begin{pmatrix} \Phi_{n,m} & \mathbf{0} \end{pmatrix}, & n = 1, \ldots, C. \end{cases} \quad (12)$$

The superdiagonal blocks of $\mathbf{A}$ correspond to transitions related to entertainment session arrivals. They are block-diagonal non-square matrices of the form

$$\mathbf{G}_n = \begin{pmatrix} \mathrm{diag}(\Lambda_{n,m}) \\ \mathbf{0} \end{pmatrix}, \quad m = 0, \ldots, M(n+1), \quad (13)$$

where $\Lambda_{n,m}$ are of size $(L(n,m)+1) \times (L(n+1,m)+1)$ and

$$\Lambda_{n,m} = \begin{pmatrix} \mathrm{diag}(\lambda\mathbf{I}) \\ \mathbf{0} \end{pmatrix}. \quad (14)$$

Finally, the subdiagonal blocks of $\mathbf{A}$ correspond to transitions related to entertainment session departures. These are block-diagonal non-square matrices given by

$$\mathbf{H}_n = \begin{pmatrix} \mathsf{M}_{n,0} & \Theta_{n,0} & & & \\ & \ddots & \ddots & & \\ & & \mathsf{M}_{n,M(n-1)} & \Theta_{n,M(n-1)} & \mathbf{0} \end{pmatrix}, \quad (15)$$

where blocks $\mathsf{M}_{n,m}$ are of block size $(L(n,m)+1) \times (L(n-1,m)+1)$ with their form depending on the considered preemption scheme. Let us denote $\mathsf{M}_{n,m} = \mathsf{M}^{PS}_{n,m}$ for *PSpreempt* scheme and $\mathsf{M}_{n,m} = \mathsf{M}^{ES}_{n,m}$ *ESpreempt* scheme. Then, for all $n = 1, \ldots, C$ and for *PSpreempt* scheme we have

$$\mathsf{M}^{ES}_{n,m} = \begin{pmatrix} n\mu\mathbf{I} & & & \\ & \ddots\ \mathbf{0} & \mathbf{0} & \\ & n\mu\mathbf{I} & \mathbf{Q}_1 & \mathbf{0} \end{pmatrix}, \ m = 0, \ldots, M(n-1), \quad (16)$$

whereas in the case of *PSpreempt* we arrive at

$$\mathsf{M}^{PS}_{n,m} = \begin{cases} \mathsf{M}^{ES}_{n,0}, & m = 0, \\ \begin{pmatrix} \mathrm{diag}(n\mu\mathbf{I}) & \mathbf{0} \end{pmatrix}, & m = 1, \ldots, M(n-1). \end{cases} \quad (17)$$

Matrices $\Theta_{n,m}$ correspond to transitions related to the preemption of entertainment sessions by autonomous driving sessions. They are of size $(L(n,m)+1) \times (L(n-1, m+1)+1)$ and have only one non-zero block $\alpha_{PS}\mathbf{I}$ located at the intersection of $(L(n,m)+1)$-th row and $(L(n,m)+1)$-th column.

Let $\mathbf{p}_{n,m,l} = (p_{n,m,l,1}, p_{n,m,l,2}, \ldots, p_{n,m,l,K})$, where $0 \le n \le C, 0 \le m \le M(n)$ and $0 \le l \le L(n,m)$, denote the stationary distribution of the Markov chain $\{X(t), t \ge 0\}$, where its elements are given by

$$\begin{aligned} \mathbf{p} &= (\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_C), \\ \mathbf{p}_n &= (\mathbf{p}_{n,0}, \mathbf{p}_{n,1}, \ldots, \mathbf{p}_{n,M(n)}), \\ \mathbf{p}_{n,m} &= (\mathbf{p}_{n,m,0}, \mathbf{p}_{n,m,1}, \ldots, \mathbf{p}_{n,m,L(n,m)}). \end{aligned} \quad (18)$$

Then, the stationary distribution is the solution of

$$\begin{cases} \mathbf{pA} = \mathbf{0}, \\ \mathbf{p1} = 1, \end{cases} \quad (19)$$

which can be found by using, e.g., Gauss elimination [45].

C. PERFORMANCE MEASURES FOR PSPREEMPT

Consider now the metrics of interest. First we determine the system resource utilization defined as $U = \bar{c}/C$, where C is the overall number of RUs, $\bar{c}$ is the mean number of occupied RUs. The latter can be obtained by summing up the number of RUs in all the system states and weighing them with the corresponding stationary state probabilities, $\mathbf{p}$, i.e.,

$$\bar{c} = \sum_{n=0}^{C} \sum_{m=0}^{M(n)} \sum_{l=0}^{L(n,m)} (n + c(m+l))\mathbf{p}_{n,m,l}\mathbf{1}, \quad (20)$$

where $\mathbf{1}$ is the unit vector.

The mean number of active autonomous driving and platooning sessions currently in service can be obtained similarly. The key difference is that we now need to account for the specific types of sessions, i.e.,

$$\begin{aligned} N_{AD} &= \sum_{n=0}^{C} \sum_{m=0}^{M(n)} \sum_{l=1}^{L(n,m)} l\mathbf{p}_{n,m,l}\mathbf{1}, \\ N_{PS} &= \sum_{n=0}^{C} \sum_{m=1}^{M(n)} \sum_{l=0}^{L(n,m)} m\mathbf{p}_{n,m,l}\mathbf{1}. \end{aligned} \quad (21)$$

Consider now the offloading-related metrics. Since autonomous driving sessions are associated with the highest priority, they are never preempted. Hence, these sessions are only offloaded when there are insufficient resources available upon their arrival. Therefore, the autonomous driving session offloading probability is given by

$$B_{AD} = \frac{1}{\alpha_{AD}}\mathbf{p}_{0,0,L(0,0)}\mathbf{Q}_1\mathbf{1} = 1 - \frac{c_{AD}\beta_{AD}}{\alpha_{AD}}. \quad (22)$$

Both platooning and entertainment sessions can be preempted during service. Hence, their total offloading probabilities can be written as a sum of the two terms corresponding to offloading probability upon arrival, B^{arr}_{PS} and B^{arr}_{ES}, and preemption probability, B^{pr}_{PS} and B^{pr}_{ES}, i.e.,

$$B_{PS} = B^{arr}_{PS} + B^{pr}_{PS}, \ B_{ES} = B^{arr}_{ES} + B^{pr}_{ES}. \quad (23)$$

The platooning session preemption probability is different for the considered schemes. Accounting for higher priority

session arrivals during the service of platooning sessions, we arrive at the following for *PSpreempt*:

$$B_{PS}^{pr} = \frac{1}{\alpha_{PS}} \sum_{n=0}^{C-1} \sum_{m=1}^{M(n)} \mathbf{p}_{n,m,L(n,m)} \mathbf{Q}_1 \mathbf{1}. \tag{24}$$

The platooning session offloading probability upon arrival is

$$B_{PS}^{arr} = \sum_{m=0}^{M(0)} \mathbf{p}_{0,m,L(0,m)} \mathbf{1}. \tag{25}$$

Substituting (24) and (25) into (23) and simplifying, we obtain the total offloading probability of platooning sessions as

$$B_{PS} = B_{PS}^{arr} + B_{PS}^{pr} = 1 - \frac{c_{PS} \beta_{PS}}{\alpha_{PS}}. \tag{26}$$

Consider now the entertainment session preemption probability. Recall that these sessions can be preempted by both autonomous driving sessions and platooning sessions. Denote these probabilities by $B_{ES}^{pr,AD}$ and $B_{ES}^{pr,PS}$, respectively. Hence, the total preemption probability of entertainment sessions can be written as $B_{ES}^{pr} = B_{ES}^{pr,PS} + B_{ES}^{pr,AD}$, where its components can be obtained by accounting for platooning and autonomous driving session arrivals in the states having non-zero numbers of entertainment sessions. Considering the specifics of *PSpreempt*, as detailed in subsection IV-A, we arrive at the following:

$$\begin{aligned} B_{ES}^{pr,PS} &= \frac{\alpha_{PS}}{\lambda} \sum_{n=1}^{C} \sum_{m=0}^{M(n)} \mathbf{p}_{n,m,L(n,m)} \mathbf{1}, \\ B_{ES}^{pr,AD} &= \frac{1}{\lambda} \sum_{n=1}^{C} \mathbf{p}_{n,0,M(n)} \mathbf{Q}_1 \mathbf{1}. \end{aligned} \tag{27}$$

The entertainment session offloading probability upon arrival is

$$B_{ES}^{arr} = \sum_{n=0}^{C} \sum_{m=0}^{M(n)} \sum_{l=\max\{L(n,m)-M+1,0\}}^{L(n,m)} \mathbf{p}_{n,m,l} \mathbf{1}. \tag{28}$$

Substituting (27) and (28) into (23) and simplifying, we obtain the total offloading probability of entertainment sessions as

$$B_{ES} = B_{ES}^{arr} + B_{ES}^{pr} = 1 - \frac{\mu N_{ES}}{\lambda}. \tag{29}$$

D. PERFORMANCE MEASURES FOR ESPREEMPT

In principle, the performance indicators for *PSpreempt* are derived similarly to the ones for *ESpreempt*. However, there are some important differences, as detailed below. Firstly, as now $\mathsf{B}_{n,m} = \mathsf{B}_{n,m}^{ES}$, equation (12) is used for $\mathsf{B}_{n,m}$ and, consequently, for all the expressions having $\mathsf{B}_{n,m}$. Secondly, $\mathsf{M}_{n,m}$ is also calculated differently for *ESpreempt*. Hence, all the equations in subsection IV-B using $\mathsf{M}_{n,m}$ should apply $\mathsf{M}_{n,m} = \mathsf{M}_{n,m}^{ES}$, where $\mathsf{M}_{n,m}^{ES}$ is given in (17). The rest of the matrix analysis in subsection IV-B holds for both schemes.

TABLE 2. Key parameters for our numerical study.

Carrier frequency of primary radio access technology	28 GHz
Bandwidth allocated for autonomous vehicles with network slicing	200 MHz
Intensity of entertainment traffic	100 Mbit/s
Intensity of platooning traffic	20 Mbit/s
Intensity of autonomous driving traffic	0...100 Mbit/s
Autonomous driving traffic pattern (MAP with two states, $K = 2$)	*Low-intensity regime* and *High-intensity regime*
Data rate of a single AD session	10 Mbit/s

Another set of changes is related to the way how certain individual performance indicators are derived. First of all, the platooning session preemption probability, B_{PS}^{pr}, is different for the considered schemes. We determine it by accounting for higher priority session arrivals during the service of platooning sessions. Considering the states with non-zero platooning sessions, we arrive at the following for *ESpreempt* to replace (24):

$$B_{PS}^{pr} = \frac{1}{\alpha_{PS}} \sum_{m=1}^{M(n)} \mathbf{p}_{0,m,L(n,m)} \mathbf{Q}_1 \mathbf{1}. \tag{30}$$

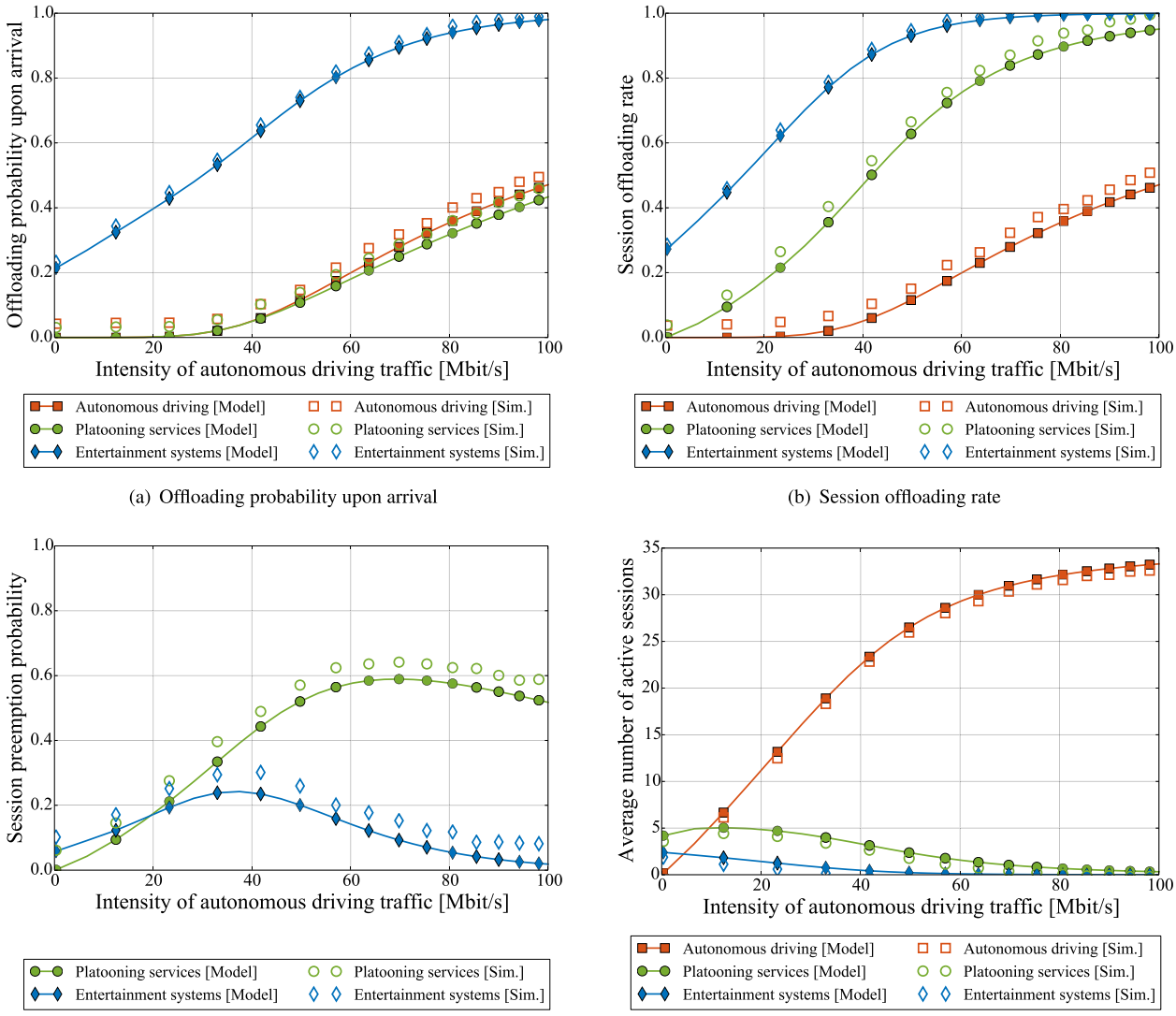
Consider now the entertainment session preemption probability. Recall that these sessions can be preempted by both autonomous driving sessions and platooning sessions and that the total preemption probability of entertainment sessions can be written as $B_{ES}^{pr} = B_{ES}^{pr,PS} + B_{ES}^{pr,AD}$. Here, the second term is different for *ESpreempt* as an entertainment session has now greater chances to be preempted by an autonomous driving session (see Fig. 3). Hence, the second part of (27) is no longer in use, and instead we have:

$$B_{ES}^{pr,AD} \frac{1}{\lambda} \sum_{n=1}^{C} \sum_{m=0}^{M(n)} \mathbf{p}_{n,m,L(n,m)} \mathbf{Q}_1 \mathbf{1}. \tag{31}$$

The rest of the analysis detailed in subsection IV-C also holds for *ESpreempt*. We numerically elaborate on the selected important characteristics in the following section.

V. MAIN NUMERICAL RESULTS

In this section, we report on the illustrative numerical results that characterize the main metrics of interest in the considered system, where multi-service data streams associated with the fleet of intelligent autonomous vehicles are handled simultaneously by the 5G mmWave cellular network. We first discuss the impact of the intensity of the mission-critical autonomous driving sessions. We then proceed with quantifying the impact of dynamic resource reallocation by comparing the key performance indicators for the two introduced prioritization schemes. Further, we investigate the effects of the maximum number of sessions in a resource unit. Finally, we assess the implications of the temporal variations in the mission-critical autonomous driving sessions for the considered vehicular network.



(a) Offloading probability upon arrival

(b) Session offloading rate

(c) Session preemption probability

(d) Average number of active sessions

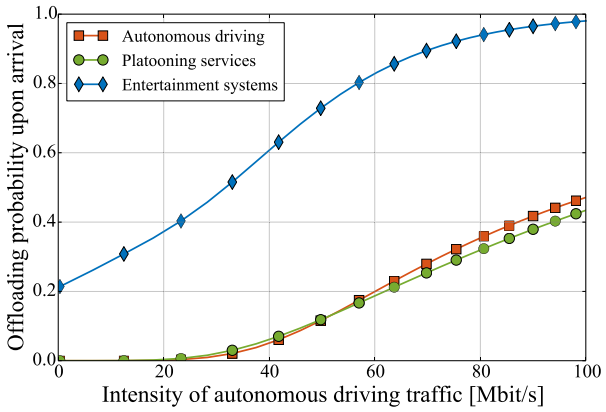
FIGURE 5. Offloading and performance metrics for *PSpreempt* strategy.

To validate our mathematical framework and confirm the accuracy of the produced numerical results, we verify the findings obtained with our analytical methodology with those delivered by system-level simulations. For this purpose, we adapt our in-house system-level evaluation software that carefully accounts for the relevant features of the target vehicular scenario as well as includes the essential components of the emerging 5G cellular networks [46]. Following Section III, we model a single-cell scenario and assume that the network is properly configured to maintain reliable session-level communications as long as there are sufficient radio resources provisioned. The major parameters of our numerical study are summarized in Table 2.

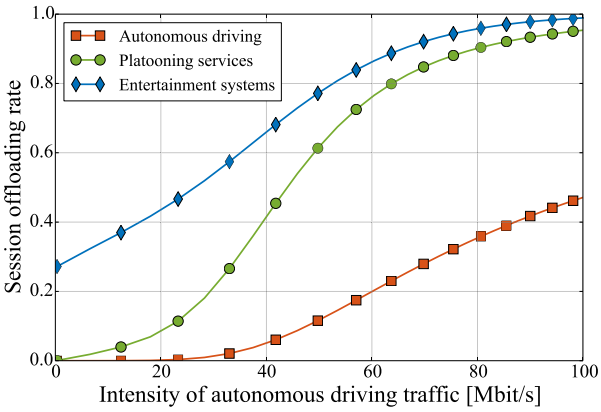
1) EFFECTS OF AUTONOMOUS DRIVING TRAFFIC

We start with Fig. 5 that presents the offloading probabilities and performance metrics for the *PSpreempt* scheme as a function of the average intensity of the mission-critical vehicular communication. First, Fig. 5(a) shows a positive effect of mission-critical traffic for autonomous driving on the offloading probability upon arrival for all the data streams. As Fig. 5(b) illustrates the session offloading rate, we clearly observe the effect of data stream prioritization: the curve for the mission-critical vehicular communication grows much slower than those for other categories of data transmissions. In contrast, PS curve grows much faster than that in Fig. 5(a), as platooning sessions are the first candidate to be preempted during service in the case where a session with a higher priority appears.

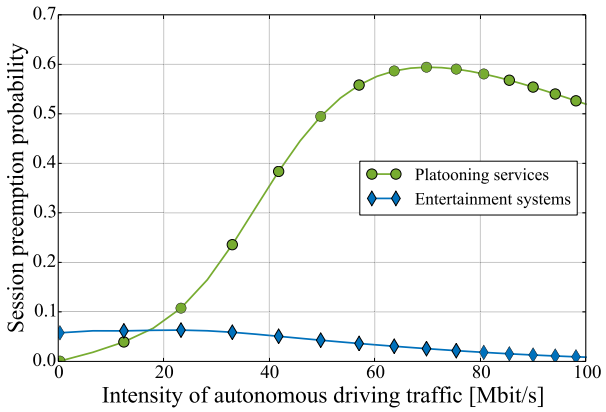
It is also important to note that the instantaneous traffic volume in mission-critical transmissions may temporarily exceed the network capacity, so the offloading rate for the mission-critical data becomes greater than zero once the session intensity exceeds a certain limit. Moreover, at higher intensities, the critical communication dominates within the



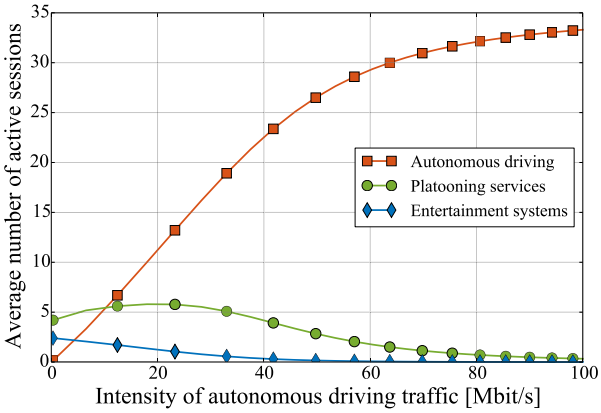
(a) Offloading probability upon arrival



(b) Session offloading rate



(c) Session preemption probability



(d) Average number of sessions

FIGURE 6. **Offloading and performance metrics for *ESpreempt* scheme.**

network having the session offloading rate of around 0.5 vs. around 0.98 for other data streams.

The above conclusions are further complemented by Fig. 5(c), which demonstrates the preemption probability using the same axes (since the AD traffic has the highest priority; hence, its preemption probability is strictly zero). It is also worth mentioning that the curves for both non-priority data streams have their maxima at a certain intensity of the AD traffic. This introduces a transition from (i) *low-intensity regime*, where a non-priority session is most likely admitted for service but will probably be offloaded to release the radio resources for a newly-arrived priority session, to (ii) *high-intensity regime*, where non-priority sessions are highly unlikely to be admitted into the system as the resources are occupied by the autonomous driving traffic.

These observations are also confirmed by Fig. 5(d), which studies the average number of sessions in the network individually for each of the traffic categories. Hence, Fig. 5(d) clearly highlights that at higher intensities of the autonomous driving data, this type of traffic tends to dominate in the system. Fig. 5 (as well as the following figures) also confirm a tight agreement between the analytical results and those obtained with computer simulations, thus advocating for the accuracy

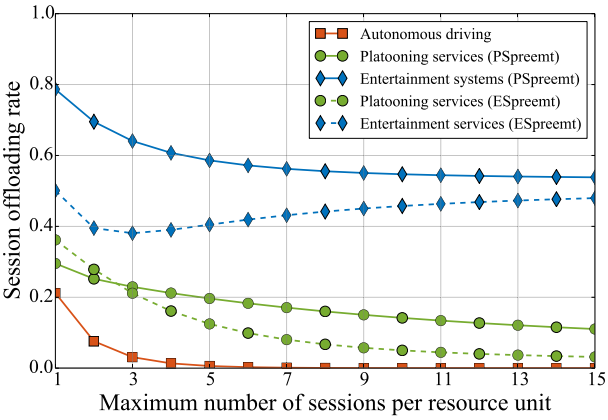
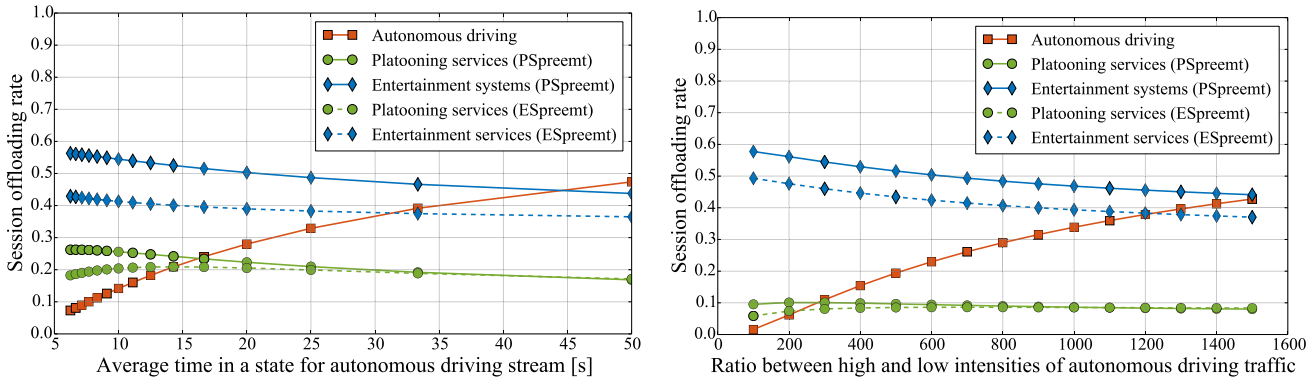


FIGURE 7. **Session offloading rate as a function of *M*.**

of our contributed mathematical framework. A slight mismatch in the numerical results is explained by the simplifying assumptions on the radio channel operation introduced by the analytical framework (see Sections III and IV) for the sake of modeling tractability. A similar agreement is observed across a wide range of network conditions for all the configurations of interest. Therefore, for the simplicity of further



(a) Average time spent in a single state

(b) Ratio of intensities

FIGURE 8. Session offloading rate as a function of autonomous driving traffic characteristics.

interpretation, we resort to discussing the analytical results in what follows.

2) EFFECTS OF TRAFFIC PRIORITIZATION IN 5G CELLULAR NETWORKS

We now continue with Fig. 6, which offers a similar analysis to that in Fig. 5, but for the second priority strategy discussed in Section IV – *ESpreempt*. The major trends and core dynamics observed in Fig. 6 for the most part repeat those visible in Fig. 5; however, the non-priority data streams are susceptible to lower session offloading rates and lower preemption probabilities, since the network resources are utilized more efficiently. In contrast, the indicators for the autonomous driving traffic are nearly identical to those displayed in Fig. 5.

3) EFFECTS OF THE NUMBER OF SESSIONS IN A RESOURCE UNIT

We also assess the effects of the number of continuous sessions supported by a single resource unit, M, on the session success probability in the considered 5G vehicular deployment. To this end, Fig. 7 quantifies the average session success rate as a function of M for the fixed intensities of all the data streams. We first notice that larger M leads to lower session offloading rate as more sessions are packed into a single radio resource unit on average. We further observe that for a given set of intensities, any value of M above 7 leads to the extremely high reliability of the mission-critical autonomous driving data transmissions (as session offloading rate tends to 0). Hence, for any particular set of traffic intensities, it is possible to find the respective M satisfying the target QoS requirements for the offloading rate of the autonomous driving sessions.

In addition, Fig. 7 allows us to better compare the prioritization strategies *PSpreempt* and *ESpreempt*. We specifically note that for lower values of M (namely, $M = 1$ and $M = 2$), the first strategy brings a distinct advantage for the entertainment data stream, while resulting in a slightly increased offloading rate for platooning transmissions. In its turn, *ESpreempt* shows the opposite effect by preferring the platooning traffic over the entertainment service. Meanwhile, with a growing value of M the offloading rate for the platooning data decreases faster for *PSpreempt*, while complemented by a slower decrease in the reliability of entertainment applications. At the same time, the *ESpreempt* strategy is fairer at lower values of M since the difference between the offloading rates for the non-priority streams is not drastic.

4) EFFECTS OF AUTONOMOUS DRIVING TRAFFIC VARIATIONS

We finally highlight the effects of the variations in the arrival intensity of the AD traffic. We thus analyze the session offloading rate in Fig. 8(a) as a function of the average time that the AD data stream spends in *low* and *high* intensity regimes. Here, we observe that any instability in the autonomous driving data streams has a profoundly negative impact on the AD data transmissions themselves: the session offloading rate increases from 0.07 up to 0.48. At the same time, increased durations of *low* and *high* intensities have a positive effect on both entertainment and platooning streams.

The above behavior can be explained by the fact that the intensity of mission-critical data flows in *high* rate regime is sufficient to preempt most of the non-priority transmissions. Meanwhile, as the time duration in each of the states increases, the non-priority sessions have a certain chance of being served in the *low* rate regime. We also note the effect of the prioritization strategy, where *ESpreempt* demonstrates a notable advantage over *PSpreempt* due to the fact that offloading a single long and bandwidth-hungry entertainment session releases much more resources than offloading a short and lightweight platooning transmission. We also note that the behavior of the session offloading rate is similar for platooning and entertainment sessions, that is, the session offloading probabilities will coincide for entertainment sessions as well if one considers higher values of the mean time spent in a state. The reason is that ultimately autonomous driving session will occupy all the provided resources and both schemes, *ESpreempt* and *PSpreempt*, should thus demonstrate similar performance.

In more detail, Fig. 8(b) illustrates the implications of varying the ratio between the *high* and the *low* state intensities of the mission-critical traffic while keeping the average intensity constant. Here, we learn that the trends are similar to those in Fig. 8(a), as a higher ratio between the said intensities leads to massive offloading of the autonomous driving sessions once the state of the *high* intensity is reached. Moreover, the higher ratio also imposes more favorable conditions for other categories of traffic as they can, for the most part, become served during the *low* intensity durations. At the same time, entertainment and platooning sessions are likely to become offloaded during the *high* intensity durations.

VI. CONCLUSION

In this article, we assessed in detail the 5G mmWave network performance when serving multi-service data streams pertaining to the fleets of intelligent autonomous vehicles. For this analysis, a comprehensive mathematical framework has been developed based on the queuing system with multiple arrival flows of different priorities and a preemption mechanism. The framework is capable of accounting for the heterogeneous nature of competing data flows coming to/from the connected vehicles as well as for the dynamic resource (re-)allocation in the 5G mmWave cellular networks to prioritize mission-critical data transmissions over consumer-centric communications. Our framework also allows to estimate the number of data sessions that need to be offloaded onto other RATs or dropped in the presence of intelligent admission control and dynamic prioritization mechanisms. The latter enable smarter provisioning of radio resources in prospective 5G deployments.

The performed numerical study particularly reveals that a high time-variance in the intensity of the autonomous driving sessions is one of the primary challenges leading to a notable decrease in system performance. The evaluation confirms that dynamic prioritization of the mission-critical autonomous driving sessions offers a notable positive impact on the reliability of vehicular communications but at the same time leads to a degradation in the performance levels for other traffic categories. Finally, our investigation also illustrates that the *ESpreempt* scheme is generally characterized by a 5–30% lower offloading rate and thus becomes preferable over the *PSpreempt* approach.

The mathematical framework contributed in the paper is flexible and can be extended to capture other traffic types, different from the ones considered in this work. Similarly, the order and the type of preemption rules can be altered to allow for capturing various use cases of interest. However, we note that the solution complexity of the model increases as more traffic classes are added to the system.

ACKNOWLEDGMENT

The authors would like to offer special thanks to Prof. Mario Gerla, University of California, Los Angeles (UCLA), who, although no longer with us, contributed his valuable input to this article.

REFERENCES

[1] G. Fodor, H. Do, S. A. Ashraf, R. Blasco, W. Sun, M. Belleschi, and L. Hu, "Supporting enhanced vehicle-to-everything services by LTE release 15 systems," *IEEE Commun. Standards Mag.*, vol. 3, no. 1, pp. 26–33, Mar. 2019.

[2] H. Peng, L. Liang, X. Shen, and G. Y. Li, "Vehicular communications: A network layer perspective," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1064–1078, Feb. 2019.

[3] S. A. A. Shah, E. Ahmed, M. Imran, and S. Zeadally, "5G for vehicular communications," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 111–117, Jan. 2018.

[4] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, Jr., "Millimeter wave vehicular communications: A survey," *Found. Trends Netw.*, vol. 10, no. 1, pp. 1–113, 2016.

[5] J. A. F. F. Dias, J. J. P. C. Rodrigues, and L. Zhou, "Cooperation advances on vehicular communications: A survey," *Veh. Commun.*, vol. 1, no. 1, pp. 22–32, Jan. 2014.

[6] "5G Automotive Vision," 5G-PPP, White Paper, Oct. 2015. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf

[7] E. Lee, E.-K. Lee, M. Gerla, and S. Oh, "Vehicular cloud networking: Architecture and design principles," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 148–155, Feb. 2014.

[8] A. Kousaridas, D. Medina, S. Ayaz, and C. Zhou, "Recent advances in 3GPP networks for vehicular communications," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Sep. 2017, pp. 91–97.

[9] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[10] N. Kumar, R. Iqbal, S. Misra, and J. J. P. C. Rodrigues, "Bayesian coalition game for contention-aware reliable data forwarding in vehicular mobile cloud," *Future Gener. Comput. Syst.*, vol. 48, pp. 60–72, Jul. 2015.

[11] R. Molina-Masegosa and J. Gozalvez, "LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range Vehicle-to-Everything communications," *IEEE Veh. Technol. Mag.*, vol. 12, no. 4, pp. 30–39, Dec. 2017.

[12] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, Jr., "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[13] E. Ahmed, I. Yaqoob, A. Gani, M. Imran, and M. Guizani, "Internet-of-Things-based smart environments: State of the art, taxonomy, and open research challenges," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 10–16, Oct. 2016.

[14] S. Kassir, G. D. Veciana, N. Wang, X. Wang, and P. Palacharla, "Enhancing cellular performance via vehicular-based opportunistic relaying and load balancing," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Apr. 2019, pp. 91–99.

[15] A. Argyriou, K. Poularakis, G. Iosifidis, and L. Tassiulas, "Video delivery in dense 5G cellular networks," *IEEE Netw.*, vol. 31, no. 4, pp. 28–34, Jul. 2017.

[16] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[17] C. Quadros, A. Santos, M. Gerla, and E. Cerqueira, "QoE-driven dissemination of real-time videos over vehicular networks," *Comput. Commun.*, vols. 91–92, pp. 133–147, Oct. 2016.

[18] V. Petrov, M. A. Lema, M. Gapeyenko, K. Antonakoglou, D. Moltchanov, F. Sardis, A. Samuylov, S. Andreev, Y. Koucheryavy, and M. Dohler, "Achieving end-to-end reliability of mission-critical traffic in softwarized 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 485–501, Mar. 2018.

[19] T. Taleb, K. Samdanis, and A. Ksentini, "Supporting highly mobile users in cost-effective decentralized mobile operator networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 7, pp. 3381–3396, Sep. 2014.

[20] A. E. Kamal, M. Imran, H.-H. Chen, and A. V. Vasilakos, "Survivability strategies for emerging wireless networks," *Comput. Netw.*, vol. 128, pp. 1–4, Dec. 2017.

[21] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3850–3860, Jun. 2018.

[22] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for low-latency vehicular communications: An effective capacity perspective," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 905–917, Apr. 2019.

[23] A. Ksentini, T. Taleb, and K. B. Letaif, "QoE-based flow admission control in small cell networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2474–2483, Apr. 2016.
[24] R. Tang, J. Zhao, H. Qu, and Z. Zhang, "User-centric joint admission control and resource allocation for 5G D2D extreme mobile broadband: A sequential convex programming approach," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1641–1644, Jul. 2017.
[25] M. Simsek, D. Zhang, D. Ohmann, M. Matthe, and G. Fettweis, "On the flexibility and autonomy of 5G wireless networks," *IEEE Access*, vol. 5, pp. 22823–22835, Jun. 2017.
[26] R. Vannithamby and S. Talwar, *Towards 5G: Applications, Requirements and Candidate Technologies*. Hoboken, NJ, USA: Wiley-Blackwell, 2017.
[27] S. Jangsher and V. O. K. Li, "Resource allocation in cellular networks employing mobile femtocells with deterministic mobility," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 819–824.
[28] V. Petrov, A. Samuylov, V. Begishev, D. Moltchanov, S. Andreev, K. Samouylov, and Y. Koucheryavy, "Vehicle-based relay assistance for opportunistic crowdsensing over narrowband IoT (NB-IoT)," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3710–3723, Oct. 2018.
[29] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, May 2017.
[30] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.
[31] V. Begishev, V. Petrov, A. Samuylov, D. Moltchanov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Resource allocation and sharing for heterogeneous data collection over conventional 3GPP LTE and emerging NB-IoT technologies," *Comput. Commun.*, vol. 120, pp. 93–101, May 2018.
[32] A. S. Shafigh, S. Glisic, and B. Lorenzo, "Dynamic network slicing for flexible radio access in Tactile Internet," in *Proc. GLOBECOM-IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–7.
[33] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 80–87, May 2017.
[34] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN—Key technology enablers for 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2468–2478, Nov. 2017.
[35] Y. Zhong, X. Ge, H. H. Yang, T. Han, and Q. Li, "Traffic matching in 5G ultra-dense networks," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 100–105, Aug. 2018.
[36] *Dynamic End-To-End Network Slicing For 5G*, Nokia, Espoo, Finland, 2017.
[37] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 210–217, Mar. 2018.
[38] L. Pierucci, "The quality of experience perspective toward 5G technology," *IEEE Wireless Commun.*, vol. 22, no. 4, pp. 10–16, Aug. 2015.
[39] A. Klemm, C. Lindemann, and M. Lohmann, "Modeling IP traffic using the batch Markovian arrival process," *Perform. Eval.*, vol. 54, no. 2, pp. 149–173, Oct. 2003.
[40] S. R. Chakravarthy, "Markovian arrival processes," in *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken, NJ, USA: Wiley, 2010.
[41] W. Cai and V. C. M. Leung, "Multiplayer cloud gaming system with cooperative video sharing," in *Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci. (IEEE CloudCom)*, Dec. 2012, pp. 640–645.
[42] H. Feng, J. Llorca, A. M. Tulino, and A. F. Molisch, "On the delivery of augmented information services over wireless computing networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.
[43] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia, PA, USA: SIAM, 1999.
[44] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*. Hoboken, NJ, USA: Wiley, 1976.
[45] C. D. Meyer, *Applied Linear Algebra and Matrix Analysis*. Philadelphia, PA, USA: SIAM, 2000.
[46] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5G multi-RAT LTE-WiFi ultra-dense small cells: Performance dynamics, architecture, and trends," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1224–1240, Jun. 2015.

VITALY PETROV (Student Member, IEEE) received the M.Sc. degree in information systems security from SUAI University, St Petersburg, Russia, in 2011, and the M.Sc. degree in communications engineering from the Tampere University of Technology, Tampere, Finland, in 2014. He is currently pursuing the Ph.D. degree with Tampere University, working on enabling mmWave and terahertz band communications for beyond 5G wireless networks. His research interests include the Internet of Things, mmWave/THz band communications, and network reliability and security. He was a recipient of the IEEE VTC-Fall 2015 Best Student Paper Award, the IEEE WCNC 2017 Best Poster Award, and the IEEE Finland Best Student Journal Paper Award 2018.

NATALIA YARKINA received the M.Sc. degree (Hons.) in applied mathematics and informatics and the Cand.Sc. degree in physics and mathematics from RUDN University, Moscow, Russia, in 2004 and 2007, respectively, and the M.Sc. degree (Hons.) in translation studies from the University of Aberdeen, U.K., in 2016. Her research interests include the application of queueing and teletraffic theory to performance evaluation of communication networks and business processes analysis, as well as telecommunication business management issues.

DMITRI MOLTCHANOV received the M.Sc. and Cand.Sc. degrees from the St. Petersburg State University of Telecommunications, Russia, in 2000 and 2003, respectively, and the Ph.D. degree from the Tampere University of Technology, in 2006. He is currently a University Lecturer with the Faculty of Information Technology and Communication Sciences, Tampere University, Finland. He has (co)authored over 150 publications. His current research interests include 5G/5G+ systems, ultra-reliable low-latency service, the industrial IoT applications, mission-critical V2V/V2X systems, and blockchain technologies.

SERGEY ANDREEV (Senior Member, IEEE) received the Specialist and Cand.Sc. degrees from SUAI, in 2006 and 2009, respectively, the Ph.D. degree from TUT, in 2012, and the Dr.Habil. degree from SUAI, in 2019. He was a Visiting Postdoctoral Researcher with the University of California, Los Angeles, USA, from 2016 to 2017, and a Visiting Senior Research Fellow with the King's College London, U.K., from 2018 to 2020. He is currently an Associate Professor of communications engineering and an Academy Research Fellow with Tampere University, Finland. He has (co)authored more than 200 published research works on the intelligent IoT, mobile communications, and heterogeneous networking.

KONSTANTIN SAMOUYLOV received the Ph.D. degree from Moscow State University and the Doctor of Sciences degree from the Moscow Technical University of Communications and Informatics. In 1996, he became the Head of the Telecommunication Systems Department, RUDN University, Russia, and later, in 2014, he became the Head of the Department of Applied Informatics and Probability Theory. He has written more than 150 scientific and technical articles and five books. His current research interests include performance analysis of 4G networks (LTE, WiMAX), signaling network (SIP) planning, and cloud computing.