A Method for Cross-Layer Analysis of Transmit Buffer Delays in Message Index Domain

Sami Akın and Markus Fidler, *Senior Member, IEEE*

***Abstract*—In data transmission systems, quality-of-service constraints are commonly defined in the form of buffer overflow probability or delay violation probability at a transmitter buffer. Some of the studies that employ the large-deviation principle have taken the buffer overflow probability as the quality-of-service constraint and performed the associated analyses in the time domain. The delay violation probability has been investigated through the buffer overflow probability given that there exists a constant service rate from or a constant data arrival rate at a buffer. These studies cultivated the concepts of effective bandwidth and effective capacity, respectively. Different from the existing studies, we investigate the performance of a transmitter buffer in the message index domain rather than the time domain by taking the waiting time (buffering delay) as the primary quality-of-service constraint. We characterize the waiting time violation probability when both the data arrival and service processes are stochastic, and provide two new concepts: effective interarrival time and effective service time, which are the duals of effective bandwidth and effective capacity, respectively, in the message index domain. The effective interarrival time of a data arrival process determines the maximum constant service time for a message that can sustain the arrival process under a stochastic waiting time constraint, and the effective service time of a data service process determines the minimum constant interarrival time between successive messages arriving at a buffer that the service process can sustain. We show that we can obtain the effective capacity of a service process or the effective bandwidth of an arrival process through the effective service time or the effective interarrival time of the corresponding process, respectively, in cases where it is difficult to formulate the effective capacity and the effective bandwidth without numerical techniques or particular assumptions. Noting that our proposed techniques can be applied in vehicular communication scenarios, e.g., highways, urban areas, and rural areas, we finally analyze a typical data dissemination and collection task in vehicular networks using a broadcast downlink and a slotted Aloha uplink transmission.**

***Index Terms*—Cross-layer analysis, large-deviation principle, quality-of-service, effective capacity, effective bandwidth, effective service time, effective inter-arrival time, message index domain.**

Manuscript received April 4, 2017; revised August 9, 2017 and October 11, 2017; accepted November 6, 2017. Date of publication January 1, 2018; date of current version March 15, 2018. This work was supported by the European Research Council under Starting Grant 306644. The review of this paper was coordinated by Prof. C. Assi. *(Corresponding author: Sami Akın.)*

The authors are with the Institute of Communications Technology, Leibniz Universität Hannover, Hanover 30167, Germany (e-mail: sami.akin@ikt.uni-hannover.de; markus.fidler@ikt.uni-hannover.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2017.2772915

I. Introduction

With communication devices progressively taking part in our daily lives, the need for faster and more reliable data transmission techniques is growing. Therefore, investigating delay-intolerant transmission models is fundamental. In the view of this motivation, cross-layer performance measures that can form a bridge between the limits in the physical layer and the buffer overflow or delay violation performance in the data-link layer are important. Herein, some seminal works have focused on the large-deviation principle and proposed effective bandwidth. The effective bandwidth of a stochastic data arrival process at a transmitter buffer can be defined as the minimum necessary constant service rate from the buffer that can support the arrival process under a buffer overflow probability constraint [1]–[5]. The effective bandwidth of relevant data arrival processes has been investigated in various studies. For instance, Elwalid *et al.* showed that it is possible to assign a notional effective bandwidth to each of the general Markovian data traffic sources at a buffer, which is an explicitly identified and simply computed quantity with provably correct properties in the natural asymptotic regime of small loss probabilities [6]. Effective bandwidth was studied in bandwidth sharing mechanisms as well. Mao *et al.* studied a multi-queue generalized processor sharing system with arrivals that are Markov modulated fluid processes and derived the effective bandwidth values of these arrivals [7]. Likewise, the authors in [8] investigated variable bit rate video sources with real-time constraints and presented a consistent estimator for the effective bandwidth of an arrival process with its confidence interval. Effective bandwidth also took place in other research agendas as well. Li *et al.* established a link between effective bandwidth and network calculus, and expressed effective bandwidth within the framework of a probabilistic version of network calculus, where both data arrival and service processes are specified in terms of probabilistic bounds [9].

The effective bandwidth concept fits well into transmission scenarios where data service rates from a buffer are almost constant (or constant for a long period of time) but data arrival processes at the same buffer show a stochastic nature. In wireless communications, on the other hand, the stochastic nature of wireless service channels comes into sight more when compared to data arrival processes. At this juncture, effective capacity, the dual of effective bandwidth with respect to the buffer, emerges as a measure to indicate the achievable performance levels in wireless channels. The concept, which can be interpreted as the maximum constant data arrival rate at a

buffer that a random service process can support under certain quality-of-service constraints, was initially provided in the work of Chang [3, Example 2.5] and leveraged in wireless communications under the term *effective capacity* by Wu and Negi in [10], [11]. Following the advent of effective capacity, the wireless communications research community welcomed the notion with a significant recognition. Moreover, the effective capacity concept established a bridge between the physical and data-link layers; as a result, it was used in resource allocation problems in the physical layer while considering the performance metrics in the data-link layer such as buffer overflow and delay violation probabilities. In [12], considering an ON-OFF wireless channel model, which is independent and identically distributed across time and users, the authors studied a multi-user formulation of effective capacity, and showed that the effective capacity in the channel decreases with the increasing burstiness in the channel. Furthermore, Liu *et al.* analyzed the effective capacity of a class of multiple-antenna systems and evaluated the system performance in low signal-to-noise ratio regimes [13]. Similarly, focusing on multiple-antenna systems, the authors in [14] maximized the effective capacity of multiple-input multiple-output systems with channel covariance feedback and showed that the stricter the delay requirement is, the more the spatial degrees of freedom are used to avoid low instantaneous transmission rates. Likewise, Femenias *et al.* combined adaptive modulation and coding in the physical layer with an automatic-repeat-request protocol in the data-link layer and obtained the effective capacity of a system by employing a multidimensional physical layer Markov model [15]. Moreover, effective capacity was implemented in cognitive radio research as well. Particularly, the authors obtained the effective capacity of a network, which is composed of secondary users, and determined the power allocation policies that maximize the effective capacity of the secondary users' relay channels [16]. In addition, the authors in [17] studied the effective capacity of cognitive radios where the secondary users perform transmission in a multi-band environment. For further relevant channel models, we refer the interested readers to [18]–[22], in which effective capacity was examined.

In the aforementioned studies, the research was based on a quality-of-service constraint in the form of buffer overflow probability in steady-state. Specifically, letting $q(t)$ be the number of messages (bits or packets) in a data buffer at time instant t, the research of effective bandwidth and effective capacity provided an approximation for the steady-state buffer overflow probability $\Pr\{q(t \to \infty) \geq q_{\text{th}}\}$, where q_{th} is the desired buffer threshold. In other words, effective bandwidth was addressed as the minimum service rate that can sustain the desired buffer overflow probability constraint for a given large buffer threshold when the data arrival process has a stochastic nature. Likewise, effective capacity was introduced as the maximum data arrival rate that can guarantee the desired buffer overflow probability constraint for a given large buffer threshold when the service process has randomly varying transmission rates. Basically, employing effective bandwidth or effective capacity, we can determine certain settings for a service process or an arrival process, respectively, with a desired buffer overflow probability and a defined buffer threshold.

Meanwhile, based on the research established around the buffer overflow probability constraint, the delay violation probability of first-come first-served systems was considered under the assumption of either a constant data arrival rate or a constant service rate. For instance, considering the relation $a \cdot d(n) = q(t)$, where a is the constant data arrival rate and $d(n)$ is the delay experienced by message n that leaves the buffer at time instant t [23], the steady-state delay violation probability of a message in the buffer can be expressed as a function of the steady-state buffer overflow probability [10], [13], [19]. Similarly, when there exists a stochastic data arrival process that is served from a buffer at a constant service rate c, we know that the messages $q(t)$ that are in the buffer at t are served from the buffer in $\frac{q(t)}{c}$ time units so that the delay experienced by message n that enters the buffer at t is $d(n) = \frac{q(t)}{c}$. To the best of our knowledge, an approach that precisely treats the delay violation probability constraint when both arrival and service processes show a stochastic nature is not available in effective bandwidth and capacity studies, because there is not a straightforward linear relation between the message backlog and the delay experienced by messages in a buffer in this case. Furthermore, in order to calculate the effective bandwidth of a given arrival process or the effective capacity of a given service process, Markov processes were exploited to project the arrival process or the service process onto analytical schemes, respectively. While Markov processes and certain analytical results that already exist help ease the effective bandwidth and effective capacity depictions, it may be difficult to obtain closed-form solutions for certain data arrival and service processes. For example, the effective capacity of certain hybrid-automatic-repeat-request protocols were provided in [24]–[26], where the solutions require certain numerical techniques to be evaluated or assumptions that facilitate the analysis. Herein, the introduction of an approach that simplifies the performance measure expressions in certain transmission scenarios becomes necessary.

In this paper, different from the approach in the effective bandwidth and capacity calculations, we perform analysis in the message index domain rather than the time domain. Specifically, while the number of messages arriving at or the number of messages leaving a data buffer in a certain time frame matters in the effective bandwidth or capacity analysis, respectively, the inter-arrival time between successive messages arriving at a buffer and the service time that is spent for each message are employed in our analysis. We initially characterize the waiting time[1] (buffering delay) for each message in a buffer, and then we show that the progress of the waiting time in the message index domain takes the same form as the progress of the backlog in the same buffer in the time domain. We define the waiting time for one message as the time spent waiting in the buffer, i.e., the time between the moment the message enters the buffer and the moment it starts being served from the buffer. Therefore, the

[1]We use the terms *waiting time* and *buffering delay* interchangeably. In the literature, *waiting time* (*buffering delay*) is generally defined as the time a data packet spends in a buffer prior to service and *sojourn time* is *waiting time* plus *service time*. Throughout the paper, we refer to *waiting time*.

waiting time of a message in a buffer can be considered as the buffering delay the message is exposed to. More specifically, the contributions of the paper are the following:

1) Characterizing the waiting time for a message in a transmitter buffer, we perform the steady-state waiting time analysis and obtain an approximation for the waiting time violation probability as an exponential function of the waiting time threshold and a quality-of-service parameter.
2) Given a stochastic service process, we characterize the minimum constant inter-arrival time between successive messages arriving at the transmitter buffer, which allows to sustain the given waiting time violation probability with a given waiting time threshold. We define it as the effective service time of the stochastic service process.
3) Given a stochastic data arrival process, we administer the maximum constant service time for one message, below which the service process can guarantee the waiting time violation probability with a given waiting time threshold. We further define it as the effective inter-arrival time of the stochastic data arrival process, which can be regarded as the dual of effective service time[2] with respect to the buffer.
4) We identify that the effective capacity of a stochastic service process or the effective bandwidth of a stochastic arrival process can be obtained through the effective service time of the corresponding service process or the effective inter-arrival time of the corresponding arrival process, respectively.
5) We apply the waiting time analysis to a practical setting of message dissemination and collection schemes for vehicular ad hoc networks in both downlink and uplink scenarios.

In particular, we know that it is difficult to obtain a closed-form solution for the waiting time violation probability when both the arrival and service processes are stochastic. Employing the concepts of effective inter-arrival time and effective service time, we can obtain a closed-form solution for the waiting time violation probability at the transmitter buffer when both the data arrival and service processes are stochastic. Secondly, it is difficult to obtain the effective bandwidth of certain data arrival processes and the effective capacity of certain service processes without resorting to numerical techniques or particular assumptions. Taking advantage of the reciprocal relation between the effective bandwidth of a data arrival process and the effective inter-arrival time of the same data process and the reciprocal relation between the effective capacity of a service process and the effective service time of the same service process, we can derive the effective bandwidth from the effective inter-arrival time and the effective capacity from the effective service time. From the other perspective, for instance in the case of an aggregate of arrival processes, we can find the effective bandwidth and then obtain the effective inter-arrival time by employing the reciprocal relation between the corresponding processes, because the time domain analysis is easier in the case of an aggregate of arrival processes [27]. Moreover, we know that the radio propagation in vehicular communications is strongly influenced by the type of environment and objects, and system features; as a result, the environments where vehicular communication occurs are classified as highways, urban areas and rural areas [28]. In this paper, our mathematical solutions can easily cover a general class of vehicular communication scenarios. In summary, we provide a toolbox that system designers can use in order to understand the performance levels of vehicular communication systems under quality-of-service constraints imposed in the form of waiting time violation and buffer overflow probabilities.

The rest of the paper is as follows: In Section II, we provide the state-of-the-art regarding the effective bandwidth and capacity notions, and their relation to the buffer overflow probability constraint. We perform the buffer analysis in the message index domain in Section III. In particular, we characterize the waiting time for messages in a buffer and achieve the steady-state waiting time analysis that provides us the effective inter-arrival time of an arrival process at the buffer and the effective service time of a service process from the buffer. We exemplify our analytical results with an ON-OFF channel model and an ON-OFF data arrival process. We demonstrate the fitness of our results in message dissemination and collection schemes in vehicular ad hoc networks in Section IV and conclude the paper in Section V. We provide a list of frequently used notations in Table I.

II. State-of-the-Art

One critical concern of providing quality-of-service guarantees in communications is to keep the buffer overflow probability or delay violation probability at a transmitter below certain values. One may choose to control the data arrival process at or the data service process from the transmitter buffer, or both, in order to limit the aforementioned probabilities. However, this becomes a challenging task when the arrival and service processes vary randomly. Particularly as seen in Fig. 1, let $a_q(t)$ be the number of messages arriving at the buffer and $s_q(t)$ be the number of messages that can be served from the buffer at time instant t. The number of messages leaving the buffer is equal to $s_q(t)$ if the number of messages in the buffer is greater than or equal to $s_q(t)$. On the other hand, the number of messages leaving the buffer is equal to the number of messages in the buffer if the number of messages in the buffer is smaller than $s_q(t)$. Here, t is a discrete-time index and we describe a time-slotted system. Subsequently, assuming that the buffer size is infinite and the service process is work-conserving, we can express the backlog (queue length) at time instant t, i.e., the number of messages in the buffer, as follows:

$$q(t) = [q(t-1) + a_q(t) - s_q(t)]^+, \tag{1}$$

where $[\cdot]^+ = \max\{\cdot, 0\}$. The aforementioned expression is also known as the *Lindley* recursion [29, eq. (5.11)]. Moreover, the time-line representation of the backlog with the data arrival and service processes is shown in Fig. 2. Meanwhile, the total number of messages that depart from the buffer in the first t

[2]When we are establishing the terms *effective service time* and *effective inter-arrival time*, we follow the pattern the authors derived *effective capacity* in [10].

TABLE I
NOTATIONS

Symbol	Description
t	Time instant or time frame index
$a_q(t)$	Number of messages arriving at a buffer at time instant t
$A_q(t)$	Cumulative number of messages arriving at a buffer until time instant t
$s_q(t)$	Available service in units of messages from a buffer at time instant t
$S_q(t)$	Available cumulative service in units of messages from a buffer until time instant t
$D_q(t)$	Cumulative number of messages departing from a buffer until time instant t
$q(t)$	Number of messages in a buffer at time instant t
q_{th}	Buffer overflow threshold
θ_q	Decay rate of the tail distribution of $q(t)$
$\Lambda_{A_q}(\theta_q)$	Log-moment generating function of $A_q(t)$
$\Lambda_{S_q}(\theta_q)$	Log-moment generating function of $S_q(t)$
$\bar{a}_q(\theta_q)$	Effective bandwidth of $A_q(t)$ (minimum constant service rate from a buffer)
$\bar{s}_q(-\theta_q)$	Effective capacity of $S_q(t)$ (maximum constant arrival rate at a buffer)
n	Message index
M_n	Message n
$a_d(n)$	Inter-arrival time between messages M_{n-1} and M_n
$A_d(n)$	Time when message M_n arrives at a buffer
$s_d(n)$	Service time of message M_{n-1}
$S_d(n)$	Cumulative service time of messages M_1 to M_{n-1}
$D_d(n)$	Time when message M_n departs from a buffer
$U_d(n)$	Time when message M_n starts service from a buffer
$d(n)$	Waiting time of message M_n in a buffer
d_{th}	Waiting time threshold
θ_d	Decay rate of the tail distribution of $d(n)$
$\Lambda_{A_d}(\theta_d)$	Log-moment generating function of $A_d(n)$
$\Lambda_{S_d}(\theta_d)$	Log-moment generating function of $S_d(n)$
$\bar{a}_d(-\theta_d)$	Effective inter-arrival time of $A_d(n)$ (maximum constant service time)
$\bar{s}_d(\theta_d)$	Effective service time of $S_d(n)$ (minimum constant inter-arrival time)

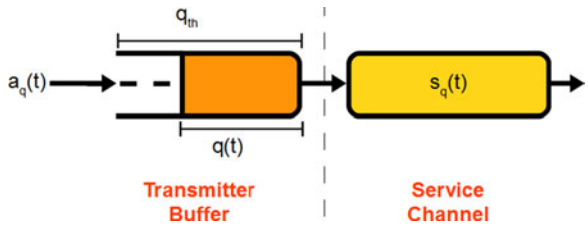

Fig. 1. Transmitter buffer and service channel.

time instants is given as

$$D_q(t) = \min\{A_q(t), D_q(t-1) + s_q(t)\},$$

where $A_q(t) = \sum_{\tau=1}^{t} a_q(\tau)$ is the total number of messages arriving at the buffer in t time instants. We show the cumulative arrival and departure processes, i.e., $A_q(t)$ and $D_q(t)$, and the backlog, which is the difference between $A_q(t)$ and $D_q(t)$, i.e., $q(t)$, in Fig. 3.

It is further known that $q(\infty)$ will converge in distribution to a finite random variable when both $a_q(t)$ and $s_q(t)$ are stationary

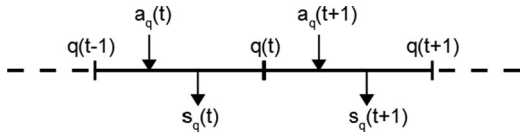

Fig. 2. Time frame representation of the backlog at the transmitter buffer.

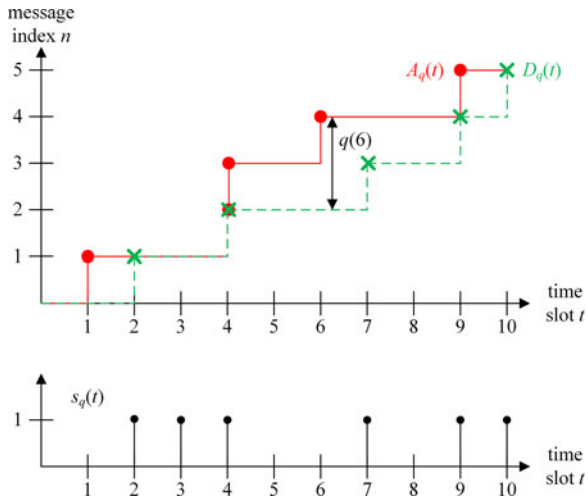

Fig. 3. Time-domain representation. $A_q(t)$ and $D_q(t)$ are the cumulative data arrival and departure processes, respectively, and $s_q(t)$ is the number of messages that can be served in the channel at time instant t. Messages 1, 2, 3, 4 and 5 arrive at the buffer at time instants $t = 1, 4, 4, 6$ and 9 and depart from the buffer at time instants $t = 2, 4, 7, 9$ and 10, respectively.

and ergodic, and when $\mathbb{E}\{a_q(t)\} < \mathbb{E}\{s_q(t)\}$ [3]. Now, given that $a_q(t)$ and $s_q(t)$ are independent of each other, we have a unique $\theta_q^\star > 0$ such that

$$\Lambda_{A_q}(\theta_q^\star) + \Lambda_{S_q}(-\theta_q^\star) = 0 \tag{2}$$

and

$$\theta_q^\star = -\lim_{q_{\text{th}}\to\infty} \frac{\log \Pr\{q(\infty) \geq q_{\text{th}}\}}{q_{\text{th}}}, \tag{3}$$

where q_{th} is the threshold and $\theta_q^\star$ is the decay rate of the tail distribution of the backlog, $q(t)$ [3, Th. 2.1]. Above, $\Lambda_{A_q}(\theta_q) = \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\left\{e^{\theta_q A_q(t)}\right\}$ and $\Lambda_{S_q}(\theta_q) = \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\left\{e^{\theta_q S_q(t)}\right\}$ are the asymptotic log-moment generating functions of the arrival and service processes, respectively, that are the Gärtner-Ellis limits and differentiable for all $\theta_q \in \mathbb{R}$. Here, $S_q(t) = \sum_{\tau=1}^{t} s_q(\tau)$ is the available total service provided in t time instants. As for the physical meaning of $\theta_q^\star$, the steady-state backlog is expressed with an exponential function of $\theta_q^\star$ and q_{th} [4], i.e.,

$$\Pr\{q(\infty) \geq q_{\text{th}}\} \approx e^{-\theta_q^\star q_{\text{th}}}.$$

Specifically, larger $\theta_q^\star$ implies stricter quality-of-service constraints, while smaller $\theta_q^\star$ means looser quality-of-service constraints.

Now, setting a constant service rate, i.e., $s_q(t) = s_q$ for all $t > 0$, we have $\Lambda_{S_q}(\theta_q) = \theta_q s_q$. Then, using the relation (2),

we obtain

$$s_q = \frac{\Lambda_{A_q}(\theta_q^\star)}{\theta_q^\star} = \lim_{t\to\infty} \frac{1}{t\theta_q^\star} \log \mathbb{E}\left\{e^{\theta_q^\star A_q(t)}\right\} = \bar{a}_q(\theta_q^\star) \quad (4)$$

for $\theta_q^\star > 0$. Specifically, given an arrival process, $A_q(t)$, the term $\bar{a}_q(\theta_q)$ denotes the effective bandwidth of the same process for any given $\theta_q > 0$, where θ_q is the quality-of-service metric. Similarly, setting a constant data arrival rate, i.e., $a_q(t) = a_q$ for all $t > 0$, we have $\Lambda_{A_q}(\theta_q) = \theta_q a_q$. Then, using again the relation (2), we obtain

$$a_q = \frac{\Lambda_{S_q}(-\theta_q^\star)}{-\theta_q^\star} = \lim_{t\to\infty} \frac{-1}{t\theta_q^\star} \log \mathbb{E}\left\{e^{-\theta_q^\star S_q(t)}\right\} = \bar{s}_q(-\theta_q^\star) \quad (5)$$

for $\theta_q^\star > 0$. Here, $\bar{s}_q(-\theta_q)$ denotes the effective capacity of the given service process, $S_q(t)$, for any given $\theta_q > 0$.

It is worth mentioning that another approximation for the buffer overflow probability for given $\theta_q^\star > 0$ when q_{th} is small is provided in [10], [30] as

$$\Pr\{q(\infty) \geq q_{\text{th}}\} \approx \varepsilon e^{-\theta_q^\star q_{\text{th}}},$$

where ε is the probability that the buffer is not empty. Further, for the special case of a constant arrival rate, the delay bound, i.e., $\Pr\{d(\infty) \geq d_{\text{th}}\}$, for a given delay threshold d_{th} is expressed as [10]

$$\Pr\{d(\infty) \geq d_{\text{th}}\} \approx \varepsilon e^{-\theta_q^\star \bar{s}_q(-\theta_q^\star) d_{\text{th}}}.$$

The delay probability approximation comes from the fact that the data arrival rate at the buffer is constant and set to the effective capacity (the maximum sustainable constant data arrival rate). Particularly, $d_{\text{th}} = \frac{q_{\text{th}}}{\bar{s}_q(-\theta_q^\star)}$ and $d(\infty) = \frac{q(\infty)}{\bar{s}_q(-\theta_q^\star)}$.

Notice that by employing the queue balance equation in (2), we can obtain the buffer overflow probability as an exponential function of the buffer threshold, q_{th}, and the decay rate, $\theta_q^\star$, when either the arrival process or the service process is stochastic, or when both processes are stochastic. However, it is difficult to obtain a closed-form solution for the delay violation probability when both the arrival and service processes are stochastic. In most practical systems, both the data arrival and service rates vary over time. Therefore, it is necessary to have a methodology that provides the delay violation probability in a buffer when both the arrival and service processes are stochastic. Further notice that in order to obtain the effective bandwidth (4) and the effective capacity (5) in some communications models, we resort to either numerical techniques or particular assumptions. For instance, the effective capacity of a general class of hybrid-automatic-repeat-request protocols can be obtained by invoking numerical search techniques [24]. Or, a closed-form solution for the effective capacity of the aforementioned system can be obtained under less-strict quality-of-service constraints only [25]. However, if we perform analysis in the domain spanned by message indices rather than time instants, we can characterize the waiting time violation probability when both the arrival and service processes are stochastic, and we can formulate the system performance with closed-form solutions in systems where it is difficult to establish straightforward analytical expressions for effective bandwidth and capacity. In the next section, we show the details of our approach and methodology in which we obtain a waiting time violation probability in the form of an exponential function by performing our analysis in the message index domain.

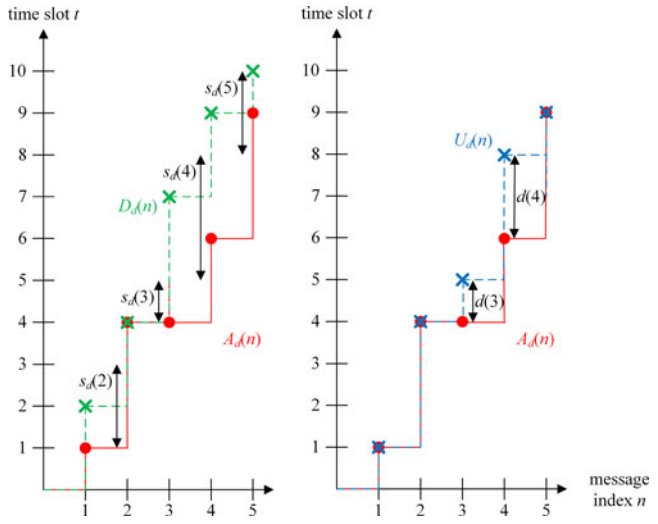

Fig. 4. Message index-domain representation. $A_d(n)$ and $D_d(n)$ are the time instants at which message n arrives at and departs from the buffer, respectively. $U_d(n)$ is the time instant at which message n starts service and $s_d(n)$ is the service time of message $n-1$, after which service of message n (if already in the buffer) can start. E.g., message 1 starts service when it arrives at $t = 1$ and finishes service at $t = 2$, so that service of message 2 can start earliest after $s_d(2) = 2$ time frames at $t = 3$. Messages 1, 2, 3, 4 and 5 start service at time instants $t = 1$, 4, 5, 8 and 9, respectively.

III. Analysis in Message Index Domain

As seen in Fig. 3, we analyze the effective bandwidth of a given data arrival process and the effective capacity of a given service process by positioning the time and the number of messages on the x-axis and y-axis, respectively, and perform analysis in the time domain. However, we can resort to a different approach by invoking the pseudo-inverse functions of the cumulative arrival and service processes. This approach was used, e.g., in deterministic[3] data traffic regulation [31, Sec. 6.2.3]. Intuitively, the pseudo-inverse corresponds to rotating and flipping the graph in Fig. 3. Particularly, the number of messages and the number of time instants are positioned on the x-axis and y-axis, respectively, as shown in Fig. 4. Suitably, considering that the messages arrive at the buffer from a source one by one, let each message be denoted by M_n for $n \in \{1, 2, \cdots\}$, where n is the message index. As seen in Fig. 5, message M_{n+1} arrives at the buffer in $a_d(n+1)$ time units after message M_n arrives at the buffer. Similarly, message M_{n+2} arrives at the buffer in $a_d(n+2)$ time units after message M_{n+1} arrives at the buffer. Therefore, $a_d(n+1)$ and $a_d(n+2)$ can be referred

[3]The author in [31, Sec. 6.2.3] used the aforementioned approach considering deterministic bounds on the data traffic while we perform our analysis considering stochastic bounds on the data arrival and service processes. We also refer the interested reader to [27] for a more recent investigation on deterministic bounds on the data traffic.

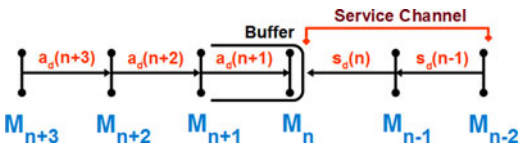

Fig. 5. Transmitter buffer with delay representation.

to as the inter-arrival time between messages M_n and M_{n+1} and the inter-arrival time between messages M_{n+1} and M_{n+2}, respectively.

We define the waiting time of a message as the time difference between the arrival of the message at the buffer and the start of the service of that message. If the message arrives at an idle system (the buffer is empty and there is no other message in service), it is immediately served from the buffer and the waiting time of the message is zero. On the other hand, if the system is busy, the message waits for the previous messages (in service and possibly in the buffer) to complete service in the first-come first-served order. We assume that the service becomes available for message M_n in $s_d(n)$ time units after message M_{n-1} starts service from the buffer. Likewise, the service becomes available for message M_{n-1} in $s_d(n-1)$ time units after message M_{n-2} starts service. Since we assume that the service process has a work-conserving policy, i.e., the system does not idle when there are messages in the system, we can regard $s_d(n)$ and $s_d(n-1)$ as the service time of message M_{n-1} and the service time of message M_{n-2}, respectively.

Hence, following the aforementioned system description in the sequel, we initially characterize the waiting time a message experiences in the buffer, and then perform the steady-state waiting time analysis and provide the effective service and inter-arrival time expressions for given data service and arrival processes, respectively.

A. Waiting Time Characterization

Although the following analysis can be performed with an initial backlog in a buffer or an initial waiting time for a message, we assume that there are no messages in the buffer at time $t = 0$ and message M_1 arrives at the buffer in $a_d(1)$ time units after time $t = 0$. Because the buffer is initially empty, the waiting time of M_1 is equal to zero, i.e., $d(1) = 0$ where $d(n)$ represents the waiting time of message M_n. Message M_1 starts service immediately after arriving at the buffer and finishes service after $s_d(2)$ time units so that service becomes available for message M_2, which comes to the buffer in $a_d(2)$ time units after message M_1. If the service becomes available for message M_2 before message M_2 comes to the buffer, i.e., $s_d(2) \leq a_d(2)$, message M_2 does not wait in the buffer to be served. Basically, the waiting time for M_2 is zero. On the other hand, if M_2 is in the buffer before the service is available for M_2, i.e., $s_d(2) > a_d(2)$, message M_2 waits in the buffer for $s_d(2) - a_d(2)$ time units. Hence, the waiting time for M_2 is expressed as $d(2) = [s_d(2) - a_d(2)]^+$. As for message M_3, we know that M_3 arrives at the buffer in $a_d(3)$ time units after M_2 comes to the buffer and that the service is ready for M_3 in $s_d(3)$ time units after M_2 starts service from the buffer. Accordingly, if M_2 does not wait in the buffer because $s_d(2) \leq a_d(2)$, the waiting time for M_3 is $d(3) = [s_d(3) - a_d(3)]^+$. On the other hand, if M_2 waits in the buffer for $s_d(2) - a_d(2)$ time units because $s_d(2) > a_d(2)$, the service becomes available for M_3 in $s_d(2) - a_d(2) + s_d(3)$ time units. As a result, the waiting time for M_3 becomes $[s_d(2) - a_d(2) + s_d(3) - a_d(3)]^+$. Combining the two cases, the waiting time for M_3 can be expressed as $[[s_d(2) - a_d(2)]^+ + s_d(3) - a_d(3)]^+ = [d(2) + s_d(3) - a_d(3)]^+$. Generalizing for message M_n, we can characterize the waiting time as follows:

$$d(n) = [d(n-1) + s_d(n) - a_d(n)]^+ \text{ for } n \geq 1, \tag{6}$$

where $d(0) = 0$ and $s_d(1) = 0$. Clearly, the waiting time increases with the service time and decreases with the inter-arrival time. The waiting time, $d(n)$, is shown graphically in Fig. 4.

Above, we have developed the waiting time characterization in (6) inductively. Considering a formal proof, let $U_d(n)$ be the time instant at which message M_n starts service, and it is expressed as

$$U_d(n) = \max\{A_d(n), U_d(n-1) + s_d(n)\}, \tag{7}$$

where

$$A_d(n) = \sum_{i=1}^{n} a_d(i) \tag{8}$$

is the time instant at which message M_n arrives at the buffer. Therefore, the waiting time for M_n is given as

$$\begin{aligned} d(n) &= U_d(n) - A_d(n) \\ &= \max\{A_d(n), U_d(n-1) + s_d(n)\} - A_d(n) \\ &= \max\{0, U_d(n-1) - A_d(n) + s_d(n)\} \\ &= [U_d(n-1) - A_d(n-1) - a_d(n) + s_d(n)]^+ \\ &= [d(n-1) + s_d(n) - a_d(n)]^+, \end{aligned}$$

which confirms the characterization in (6). The aforementioned waiting time principle can also be related to the *Lindley* recursion [32].

B. Steady-State Waiting Time Analysis

Analogous to the time domain model in (1) that defines the backlog in data units, the message index domain model (6) evaluates the waiting time in time units. This is a consequence of the first-come first-served assumption that implies that the entire backlog $q(A_d(n))$ that exists at the time of arrival of message M_n is cleared during the waiting time $d(n)$ of that message. Corresponding to the backlog analysis [3], we can conclude from (6) that the steady-state waiting time $d(\infty)$ converges in distribution to a finite random variable when both $s_d(n)$ and $a_d(n)$ are stationary and ergodic, and when $\mathbb{E}\{s_d(n)\} < \mathbb{E}\{a_d(n)\}$. Recalling that the service process is work-conserving and that the arrival and service processes are independent of each other, we have a unique $\theta_d^\star > 0$ such that

$$\Lambda_{S_d}(\theta_d^\star) + \Lambda_{A_d}(-\theta_d^\star) = 0 \tag{9}$$

and

$$\theta_d^\star = -\lim_{d_{\text{th}}\to\infty} \frac{\log \Pr\{d(\infty) \geq d_{\text{th}}\}}{d_{\text{th}}},$$

where d_{th} is the waiting time threshold and $\theta_d^\star$ is the decay rate of the tail distribution of the waiting time, $d(n)$ [3, Th. 2.1]. The Gärtner-Ellis limits that are differentiable for all $\theta_d \in \mathbb{R}$ are given as follows: $\Lambda_{S_d}(\theta_d) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left\{e^{\theta_d S_d(n)}\right\}$ and $\Lambda_{A_d}(\theta_d) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left\{e^{\theta_d A_d(n)}\right\}$, where $S_d(n) = \sum_{m=1}^{n} s_d(m)$, and $A_d(n)$ is given in (8). Hence, for a large[4] d_{th}, we have the following approximation for the waiting time of a message in steady-state:

$$\Pr\{d(\infty) \geq d_{\text{th}}\} \approx e^{-\theta_d^\star d_{\text{th}}}. \tag{10}$$

Now, considering a constant inter-arrival time between consecutive messages, i.e., $a_d(n) = a_d$, and then benefiting from the relation in (9) along with $\Lambda_{A_d}(\theta_d) = \theta_d a_d$, we obtain

$$a_d = \frac{\Lambda_{S_d}(\theta_d^\star)}{\theta_d^\star} = \lim_{n\to\infty} \frac{1}{n\theta_d^\star} \log \mathbb{E}\left\{e^{\theta_d^\star S_d(n)}\right\}, \tag{11}$$

which is the minimum constant inter-arrival time between consecutive messages that the defined service process can sustain for $\theta_d^\star > 0$. We note that $\theta_d^\star$ is different from the decay rate, $\theta_q^\star$, of the tail distribution of the queue length, $q(t)$. In particular, $\theta_d^\star$ expresses the strictness of the waiting time violation probability constraint, while $\theta_q^\star$ shows the strictness of the buffer overflow probability constraint.

Definition 1 (Effective Service Time): Given a service process $S_d(n)$ that is work-conserving and stationary with randomly distributed service time for a message, the minimum constant inter-arrival time between consecutive messages arriving at the buffer that the service process can sustain under the quality-of-service constraints defined by the decay rate of the tail distribution of the waiting time, $\theta_d > 0$, is called effective service time, and it is expressed as

$$\bar{s}_d(\theta_d) = \frac{\Lambda_{S_d}(\theta_d)}{\theta_d} = \lim_{n\to\infty} \frac{1}{n\theta_d} \log \mathbb{E}\left\{e^{\theta_d S_d(n)}\right\}. \tag{12}$$

In (12), if the service time samples are independent and identically distributed, the effective service time of the given data service process becomes

$$\bar{s}_d(\theta_d) = \frac{1}{\theta_d} \log \mathbb{E}\left\{e^{\theta_d s_d(n)}\right\} \text{ for any } n \geq 2. \tag{13}$$

Notice that when θ_d goes to zero in (13), i.e., when there are no quality-of-service constraints, the effective service time approaches the average service time in the channel [1]. Basically,

$$\lim_{\theta_d\to 0} \bar{s}_d(\theta_d) = \lim_{\theta_d\to 0} \frac{\Lambda_{S_d}(\theta_d)}{\theta_d} = \mathbb{E}\{s_d(n)\}.$$

On the other hand, when θ_d goes to infinity, i.e., when there are the strictest quality-of-service constraints, the effective service time becomes the maximum service time [1], i.e.,

$$\lim_{\theta_d\to\infty} \bar{s}_d(\theta_d) = \lim_{\theta_d\to\infty} \frac{\Lambda_{S_d}(\theta_d)}{\theta_d} = \max\{s_d(n)\}.$$

Above, $\max\{s_d(n)\}$ is the (possibly infinite) essential supremum, i.e, $\max\{s_d(n)\} = \sup\{s : \Pr\{s_d(n) > s\} > 0\}$.

Similarly, considering a constant service time for each message, i.e., $s_d(n) = s_d$, and benefiting from the relation in (9) along with $\Lambda_{S_d}(\theta_d) = \theta_d s_d$, we obtain

$$s_d = \frac{\Lambda_{A_d}(-\theta_d^\star)}{-\theta_d^\star} = \lim_{n\to\infty} \frac{-1}{n\theta_d^\star} \log \mathbb{E}\left\{e^{-\theta_d^\star A_d(n)}\right\},$$

which is the maximum allowable constant service time for a message that sustains the defined arrival process for $\theta_d^\star > 0$.

Definition 2 (Effective Inter-arrival Time): Given an arrival process $A_d(n)$ that is stationary with randomly distributed inter-arrival time between consecutive messages, the maximum allowable constant service time for a message being served from the buffer that supports the arrival process under the quality-of-service constraints defined by the decay rate of the tail distribution of the waiting time, $\theta_d > 0$, is called effective inter-arrival time, and it is expressed as

$$\bar{a}_d(-\theta_d) = \frac{\Lambda_{A_d}(-\theta_d)}{-\theta_d} = \lim_{n\to\infty} \frac{-1}{n\theta_d} \log \mathbb{E}\left\{e^{-\theta_d A_d(n)}\right\}. \tag{14}$$

In (14), if the inter-arrival time samples are independent and identically distributed[5], the effective inter-arrival time of the given data arrival process becomes

$$\bar{a}_d(-\theta_d) = \frac{-1}{\theta_d} \log \mathbb{E}\left\{e^{-\theta_d a_d(n)}\right\} \text{ for any } n \geq 2. \tag{15}$$

Notice again that when θ_d in (15) goes to zero, i.e., when there are no quality-of-service constraints, the effective inter-arrival time approaches the average inter-arrival time of the process, i.e.,

$$\lim_{\theta_d\to 0} \bar{a}_d(-\theta_d) = \lim_{\theta_d\to 0} \frac{-1}{\theta_d} \log \mathbb{E}\left\{e^{-\theta_d a_d(n)}\right\} = \mathbb{E}\{a_d(n)\}.$$

On the other hand, when θ_d goes to infinity, i.e., when there are the strictest quality-of-service constraints, the effective inter-arrival time approaches the minimum inter-arrival time of the process, i.e.,

$$\lim_{\theta_d\to\infty} \bar{a}_d(-\theta_d) = \lim_{\theta_d\to\infty} \frac{-1}{\theta_d} \log \mathbb{E}\left\{e^{-\theta_d a_d(n)}\right\} = \min\{a_d(n)\},$$

where $\min\{a_d(n)\}$ is the (possibly zero) essential infimum, i.e., $\min\{a_d(n)\} = \inf\{a : \Pr\{a_d(n) < a\} > 0\}$.

[4]Throughout the paper, we consider that d_{th} is relatively large when compared to the inter-arrival time and service time values. Moreover, θ_d shows how fast the waiting time violation probability decays as a function of d_{th}. For interested readers regarding different threshold values, we refer to [33] where the author provides upper bounds on the backlog and delay of a dynamic server and [34] where the authors show an upper bound on the waiting time violation probability in fork-join systems.

[5]As for the effective service time and the effective inter-arrival time when there exists a temporal correlation among the data packet (message) service time samples and the inter-arrival time samples, respectively, which is one of the properties of vehicular communication channels [28], we refer to [31, Chap. 7, Example 7.2.7]. Since we focus on providing a toolbox for performance analysis in a general class of vehicular communication scenarios, we consider temporally uncorrelated service time samples and temporally uncorrelated inter-arrival time samples.

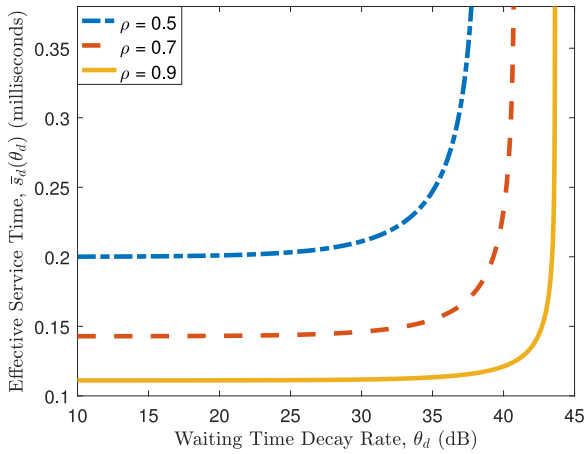

Fig. 6. Effective service time, $\bar{s}_d(\theta_d)$, vs. waiting time decay rate, $\theta_d(\text{dB}) = 10\log_{10}(\theta_d)$.

To illustrate the concepts of effective service time and effective inter-arrival time, we consider the basic scenario of ON-OFF processes. ON-OFF processes are a relevant model to characterize random outages, e.g., of wireless block fading channels. Further, the ON-OFF model has the convenient property that it can be defined equivalently in the time domain as well as in the message domain permitting us to draw important conclusions on the connection of the two domains.

1) ON-OFF Service Channel Model: Let us consider a discrete-time ON-OFF channel model with a frame duration of T time instants as the service process. The channel is ON with probability ρ and it is OFF with probability $(1-\rho)$. The channel state changes from one frame to another independently. In each frame, if the channel is ON, one message is successfully served from the buffer. Otherwise, the transmission of the message fails, and then the message is re-transmitted in the next frame. Therefore, the probability distribution of the service time of message M_n is given as $\Pr\{s_d(n) = iT\} = \rho(1-\rho)^{i-1}$ for $i \in \{1, 2, \cdots\}$. As a result, the effective service time of the process (the minimum constant inter-arrival time that can be sustained) is characterized as follows:

$$\begin{aligned}\bar{s}_d(\theta_d) &= \frac{1}{\theta_d}\log\mathbb{E}\left\{e^{s_d(n)\theta_d}\right\}\\ &= \frac{1}{\theta_d}\log\left\{\sum_{i=1}^{\infty}\rho(1-\rho)^{i-1}e^{iT\theta_d}\right\}\\ &= \frac{1}{\theta_d}\log\left\{\frac{\rho e^{T\theta_d}}{1-(1-\rho)e^{T\theta_d}}\right\} \qquad (16)\end{aligned}$$

when $0 < \theta_d < -\frac{\log\{1-\rho\}}{T}$ and $\bar{s}_d(\theta_d) = \infty$ when $-\frac{\log\{1-\rho\}}{T} \leq \theta_d$. Notice that the effective service time approaches $\frac{T}{\rho}$ when θ_d goes to zero, which is the average service time and hence the reciprocal of the average transmission rate of the channel.

In Fig. 6, we set the transmission frame to 0.1 milliseconds, i.e., $T = 0.1$, and plot the effective service time as a function of the decay rate parameter, θ_d. The effective service time approaches the average service time, which is $\frac{0.1}{\rho}$, as θ_d decreases to zero, whereas the effective service time goes to infinity with increasing θ_d. Herein, for a better understanding of the results in Fig. 6, let us consider that we are given a waiting time violation probability constraint, $\Pr\{d(n) \geq d_{\text{th}}\}$ for large n, and a waiting time threshold, d_{th}. Then, we can use (10) to calculate the decay rate of the tail distribution of the waiting time, θ_d, as follows: $\theta_d \approx -\frac{\log\Pr\{d(n)\geq d_{\text{th}}\}}{d_{\text{th}}}$. Now, having θ_d, we can easily determine the effective service time of the defined ON-OFF service process through (16), which is the minimum inter-arrival time between consecutive messages that arrive at the transmitter buffer such that the waiting time violation probability constraint for the desired waiting time threshold is guaranteed.

Connection to the queue decay rate, θ_q: Let us again consider the aforementioned scenario and assume that the constant data arrival rate is a messages per time instant, and hence, the constant inter-arrival time between two consecutive messages is $\frac{1}{a}$ time instants. Noting the relations in (11) and (16), we have a unique waiting time decay rate such that

$$\frac{1}{a} = \frac{1}{\theta_d^\star}\log\left\{\frac{\rho e^{T\theta_d^\star}}{1-(1-\rho)e^{T\theta_d^\star}}\right\}. \qquad (17)$$

Moreover, noting the relation given in (5), and that the service rate is $s_q(t) = 1$ message per one frame of T time instants if the service channel is ON with probability ρ and $s_q(t) = 0$ messages per time frame if the service channel is OFF with probability $(1-\rho)$, we have a unique queue decay rate such that

$$a = \lim_{t\to\infty}\frac{-1}{tT\theta_q^\star}\log\mathbb{E}\left\{e^{-\theta_q^\star S_q(t)}\right\} \qquad (18)$$

$$= -\frac{1}{T\theta_q^\star}\log\mathbb{E}\left\{e^{-\theta_q^\star s_q(t)}\right\} \qquad (19)$$

$$= -\frac{1}{T\theta_q^\star}\log\left\{\rho e^{-\theta_q^\star} + 1 - \rho\right\}. \qquad (20)$$

Notice that t is the time frame number and T is the frame duration in (18). Because a is given in *messages per time instant*, we divide the expression in (18) by T. Moreover, (19) follows from the fact that the channel state changes from one frame to another independently. Hence, noting that the moment generating function of the service is $\mathbb{E}\{e^{\theta s_q(t)}\} = \rho e^{\theta_q} + 1 - \rho$, we obtain the result in (20). Above, the right-hand-side of (20) for given θ_q provides us the effective capacity of the service process. Then, solving (17) and (20), we obtain

$$0 = 1 - \rho + \rho e^{-\frac{\theta_d^\star}{a}} - e^{-\theta_d^\star T}$$

and

$$0 = 1 - \rho + \rho e^{-\theta_q^\star} - e^{-a\theta_q^\star T},$$

respectively. Since the aforementioned balance expressions are obtained when the buffer is in steady-state, we verify a linear relation between the decay rate of the waiting time and the decay rate of the backlog such that $\theta_d^\star = a\theta_q^\star$.

In the following theorem, given any stationary and ergodic data service process or data arrival process, we generalize the aforementioned result and provide the relation between the

effective service time and capacity of a given service process, and the relation between the effective inter-arrival time and bandwidth of a given arrival process.

Theorem 1: Given that the effective service time of a data service process is $\bar{s}_d(\theta_d)$ for $\theta_d > 0$ and the effective capacity of the same service process is $\bar{s}_q(-\theta_q)$ for $\theta_q > 0$, there exists a reciprocal relation between $\bar{s}_d(\theta_d)$ and $\bar{s}_q(-\theta_q)$ such that

$$\bar{s}_d(\theta_d)\bar{s}_q(-\theta_q) = 1 \quad \text{and} \quad \theta_q = \bar{s}_d(\theta_d)\theta_d. \tag{21}$$

Likewise, given that the effective inter-arrival time of a data arrival process is $\bar{a}_d(-\theta_d)$ for $\theta_d > 0$ and the effective bandwidth of the same arrival process is $\bar{a}_q(\theta_q)$, there exists a reciprocal relation between $\bar{a}_d(-\theta_d)$ and $\bar{a}_q(\theta_q)$ such that

$$\bar{a}_d(-\theta_d)\bar{a}_q(\theta_q) = 1 \quad \text{and} \quad \theta_q = \bar{a}_d(-\theta_d)\theta_d. \tag{22}$$

Proof. See appendix. ∎

Connection of message index and time domain models: Theorem 1 lays ground for a method for the construction of time domain models from message index domain models and vice versa. First, notice that the decay rate of the tail distribution of the queue length in steady-state, θ_q, in (21) is the log-moment generating function of the service process in the message index domain. In particular, we have with (12) that

$$\theta_q = \bar{s}_d(\theta_d)\theta_d = \Lambda_{S_d}(\theta_d). \tag{23}$$

Now, define $f(\theta_d) = \Lambda_{S_d}(\theta_d)$ to be a function of the decay rate of the tail distribution of the waiting time in steady-state, θ_d. Then, we can solve $\theta_q = f(\theta_d)$ for θ_d to express θ_d as a function of θ_q as follows: $\theta_d = f^{-1}(\theta_q)$, where $f^{-1}(\theta_q)$ is the inverse function. In general, when we are given the effective service time of a service process in the message index domain, we can use that $\bar{s}_d(\theta_d)\bar{s}_q(-\theta_q) = 1$ from (21) to reach the effective capacity of the same service process in the time domain by taking the reciprocal of the effective service time and substituting $f^{-1}(\theta_q)$ for θ_d, i.e.,

$$\bar{s}_q(-\theta_q) = \frac{1}{\bar{s}_d(f^{-1}(\theta_q))}.$$

For instance, noting the log-moment generating function of the ON-OFF service channel model in (16) as $f(\theta_d) = \log\left\{\frac{\rho e^{T\theta_d}}{1-(1-\rho)e^{T\theta_d}}\right\}$, and letting $\theta_q = f(\theta_d)$ we can solve for $\theta_d = f^{-1}(\theta_q)$ to show that

$$\begin{aligned}\theta_d &= \frac{1}{T}\log\left\{\frac{e^{\theta_q}}{\rho + (1-\rho)e^{\theta_q}}\right\} \\ &= -\frac{1}{T}\log\left\{\rho e^{-\theta_q} + 1 - \rho\right\}.\end{aligned} \tag{24}$$

Then, plugging (24) into the effective service time (16) and taking the reciprocal of (16), we express the effective capacity of the ON-OFF service channel model as

$$\bar{s}_q(-\theta_q) = \frac{1}{\bar{s}_d(\theta_d)} = -\frac{1}{T\theta_q}\log\left\{\rho e^{-\theta_q} + 1 - \rho\right\}, \tag{25}$$

which also confirms (20).

Overall, we know that achieving closed-form solutions for the effective capacity or bandwidth of certain processes in the time domain may be intractable. On the other hand, we can carry out an analysis in the message index domain, and then we can arrive at solutions for the effective capacity or bandwidth by invoking Theorem 1. Similarly, if results in the message index domain cannot be achieved, we can acquire solutions through an analysis in the time domain as well.

2) ON-OFF Data Arrival Process: In addition to the ON-OFF channel model described in Section III-B1, let us also consider an ON-OFF arrival process[6] with an ON probability λ. In particular, a message arrives at the buffer with probability λ in a frame of T time instants. Noting that the arrival process has two states, i.e., ON and OFF states, we further assume that the arrival process changes its state from one frame to another independently. Therefore, the probability distribution for the inter-arrival time between message M_{n-1} and M_n is given as $\Pr\{a_d(n) = iT\} = \lambda(1-\lambda)^{i-1}$ for $i \in \{1, 2, \cdots\}$. Now, regarding independent and identically distributed inter-arrival time samples, we formulate the asymptotic log-moment generating function of the arrival process as follows: $\Lambda_{A_d}(\theta_d) = \log\left\{\frac{\lambda e^{T\theta_d}}{1-(1-\lambda)e^{T\theta_d}}\right\}$ for $0 < \theta_d < -\frac{\log\{1-\lambda\}}{T}$ and $\Lambda_{A_d}(\theta_d) = \infty$ for $-\frac{\log\{1-\lambda\}}{T} \leq \theta_d$. Hence, we invoke the relation (9) and have the following characterization:

$$\log\left\{\frac{\rho e^{T\theta_d^\star}}{1-(1-\rho)e^{T\theta_d^\star}}\right\} + \log\left\{\frac{\lambda e^{-T\theta_d^\star}}{1-(1-\lambda)e^{-T\theta_d^\star}}\right\} = 0,$$

which provides the decay rate of the tail distribution of the waiting time as

$$\theta_d^\star = \frac{1}{T}\log\left\{\frac{1-\lambda}{1-\rho}\right\} \tag{26}$$

where $\lambda < \rho$. Notice that the decay rate increases with the increasing ON probability in the service channel while it decreases with the increasing ON probability in the arrival process.

Non-linear relation between θ_d and θ_q: Let us again consider the communication scenario in Section III-B2. We know that the decay rate of the tail distribution of the waiting time is given in (26). As for the decay rate of the tail distribution of the queue, we start with defining the service rate and the arrival rate in one time frame. In particular, the service rate is $s_q(t) = 1$ message per time frame if the service channel is ON with probability ρ and $s_q(t) = 0$ messages per time frame if the service channel is OFF with probability $(1-\rho)$. Similarly, the arrival rate is $a_q(t) = 1$ message per time frame if the data arrives at the buffer with probability λ and $a_q(t) = 0$ messages per time frame if the data does not arrive at the buffer with probability $(1-\lambda)$. Noting that the state transitions occur independently in both the service and arrival process, and invoking the relation in (2), we can show that

$$\frac{1}{T}\log\left\{\lambda e^{\theta_q^\star} + 1 - \lambda\right\} + \frac{1}{T}\log\left\{\rho e^{-\theta_q^\star} + 1 - \rho\right\} = 0.$$

[6]Our aims are twofold in this analysis. First, in practical systems, both the arrival and service processes are generally stochastic. Second, ON-OFF arrival and service process models provide a clear insight into the analytical findings from a practical perspective to understand the buffer dynamics.

Given that $0 < \theta_q^\star$, we can obtain the decay rate of the tail distribution of the queue length as

$$\theta_q^\star = \log\left\{\frac{\rho}{\lambda}\right\} + \log\left\{\frac{1-\lambda}{1-\rho}\right\} \tag{27}$$

$$= \log\left\{\frac{\rho}{\lambda}\right\} + T\theta_d^\star. \tag{28}$$

Specifically, (28) shows that there is not always a linear relationship between $\theta_d^\star$ and $\theta_q^\star$ when both the arrival and service processes are stochastic. Following this result, we can see that by taking the buffer overflow probability constraint as the primary quality-of-service metric, we may not easily reach a conclusion on the buffering delay performance because of a possible nonlinear relation between $\theta_d^\star$ and $\theta_q^\star$. Therefore, the results in this paper provide a straightforward method to obtain a performance measure regarding the waiting time in a transmitter buffer.

IV. Vehicular Ad Hoc Networks

In this section, we substantiate our analytical results in practical settings. We consider first a reliable broadcast in a downlink scenario, where a transmitter employs a fountain encoder and receivers send feedback to the transmitter only when they stop receiving the transmitted data packets, e.g., when they have successfully decoded the message. Then, we employ the slotted Aloha transmission protocol as a relevant multi-access protocol in an uplink scenario. These two scenarios can be embedded in message dissemination and collection schemes for vehicular ad hoc networks. For instance, in order to minimize packet arrival time and bandwidth occupancy, the authors studied fountain codes-based data dissemination techniques in vehicular ad hoc networks and showed the effectiveness of their techniques when compared with traditional data dissemination approaches [35]. Furthermore, the authors in [36] studied mobile slotted Aloha protocols in vehicular ad hoc networks and validated their solutions through simulations that confirm a dramatic improvement in the scalability of the solution and in the overall protocol performance. Besides, we consider a time-constrained[7] communication scenario. Particularly, in the downlink scenario, a base station broadcasts a time-constrained message to the vehicles that are in certain proximity to the base station. In the uplink scenario, the vehicles in close proximity pass their time-constrained messages to the base station in a multi-access manner. We also note that while we analyze these two protocols for numerical presentations, we emphasize that our analytical methods are for a general class of communication scenarios.

A. Downlink Broadcast Transmission

We consider a downlink scenario in which a transmitter (base station) sends a message to L receivers as seen in Fig. 7. The transmitter partitions the message into K encoding (data) packets and composes N encoded packets by randomly forming a combination of the encoding packets, where $K < N$. Subsequently, the transmitter broadcasts the encoded packets one-by-one in frames of T time units. During the transmission of each message, when a receiver is able to collect $K^\star$ of the encoded packets successfully, where $K \leq K^\star < N$, it can decode the message. On the other hand, when the receiver realizes in any time frame that it is not possible to recover the message even if it receives the future encoded packets, it stops the reception of the encoded packets. In both cases, the receiver sends an acknowledgment to the transmitter to inform about its status. Subsequently, the transmitter stops the transmission when it collects acknowledgments from all the receivers or when it completes the transmission of all the N encoded packets regardless of the number of received acknowledgments. The transmitter in this scenario can be associated with a fountain encoder [37] or a Luby transform (LT) encoder [38], [39], where the only difference comes from the acknowledgments that the receivers send when they are done with the reception. On the other hand, the transmitter scenario is different than the conventional automatic-repeat-request transmission protocols in such a way that the receivers do not confirm every successful packet reception but they send acknowledgments only once when they either decode the whole message or fail to decode the whole message.

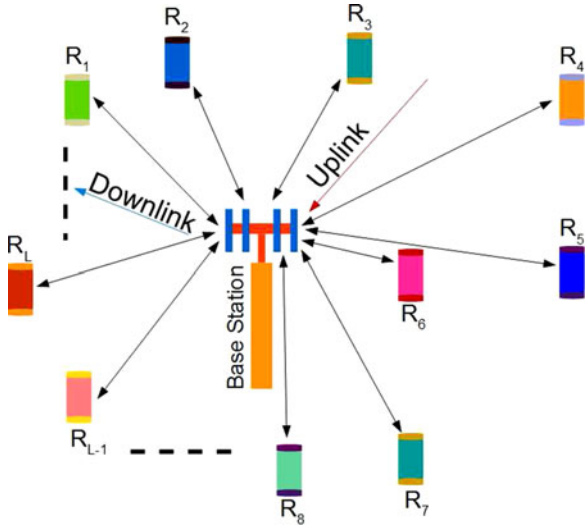

Fig. 7. System model and communication scenario.

Accordingly, let ρ_l denote the probability of successful reception of an encoded packet by the lth receiver for $l \in \{1, \cdots, L\}$. Principally, we consider an ON-OFF channel model between the transmitter and the lth receiver, i.e., the channel is ON with probability ρ_l and OFF with probability $1 - \rho_l$. Let us initially assume $2K^\star \leq N + 1$. Then, noting that one receiver needs at least $K^\star$ encoded packets, we realize that a receiver may declare a success in decoding and then send an acknowledgment to the transmitter in one of the time frames starting from the $(K^\star)$th time frame. Therefore, the probability of successful decoding in the jth time frame is

$$\binom{j-1}{K^\star-1} \rho_l^{K^\star} (1-\rho_l)^{j-K^\star}$$

[7]We assume messages have a limited lifetime as, e.g., in the case of periodic updates where each message supersedes the previous ones. The lifetime is reflected by a delivery deadline. Hence, messages are time-constrained in the sense that delivery is important before the deadline but not afterwards.

for $j \in \{K^\star, \cdots, N\}$. On the other hand, one receiver may declare a failure after the $(N-K^\star)$th time frame because the receiver may realize that it is not possible to obtain at least $K^\star$ encoded packets at the end of the Nth frame. For instance, let us assume that a receiver has collected zero encoded packets at the end of the $(N-K^\star)$th frame. Hence, it missed $(N-K^\star)$ encoded packets due to transmission errors. Basically, the receiver has to obtain the encoded packets in the following $K^\star$ time frames correctly so that it can recover the message. If the receiver fails to receive one encoded packet in the last $K^\star$ time frames, it declares a failure and sends a negative acknowledgment to the transmitter. Hence, the probability of decoding failure is expressed as

$$\binom{j-1}{j-1-N+K^\star} \rho_l^{j-1-N+K^\star} (1-\rho_l)^{N-K^\star+1}$$

for $j \in \{N-K^\star+1, \cdots, N\}$. Consequently, we formulate (29) shown at the bottom of this page, that provides us the probability of sending an acknowledgment to the transmitter for the lth receiver in the jth time frame. Furthermore, when we have $N+1 < 2K^\star$, the probability of sending the acknowledgment is given in (30) shown at the bottom of this page.

Recall that the transmitter stops the transmission of a message when either it receives acknowledgments from all the receivers or it reaches the transmission deadline, which is set to the total transmission duration of the N encoded packets. Now, let p_j be the probability that the transmitter stops the transmission of the message that is in transition at the end of the jth time frame. Herein the probability that some of the receivers send their acknowledgments before the jth time frame and the last receiver sends its acknowledgment or the last receivers send their acknowledgments in the jth time frame is expressed as

$$p_1 = \prod_{l=1}^{L} p_{1l} \quad \text{and} \quad p_j = \prod_{l=1}^{L} \sum_{i=1}^{j} p_{il} - \prod_{l=1}^{L} \sum_{i=1}^{j-1} p_{il},$$

where $2 \le j \le N-1$. Moreover, because the transmitter stops the transmission at the end of the Nth transmission frame due to the transmission deadline, we have

$$p_N = 1 - \sum_{j=1}^{N-1} p_j.$$

If the transmitter stops the transmission of a message and removes it from the buffer after the jth time frame, we can declare that the service time of the message is jT time units. Herein, because the ON-OFF process in each channel between the transmitter and the receivers is composed of independent state transitions, the effective service time becomes

$$\bar{s}_d(\theta_d) = \frac{1}{\theta_d} \log \left\{ \sum_{j=1}^{N} p_j e^{jT\theta_d} \right\} \tag{31}$$

for $0 < \theta_d$. Notice that when θ_d goes to zero, the effective service time approaches the average service time, which is $T\sum_{j=1}^{N} jp_j$. On the other hand, when θ_d goes to infinity, the effective service time approaches the maximum service duration, which is NT time units.

For numerical presentations, we assume that the probability of successful reception of an encoded packet by one receiver is uniformly distributed[8] between $\rho_{\min}$ and 1, i.e., $\rho_{\min} \le \rho_l \le 1$ and

$$f_{\rho_l}(\rho_l) = \begin{cases} \frac{1}{1-\rho_{\min}}, & \rho_{\min} \le \rho_l \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

We set the number of receivers to 100 and 1000, i.e., $L = 100$ and $L = 1000$, in Figs. 8 and 9, respectively, and plot the effective service time as a function of the waiting time decay rate, θ_d.

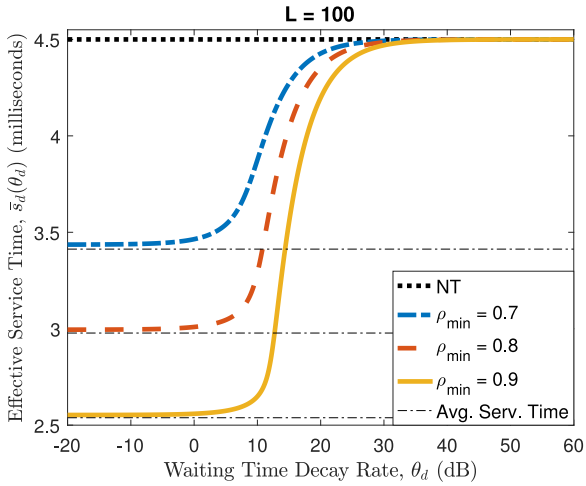

Fig. 8. Effective service time, $\bar{s}_d(\theta_d)$, vs. waiting time decay rate, θ_d (dB).

[8]We consider a highway scenario and assume that cars move from the coverage area of one base station to the coverage area of another. The probability of successful packet reception depends on the signal-to-noise ratio, and hence the distance between the base station and the cars on the highway. In order to consider this, we model the distribution of the successful packet reception probability among cars as uniformly distributed. However, regardless of the distribution, we can plug any given packet reception probability into (29) and (30) and apply our analytical results in any setting.

$$p_{jl} = \begin{cases} 0, & 1 \le j \le K^\star - 1 \\ \binom{j-1}{K^\star-1} \rho_l^{K^\star} (1-\rho_l)^{j-K^\star}, & K^\star \le j \le N-K^\star \\ \binom{j-1}{j-1-N+K^\star} \rho_l^{j-1-N+K^\star} (1-\rho_l)^{N-K^\star+1} + \binom{j-1}{K^\star-1} \rho_l^{K^\star} (1-\rho_l)^{j-K^\star}, & N-K^\star+1 \le j \le N \end{cases} \tag{29}$$

$$p_{jl} = \begin{cases} 0, & 1 \le j \le N-1 \\ \binom{j-1}{N-K^\star} \rho_l^{j-1-N+K^\star} (1-\rho_l)^{N-K^\star+1}, & N-K^\star+1 \le j \le K^\star - 1 \\ \binom{j-1}{N-K^\star} \rho_l^{j-1-N+K^\star} (1-\rho_l)^{N-K^\star+1} + \binom{j-1}{K^\star-1} \rho_l^{K^\star} (1-\rho_l)^{j-K^\star}, & K^\star \le j \le N \end{cases} \tag{30}$$

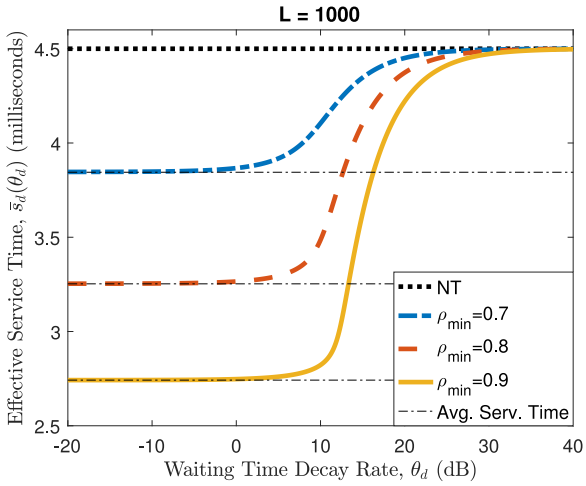

Fig. 9. Effective service time, $\bar{s}_d(\theta_d)$, vs. waiting time decay rate, θ_d (dB).

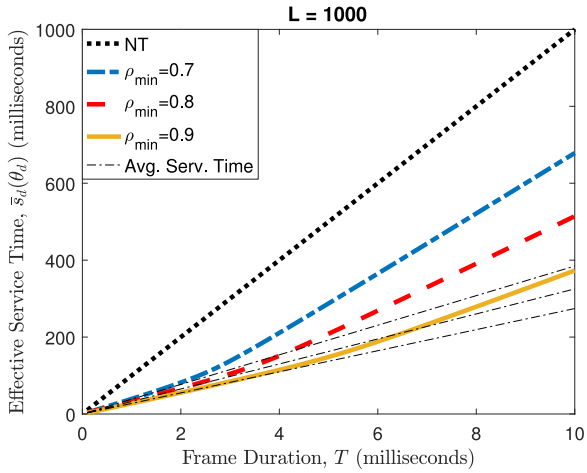

Fig. 10. Effective service time, $\bar{s}_d(\theta_d)$, vs. frame duration, T.

Given that the transmission time frame is 1 milliseconds, i.e., $T = 1$, we observe that the effective service time goes to the average service time in the channel with decreasing θ_d, while it approaches the maximum service time for one message in the channel with increasing θ_d, which is NT milliseconds. We further observe that the number of receivers does not matter with increasing θ_d. Specifically, larger θ_d means that a message arriving at the buffer is to be transmitted almost without waiting in the buffer. Therefore, the inter-arrival time (effective service time) increases to the maximum, which is NT seconds, and the value of L does not matter anymore. Notice that when the inter-arrival time is NT, the data arrival rate decreases but each message arriving at the buffer is served without waiting in the buffer. On the other hand, with decreasing θ_d, the waiting time constraint becomes loose and the inter-arrival time decreases to the average service time in the channel. Since the average service time is a function of L because the state transition probabilities depend on L, the effective service time decreases with decreasing L as seen in Figs. 8 and 9, where the average service time is higher when L is higher. Moreover, we plot the effective service time as a function of the time frame, T, in Fig. 10 when $\theta_d = -5$ dB and $L = 1000$. With increasing T, the effective service time increases and the gap between the effective service time and the average service time increases.

B. Slotted Aloha

We consider a multi-access scenario with L transmitters sending messages to a receiver in frames of T time units as seen in Fig. 7. Each transmitter enters the channel and sends its message with probability ρ, and each message is received successfully by the receiver when there is no collision. Following a successful reception of a message, the receiver sends an acknowledgment to the respective transmitter. Subsequently, the transmitter removes the message from its queue. On the other hand, if the receiver fails receiving any message, it does not send any acknowledgment. Meanwhile, the transmitters that do not get any acknowledgment enter the channel again with probability ρ and repeat the transmission of the messages that are not acknowledged. In particular, the probability of removing a message from a transmitter buffer as a result of its successful reception by the receiver in any time frame is $p = \rho(1-\rho)^{L-1}$. Hence, p becomes the probability of the channel being ON for the transmitter. Now, noting that the channel is ON with probability p and OFF with probability $(1-p)$ for one transmitter and that the transmitters opt entering the channel randomly, we consider that the channel states change independently from one time frame to another. Therefore, we can refer to Section III-B1 and express the effective service time of the channel for one transmitter as follows:

$$\bar{s}_d(\theta_d) = \frac{1}{\theta_d} \log\left\{ \frac{p e^{T\theta_d}}{1-(1-p)e^{T\theta_d}} \right\} \tag{33}$$

for $0 < \theta_d < -\frac{\log\{1-p\}}{T}$, and $\bar{s}_d(\theta_d) = \infty$ for $-\frac{\log\{1-p\}}{T} \leq \theta_d$. Let us also assume that there exists a transmission deadline over the transmission duration of a message from a transmitter. Noting that a transmitter can attempt to send a message for maximum N times, the effective service time becomes

$$\bar{s}_d(\theta_d) = \frac{1}{\theta_d} \log\left\{ (1-p)^{N-1} e^{NT\theta_d} + p e^{T\theta_d} \frac{1-(1-p)^{N-1} e^{(N-1)T\theta_d}}{1-(1-p)e^{T\theta_d}} \right\}. \tag{34}$$

The effective service time expressions in (33) and (34) provide us the minimum constant inter-arrival time between the consecutive messages arriving at one transmitter buffer such that the quality-of-service constraint in the form of steady-state waiting time violation probability is sustained by the defined service process. Moreover, notice in (33) that the range of probability of entering the channel under the quality-of-service constraint is bounded as

$$1 - e^{-T\theta_d} < p \leq \max_{\rho} \left\{ \rho(1-\rho)^{L-1} \right\}. \tag{35}$$

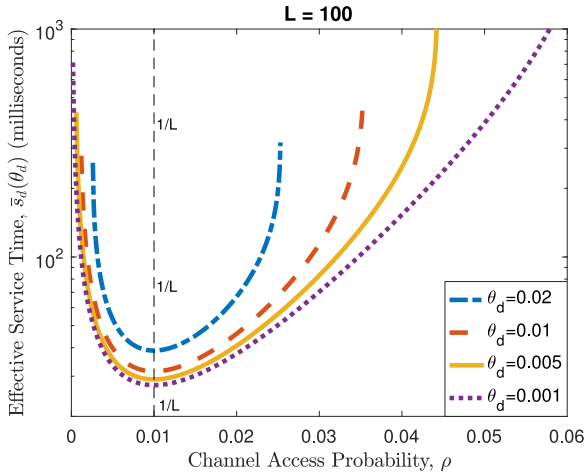

Fig. 11. Effective service time, $\bar{s}_d(\theta_d)$, vs. channel access probability, ρ.

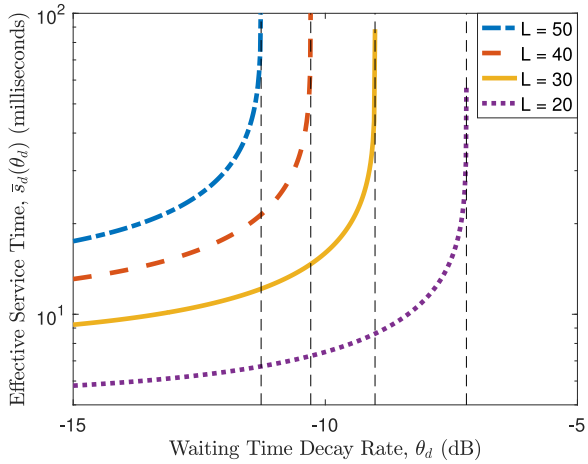

Fig. 12. Effective service time, $\bar{s}_d(\theta_d)$, vs. waiting time decay rate, θ_d (dB).

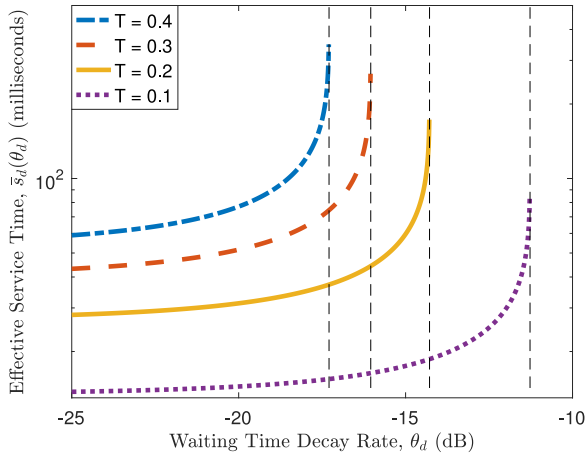

Fig. 13. Effective service time, $\bar{s}_d(\theta_d)$, vs. waiting time decay rate, θ_d (dB).

Basically, given a decay rate constraint, θ_d, the maximum number of transmitters is limited as follows:

$$L_{\max} = \max\left\{L : \max_{\rho}\left\{\rho(1-\rho)^{L-1}\right\} > 1 - e^{-T\theta_d}\right\}. \tag{36}$$

Thus, we can conclude that when the number of transmitters is greater than $L_{\max}$ for given θ_d, the quality-of-service constraint in the form of waiting time violation probability is generally not sustainable. Additionally, because we need to minimize the effective service time such that the number of messages arriving at one transmitter buffer increases, we have to maximize the ON probability, p. We can show this by taking the first derivative of the expression in (33) with respect to p, which is always negative. Therefore, we have to find the channel access probability, ρ, that maximizes the ON probability, p. By taking the derivative of p with respect to ρ and setting it to zero, we find that the optimal channel access probability that minimizes the effective service time is $\rho^\star = \frac{1}{L}$. The probability $\rho^\star = \frac{1}{L}$ is also the probability that maximizes the average service rate in the channel. Finally, when there is not a transmission deadline, the effective service time in (33) approaches infinity as T goes to $\log(1-p)^{\frac{-1}{\theta_d}}$ and the transmitter buffer becomes unstable when $T \geq \log(1-p)^{\frac{-1}{\theta_d}}$. When there is a transmission deadline, the effective service time in (34) approaches infinity as T goes to infinity.

In numerical presentations, we plot the effective service time as a function of the channel access probability for given waiting time decay rate values in Fig. 11, and as a function of the waiting time decay rate parameter in Figs. 12 and 13 given that the optimal channel access probability is employed. We set the transmission frame to 1 millisecond, i.e., $T = 1$, in Figs. 11 and 12, and the number of transmitters to 50, i.e., $L = 50$, in Fig. 13. Regardless of θ_d, the effective service time is minimized when the channel access probability is set to one over the number of transmitters, i.e., $\rho = \frac{1}{L}$, as seen in Fig. 11. Another observation is the increase of the effective service time with increasing θ_d. Moreover, the effective service time goes to infinity with θ_d increasing beyond $-\frac{\log\{1-p\}}{T}$, which is -11.27, -10.29, -9.01 and -7.20 dB for $L = 50$, 40, 30 and 20, respectively, as seen in Fig. 12, and -17.29, -16.04, -14.28 and -11.27 dB for $T = 0.4$, 0.3, 0.2 and 0.1, respectively, as seen in Fig. 13. In addition, the effective service time goes to the average service time in the channel with decreasing θ_d in all cases.

Remark 1: As seen in (31), a closed-form expression exists for the effective service time of the aforementioned downlink scenario. On the other hand, a simple closed form solution for the effective capacity of the same scenario may not be obtained. Noting that the hybrid-automatic-repeat-request with incremental redundancy protocol is same with the aforementioned downlink scenario when the number of users is set to $L = 1$, we refer to [24, Th. 1] for the effective capacity of the downlink scenario with one user, which requires extensive numerical calculations. Basically, we observe that an analytical closed-form solution can be obtained for the effective service time of a system while it is difficult to have a simple expression for the effective capacity of the same system. Similarly, as for the effective capacity of the uplink scenario, because there exists an analogy between the aforementioned uplink scenario and the hybrid-automatic-repeat-request type-I protocol, we again refer to [24, Th. 1]. We can see that a simple closed-form solution for the

effective capacity of one service process may not be possible when there exists a transmission deadline, whereas the effective service time of the same service process can be obtained easily as seen in (34). Therefore, invoking Theorem 1, we can obtain the effective capacity of a service process through the effective service time of the same service process and the effective bandwidth of an arrival process through the effective inter-arrival time of the same arrival process. Similarly, when it becomes difficult to obtain the effective service time of a service process or the effective inter-arrival time of an arrival process, we can again resort to Theorem 1 to obtain them through the effective capacity of the corresponding service process and the effective bandwidth of the corresponding arrival process, respectively.

V. Conclusion

In this paper, we have approached the data queueing problem by employing the steady-state waiting time violation probability rather than the buffer overflow probability as the primary quality-of-service constraint. Different than the existing studies, we have not operated in the time domain but the message index domain. We have characterized the waiting time for one message in a buffer. Subsequently, we have formulated the steady-state analysis between the data arrival and service processes, and then we have identified the effective service time and effective inter-arrival time. Under the waiting time violation probability constraint, the effective service time of a service process yields the minimum constant inter-arrival time between consecutive messages arriving at a buffer and the effective inter-arrival time of an arrival process yields the maximum constant service time for a message that is in transition in the service channel. In particular, we have shown that we can sustain a desired waiting time violation probability by controlling the inter-arrival time between the messages arriving at a buffer and the service time spent for a message during its transition in a service channel. Moreover, we have identified the reciprocal relation between the effective capacity and service time of a service process and the reciprocal relation between the effective bandwidth and inter-arrival time of an arrival process. Basically, we have provided a mathematical toolbox that system designers can use in order to understand the performance levels of a general class of vehicular communication scenarios under quality-of-service constraints imposed in the form of waiting time violation and buffer overflow probabilities. Finally, as an example, we have substantiated our analytical results in numerical demonstrations where we have employed a message dissemination and collection scenario as it is common, e.g., in vehicular ad hoc networks with a downlink broadcast channel and an uplink multi-access slotted Aloha protocol.

Appendix
Proof of Theorem 1

Let $\bar{s}_d(\theta_d)$ be the effective service time of the service process from a data buffer for $\theta_d > 0$ and the inter-arrival time between messages arriving at the buffer be constant. Now, recall that the unique $\theta_d^\star$ that is the decay rate of the tail distribution of the waiting time in (9) is

$$\theta_d^\star = -\lim_{d_{\text{th}}\to\infty} \frac{\Pr\{d(\infty) \geq d_{\text{th}}\}}{d_{\text{th}}}. \tag{37}$$

Noting that $\bar{s}_d(\theta_d^\star)$ equals the constant inter-arrival time between messages arriving at the buffer, we re-write (37) as

$$\theta_d^\star = -\lim_{d_{\text{th}}\to\infty} \frac{\Pr\left\{\frac{d(\infty)}{\bar{s}_d(\theta_d^\star)} \geq \frac{d_{\text{th}}}{\bar{s}_d(\theta_d^\star)}\right\}}{\bar{s}_d(\theta_d^\star)\frac{d_{\text{th}}}{\bar{s}_d(\theta_d^\star)}}. \tag{38}$$

Moreover, let us consider that message M_n enters the service at time instant $U_d(n)$ and the waiting time of message M_n is $d(n)$. Hence, the backlog in the buffer at time instant $U_d(n)$ is $q(U_d(n)) = \frac{d(n)}{\bar{s}_d(\theta_d^\star)}$ when message M_n enters the service. Subsequently, in steady-state, i.e., when n goes to infinity, we have $q(\infty) = \frac{d(\infty)}{\bar{s}_d(\theta_d^\star)}$. Then, we re-organize (38) as

$$\bar{s}_d(\theta_d^\star)\theta_d^\star = -\lim_{q_{\text{th}}\to\infty} \frac{\Pr\{q(\infty) \geq q_{\text{th}}\}}{q_{\text{th}}}, \tag{39}$$

where $q_{\text{th}} = \frac{d_{\text{th}}}{\bar{s}_d(\theta_d^\star)}$ is the buffer overflow threshold. With (39) the decay rate of the tail distribution of the backlog (3) follows as $\theta_q^\star = \bar{s}_d(\theta_d^\star)\theta_d^\star$. Further, the constant inter-arrival time between consecutive messages $\bar{s}_d(\theta_d^\star)$ implies a constant arrival rate at the buffer of $\frac{1}{\bar{s}_d(\theta_d^\star)}$. At the same time, the constant arrival rate is equal to the effective capacity, so that we have $\bar{s}_q(-\theta_q^\star) = \frac{1}{\bar{s}_d(\theta_d^\star)}$ which completes the proof of (21).

Now, let $\bar{a}_d(-\theta_d)$ be the effective inter-arrival time of the arrival process at a data buffer for $\theta_d > 0$ and the service time for one message from the buffer be constant. Then, we re-write (37) as

$$\theta_d^\star = -\lim_{d_{\text{th}}\to\infty} \frac{\Pr\left\{\frac{d(\infty)}{\bar{a}_d(-\theta_d^\star)} \geq \frac{d_{\text{th}}}{\bar{a}_d(-\theta_d^\star)}\right\}}{\bar{a}_d(-\theta_d^\star)\frac{d_{\text{th}}}{\bar{a}_d(-\theta_d^\star)}}, \tag{40}$$

where $\bar{a}_d(-\theta_d^\star)$ equals the constant service time of messages. Note that when message M_n enters the buffer at time instant $A_d(n)$, the backlog is $q(A_d(n))$. Because the service time for one message, i.e., $\bar{a}_d(-\theta_d^\star)$, is constant, the waiting time that message M_n experiences is $d(n) = q(A_d(n))\bar{a}_d(-\theta_d^\star)$. Then, we have $d(\infty) = q(\infty)\bar{a}_d(-\theta_d^\star)$ in steady-state and re-organize (40) as

$$\bar{a}_d(-\theta_d^\star)\theta_d^\star = -\lim_{q_{\text{th}}\to\infty} \frac{\Pr\{q(\infty) \geq q_{\text{th}}\}}{q_{\text{th}}}, \tag{41}$$

where $q_{\text{th}} = \frac{d_{\text{th}}}{\bar{a}_d(-\theta_d^\star)}$. With (41), the decay rate of the tail distribution of the backlog is $\theta_q^\star = \bar{a}_d(-\theta_d^\star)\theta_d^\star$. Further, since the service time for each message from the buffer is constant, the effective bandwidth (constant service rate) equals $\bar{a}_q(\theta_q^\star) = \frac{1}{\bar{a}_d(-\theta_d^\star)}$, which confirms the result in (22).

References

[1] F. Kelly, "Notes on effective bandwidths," in *Royal Statistical Society Lecture Notes Series*, vol. 4. London, U.K.: Oxford Univ. Press, 1995, ch. 8.

[2] A. J. Ganesh, N. O'Connell, and D. J. Wischik, *Big Queues*. New York, NY, USA: Springer, 2004.

[3] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *Proc. IEEE Conf. Comput. Commun.*, 1995, pp. 1001–1009.
[4] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.
[5] A. W. Berger and W. Whitt, "Extending the effective bandwidth concept to networks with priority classes," *IEEE Commun. Mag.*, vol. 36, no. 8, pp. 78–83, Aug. 1998.
[6] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 329–343, Jun. 1993.
[7] S. Mao, S. S. Panwar, and G. Lapiotis, "The effective bandwidth of markov modulated fluid process sources with a generalized processor sharing server," in *Proc. Global Telecommun. Conf.*, 2001, vol. 4, pp. 2341–2346.
[8] J. Pechiar, G. Perera, and M. Simon, "Effective bandwidth estimation and testing for markov sources," *Perform. Eval.*, vol. 48, no. 1, pp. 157–175, 2002.
[9] C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1442–1453, Dec. 2007.
[10] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
[11] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," *Mobile Netw. Appl.*, vol. 11, no. 1, pp. 91–99, 2006.
[12] S. Shakkottai, "Effective capacity and qos for wireless scheduling," *IEEE Trans. Automat. Control*, vol. 53, no. 3, pp. 749–761, Apr. 2008.
[13] L. Liu and J.-F. Chamberland, "On the effective capacities of multiple-antenna gaussian channels," in *Proc. IEEE Int. Symp. Inf. Theory*, 2008, pp. 2583–2587.
[14] E. A. Jorswieck, R. Mochaourab, and M. Mittelbach, "Effective capacity maximization in multi-antenna channels with covariance feedback," *IEEE Trans. Wireless Commun.*, vol. 9, no. 10, pp. 2988–2993, Oct. 2010.
[15] G. Femenias, J. Ramis, and L. Carrasco, "Using two-dimensional markov models and the effective-capacity approach for cross-layer design in amc/arq-based wireless networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4193–4203, Oct. 2009.
[16] L. Musavian, S. Aïssa, and S. Lambotharan, "Effective capacity for interference and delay constrained cognitive radio relay channels," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1698–1707, May 2010.
[17] S. Akin and M. C. Gursoy, "Cognitive radio transmission under qos constraints and interference limitations," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, no. 1, pp. 1–15, 2012.
[18] K. T. Phan, T. Le-Ngoc, and L. B. Le, "Optimal resource allocation for buffer-aided relaying with statistical qos constraint," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 959–972, Mar. 2016.
[19] W. Yu, L. Musavian, and Q. Ni, "Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3339–3353, May 2016.
[20] S. Efazati and P. Azmi, "Cross layer power allocation for selection relaying and incremental relaying protocols over single relay networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4598–4610, Jul. 2016.
[21] L. Zhao, X. Chi, and S. Yang, "Optimal aloha-like random access with heterogeneous qos guarantees for multi-packet reception aided visible light communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7872–7884, Nov. 2016.
[22] G. Ozcan, M. Ozmen, and M. C. Gursoy, "Qos-driven energy-efficient power control with random arrivals and arbitrary input distributions," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 376–388, Jan. 2017.
[23] R. Lübben, M. Fidler, and J. Liebeherr, "Stochastic bandwidth estimation in networks with random service," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 484–497, Apr. 2014.
[24] S. Akin and M. Fidler, "Backlog and delay reasoning in harq system," in *Proc. 27th Int. Teletraffic Cong.*, 2015, pp. 185–193.
[25] Y. Li, M. C. Gursoy, and S. Velipasalar, "On the throughput of hybrid-arq under statistical queuing constraints," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2725–2732, Jun. 2015.
[26] R. Sassioui, L. Szczecinski, L. Le, and M. Benjillali, "Amc and harq: Effective capacity analysis," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2016, pp. 1–7.
[27] J. Liebeherr, "Duality of the max-plus and min-plus network calculus," *Found. Trends Netw.*, vol. 11, no. 3/4, pp. 139–282, 2017. [Online]. Available: http://dx.doi.org/10.1561/1300000059
[28] W. Viriyasitavat, M. Boban, H.-M. Tsai, and A. Vasilakos, "Vehicular communications: Survey and challenges of channel and propagation models," *IEEE Veh. Technol. Mag.*, vol. 10, no. 2, pp. 55–66, Jun. 2015.
[29] S. Asmussen, *Applied Probability and Queues*, vol. 51. New York, NY, USA: Springer, 2008.
[30] A. Elwalid, D. Heyman, T. Lakshman, D. Mitra, and A. Weiss, "Fundamental bounds and approximations for atm multiplexers with applications to video teleconferencing," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1004–1016, Aug. 1995.
[31] C.-S. Chang, *Performance Guarantees in Communication Networks*. New York, NY, USA: Springer, 2000.
[32] Y. Jiang, "Network calculus and queueing theory: Two sides of one coin," in *Proc. 4th Int. ICST Conf. Perform. Eval. Methodologies Tools.*, 2009, Paper 37.
[33] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. 14th IEEE Int. Workshop Quality Servi.*, 2006, pp. 261–270.
[34] M. Fidler and Y. Jiang, "Non-asymptotic delay bounds for (k, l) fork-join systems and multi-stage fork-join networks," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.
[35] V. Palma, E. Mammi, A. M. Vegni, and A. Neri, "A fountain codes-based data dissemination technique in vehicular ad-hoc networks," in *Proc. Int. Conf. ITS Telecommun.*, 2011, pp. 750–755.
[36] R. Scopigno and H. A. Cozzetti, "Mobile slotted aloha for vanets," in *Proc. IEEE Veh. Technol. Conf. Fall*, 2009, pp. 1–5.
[37] D. J. MacKay, "Fountain codes," *IEE Proc. Commun.*, vol. 152, no. 6, pp. 1062–1068, 2005.
[38] M. Luby, "Lt codes," in *Proc. IEEE Symp. Found. Comput. Sci.*, 2002, pp. 271–280.
[39] A. Khisti, "Tornado codes and luby transform codes," 2003. [Online]. Available: http://Web.mit.edu/6.454/www/www fall 2003/khisti/tor_summary, pdf

Sami Akın received the B.S. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2005, and the Ph.D. degree in electrical engineering from the University of Nebraska—incoln, Lincoln, NE, USA, in 2011. Since December 2011, he has been with the Institute of Communications Technology, Leibniz Universität Hannover, Hanover, Germany, as a Research Scientist. He was the technical group leader of the Cognitive Radio for Audio Systems project funded by Lower Saxony Ministry of Science and Culture, and worked in the Towards a Unified Information and Queueing Theory project funded by the European Research Council Starting Grant. He is currently a Research Member of the Feedback-Less Machine-Type Communications project funded by the German Research Foundation (DFG). His research interests include wireless communications, signal processing, information theory, queueing theory, network calculus, and energy harvesting with a focus on wireless communications and networks under quality of service constraints.

Markus Fidler (M'04–SM'08) received the Doctoral degree in computer engineering from RWTH Aachen University, Aachen, Germany, in 2004. He was a Postdoctoral Fellow with the Norwegian University of Science and Technology, Trondheim, Norway, in 2005 and with the University of Toronto, Toronto, ON, Canada, in 2006. During 2007 and 2008, he was an Emmy Noether Research Group Leader with Technische Universität Darmstadt, Darmstadt, Germany. Since 2009, he has been a Professor in communications networks with Leibniz Universität Hannover, Hanover, Germany. He was the recipient of the Starting Grant of the European Research Council in 2012.