IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 6, 2020, accepted February 24, 2020, date of publication March 18, 2020, date of current version March 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981838

Modeling of Real-Time Multimedia Streaming in Wi-Fi Networks With Periodic Reservations

EVGENY KHOROV[1,2], (Senior Member, IEEE), ANDREY LYAKHOV[1,2], (Member, IEEE), ALEXANDER IVANOV[1], AND IAN F. AKYILDIZ[1,3], (Fellow, IEEE)

[1]Institute for Information Transmission Problems, Russian Academy of Sciences, 127051 Moscow, Russia
[2]Moscow Institute of Physics and Technology, 141701 Moscow, Russia
[3]Georgia Institute of Technology, Atlanta, GA 30332, USA

Corresponding author: Evgeny Khorov (e@khorov.ru)

This work has been carried out at IITP RAS and supported by the Russian Government under Contract 14.W03.31.0019.

ABSTRACT An important problem of modern Wi-Fis is the interferences caused by hidden stations active in the same area, or in multihop communications. All these issues significantly degrade the efficiency of the random channel access methods. Recent standardization and research activities are focused on solving coordination problems between various Wi-Fi devices. For example, the ongoing development of Wi-Fi 7 includes a coordinated schedule between the access points as a candidate solution. Consequently, Wi-Fi has many deterministic channel access mechanisms, which schedule channel time in a periodic manner well in advance and, thus, are utilized for streaming QoS sensitive data. However, both random traffic intensity and error-prone nature of the wireless channel complicate choosing such reservation parameters, i.e., the duration and the period of the reserved time intervals, that satisfy QoS requirements while minimizing channel time consumption. This paper introduces a general mathematical framework to solve the problem of choosing appropriate reservations parameters. The comparison of the analytical and simulation results show the high accuracy of the proposed framework. Finally, the paper gives an example of how to use the developed framework to maximize the network capacity.

INDEX TERMS Wi-Fi, QoS, IEEE 802.11s, IEEE 802.11aa, IEEE 802.11ad, IEEE 802.11ah, IEEE 802.11ax, IEEE 802.11be, MCCA, RAW, periodic reservations, mathematical model.

I. INTRODUCTION

Communication technologies are continuously evolving at a fast pace. While users enjoy mobile Internet access anytime and anywhere, system designers are anxious about the exponentially increasing number of wireless devices, most of which are battery supplied, and the exponential growth of traffic volume. Moreover, a considerable part of the traffic is real-time multimedia data sent by popular YouTube, IPTV, VoIP, online gaming, and video conferencing applications. Such real-time applications impose Quality of Service (QoS) requirements, which means that the network shall deliver the data within strict delay bounds and packet loss ratios (PLR). These challenges are reflected in the latest activities of wireless standardization bodies, in particular, in the recent amendments (IEEE 802.11s, 11aa, 11ad, 11ah, 11ax, 11ay, and 11be) to the Wi-Fi standard [1]. For example, 11aa targets at robust audio and video streaming, 11ad and 11ay enable transmission of uncompressed multimedia flows through mmWave communications, while the latest 11be focuses on support of the real-time applications and Virtual Reality.

The reliability of the Wi-Fi random channel access dramatically degrades due to the hidden station (STA) problem, which affects the satisfaction of QoS requirements. This is why the IEEE 802.11 Working Group actively uses scheduled access in the recent progress. Although even the earliest versions of the Wi-Fi standard include optional polling-based contention-free channel access mechanisms, they can hardly be efficient in emerging scenarios with overlapping or multihop networks and energy-limited STAs in the power-save mode.

We can improve the scheduled channel access by introducing some coordination between various Wi-Fi devices in a centralized or distributed manner. The distributed coordination is widely used in Mesh Wi-Fi (11s) and mmWave

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

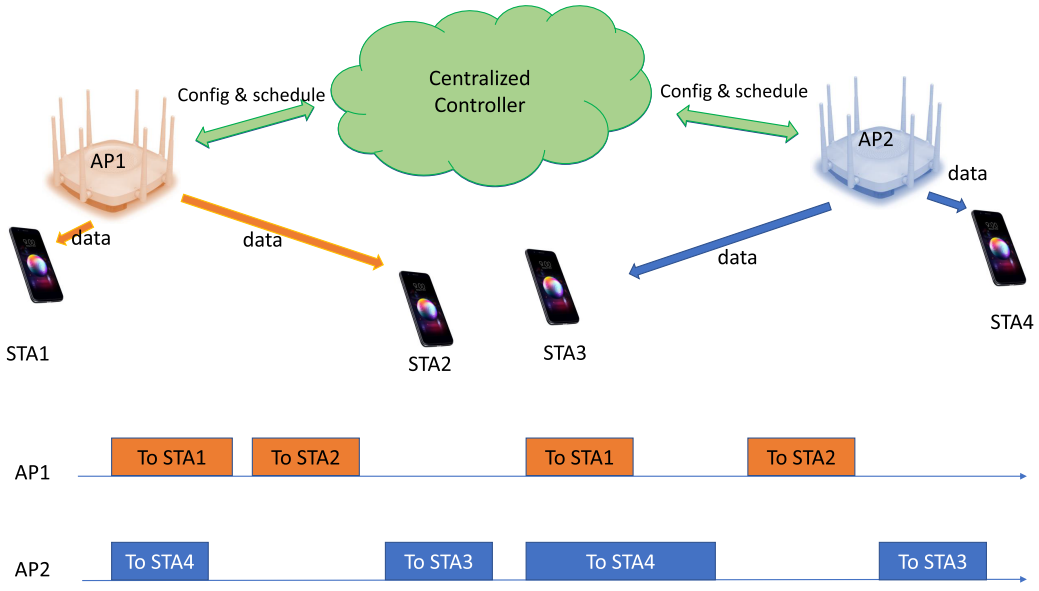


FIGURE 1. Cloud Wi-Fi architecture.

Wi-Fi (11ad/ay). In 11aa, the coordination between access points (APs) was achieved in a distributed manner, too. However, modern Wi-Fi deployments may have some centralized controller that manages the parameters of the APs, see Fig. 1. Such a controller also provides new opportunities to coordinate the channel access, including synchronization of transmission times, power, and frequencies. The coordination between APs is a widely discussed topic in the Task Group TGbe, which develops 11be a new amendment to the Wi-Fi standard, aka Wi-Fi 7 [2]. Moreover, the activities of 11be developers towards centralized AP coordination support the idea that Wi-Fi 7 may become the first Wi-Fi technology relying on such a centralized controller.

Both centralized and distributed coordination cannot be instant and reliable, which complicates their usage. To deal with these challenges, the above-mentioned amendments exploit the same idea based on periodic channel reservations. According to this idea, the STAs (or APs) can reserve and distribute *well in advance* the time intervals, during which only one STA is permitted to access the channel, while all neighboring STAs are forbidden. Such an approach eliminates collisions during the data transmission and helps to satisfy QoS requirements.

Although the channel reservations reduce the collision probability with the neighboring devices or hidden terminals, they cannot guarantee that the packet sent within the reservation is delivered for several reasons. First, the devices that do not support/respect the reservations (e.g., legacy ones) may transmit in the time intervals reserved by other devices. Second, because of the fast fading, the signal to noise ratio (SNR) at the receiver unpredictably varies in a high range. The selected modulation and coding scheme may be non-reliable enough to cope with instant channel quality degradation. Third, the reservation mechanisms cannot

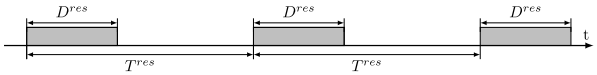


FIGURE 2. A periodic reservation.

protect transmissions from the interference even if all the STAs in the network support them. The information about reservations is disseminated among the one-hop neighborhood of both the sender and receiver. So, the reservation cannot protect transmissions from the interference induced by the distant STAs, i.e., the STAs outside the one-hop neighborhood. Although such STAs are rather far, their cumulative interference can be significant. In [3], it is shown that in a Manhattan scenario, such interference leads to packet error rates of up to 40% within the reservations. In the paper, we consider transmission failures caused by *random noise*. Because of random noise, some retries may be needed. It means that the STAs shall reserve more channel resources taking into account both initial packet transmissions and retries, although they do not know the exact number of needed retries. Moreover, the existing applications typically generate variable bit-rate (VBR) data flows, which make STAs reserve extra channel time just in case when the flow bit rate increases. Deterministic channel access mechanisms specified in the Wi-Fi amendments restrict the possible location of the reserved time intervals. To minimize the signaling overhead, they allow only reservation of periodic time intervals with some period T^{res} same duration D^{res} (Fig. 2). All these issues complicate the choice of the reservation parameters which satisfy QoS requirements of a given flow. This choice directly affects network performance because a too high amount of the reserved channel time decreases the number of concurrent data flows in the network, degrading the overall network capacity. At the same time, a too low amount may lead to a violation of QoS requirements.

The paper studies numerous channel access mechanisms defined in the aforementioned amendments to the Wi-Fi standard. Despite the different purposes of the amendments and different signaling, these mechanisms have much in common. Specifically, they allow reserving periodic channel times in advance.

While the benefits of various channel reservation mechanisms have been studied in detail in the literature (e.g., see [3]–[7]), this paper provides a detailed description of a mathematical framework to model data streaming with periodic reservations in Wi-Fi networks.

The framework can be used for choosing such reservation parameters that satisfy QoS requirements with minimal channel time consumption.

Specifically, the simplest model for individual transmission of a Constant Bit Rate (CBR) flow is initially proposed in [8]. In the present paper, we extend the study of this case by deep analysis of how the allocation of the reservations with respect to traffic affects PLR. Additionally, in contrast to all previous simulation studies, we study the impact of jitter in packet inter-arrival time on the performance. The model is extended to a VBR flow in [9], which we also include in this paper. The core contribution of the paper is threefold. First, we develop original models for more complex cases, e.g., described in Sections II-A and V-E block transmission. While developing these new models, we balance the model complexity and accuracy. Specifically, we consider the case when the transmission failures within a block are independent. Second, since the model is based on the Markov chains, for each chain, we prove that there is a unique stationary distribution. Third, we provide a solid solution for how to find appropriate reservation parameters of a transmission mode to meet QoS requirements and minimize the channel resource consumption while streaming QoS-sensitive data.

The structure of the paper is as follows. In Section II, we classify the possible transmission methods available in Wi-Fi and review the channel access mechanisms introduced recently, including those that are under development. In Section III, we state the formal problem. Section IV is devoted to existing works. In Section V-A, we develop a general approach to analytical modeling of QoS sensitive data streaming via periodic reservations. Following the proposed approach, we consider several data streaming scenarios and build their analytical models in Sections V-B–V-E. We start with simple cases. We explicitly indicate which results have been obtained before and are included in the paper for consistency and which are new. Then, we develop original models for more complex cases, e.g., block transmission, which requires some tricky techniques described in Section V-E. Here, we need to find a trade-off between model accuracy and complexity. Apart from that, we contribute with new theorems. In Section VI, we use the developed analytical models to find such reservation parameters that guarantee the satisfaction of the QoS requirements with the minimal amount of the reserved channel time. Additionally, we study the efficiency of the presented transmission methods under various conditions. We also extend the results obtained earlier for simple cases with the study of the influence of reservation positions with respect to the traffic arrival times. Apart from that, we compare the transmission methods and study which of them requires the minimal channel time to satisfy QoS requirements for a given flow. In Section VII, we conclude the paper, generalize its results, and discuss further work.

TABLE 1. List of acronyms.

ACK	Acknowledgement
A-MPDU	Aggregated MAC Packet Data Unit
A-MSDU	Aggregated MAC Service Data Unit
AP	Access Point
BlockAck	Block Acknowledgement
BT	Block Transmission
CBAP	Contention-Based Access Period
CBR	Constant Bit Rate
CF-End	Contention Free (period) End
DTIM	Delivery Traffic Indication Map
EDCA	Enhanced Distributed Channel Access
IPTV	Internet Protocol Television
IT	Individual Transmission
HCCA	Hybrid coordination function Controlled Channel Access
HoL	Head Of Line
MAC	Medium Access Control
MCCA	Mesh Coordination Function Controlled Channel Access
MCCAOP	MCCA Opportunity
OT	Ordered Transmission
PHY	Physical (layer)
PLR	Packet Loss Ratio
RTS/CTS	Request To Send / Clear To Send
QoS	Quality Of Service
QTP	Quiet Time Period
RAW	Restricted Access Window
RTP	Real Time Protocol
SIFS	Short Interframe Space
SP	Service Period
STA	Station
TDMA	Time Division Multiple Access
TBTT	Target Beacon Transmission Time
TX	Transmission
TXOP	Transmission Opportunity
UR	Unsolicited Retries
VANET	Vehicular Ad Hoc Network
VBR	Variable Bit Rate
VOIP	Voice Over IP
WPAN	Wireless Personal Area Network
WSN	Wireless Sensor Network

II. PACKET TRANSMISSION WITH DETERMINISTIC CHANNEL ACCESS IN Wi-Fi

To avoid collisions, many amendments for the Wi-Fi standard introduce the deterministic channel access mechanisms

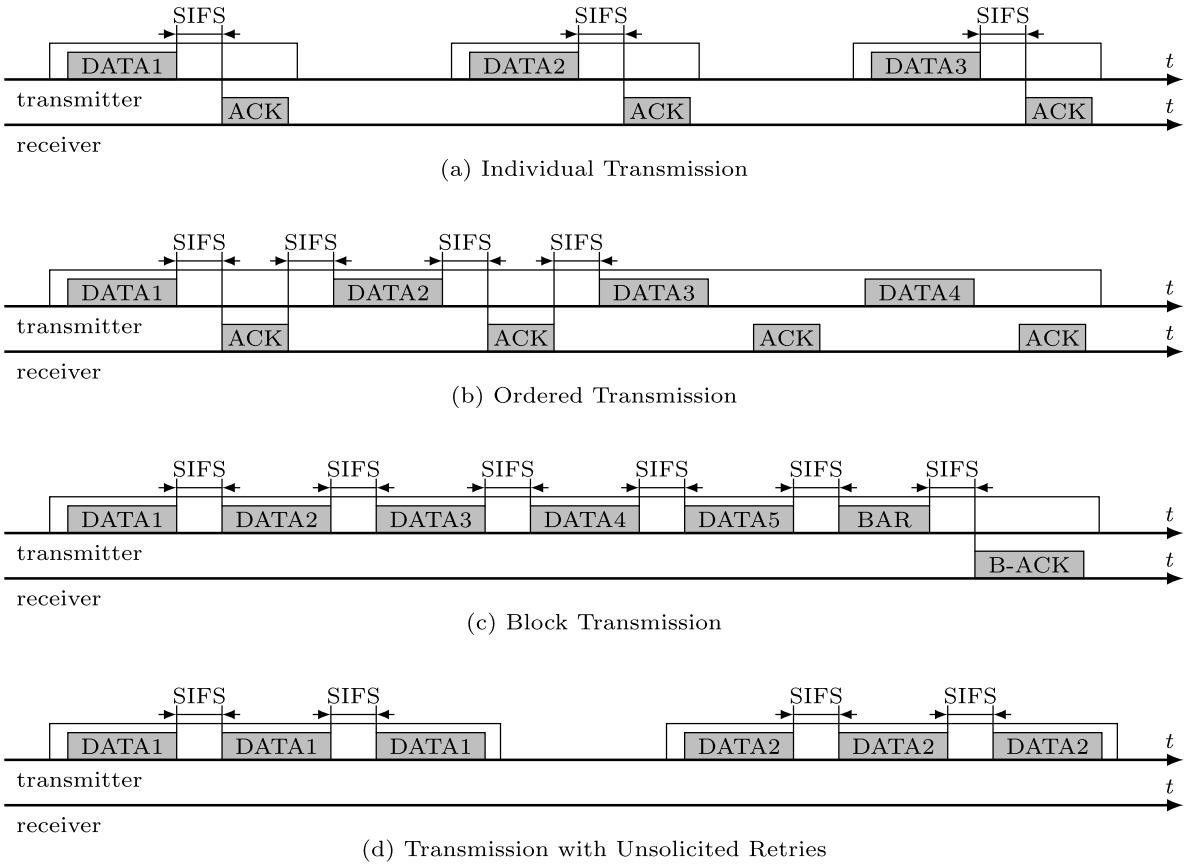


FIGURE 3. Various transmission methods.

which were designed for different use cases. So, they have different functionality and signaling. In spite of that, they implement the same idea. All the mechanisms reserve the channel (i) well in advance and (ii) in a periodic manner, i.e., the reserved time intervals are equidistant from each other and are of the same duration. Such approach decreases the overhead related to reservation parameters negotiation and advertisement of the reserved intervals to neighboring STAs since the whole series of reserved time intervals (further referred to as a *reservation*) can be described by just three parameters: the beginning of the first reserved interval, the duration of each interval, and the period, i.e., the time between beginnings of two consecutive reserved intervals (see Fig. 2). The possible number of packet transmissions in a reserved interval depends on its length and the used transmission method.

A. TRANSMISSION METHODS

Let us consider how the devices transmit data packets.

First of all, the packet reception requires an acknowledgment (ACK). The transmitter repeats undelivered packets until they are delivered or discarded. According to the standard, a STA discards a packet when the packet retry counter reaches some limit or when the packet lifetime expires. In this paper, we consider that the retry limit is set to a rather high value (thus, it is not used), while the lifetime is measured from the time moment when the packet appears at the transmitter.

The first Wi-Fi standard allows the devices to transmit packets one by one. To minimize the overhead induced by the acknowledgment frames, 11e introduces Block Acknowledgment (BlockAck), which allows sending a series of packets and then replying with a BlockAck carrying a bitmap indicating received packets. The high-throughput 11n amendment enables frame aggregation. It allows a STA to send several packets as a single long frame.

These options can be used with periodic reservations as follows, see Fig. 3.

With **Individual Transmission** (IT), a transmitter makes only one packet transmission attempt in a reserved time interval (Fig. 3(a)). When an intended receiver gets the frame, it immediately responses with a short ACK in the same reserved interval. The receiver sends ACKs since even in the case of deterministic channel access, the wireless channel is unreliable. If the transmitter receives the ACK, it considers the packet to be delivered. Otherwise, it retransmits the packet in the next reserved time interval.

If a reserved interval is long enough, the transmitter can make several transmission attempts (of one or more packets) in this interval. In terms of Wi-Fi, such behavior is called TXOP (Transmission Opportunity) operation. In this paper, we consider several transmission methods exploiting TXOP.

With **Ordered Transmission** (OT), the transmitting STA waits for an ACK after each transmission attempt (Fig. 3(b)). The STA retransmits a packet until it is delivered. Only after that, it starts the transmission of the next packet.

The ordered transmission has too high ACK-induced overhead. Although ACKs contain little useful information, they take much channel time because of the irreducible duration of PHY headers, interframe spaces, and a rather low rate of the mandatory modulation/coding schemes used for ACK transmission. To exclude ACKs from the frame exchange and, thus, to reduce the overhead, the Wi-Fi standard provides a flexible mechanism called Block-ACK. In this paper, we consider only one way of Block-ACK usage, which we refer to as **Block Transmission** (BT) (Fig. 3(c)). With BT, the STA transmits several packets in a row. Then it transmits a short Block-ACK request, and the receiver immediately replies with a Block-ACK frame, which contains a bitmap indicating delivered packets. Non-delivered packets are retransmitted during the next reserved time interval (together with several new packets if the room is enough).

Note that the standard allows aggregating several packets together. Such an aggregated frame has a single preamble and avoids any inter-frame spaces between the aggregated packets. In the case of an Aggregated MAC Packet Data Unit (A-MPDU), each packet within the aggregated frame has its sequence number and a checksum. A series of bits, called delimiter, separates the packets. If some of the aggregated packets are lost, the receiver may identify the start of the other frames and try to decode them. Thus, the transmission of A-MPDUs is similar to the shown block transmission, with the only exception related to zero inter-frame space.

In contrast to A-MPDU, the Aggregated MAC Service Data Unit (A-MSDU) has a single MAC header, sequence number, and checksum. Thus, the whole aggregated frame is either lost or delivered. So, the transmission of A-MSDUs can be modeled as an individual or ordered transmission depending on the duration of the frames and the reserved time intervals.

Enabled by 11aa, the transmission with **Unsolicited Retries** (UR) eliminates ACK-induced overhead, i.e., it uses no feedback information about packets reception. In this case, to achieve the required probability of delivery, a transmitter makes several successive transmission attempts of each packet (Fig. 3(d)).

In this paper, we consider QoS-sensitive data streaming with restrictions on the packet loss ratio PLR^{QoS}, and the delivery delay D^{QoS}. To satisfy the QoS requirements, the transmitter and receiver set up a periodic reservation. With any considered transmission method, within the reserved intervals, the packets are served in the First-In-First-Out order, and the transmitter repeats the packet until the packet is delivered (i.e., the transmitter receives an ACK) or the packet lifetime exceeds the delay bound D^{QoS}. If the packet is not delivered within D^{QoS} from the instant when it appears at the transmitter, it becomes outdated and gets discarded, which increases PLR.

The transmitter can reserve channel time using one of the Wi-Fi channel access mechanisms described below.

B. IEEE 802.11s (Wi-Fi MESH) –MCCA

The first mechanism with periodic channel reservation appears in the 11s amendment, which describes the functionality to support mesh networking. Apart from the routing framework, the amendment introduces an optional deterministic channel access mechanism called MCCA (Mesh Coordination Function Controlled Channel Access) [10]. The main reason for developing MCCA is the dramatic performance degradation of the Wi-Fi random channel access in a multihop environment caused by the hidden station effect.

MCCA allows a STA (called the owner) to reserve some time intervals (called MCCAOPs, MCCA OPportunities) during which it can transmit packets to another STA (called the responder), while all owner's or responder's neighbors are forbidden to access the channel.

Each reservation is a series of periodic MCCAOPs of the same duration, the positions of which are determined with respect to a series of so-called Delivery Traffic Indication Map (DTIM) beacons. A STA defines positions of both own and alien MCCAOPs with only three parameters: the *Offset* of the first MCCAOP from the DTIM beacon, the *Duration* of each MCCAOP, and the *Periodicity*, i.e., the number of MCCAOPs between two consequent DTIM beacons. Thus, the reservation period equals the time interval between consecutive DTIM beacons, divided by the value of the *periodicity*.

To establish a new reservation, a future owner and receiver carry out the MCCAOP setup handshake, by which they ensure that the new reservation does not overlap with the existing reservations in their neighborhood. After that, they advertise (e.g., by means of beacons) the established MCCAOP reservation among their neighborhood to prevent neighbors' transmissions in the reserved MCCAOPs. Thus, the MCCAOP protects the transmissions from collisions. However, the long reservation establishment procedure does not allow us to change the amount of reserved channel time on the fly.

C. IEEE 802.11aa (ROBUST AUDIO AND VIDEO)–HCCA TXOP NEGOTIATION

The HCCA (Hybrid coordination function Controlled Channel Access) TXOP Negotiation is another mechanism that uses periodic reservations. It is introduced in the 11aa amendment, which improves real-time audio/video streaming by extending basic 11e functionality. The HCCA was proposed in 11e as a polling-based contention-free channel access mechanism. It allows an AP to *instantly* allocate a channel time interval called TXOP for a single STA. However, because of the high density of APs, overlapping networks have become a typical scenario considered in many recent Wi-Fi amendments. In those cases, two neighboring APs may erroneously allocate the same time intervals to STAs transmissions, which may result in collisions.

With HCCA TXOP Negotiation, the APs can proactively negotiate the scheduled time intervals. To minimize the negotiation overhead, an AP reserves a periodic series of time intervals, called an HCCA TXOP reservation. The entire HCCA TXOP reservation can be described by three parameters: the duration of each Service Period, the Service Interval, i.e., the period of the reserved intervals, and the Offset from the next target beacon transmission time (TBTT) to the start of the first interval.

D. IEEE 802.11ad/ay (mm-WAVE Wi-Fi)—SERVICE PERIOD

11ad and 11ay [5] bring Wi-Fi to the market of multigigabit wireless personal area networks (WPANs) by exploiting the 60 GHz frequency band. On the one hand, the operation in 60 GHz band, achieves high throughput (up to 8 Gbps for 11ad and over 270 Gbps for 11ay), results in significant signal strength attenuation and transmission range decreasing on the other hand. To extend the coverage of 11ad networks, the amendment allows devices to exploit directional communication. Directional transmissions exacerbate the hidden terminal problem and additionally affects the random access.

To improve the channel resource utilization, one of STAs in an 11ad/ay network may be chosen as a central coordinator. This STA divides the medium time into beacon intervals of equal duration and sends beacons (in one or several directions) at the beginning of each beacon interval. It also splits each beacon interval into CBAPs (Contention-Based Access Periods) and SPs (Service Periods) and advertises the schedule via beacons. During CBAPs, any STA in the network can transmit using random access, whereas SPs are granted to the STAs for dedicated usage.

Again, to minimize the advertisement overhead and to simplify the SP management, the central STA allocates SPs in a periodic manner. The SP reservation is a sequence of periodic groups of SPs so that groups follow periodically (with some network-wide period). Each group is described by the start time of the first SP, the duration of the SPs, the period between two consecutive SPs, and the number of SPs in the group.

As previously mentioned, on the one hand, the channel reservation is favorable for directional transmissions. On the other hand, directional transmissions allow the coordinator to overlap SPs belonging to different links if the directional transmissions over these links do not interfere. This feature helps to combat the effect of the exposed terminals since neighboring STAs can share the same channel resource.

E. IEEE 802.11ah (Wi-Fi HaLoW) —PERIODIC RESTRICTED ACCESS WINDOW

11ah is a recent Wi-Fi amendment for the Internet of Things scenarios [11]. It works in $\leq$1 GHz frequency bands, which allows setting up radio links up to 1 km at the default power of 200 mW. The amendment aims to gather information from up to 8000 sensor STAs and to provide offloading to mobile devices. For the latter case, 11ah supports rather high data rates: up to 350 Mbps in a 16 MHz channel.

To limit the contention of such a large number of STAs, 11ah introduces a novel channel access mechanism called Restricted Access Window (RAW). RAW allows an AP to limit the set of STAs accessing the channel and to spread access attempts over a long time. Briefly, RAW operates as follows [11]:

The AP defines a time interval called RAW and splits it into slots. After that, it assigns each slot to a STA or a group of STAs. The STAs can only transmit in assigned slots. The AP can define a periodic series of slots, which can be used for long-term regular flows. The AP defines the entire series of reserved time intervals, similar to the mechanisms mentioned above.

F. IEEE 802.11ax (Wi-Fi 6) —QUIET TIME PERIOD AND PERIODIC TRIGGER FRAMES

To protect from interferences caused by ad hoc networks or direct link transmissions, the 11ax amendment introduces the Quiet Time Period (QTP) mechanism. QTP allows a STA to request the AP for a series of periodic time intervals during which all other STAs are not permitted to access the channel. As explained above, the AP defines a series of time intervals by the position (or offset) of the first reserved interval, the duration of each interval, their period, and the total number of requested intervals. QTP can also be used to eliminate the interference between neighboring networks operating in the same channel. For that, a centralized controller may split the channel time between the neighboring networks.

Apart from that, 11ax introduces trigger-based multi-user scheduled access, which can be carried out periodically [12].

G. IEEE 802.11be (Wi-Fi 7)—COORDINATED SCHEDULE

Before 11be, the amendments for the Wi-Fi standard were just limited to MAC and PHY issues. That means the coordination between the APs was possible only through the air interface. Meanwhile, many Wi-Fi vendors (including Cisco/Miraki [13], Huawei [14], HP/Aruba Networks [15], Quantenna Communications [16], and others) have developed special software systems that can gather statistics and manage large campus Wi-Fi deployments. Such systems are used in many office and residential scenarios where a single operator deploys the APs and remotely controls their behavior. Apart from tuning the configuration parameters of the APs, such coordination provides new opportunities.

11be is the first standard that considers such an opportunity to provide real-time applications and provide efficient channel usage in dense deployments. Although, by now, the amendment is at the initial stage without any draft specification, several ideas on the coordinated schedule are being discussed in the IEEE 802.11 Working Group.

Consider two APs, each of which has two associated STAs, see Fig. 1. STA1 and STA4 are much closer to their own APs than to the interfering APs. So the APs can use the same time-frequency channel resources to transmit data via different links. At the same time, STA2 and STA3 are far from their

own APs and may experience strong interference from the neighboring APs. Thus, the AP can only use different channel resources (frequency or time) to transmit them data [2]. The exact approach of how to design such a coordinated schedule is under development in the IEEE 802.11 Working Group. Also, it is not clear whether the Wi-Fi 7 standard will allow coordinating the APs that belong to different operators. The lack of such a feature is not a significant limitation because there is a single operator in many enterprise deployments.

Naturally, the centralized controller may have a solid view of the interference in the network. It can split the channel resources between these transmissions in order to eliminate inter-cell interference if it is notable or to improve spatial reuse if the interference is negligible. In practice, the scheduling problem is difficult. First of all, the communication between the APs and the remote controller is not instant. Second, the AP needs some time to block channel access by its associated STAs during the time intervals allocated for another AP only. These issues can be solved again by allocating periodic series of channel times.

Although, by now, no scheme for communication between the AP and the centralized controller is standardized, a possible solution can be as follows [17]. Each AP measures the parameters of the stream (e.g., the rate and its fluctuation) and the channel quality. In each time window, the AP reports information about the streams and channel quality to the centralized controller. For example, with Real-Time Protocol (RTP) [18], a video is streamed frame-by-frame, and the traffic may look like a sequence of batches of random size. The channel quality can be described by the Modulation and Coding Schemes (MCS) used to transmit the flow and the probability of transmission attempt failure.

Given the reported values, the controller calculates how much channel time the AP needs to stream a particular video flow and which reservation parameters and which transmission method can be used. Then, it sends these results to APs. In this paper, we develop a model that can be used by the controller to select appropriate reservation parameters and a transmission method. We state the formal problem more precisely in Section III.

III. PROBLEM STATEMENT

In this paper, we study QoS-sensitive data streaming between two STAs with a periodic reservation over a noisy channel. Despite the reservation, the packet transmissions in a reserved time interval may not be protected from the random noise and interference from distant and legacy STAs. Let q be the probability of an unsuccessful transmission attempt. For the sake of simplicity, we assume that the consequent transmission attempts fail independently from each other, and short ACK frames are reliably delivered.

We consider the following types of input flows.

- **Constant Bit Rate (CBR)** shown in Fig. 4(a). The packets arrive at the transmitter at a certain time within the period T^{in}. For example, such a flow can be generated by the G.711 [19] or G.729 [20] voice codecs.
- **Bursty Variable Bit Rate (VBR) flow** shown in Fig. 4(b). The packets arrive periodically in batches of random size, that is, the time interval between consequent batches is constant and equals T^{in}, while the batch sizes are independent random variables, and with the probability p_i^{in} the batch size equals i, where $i \in \{1, \ldots, M\}$. Such a flow is typical for compressed video streaming via RTP [18] when video frames are not of the same size, and each frame may contain different numbers of packets. Obviously, the CBR flow is a particular case of the VBR flow with $p_1^{in} = 1$ and $p_i^{in} = 0, i > 1$.

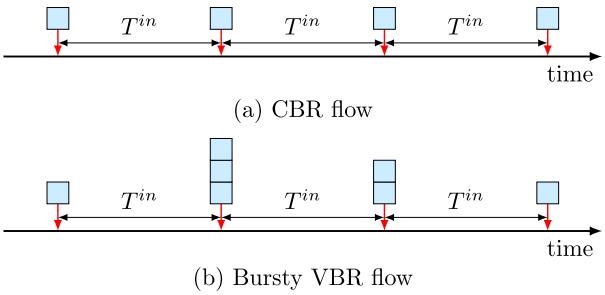


FIGURE 4. Various traffic patterns.

To transmit a flow, we set up a reservation with the period T^{res} and the duration D^{res} in the reserved time intervals. Then we select some transmission method (see Section II-A). The problem solved in this paper is *how to choose such reservation parameters and the transmission method that guarantee the QoS requirements (expressed by values D^{QoS} and PLR^{QoS}) utilizing the minimum part of reserved channel time C*:

$$C = \frac{D^{res}}{T^{res}}. \tag{1}$$

Note that if the STAs reserve the channel time too long, then the common channel time will be wasted. It can happen both when the reserved time interval starts, but the STA has no data to transmit or when the amount of data to transmit is less than that which can be transmitted within the reserved time interval. Although the standard allows releasing such time by sending a CF-End frame, this may be inefficient for several reasons. First, some of the STAs may not receive the CF-End since they are located too far from the transmitter but close to the receiver of the corresponding reservation. Second, they may ignore this CF-End according to the latest rules of virtual carrier sense (namely double NAV) specified in 11ax. Third, at the beginning of the reserved time interval, the neighboring STAs may switch to the doze state, so that they may not receive this CF-End. For these reasons, we consider that the channel time is consumed once it is reserved, regardless of whether any transmission occurs in the reserved time interval or not.

TABLE 2. Input and output parameters.

Input parameters	
t^{in}	Inter-arrival time of batches of packets
$\{p_i^{in}\}$	Distribution of sized of batches
D^{QoS}	Maximal affordable packet delay
PLR^{QoS}	Maximal affordable portion of packets not delivered within D^{QoS}
q	The probability of a packet transmission failure inside a reserved time interval
L	Packet length
MCS	MCS at which packets are transmitted within the reservation
Output parameters	
$TXMETHOD$	The best transmission method out of those specified in Section II-A
t^{res}	Period of the reservation
B	The number of transmission attempts within one reserved time interval (in case of IT, $B = 1$)

To find the optimal reservation parameters, we develop a solid analytical framework by modeling of QoS sensitive streaming via periodic reservations. To describe the solution, we consider several combinations of input flows and transmission methods, from the simplest one with a CBR flow and the individual transmission to the most complicated one with a VBR flow and block transmission.

The framework allows us to obtain PLR (packet loss ratios) and the channel resource consumption for a given flow, the transmission method, and the reservation parameters. Thus, by iterating over different parameters, we can find reservation parameters and transmission methods that minimize the channel time consumption while satisfying the predefined QoS requirements. In other words, we can find optimal output parameters for any given input parameters listed in Table 2.

IV. RELATED PAPERS

Periodic reservations are a type of Time Division Multiple Access (TDMA) which is a part of almost every existing wireless technology. Thus, the problem of channel reservation has been widely studied in the literature in the past. For example, WiMedia is one of the early wireless technologies used TDMA-based reservations. However, the WiMedia standard does not specify any reservation algorithms. For example, to deal with quite intricate WiMedia MAC allocation restrictions, a reservation algorithm is proposed in [21] that decomposes an incoming data flow into several sub-flows with consecutive scheduling of the sub-flows. As for video streaming, a dynamic resource allocation algorithm is proposed in [22] based on a specially developed VBR flows rate predictor. Although WiMedia is now out of date, the ideas are still noteworthy.

TDMA-based reservations are often studied as a part of routing protocols in multihop networks. More specifically, the TDMA-based reservation and dynamic channel reservation schemes are proposed in [23] and [24].

The concept of Pseudo TDMA (PTDMA) [25] for distributed systems has gained the interest of the research community in recent years. In PTDMA, each STA starts its transmission using random access, but after the first successful transmission, it switches to regular TDMA-based scheduling of packet transmissions, in order to repeat the sequence of successful channel accesses. A dynamic self-organizing TDMA-based MAC is proposed in [26], which outperforms simple PTDMA in terms of throughput and QoS provisioning. In [27], a distributed PTDMA scheme for mesh networks is introduced where devices can transmit to, or receive from multiple neighbors simultaneously, i.e., they exploit the so-called Multi-Transmit-Receive capability.

Lately, TDMA receives much attention due to vehicular communications and industrial wireless sensor networks (WSN). A good survey on TDMA-based MAC protocols for Vehicular Ad Hoc Networks (VANETs) can be found in [28]. For example, a TDMA protocol is proposed in [29] for vehicular communications, which is based on the prediction of encounter collisions and accordingly trying to avoid them in advance. A hybrid TDMA is proposed in [30] as a multichannel MAC protocol for VANETs that allows efficient broadcasting of messages and increases the throughput on the control channel.

As for sensor networks, TDMA is adopted by many modern industrial wireless standards like Wireless HART [31], ISA 100.11a [32], and IEEE 802.15.4e [33] which combine TDMA scheme with deadline constrained transmission scheduling policies to enhance the transmission reliability over lossy wireless channels.

An example of such a transmission policy based on periodic slots allocation is given in [34], and some results are obtained for some specific topologies. For example, a TDMA scheduling algorithm for tree topology WSNs is introduced in [35], where the proposed algorithm determines a periodic and collision-free allocation of time slots to sensors such that the end-to-end deadline of each data flow is satisfied in the tree topology WSNs.

Wi-Fi also adopts TDMA and defines a palette of channel reservation methods but without specifying how much channel time should be reserved to provide QoS. As a consequence, this problem attracted much attention from the research community recently. One of the solutions for

channel reservation is a polling-based HCCA (Hybrid coordination function Controlled Channel Access) introduced in the 11e amendment along with random EDCA (Enhanced Distributed Channel Access) mechanism. It allows an AP to instantly allocate TXOPs to STAs desiring to communicate. The performance of the HCCA depends on the HCCA scheduler, the standard version of which is described in Annex K in the 11e amendment. Using the standard scheduler, the AP polls all STAs with the same period, and durations of the granted TXOPs calculated based on the STAs' traffic intensities. This scheduler is simple and can be efficient if the traffic is strictly CBR. However, when real-time applications such as video conferencing generate VBR traffic, this scheduler suffers from unnecessary polling of STAs and resource wastage, as shown in [36]–[40]. These papers emphasize deficiencies of the standard scheduler and propose their own schedulers, which outperform the standard one in terms of PLR, average delays, the number of admitted flows, etc.

One of the ways to enhance HCCA is considered in [36] and consists of polling STAs with different periods. The proposed Scheduling Based on Estimated Transmission Times - Earliest Due Date (SETT-EDD) scheduler provides lower PLR and average delays than the standard scheduler. The FHCF (Fair Hybrid Coordination Function) is proposed in [37], which tries to be efficient for both CBR and VBR flows under delay constraints. FHCF uses queue length estimations to tune AP's TXOP allocations to STAs. This idea is further evolved in [38], where an ARROW (Adaptive Resource Reservation Over WLANs) algorithm is proposed. ARROW performs channel allocations based on the actual traffic buffered in the various STAs, i.e., on the exact transmission requirements, which improves its performance in comparison with SETT-EDD in terms of PLR, average delay and TXOP utilization (less TXOPs are wasted). The scheduler proposed in [39] is based on dividing activities into online and offline. Offline activities are performed at the admission control time scale and consist of computing polling schedules based on the earliest deadline first policy (high computational cost). On the other hand, online activities consist of reading the pre-computed schedule and are performed at the frame transmission timescale (little or no computational cost). An analytical model of standard HCCA scheduler is developed in [40]. The analysis reveals the already mentioned performance drawbacks of the standard scheduler, i.e., high delays and PLR. Also, a scheduler is proposed in [40], which i) takes into account the different delay bounds of data flows and ii) tries to predict VBR flows intensities. A new analytical model is proposed in [41] for performance analysis of channel access in 802.11ad, which takes reservations and directional antennas into account.

While the proposed schedulers are shown to be efficient in the case of a perfect channel, their performance regarding packet errors is questionable since they do not consider the error-prone nature of the wireless channel. The imperfectness of the channel complicates the choice of HCCA TXOPs parameters (polling periods and durations) required to satisfy the QoS requirements. Additional complications come from VBR flows, i.e., their bursty nature additionally makes it difficult to predict the proper amount of resources to allocate.

The simplest way to consider packet errors is described in Annex N of the Wi-Fi standard [1]. It suggests to reserve time for additional transmission attempts. In particular, given the packet error rate q, the standard suggests reserving channel time $\frac{1}{1-q}$ times more than it is needed to transmit each packet once. Obviously, and the standard emphasizes it, such amount of channel time is enough to guarantee zero PLR for a delay-tolerant flow. However, it may not be enough to satisfy QoS requirements for delay-sensitive traffic, e.g., real-time audio and video.

Apart from the multimedia traffic, periodic reservations have been studied in the context of Wi-Fi HaLow and Industrial Internet of Things applications recently. An analytical model is presented in [42] explaining how to determine a set of proper values of RAW parameters to satisfy the requirements of real-time applications. Among the considered transmission methods, the Unsolicited Retries is the easiest one to account for packet transmission errors in streaming QoS sensitive data [43]. If a transmitter makes N successive transmission attempts for each packet and all N attempts are made before the packet lifetime reaches D^{QoS}, then PLR equals q^N. The QoS requirements of the flow are satisfied if $N \geq \lceil \log_q PLR^{QoS} \rceil$. Despite the simplicity of such an approach, it has a significant drawback. If the transmitter delivers a packet during one of its N transmission attempts, then the rest of the retries cannot be used to transmit other packets present in the queue. On the contrary, individual, ordered, and block transmissions do not suffer from this problem.

For individual transmissions of a CBR flow, the transmission errors are considered in the reserved time intervals, and an analytical model is developed in [8]. To the best of our knowledge, it is the first model that allows determining the reservation parameters which guarantee the satisfaction of the delay and PLR requirements. In [44], the model is extended to a non-bursty VBR flow where packets in each flow may not arrive at the transmitter periodically. The model in [8] is extended to a bursty VBR flow in [45].

Our paper extends the results obtained earlier for individual [8], [17], [45] and ordered [9] transmission methods to *block transmission* and *transmission with unsolicited retries*. It also provides a solid solution on how to find appropriate reservation parameters for a transmission method to satisfy QoS requirements and how to minimize the channel resource consumption, while streaming QoS-sensitive data.

V. PROPOSED MATHEMATICAL FRAMEWORK

A. GENERAL IDEA

Whatever input flow is transmitted and the transmission method is used, we represent the streaming process as a discrete-time Markov chain with time unit T^{res}, so that the time instances t and $t+1$ correspond to the beginnings of consecutive reserved time intervals. One or several integer

values determine the state of the Markov chain. The number of these values and their meanings depend on a particular flow and transmission method. In any case, they should allow obtaining the age of the Head of Line (HoL) packet at any instance t. After determining the transition probabilities and the stationary distribution of the probabilities of the states, we can then calculate the PLR values for the given input parameters as

$$PLR = \frac{I^{dis}}{I^{in}}, \tag{2}$$

where I^{in} and I^{dis} are the average number of the packets appeared at the transmitter and the average number of discarded packets within T^{res}.

An essential feature of the proposed approach is that the ages of packets are expressed in specially introduced time units, which we refer to as slots. The duration of a slot is defined as follows

$$\tau = \gcd(T^{in}, T^{res}).$$

Although $\gcd(\cdot)$ is defined only for integer arguments, one can find values T^{in} and T^{res} as non-integer. However, by expressing them in appropriate units (like μs or ns) and then rounding them, we can obtain their integer approximation with any given degree of accuracy.

Next, we express all time values in slots:

$$t^{in} = \frac{T^{in}}{\tau} \in \mathbb{N}, \quad t^{res} = \frac{T^{res}}{\tau} \in \mathbb{N}.$$

We split the time axis into slots so that each reserved interval starts at some slot boundary (see Fig. 5). It is feasible since the period T^{res} is a multiple of τ. Note that in reality, the moments when packets appear and the reservations positions are subject to random shifts and drifts so that the performed alignment is not possible in practice. However, for the sake of simplicity, we do not consider these effects during the development of the mathematical model of the QoS sensitive data streaming.

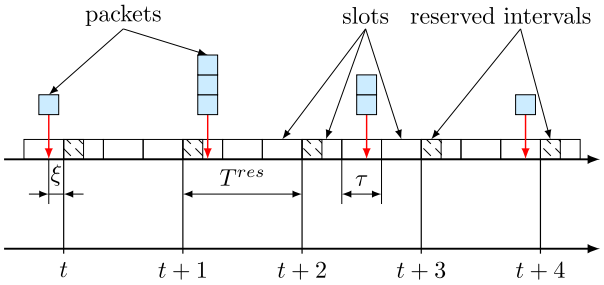


FIGURE 5. The Markov chain time instances.

Let ξ be the time interval between two time instances: when a packet appears at the transmitter and when the next slot begins ($0 \leqslant \xi < \tau$). Note that for all packets, ξ is the same. Let us virtually shift all packets generation instances forward by ξ and simultaneously reduce the delay bound D^{QoS} by ξ. Such modification leaves the transmission process unchanged, though it eliminates the parameter ξ from further derivations. In Section VI, we consider this value again to study its influence on the transmission process.

Let the age of a packet (or a batch) be the time elapsed from the instant when the packet appeared. According to this formal definition, packets in the queue have a non-negative age, and those which will appear in the future have a negative age. We define *Head of Line (HoL) packet* of the queue as the oldest packet among those which are now in the queue or will appear in the future. Similarly, the *HoL batch* is the batch to which the HoL packet belongs. Unless the opposite is stated explicitly, we use parameter $a(t)$ in the models presented in this paper to express the age of the current HoL packet (in time slots) at the observation moment t. Due to the performed shifts, at any observation moment $(t, t+1, t+2, \dots)$ the age $a(t)$ of the HoL packet at the observation moment t equals an integer number of slots. Specifically:

- If the queue is not empty at moment t, then $a(t)$ is the number of whole slots the HoL packet spent in the queue. In this case, $a(t) \geq 0$ and $a(t) \cdot \tau > 0$ equals the time the HoL packet spent in the queue.
- If the queue is empty, then $a(t) < 0$ and $|a(t) \cdot \tau|$ is the time to the next packet appearance.

Since all outdated packets are immediately dropped, the age of the HoL packet cannot exceed D^{QoS}, i.e., at any moment t holds $a \cdot \tau \leq D^{QoS}$ and the greatest value of $a(t)$ equals $d = \left\lfloor \frac{D^{QoS}}{\tau} \right\rfloor$. As for the least possible value of a, it is observed when immediately after a batch arrival, a reservation occurs, and all the packets from the queue are delivered. Thus, immediately after the reservation, the queue is empty, and the next batch arrival occurs in t^{in} slots. However, the next time instance of the observation of the chain occurs in t^{res}. At this time instant, the age of the HOL packet in the queue is $\max\{t^{res} - t^{in}, 0\}$.

Above, we assume that a packet is not considered to be outdated and can be transmitted in a reserved time interval if its age is not higher than D^{QoS} at the beginning of this interval. However, for a packet not to be discarded by both the transmitter and the receiver, its *overall delivery time* must be lower than D^{QoS}. The *overall packet delivery time* is composed of two components: a) the time the packet spends in the queue until the beginning of a reserved interval in which it is delivered and b) the time from the beginning of this interval up to the time when the receiver successfully obtains the packet. In case of existing Wi-Fi network, the latter component is usually several orders of magnitude smaller than the former one[1] and does not have much influence (as shown by numerical results in Section VI). Nevertheless, one can easily account for this by decreasing the initial D^{QoS} value specified in the QoS-requirements by duration D^{res} of the reserved time interval.

[1]Typical D^{QoS} values are about hundreds of ms, while in modern Wi-Fi networks, a packet transmission time is much less than 1 ms.

In the next sections, we show how to use this approach described above to model QoS sensitive data streaming in various scenarios. We start from the most straightforward case of a CBR flow and individual transmission, focusing on those findings, which are not published yet.

B. CBR FLOW WITH INDIVIDUAL TRANSMISSION

Let us demonstrate how to use the approach proposed in Section V-A to analyze CBR flow transmission [8]. In this section, we assume that $t^{res} \le t^{in}$ and $t^{in} \le d$. Indeed, it is senseless to set up a reservation with period $T^{res} > T^{in}$ since it results in packet losses barely because of the insufficient amount of resources, even in the case of the perfect channel conditions ($q = 0$). Additionally, we do not consider flows, for which $T^{in} \ge D^{QoS}$ since they are not typical for multimedia streaming. Under these conditions $a \in \{t^{res} - t^{in}, \ldots, d\}$. The minimum value $t^{res} - t^{in}$ is achieved at instant t if a packet appears in an empty queue at instant $t-1$, and it is delivered at the first attempt.

The state of the Markov chain at instant t can be represented by the only parameter a. Indeed, since the interval between the appearance of two packets equals t^{in} (in slots), given a, we know the age of the HoL packet in the queue and the age of the next packet, if any. Thus, we can write down possible transitions and their probabilities.

- If $a < 0$, then the queue is empty at instant t and there are $|a|$ slots before the next packet appearance. If $a < -t^{res}$, then no packets appear by instant $t+1$ and the queue is still empty at this instant, though the number of slots to the nearest appearance decreases to $|a + t^{res}|$. If $a \ge -t^{res}$, then a packet appears by instant $t+1$ and its age equals $a + t^{res}$ at instant $t+1$. Thus, no matter what the value of a is, the process transits to state $a + t^{res}$ with probability 1.
- If $0 \le a \le d - t^{res}$, then the queue is not empty and the oldest packet does not become outdated by instant $t+1$. If the transmission attempt is unsuccessful in the current reserved interval (which happens with probability q), then at instant $t+1$ the age of the oldest packet equals $a + t^{res}$, i.e., the process transits to state $a + t^{res}$. Otherwise (with probability $1-q$), the process transits to state $a + t^{res} - t^{in}$ since the oldest packet is successfully transmitted and the second packet in the queue becomes the oldest one at instant $t+1$.
- If $a > d - t^{res}$, the queue is not empty and the oldest packet becomes outdated at instant $t+1$. No matter whether the transmission attempt is successful or not, the process transits to state $a + t^{res} - t^{in}$.

Having reproduced the chain from [8], let us contribute with the analyses of its properties.

Theorem 1: The chain described in Section V-B is irreducible and positive recurrent if $0 < q < 1$, $t^{in} > t^{res}$, *and* $t^{res} \le d$.

Proof: A Markov chain is said to be irreducible if it is possible to get to any state from any state.

Let us refer to the states with $a = t^{res} - t^{in}, \ldots, 2t^{res} - t^{in} - 1$ as the *basic* states. Note that the values of a of the basic states form a complete residue system modulo t^{res}.

Let a^* be a basic state. From this state, the chain can transit with non-zero probability to a state $a^* + kt^{res}$, $k = 1, 2, \ldots$ until $a^* + kt^{res} > d - t^{res}$. All these states are congruent with a^* modulo t^{res}. The union of subsets $\mathfrak{A}(a^*) = \{a^* + kt^{res} : a^* + kt^{res} < d, k = 0, 1, 2, \ldots\}$ form the full set of states of the chain. Thus, any non-basic state is reachable from some basic state.

Consider a non-basic state $\tilde{a} \in \mathfrak{A}(a^*) \backslash a^*$. Now let us prove that all basic states are reachable from any other non-basic state a. For that, we consider the following sequence of the chain transitions. At each step, if the current state $a \ge 0$, i.e., there is a packet in the queue, the chain transits to $a - t^{res} + t^{in}$, which corresponds to a successful transmission attempt. Otherwise, i.e., if the queue is empty, the chain transits to $a + t^{res}$. Thus, while the queue is not empty, every step, the value of a goes down. The states, to which the chain transits, are $a^* = a + kt^{res} + lt^{in}$, $k, l \in \mathbb{Z}$, $d - t^{in} < a^* \le d$. As $\gcd(t^{res}, t^{in}) = 1$, such states form a complete residue system modulo t^{in}. Since $t^{res} < t^{in}$, a^* can be equal to any value $d - t^{res} < a^* d \le d$, i.e., all possible values of a^* such that $d - t^{res} < a^* d \le d$ form a complete residue system modulo t^{res}.

An irreducible finite-state chain is positive recurrent. Thus the considered chain is positive recurrent, and there is a single stationary distribution. □

Thus, having obtained the transition matrix $\boldsymbol{P}$, we can easily find the stationary distribution $\boldsymbol{\pi}_a$ of the Markov chain states by solving the following system of linear equations:

$$\begin{cases} \boldsymbol{\pi}_a^T \boldsymbol{P} = \boldsymbol{\pi}_a^T, \\ \sum\limits_a \pi_a = 1. \end{cases} \tag{3}$$

Moreover, since the chain is irreducible and positive recurrent, there is the only solution of (3) and $\forall a \Rightarrow \pi_a = 1$.

Note that if $t^{in} \le t^{res}$, the chain cannot transit from any of t^{res} stages with the highest values of a to $a \le d - t^{res}$. However, even in this case, all the states still form the same communicating class. Since the chain is time-homogeneous and all the states belong to the same communicating class, there exists the only solution of (3).

To find PLR with (2), we need to obtain I^{in} and I^{dis}. For a CBR flow,

$$I^{in} = \frac{1}{T^{in}}. \tag{4}$$

To find I^{dis}, we note that a packet is only discarded after an unsuccessful transmission from such state a that $a > d - t^{res}$. So, the average rate of the discarded packets is

$$I^{dis} = \frac{q}{T^{res}} \sum_{a > d - t^{res}} \pi_a. \tag{5}$$

By substituting (4) and (5) in (2) we obtain PLR.

C. BURSTY VBR FLOW WITH ORDERED TRANSMISSION

In this Section, we consider *ordered transmission of a bursty VBR flow* [9]. The duration of reserved time intervals allows up to B successive packet transmissions using the ordered transmission. That is why we do not assume that $t^{res} \le t^{in}$ since even when $t^{res} > t^{in}$, we can compensate for the sparseness of the reserved intervals with a high value of B. The only remaining restrictions are $t^{in} \le d$ and $t^{res} \le d$. We do not consider the individual transmission of a bursty VBR flow separately since it is a particular case of ordered transmission for $B = 1$.

To be able to describe the transmission of a bursty VBR flow in the same fashion as in Section V-B, we introduce a new state variable m which equals the number of packets of age a, i.e., the number of packets in the oldest batch. So, the state of the Markov chain at instance t is defined by a pair of integer numbers (a, m).

To simplify the description of transitions, we split a transition from state (a, m) to state (a', m') at instant t into $B + 1$ intermediate transitions by introducing B intermediate states which correspond to the system states after each of B transmission attempts (see Fig. 6). We can describe each intermediate state i, $i \in \{1, \dots, B\}$, with pair $(a^{(i)}, m^{(i)})$ where $a^{(i)}$ and $m^{(i)}$ have the same meanings as defined for the main Markov chain states but related to the time moment immediately following the end of transmission attempt i. We formally put that $(a^{(0)}, m^{(0)})$ stands for (a, m) at the beginning of a reserved interval.

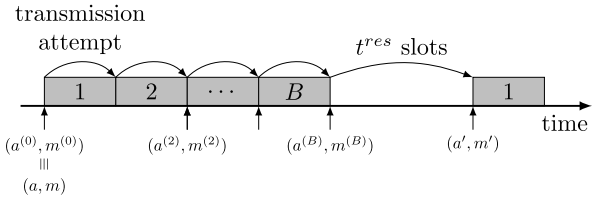


FIGURE 6. Intermediate states and transitions.

The nature of $B + 1$ intermediate transitions is as follows. Each of the first B transitions corresponds to a packet transmission attempt. Cumulatively, they result in the transition from the initial state $(a, m) \equiv (a^{(0)}, m^{(0)})$ at instant t to the last intermediate state $(a^{(B)}, m^{(B)})$. Transition $B+1$ is a time jump of duration T^{res} that makes the system finally transit from instant t to instant $t + 1$. We assume that all B transmissions occur instantly one by one: each transmission starts at the beginning of a reserved interval and has zero duration. Under such an assumption, no packets are discarded during intermediate transitions since their ages do not increase during a period of zero duration.

The assumption allows us to describe the first B intermediate transitions in the same way. Let $(a^{(i)}, m^{(i)})$ be the system state before a transmission attempt $i + 1$, $i \in \{0, \dots, B - 1\}$. We find to which intermediate states $(a^{(i+1)}, m^{(i+1)})$ the system can transit.

- If $a^{(i)} < 0$, then the queue is empty and the system remains in state $(a^{(i)}, m^{(i)})$ with probability 1.
- If $a^{(i)} \ge 0$, then the queue is not empty and a packet transmission occurs.
  - With probability q, the transmission is not successful and the system remains in state $(a^{(i)}, m^{(i)})$.
  - With probability $1 - q$, the transmission is successful. If $m^{(i)} = 1$, the system transits to state $(a^{(i)} - t^{in}, j)$ with j packets in the HoL batch. The transition probability equals $(1 - q)p_j^{in}$, $j = 1, \dots, M$. Otherwise, if $m^{(i)} > 1$, then the system transits to state $(a^{(i)}, m^{(i)} - 1)$.

The lowest possible value of $a^{(i)}$ in an intermediate state equals $-t^{in}$. It is reached if a batch appears at the beginning of the reserved interval to the empty queue and is fully transmitted by the end of this interval, so that the last intermediate state has a negative value of a equal to $-t^{in}$.

From the last intermediate state $(a^{(B)}, m^B)$ the system transits to state (a', m') at the beginning of the next reserved interval.

- If $a^{(B)} \le d - t^{res}$, then the system transits to state $(a^{(B)} + t^{res}, m^{(B)})$.
- If $a^{(B)} > d - t^{res}$, then the oldest batch of packets, which is currently of age $a^{(B)}$, becomes outdated by instant $t+1$ and gets discarded. It t^{res} is too high, then several batches following the oldest one also become outdated, thus leading to additional discarded packets. Obviously, since batches appear every t^{in} slots, the number $n_{(a^{(B)}, m^{(B)})}$ of *additionally* discarded batches is calculated as follows:

$$n_{(a^{(B)},m^{(B)})} = \begin{cases} \left\lceil \dfrac{a^{(B)} + t^{res} - d}{t^{in}} \right\rceil, & \text{if } t^{res} > t^{in}, \\ 1, & \text{if } t^{res} \le t^{in}. \end{cases}$$

Thus, the system transits to state $(a^{(B)} + t^{res} - n_{(a^{(B)},m^{(B)})}t^{in}, m^{(B)})$.

Let A be a transition matrix for each of the first B intermediate transitions and C be a transition matrix for the last intermediate transition. Thus, the transition matrix P for the original Markov chain can be found as follows:

$$P = A^B C.$$

Given P, we find stationary probability distribution $\boldsymbol{\pi}_{(a,m)} = (\dots, \pi_{(a,m)}, \dots)^T$ of the Markov chain states.

Let us analyze when the chain has a single stationary probability distribution.

Theorem 2: For a flow with batches of size M, the chain described in Section V-C has a single stationary probability distribution if $0 < q < 1$.

Proof: Every Markov Chain with a finite state space has a unique stationary distribution unless the chain has two or more closed communicating classes. A closed class is one that is impossible to leave.

Let us show that the considered Markov chain does not have more than one communicating class. For that, we show that from any state, the chain can transit with a non-zero

probability to state (d, M). Thus, if a state belongs to some communicating class, (d, M) belongs to the same communicating class.

Let the chain be in an arbitrary state (a, m). In the case of a rather long sequence of unsuccessful transmission attempts, the chain consequently transits to states $(a + kt^{res}, m)$, $k = 1, 2, \ldots, k^*$, until $a + k^* t^{res} \le d$.

After every next unsuccessful transmission attempt, the chain transits from the state (a^*, m) to

- $(a^* + t^{res}, M)$ if $a^* + t^{res} \le d$, or
- $(a^* + t^{res} - (n_{(a^*,M)} t^{in}, M)$, otherwise.

Let $a^* > d - t^{res}$ and $x = d - a^* < t^{res}$. Than, $a \ge a^* + t^{res} - n_{(a^*,M)} t^{in} = d - x + t^{res} - \left\lceil \frac{t^{res} - x}{t^{in}} \right\rceil t^{in} > d - t^{in}$. Thus, the set of the ages can be described as $\{a^* + l_1 t^{res} - l_2 t^{in}\}$ such that $l_1, l_2 = 0, 1, 2, \ldots$ and $\forall l_1, l_2 \Rightarrow d - t^{in} + 1 \le a^* + l_1 t^{res} - l_2 t^{in} \le d$. Such a set of the ages forms a complete residue system modulo t^{in} because $\gcd(t^{in}, t^{res}) = 1$. Thus, the chain can transit from any arbitrary state (a, m) to (d, M), and the chain cannot have more than one closed communicating class. So, the chain described in Section V-C has a single stationary probability distribution. □

Theorem 3: For a bursty VBR flow defined in Section III, the chain described in Section V-C has a single stationary probability distribution if $0 < q < 1$.

Proof: Since the batch sizes are independent random variables, let us consider a subchain of the chain that corresponds to the batch sizes of size M arriving in the queue, i.e., we exclude some transitions. By applying Theorem 2, we obtain that the chain cannot have more than two closed communicating classes. Thus, it has a single stationary probability distribution. □

To find PLR, we notice that packets can be discarded only during intermediate transition $B+1$. So, we find the stationary probability distribution $\tilde{\boldsymbol{\pi}}_{(a,m)}$ of the system states before this transition:

$$\tilde{\boldsymbol{\pi}}^T_{(a,m)} = \boldsymbol{\pi}^T_{(a,m)} A^B.$$

After that, PLR can be found as follows:

$$PLR = \frac{T^{in}}{T^{res}} \frac{\sum\limits_{(a,m):\, h > d - t^{res}} \tilde{\pi}_{(a,m)} (m + n_{(a,m)} \sum\limits_{j=1}^{M} j \cdot p_j^{in})}{\sum\limits_{j=1}^{M} j \cdot p_j^{in}}.$$

D. BURSTY VBR FLOW WITH UNSOLICITED RETRIES

Let us design a new model for a bursty VBR flow with UR (the model of a CBR flow with UR is straightforward and is not considered in the paper). In the case of UR, each reserved interval is fully occupied by B successive transmissions of only one packet. After these B transmissions, the packet leaves the queue. Since only one packet can be transmitted per t^{res} slots, we assume that $t^{res} \le t^{in}$. Otherwise, packet drops would occur even in case of a CBR flow being transmitted over the perfect channel. However, it is impractical.

To describe UR-based transmission, we consider the transmission process at the beginning of the reserved time intervals and describe the process state with two integer numbers (a, m) having the same meaning as in Section V-C. Let us find to which states (a', m') and with which probabilities the process can transit from state (a, m).

- If $a < 0$, then the queue is empty and the process remains in state (a, m) with probability 1.
- If $0 \le a \le d - t^{res}$, then the queue is not empty and a packet transmission occurs.
  - If no packets are delivered within B transmissions, the process transits to state $(a + t^{res} - t^{in}, j)$ with j packets in the HoL batch. The transition probability equals $q^B p_j^{in}$.
  - If at least one packet is successfully delivered, then the following situations are possible:
    - If $m = 1$, then all packets of the HoL batch have been delivered and the process transits to state $(a + t^{res} - t^{in}, j), j = 1, \ldots, M$, with the overall probability $(1 - q^B) p_j^{in}$.
    - If $m > 1$, then the process transits to state $(a + t^{res}, m - 1)$ with probability $1 - q^B$.
- If $a > d - t^{res}$, then the HoL batch anyway leaves the queue. In this case the system transits to state $(a + t^{res} - t^{in}, j), j = 1, \ldots, M$, with probability p_j^{in}.

Given the transition matrix $\boldsymbol{P}$, we find the stationary probability distribution $\boldsymbol{\pi}_{(a,m)}$.

Note that the chain introduced in this section can be considered as a special case of the chain described in Section V-C, where q is replaced by q^B. Thus, Theorem 3 is valid for the chain introduced in this Section, and there is a single stationary probability distribution $\boldsymbol{\pi}_{(a,m)}$.

Finally, PLR can be found as follows:

$$PLR = \frac{T^{in}}{T^{res}} \frac{\sum\limits_{(a,m)} \pi_{(a,m)} q^B + \sum\limits_{\substack{(a,m):\\ a > d - t^{res}}} \pi_{(a,m)} (m - 1)}{\sum\limits_{j=1}^{M} j \cdot p_j^{in}}$$

E. BURSTY VBR FLOW WITH BLOCK TRANSMISSION

1) BASIC IDEA

Now we consider *block transmission* of a bursty VBR flow so that up to B oldest packets are transmitted in a reserved time interval. If some packet transmission attempt fails, the packet is retransmitted during the next reserved interval if its lifetime does not exceed D^{QoS}. For simplicity, we assume that the transmission failures are independent within a block.

Similar to the previous models, the slotted time scale is applied, and the transmission process is modeled as a discrete-time Markov chain with time unit T^{res}, so that instants t and $t + 1$ of the Markov chain time correspond to the beginnings of two consecutive reserved intervals.

Being familiar with models developed in previous Sections V-B–V-C, one can propose to describe the process at moments $t, t + 1, \ldots$ by a vector of numbers. In comparison

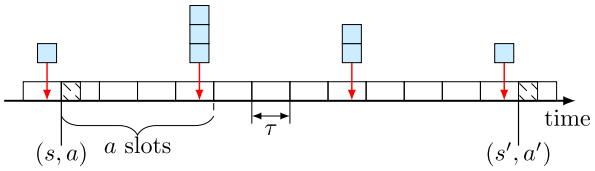


FIGURE 7. System states.

with the model developed in Section V-C, we have to keep track of B oldest packets since all of them are transmitted during a reserved time interval. Thus, the state vector must implicitly or explicitly contain information about the ages of these B packets. For example, the system can be described by vector $(a_1, a_2, \ldots, a_B, m_B)$, where $a_1, a_2, \ldots, a_B$ are the ages of the first B packets in descending order ($a_1 \geq a_2 \geq \cdots \geq a_B$) and m_B is the number of packets of age a_B, which are not among the oldest B packets. Though this approach can be used to find PLR, it leads to a huge number of states and high implementation complexity. That is why we develop another approach that significantly reduces the complexity of the analytical model.

For that, we make the following *key assumption*. We assume that *(a) a packet is discarded on appearance with probability equal to the probability of this packet not to be delivered to the receiver for time D^{QoS} if it were put into the queue, (b) all packets which are not discarded on appearance are eventually delivered to the receiver no matter how much time such delivery takes*.

Let us show how we describe the system state at some instant t. Because of the made assumption, there is no more need to control packets lifetimes, and the queue can be characterized only by its size s. However, to calculate the dropping probabilities mentioned in the assumption, we need to know when batches appear in the queue (relatively to the reserved intervals). For that purpose, we *re-define* parameter a to be the time from instant t to the nearest future batch appearance, expressed in slots and *rounded up* (see Fig. 7).[2] Thus, instead of describing the state of the Markov chain at instant t by vector $(a_1, a_2, \ldots, a_B, m_B)$, we characterize it only by a pair of integer numbers (s, a).

The absolute time before the next batch appearance equals $a \cdot \tau - \xi$ seconds. The values of a are in the range of 1 to t^{in}. If the nearest batch appears later than instant $t+1$, then $a > t^{res}$. Otherwise, $a \leq t^{res}$.

According to the key assumption, the queue size s cannot exceed some value s^{max}. Indeed, if the queue is long enough, then any appeared packet is discarded because it cannot be transmitted even once during D^{QoS} seconds after its appearance. Since during D^{QoS} seconds no more than $\lfloor D^{QoS}/T^{res} \rfloor + 1$ reserved intervals occur, the maximum number of packets that can be transmitted equals $(\lfloor D^{QoS}/T^{res} \rfloor + 1)B$. That is why a packet is surely discarded if it appears into the queue of size $(\lfloor D^{QoS}/T^{res} \rfloor + 1)B$. Hence s ranges from 0 to $s^{max} = \left(\lfloor D^{QoS}/T^{res} \rfloor + 1\right) B$.

[2]We use the same notation as we used previously to denote the age of a packet.

2) PACKET DISCARDING PROBABILITY $P_{dis}(s, a)$

Let a packet appear a slots after the beginning of the closest preceding reserved interval ($a \leq t^{res}$). We denote by $P_{dis}(s, a)$ the probability of the packet being discarded on appearance. Assuming that the packet is not discarded, we denote by $P_{wait}(s, k)$ the probability that after the appearance the packet waits for the first transmission attempt during k reserved time intervals, so, it is transmitted for the first time only in interval $(k+1)$. Let $r(a)$ be the number of reserved intervals during D^{QoS} seconds after the packet appearance. Then the probability of the packet to be discarded equals the sum of probabilities of the following disjoint events.

1) The packet is never transmitted during D^{QoS} seconds since its appearance. The probability of this event equals

$$\sum_{k=r(a)}^{+\infty} P_{wait}(s, k) = 1 - \sum_{k=0}^{r(a)-1} P_{wait}(s, k).$$

2) The packet is transmitted several times during D^{QoS} seconds since its appearance, but unsuccessfully. The probability of this event equals

$$\sum_{k=0}^{r(a)-1} P_{wait}(s, k) q^{r(a)-k},$$

where $q^{r(a)-k}$ is the probability that the first $r(a) - k$ transmissions are unsuccessful.

Thus, $P_{dis}(s, a)$ is calculated as follows:

$$\begin{aligned} P_{dis}(s, a) = &\sum_{k=0}^{r(a)-1} P_{wait}(s, k) q^{r(a)-k} + 1 \\ &- \sum_{k=0}^{r(a)-1} P_{wait}(s, k) = 1 \\ &- \sum_{k=0}^{r(a)-1} (1 - q^{r(a)-k}) P_{wait}(s, k). \end{aligned} \quad (6)$$

To calculate $P_{dis}(s, a)$, we need to know both $P_{wait}(s, k)$ and $r(a)$.

Let us find $P_{wait}(s, k)$. If $s < B$ then the packet is immediately transmitted in the nearest reserved interval, i.e., $P_{wait}(s, 0) = 1$ and $P_{wait}(s, k) = 0$ for $k \neq 0$. If $s \geq B$ then $P_{wait}(s, k)$ is the joint probability of the following events.

1) During $k-1$ reserved time intervals, the position of the considered packet in the queue changes from $s+1$ to $j+1 \in \{B, \ldots, \min\{s, 2B-1\}\}$, i.e., $s-j$ packets are successfully transmitted in the first $k-1$ reserved intervals. The probability of this event is

$$P_{tx}(s-j, (k-1)B),$$

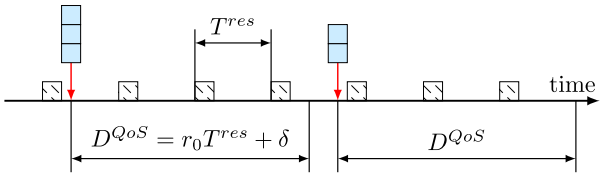


FIGURE 8. To calculation of $r(a)$.

where $P_{tx}(b, n)$ is the probability that b of n packets are successfully transmitted:

$$P_{tx}(b, n) = C_n^b (1-q)^b q^{n-b}.$$

2) During reserved interval k, the considered packet moves from position $j+1$ to one of the first B positions in the queue. The probability of this event is

$$\sum_{i=j+1-B}^{B} P_{tx}(i, B),$$

where i is the number of the transmitted packets.

Thus, $P_{wait}(s, k)$ is calculated as follows

$$P_{wait}(s, k) = \sum_{j=B}^{\min\{s, 2B-1\}} P_{tx}(s-j, (k-1)B) \cdot \sum_{i=j+1-B}^{B} P_{tx}(i, B). \tag{7}$$

3) CALCULATION OF $r(a)$

A packet that appears a slots after the beginning of a reserved interval can witness up to $r(a)$ reserved intervals while standing in the queue. To find $r(a)$, we represent D^{QoS} as $r_0 \cdot T^{res} + \delta$, where $r_0 = \lfloor D^{QoS}/T^{res} \rfloor$ and $\delta = D^{QoS} \bmod T^{res}$. If a packet appears no earlier than δ seconds before the next interval, that is, $a \cdot \tau - \xi \geq T^{res} - \delta$, then for this packet $r(a) = r_0 + 1$. Otherwise, $r(a) = r_0$. Thus,

$$r(a) = \begin{cases} r_0, & \text{if } a \in \{1, \ldots, \hbar - 1\}, \\ r_0 + 1, & \text{if } a \in \{\hbar, \ldots, t^{res}\}, \end{cases} \tag{8}$$

where $\hbar = \left\lceil \frac{T^{res} - \delta + \xi}{\tau} \right\rceil$.

Substituting (7) and (8) into (6), we can calculate probability $P_{dis}(s, k)$. It is worth to mention that if $s > s^{max}$, then $P_{dis}(s, a) = 1$ since in this case $P_{wait}(s, k) = 0$ for $k \leq r(a)$.

4) DISTRIBUTION OF THE NUMBER OF ENQUEUED PACKETS

Since we drop some packets on their appearances, the distribution of the size of an enqueued batch differs from that of the initially appeared batch. Given a batch of size n that appears a ($a \leq t^{res}$) slots after the previous reserved interval into the queue of size s, let $P_{arr}(i|n, s, a)$ be the probability that i packets from this batch are enqueued. The event "i packets are enqueued" is a union of the following disjoint events.

1) The first packet of the batch is discarded (with probability $P_{dis}(s, a)$) and i packets from the remaining $n-1$ are enqueued (with probability $P_{arr}(i|n-1, s, a)$).

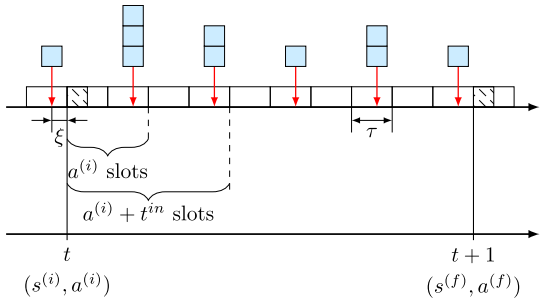


FIGURE 9. Batches appearance between two consecutive intervals.

2) The first packet of the batch is enqueued (with probability $1 - P_{dis}(s, a)$) and $i-1$ packets from the remaining $n-1$ packets are also enqueued (with probability $P_{arr}(i-1|n-1, s+1, a)$).

Thus, $P_{arr}(i|n, s, a)$ can be calculated recurrently

$$P_{arr}(i|n, s, a) = P_{dis}(s, a) P_{arr}(i|n-1, s, a) + (1 - P_{dis}(s, a)) P_{arr}(i-1|n-1, s+1, a)$$

with the following boundary conditions

$$\begin{cases} P_{arr}(i|n, s, a) = 0, \text{ if } i > n \text{ or } i < 0; \\ P_{arr}(0|0, s, a) = 1. \end{cases}$$

Let $P_{batch}(i|s, a)$ be the probability that i packets are enqueued from an appeared batch of unknown size. Given $P_{arr}(i|n, s, a)$, this probability is calculated as follows

$$P_{batch}(i|s, a) = \sum_{n=i}^{M} P_{arr}(i|n, s, a) p_n^{in}.$$

5) MARKOV CHAIN TRANSITIONS

Let us find to what state $(s^{(f)}, a^{(f)})$ and with what probability the system can transit from state $(s^{(i)}, a^{(i)})$ in one transition (see Fig. 9).[3] The transition consists of two steps.

1) b packets are successfully transmitted during the current reserved interval with probability

$$P_{tx}(b, \min\{s, B\}), \; b \in \{0, \ldots, \min\{s, B\}\}.$$

2) New packets appear before the next reserved interval and some of them are enqueued.

Let the process be in state $(s^{(i)}, a^{(i)})$. If $a^{(i)} \leq t^{res}$, then $N = \lfloor (t^{res} - a^{(i)})/t^{in} \rfloor + 1$ batches appear until the next reserved interval (Fig. 9) and the process transits to a state with $a^{(f)} = a^{(i)} + Nt^{in} - t^{res}$. Otherwise, no batches appear by the next reserved interval ($N = 0$) and the process transits to a state with $a^{(f)} = a^{(i)} - t^{res}$.

Let us consider the batch appearance between two consecutive reserved intervals. Let some batch appear $\tilde{a}$ slots after the beginning of the first reserved interval and let the queue have size $\tilde{s}$ at the appearance moment. $\tilde{a} \leq t^{res}$ since the batch appears not later than the beginning of the second interval.

[3] Upper indices (i) and (f) in $(s^{(i)}, a^{(i)})$ and $(s^{(f)}, a^{(f)})$ correspond to initial and final states, respectively.

We denote by $P_{rx}(m|\tilde{s},\tilde{a})$ the probability that m packets are enqueued from all the batches arriving from $\tilde{a}$ up to t^{res} slots after the first reserved interval inclusively. Enqueuing of m packets can be represented as a sequence of two events:

- i packets enter the queue from the batch arriving $\tilde{a}$ slots after the beginning of the first reserved interval into the queue of size $\tilde{s}$ with probability $P_{batch}(i|\tilde{s},\tilde{a})$.
- The remaining $m-i$ packets are enqueued from other batches arriving $\tilde{a}+t^{in}$ slots after the beginning of first reserved time interval until the next reserved interval. Probability of this event equals $P_{rx}(m-i|\tilde{s}+i,\tilde{a}+t^{in})$.

The probability $P_{rx}(m|\tilde{s},\tilde{a})$ can be found recurrently

$$P_{rx}(m|\tilde{s},\tilde{a})=\sum_{i=0}^{M}P_{batch}(i|\tilde{s},\tilde{a})P_{rx}(m-i|\tilde{s}+i,\tilde{a}+t^{in})$$

with the following boundary conditions:

$$P_{rx}(m|\tilde{s},\tilde{a})=\begin{cases}\delta_{m0}, & \text{if } \tilde{a}>t^{res};\\ 0, & \text{if } m>s^{max}-\tilde{s} \text{ or } m<0,\end{cases} \tag{9}$$

where $\delta_{ij}=1$, if $i=j$ and $\delta_{ij}=0$ otherwise. It can be seen that (9) handles both cases emphasized above: $a^{(i)}\le t^{res}$ and $a^{(i)}>t^{res}$. Thus, the transition probability from state $(s^{(i)},a^{(i)})$ to state $(s^{(f)}=s^{(i)}+n, a^{(f)}=a^{(i)}-Nt^{in}+t^{res})$ is calculated as follows

$$P_{tr}^{(i,f)}=\sum_{b=0}^{\min\{s^{(i)},B\}}P_{tx}(b,\min\{s^{(i)},B\})P_{rx}(b+n|s^{(i)}-b,a^{(i)}).$$

Given a matrix $\boldsymbol{P}$ of the transition probabilities, we can find stationary probabilities $\boldsymbol{\pi}_{s,a}$ of the chain states.

Theorem 4: For a bursty VBR flow defined in Section III, the chain described in Section V-E has a single stationary probability distribution if $0<q<1$.

Proof: The proof is similar to the proof of Theorem 2. But here, we show that from any state, it is possible to reach state (s^{max},t^{in}).

Let the chain be in an arbitrary state (s,a), such that $s<s^{max}$. Consider that all the transmission attempts are unsuccessful during a rather long time interval. It happens with a non-zero probability. While $s<s^{max}$ with a non-zero probability at least one packet is enqueued. Thus, in several steps, the chain transits to the state (s^{max},a^*). If transmission attempts remain unsuccessful, the value of $s=s^{max}$, while the value of a^* changes. Similar to previous theorems, since $\gcd(t^{in},t^{res})=1$ the values of a^* form a complete residue system modulo t^{in}. Thus, from an arbitrary state (s,a) the chain can transit to the state (s^{max},t^{in}) with a non-zero probability. So, it has a single stationary probability distribution. □

6) PLR CALCULATION

If the process is in state (s,a), then the probability of i packets being successfully transmitted during the reserved time interval equals $P_{tx}(i,\min\{s,B\})$. Let p_i^{out} be the probability that i packets are successfully transmitted during a reserved interval. This probability can be found as follows:

$$p_i^{out}=\sum_s P_{tx}(i,\min\{s,B\})\pi_s,$$

where $\pi_s=\sum_a\pi_{s,a}$ is the probability that the queue size is s at the beginning of the reserved interval.

I^{in} and I^{out} are calculated as follows:

$$I^{in}=\frac{1}{T^{in}}\sum_{i=1}^{M}i\cdot p_i^{in},\quad I^{out}=\frac{1}{T^{res}}\sum_{j=1}^{B}j\cdot p_j^{out}.$$

Finally, the difference of I^{in} and I^{out} divided by I^{in} gives PLR:

$$PLR=\frac{I^{dis}}{I^{in}}=1-\frac{I^{out}}{I^{in}}=1-\frac{T^{in}}{T^{res}}\frac{\sum_{j=1}^{B}j\cdot p_j^{out}}{\sum_{i=1}^{M}i\cdot p_i^{in}}.$$

VI. NUMERICAL RESULTS

A. MAIN PECULIARITIES

Let us demonstrate how to apply the mathematical framework developed in Section V to choose optimal transmission parameters, i.e., those transmission method and reservation parameters that satisfy QoS requirements for a given flow with the minimal channel time consumption.

We start with the ordered transmission of a CBR flow with $T^{in}=20$ ms that corresponds to the usage of voice codecs like G.711 [19] and G.729 [20]. Let $\xi=0$, i.e., packets appear exactly at the beginnings of some slots. Fig. 10(a) shows how PLR depends on the reservation period T^{res} for $q=0.3$ and different D^{QoS} values. These results have been obtained both analytically and by simulation (see details in [4]). In the simulation, the packets of a batch appear not simultaneously, but their appearances are scattered around the expected batch appearance moment according to the normal distribution $\mathcal{N}(\mu,\sigma)$: $\mu=0$, $\sigma=0.2T^{in}$. Such scattering models the behavior of real systems where packets experience different delays while traversing the path from a source to a sink.

In most points, the simulation and analytical results correspond to each other, so that the relative error does not exceed 5%. However, at some points, the analytical values of PLR experience significant drops. These drops are caused by the discrete nature of the system assumed by the analytical model. Specifically, at the points of drops, on average, a packet has more transmission opportunities than for neighbor points. For example, for $D^{QoS}=30$ ms and $T^{res}=9$ ms, $\tau=1$ ms and the packets can appear $k=0,1,2,\ldots,9$ slots before the nearest reserved interval. If $k\ge 3$, then the packet witnesses four reserved time intervals during its lifetime. Otherwise, the packet witnesses only three reserved intervals. In contrast, if $T^{res}=10$ ms, then $\tau=T^{res}$. Thanks to the periodic arrivals, if at least one packet appears right before some reserved interval, all other packets appear exactly before the corresponding reserved intervals. So *all*

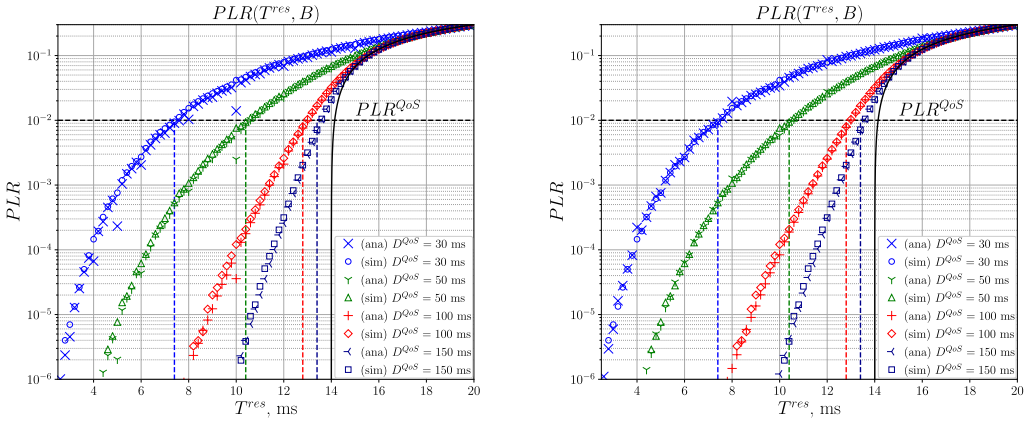


FIGURE 10. PLR as a function of T^{res} for $T^{in} = 20$ ms, $q = 0.3$, and different ξ: (a) $\xi \in [0, \gamma]$, (b) $\xi \in (\gamma, \tau)$.

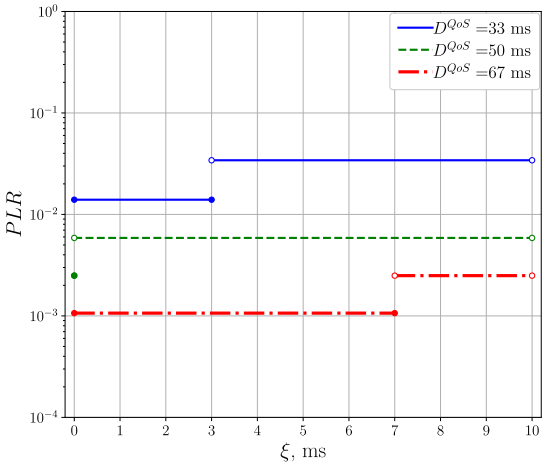


FIGURE 11. PLR as function from ξ for $T^{in} = 20$ ms, $T^{res} = 10$ ms, $q = 0.3$.

packets witness four reserved intervals. Additional reserved interval reduces the value of PLR. Generally, the magnitude of the PLR drops grows with τ and reaches local maximums when T^{in} is a multiple of T^{res}, that is, $\tau = T^{res}$ and $t^{res} = 1$. PLR drops also grow with decreasing D^{QoS} since the significance of an additional transmission attempts increases.

The presence of a drop depends also on ξ. At the beginning of this section, we set $\xi = 0$, implicitly assuming that its influence is not notable. However, it turns out not to be so. ξ is a shift from a packet appearance moment to the beginning of the following slot, which affects the maximum number of slots d a packet can spend in the queue: $d = \lfloor (D^{QoS} - \xi)/\tau \rfloor$. Varying ξ in range $[0, \tau)$ results in a single PLR value leap which occurs at point $\gamma = (D^{QoS} \bmod \tau) \in [0, \tau)$. PLR^{high}, the PLR value for $\xi \in (\gamma, \tau)$, is higher than that for $\xi \in [0, \gamma]$, i.e., PLR^{low}. The dependence of PLR from ξ is shown in Fig. 11.

In reality, packets do not appear strictly periodically and the values of ξ for various packets differ. It may happen that for some packets, $\xi \geq \gamma$, while the other packets have $\xi < \gamma$. Thus, PLR in a real system is somewhere in between PLR^{high} and PLR^{low}.

The described PLR leap affects the choice of reservation period T^{res}. Generally, given PLR^{QoS}, one can use the found dependency $PLR(T^{res})$ to find the maximum reservation period T^{res*} for which $PLR(T^{res*}) \leq PLR^{QoS}$, i.e., the QoS requirements are satisfied. For example, to provide PLR lower than 2% under $D^{QoS} = 30$ ms, one can naively choose $T^{res*} = 10$ ms by looking at Fig. 10(a). However, the simulation-based PLR value at this point is almost 6 times higher than PLR predicted by the analytical model and certainly exceeds the desired 2%. To cope with this problem, we should consider the worst-case situation $\xi \in (\gamma, \tau)$ during the choice of the reservation period. The $PLR(T^{res})$ functions for the worst-case situations are shown in Fig. 10(b). Now instead of drops, we observe PLR rises. However, the overestimation can be considered as a good effect since it reduces the risk of choosing T^{res}, which violates the QoS requirements.

Another interesting fact is that all the dependencies $PLR(T^{res})$ corresponding to different values of D^{QoS} converge to q as T^{res} approaches T^{in}. If $T^{res} = T^{in}$, then there is a one-to-one mapping of arriving packets to the reserved time intervals. In this case, the process eventually comes to the state when each packet has only one transmission attempt to be delivered, and if it is unsuccessful, the packet gets discarded. It results in PLR equal to q.

The solid curves in Fig. 10 show the lower bound of PLR. They correspond to the case when there is no delay bound on packets delivery (formally $D^{QoS} = \infty$). To find this lower bound, we notice that if there is no delay bound, packet losses do not occur unless $I^{in} > I^{out}_{max}$, where $I^{in} = \frac{1}{T^{in}}$ is the CBR input flow intensity and $I^{out}_{max} = \frac{1-q}{T^{res}}$ is the maximum possible output flow intensity. Finally, for PLR_∞ we obtain

$$PLR_\infty = \max\left\{0, \frac{I^{in} - I^{out}_{max}}{I^{in}}\right\} = \begin{cases} 0, & T^{res} \leq (1-q)T^{in}, \\ 1 - \dfrac{(1-q)T^{in}}{T^{res}}, & T^{res} > (1-q)T^{in}. \end{cases}$$

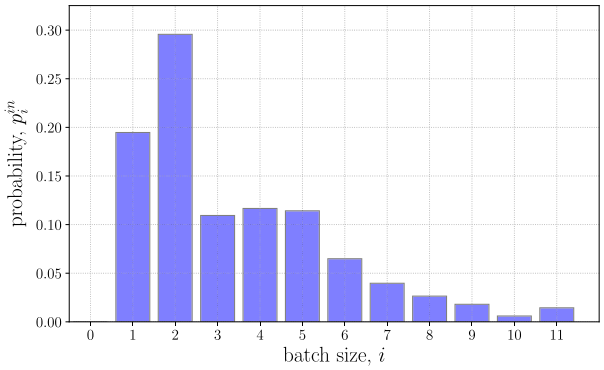


FIGURE 12. Batch size probability distribution $\{p_i^{in}\}_{i=1}^{M}$.

The lower bound shows the amount of channel resources we would need if there were not any requirement on delivery delay.

Such lower bounds can be also found for the ordered and block transmissions of a bursty VBR flow. The only difference is the expressions for I^{in} and I_{max}^{out}:

$$I^{in} = \frac{\sum_j j p_j^{in}}{T^{in}}, \quad I_{max}^{out} = \frac{(1-q)B}{T^{res}}.$$

B. COMPARISON OF TRANSMISSION METHODS

Let us compare the efficiency of the transmission methods (see Section II) considered in the paper. As a criterion for the comparison, we use (1). Let $D_{ordered}^{res}(B)$, $D_{block}^{res}(B)$ and $D_{ur}^{res}(B)$ be the durations enough for B transmission attempts with ordered transmission, block transmission and unsolicited retries. For ordered transmission

$$D_{ordered}^{res}(B) = T_{PIFS} + B(T_{DATA} + T_{SIFS} + T_{ACK} + T_{SIFS}) - T_{SIFS}.$$

In the particular case of $B = 1$ ordered transmission turns out into individual transmission. For block transmission

$$D_{block}^{res}(B) = T_{PIFS} + B(T_{DATA} + T_{SIFS}) + T_{BAR} + T_{SIFS} + T_{B-ACK}.$$

For unsolicited retries

$$D_{ur}^{res}(B) = T_{PIFS} + B(T_{DATA} + T_{SIFS}).$$

Given an input flow, we define $C^*(PLR^{QoS})$ to be the minimal channel consumption required to satisfy the QoS requirements:

$$C^*(PLR^{QoS}) = \min_{\substack{T^{res},B: \\ PLR(T^{res},B) \le PLR^{QoS}}} C(T^{res}, B). \quad (10)$$

Next, we find the dependency $C^*(PLR^{QoS})$ for a VBR flow with $T^{in} = 40$ ms, $q = 0.3$, and the batch size distribution shown in Fig. 12. This flow has been obtained experimentally by streaming a video fragment with RTP and gathering traces with `tcpdump`.

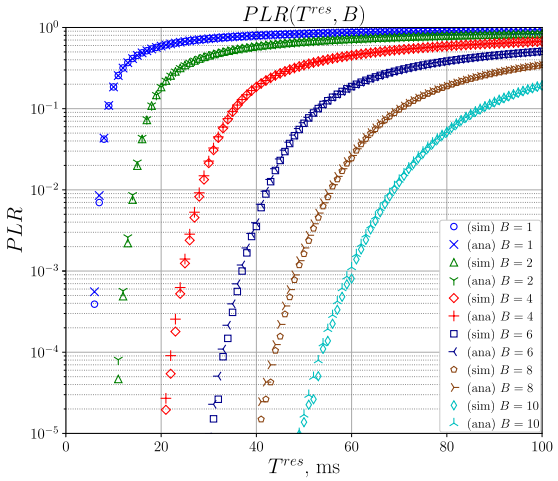


FIGURE 13. PLR for the VBR flow with $D^{QoS} = 200$ ms for ordered transmission with different values of the reservation period and duration.

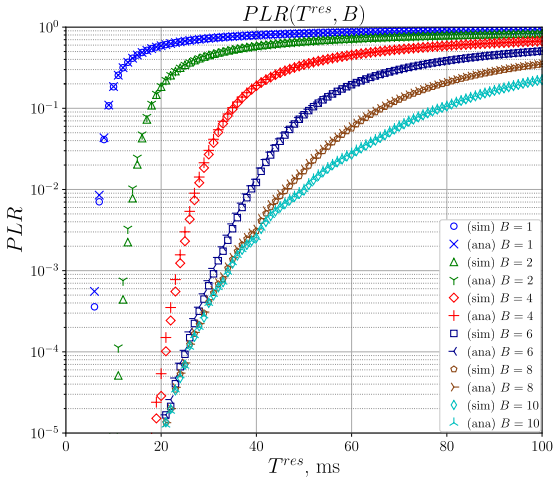


FIGURE 14. PLR for the VBR flow with $D^{QoS} = 200$ ms for block transmission with different values of the reservation period and duration.

The calculation of $C^*(PLR^{QoS})$ is based on functions $PLR(T^{res}, B)$. For ordered and block transmissions, these functions are shown in Fig. 13 and Fig. 14 respectively.

To be able to calculate D^{res} correctly, we need to take into account the appropriate parameters of the PHY layer. We suppose that the transmission occurs in a 160 MHz channel with 11ax PHY. The control frames are transmitted at the rate of 30 Mbps, and data packets of 1.5K are transmitted at the rate of 576 Mbps. As we consider only one stream, we do not take into account multi-user features of 11ax, e.g., OFDMA.

Fig. 15 shows the minimal channel time consumption $C^*(PLR^{QoS})$ that can be achieved with different transmission methods for different QoS requirements. The results show that in the whole range of typical PLR requirements ($PLR^{QoS} = 10^{-4} — 10^{-2}$), the unsolicited retries is the worst transmission method, while block transmission is the most efficient one. It is not surprising since UR transmits each packet exactly B times even if it is delivered with the first

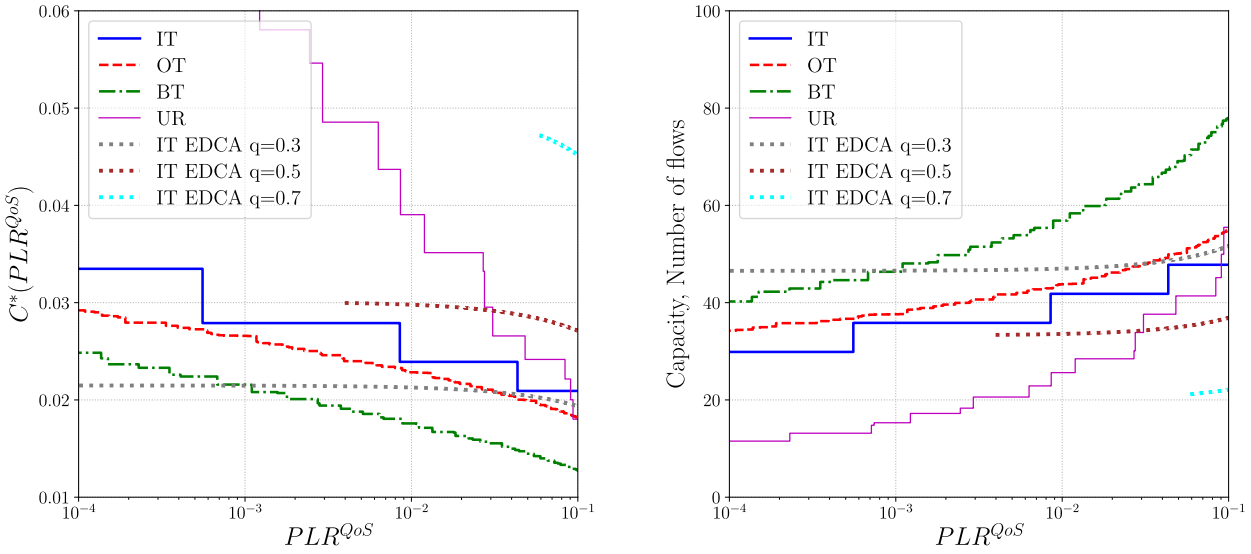


FIGURE 15. Channel time consumption C^* and the network capacity with different transmission methods used for streaming the VBR flow.

attempt, thus overloading the channel. In the case of block transmission, multiple ACK frames are aggregated into a single BlockAck frame reducing the ACK-induced overhead as compared with the ordered transmission. Lower channel time consumption means that the network can serve more flows, i.e., its capacity is higher, as shown in Fig. 15.

Fig. 15 also presents the channel time consumption and the network capacity that can be achieved without reservations, using the legacy EDCA and individual transmissions. The corresponding values depend on the PER on the channel outside the reservation. If it equals 0.3, i.e., the same as within reserved time intervals, reservations do not improve reliability and are inefficient because of high overhead induced by extra reserved time intervals. In other words, the reservations shall not be used if there is no interference and hidden stations that result in high collision probability if no transmission protection method is used.

However, the more realistic situation is that the reservations improve reliability [3], [4], [6], [7], [10]. So, the error rate outside reservations is much higher than inside the reserved time intervals. Even if outside reservations PER $= 0.5$, and inside reservations it equals 0.3, the reservation mechanisms reduce channel time consumption and increase network capacity.

High PER outside the reserved time intervals may prevent satisfying QoS requirements regardless of the channel time consumption. As in Wi-Fi networks, the number of transmission attempts is limited by default by eight, at $q = 0.5$, it is impossible to achieve PLR smaller than 0.4%. This effect is shown in Fig. 15, where some curves "start" only at rather high values of PLR^{QoS}.

VII. CONCLUSION

This paper presents mathematical modeling and analysis of QoS-sensitive data streaming in modern Wi-Fi networks. We analyze the existing channel reservation mechanisms in the latest Wi-Fi standard and show that all of them use periodic reservations to mitigate interference in various deployments. We discuss a possible way to organize a centralized control of dense Wi-Fi deployments. We also introduce a general mathematical framework to model reservation-based streaming, use this framework in various scenarios, and develop corresponding analytical models. Through numerical results, we show how the models can be used to find the minimum amount of reserved channel resources needed to satisfy QoS requirements. The results of this paper can be applied to various Wi-Fi standard amendments, both existing and under development.

REFERENCES

[1] *IEEE Standard for Information Technology—Telecommunications and Information Exchange Between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard IEEE 802.11, 2016.

[2] *AP Coordination in EHT*. Accessed: Mar. 1, 2020. [Online]. Available: https://mentor.ieee.org/802.11/dcn/19/11-19-0801-00-00be-ap-coordination-in-eht.pptx

[3] A. Krasilov, A. Lyakhov, and A. Safonov, "Interference, even with MCCA channel access method in IEEE 802.11s mesh networks," in *Proc. IEEE 8th Int. Conf. Mobile Ad-Hoc Sensor Syst.*, Oct. 2011, pp. 752–757.

[4] E. Khorov, A. Lyakhov, and A. Safonov, "Flexibility of routing framework architecture in IEEE 802.11s mesh networks," in *Proc. IEEE 8th Int. Conf. Mobile Ad-Hoc Sensor Syst.*, Oct. 2011, pp. 777–782.

[5] P. Zhou, K. Cheng, X. Han, X. Fang, Y. Fang, R. He, Y. Long, and Y. Liu, "IEEE 802.11ay-based mmWave WLANs: Design challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1654–1681, 3rd Quart., 2018.

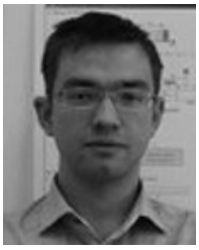
[6] G. Hiertz, T. Junge, S. Max, Y. Zang, L. Stibor, and D. Denteneer, "Mesh deterministic access (MDA)-optional IEEE 802.11s MAC scheme-simulation results (IEEE 802.11 TGs submission)," Melbourne, VIC, Australia, Tech. Rep. 11-06-1370-00-000s, IEEE 802 LMSC, Sep. 2006.

[7] L. R. Pinto, L. Almeida, and A. Rowe, "Video streaming in multi-hop aerial networks: Demo abstract," in *Proc. 16th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*. New York, NY, USA: ACM, 2017, pp. 283–284, doi: 10.1145/3055031.3055047.

[8] E. Shvets, A. Lyakhov, A. Safonov, and E. Khorov, "Analytical model of IEEE 802.11s MCCAbased streaming in the presence of noise," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 39, no. 2, p. 38, Sep. 2011.
[9] A. Ivanov, E. Khorov, and A. Lyakhov, "Analytical model of QoS-aware streaming in Wi-Fi networks via periodic TXOPs," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.
[10] E. Khorov, A. Krasilov, A. Krotov, and A. Lyakhov, "Will MCCA revive wireless multihop networks?" *Comput. Commun.*, vol. 104, pp. 159–174, May 2017.
[11] E. Khorov, A. Lyakhov, A. Krotov, and A. Guschin, "A survey on IEEE 802.11ah: An enabling networking technology for smart cities," *Comput. Commun.*, vol. 58, pp. 53–69, Mar. 2015.
[12] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, "A tutorial on IEEE 802.11ax high efficiency WLANs," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 197–216, 1st Quart., 2019.
[13] Cisco. *Meraki Datasheet Cloud Management*. Accessed: Mar. 1, 2020. [Online]. Available: https://meraki.cisco.com/lib/pdf/meraki_datasheet_cloud_management.pdf
[14] Huawei. *Wlan Access Controller*. Accessed: Mar. 1, 2020. [Online]. Available: https://e.huawei.com/en/products/enterprise-networking/wlan/access-controllers/ac6805
[15] Hewlett Packard Enterprise. *Quickspecs. Aruba Central Software*. Accessed: Mar. 1, 2020. [Online]. Available: https://h20195.www2.hpe.com/v2/getpdf.aspx/c05272678.pdf
[16] D. Takahashi. *Quantenna's Maui Can Automatically Detect Problems in Your Home's Wi-Fi Network*. [Online]. Available: https://venturebeat.com/2015/01/05/quantennas-maui-can-automatically-detect-problems-in-your-homes-wi-fi-network/
[17] E. Khorov, A. Ivanov, A. Lyakhov, and I. F. Akyildiz, "Cloud control to optimize real-time video transmission in dense IEEE 802.11aa/ax networks," in *Proc. IEEE 15th Int. Conf. Mobile Ad Hoc Sensor Syst. (MASS)*, Oct. 2018, pp. 193–201.
[18] *Rtp: A Transport Protocol for Real-Time Applications*, document RFC 3550, IETF, 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3550.txt
[19] *Pulse Code Modulation (PCM) of Voice Frequencies*, document TU-T Rec. G.711, 1988.
[20] *Coding of speech at 8 kbit/s using conjugate-structure algebraic code-excited linear prediction (CS-ACELP)*, document ITU-T Rec. G.729, 1996.
[21] M. Daneshi, J. Pan, and S. Ganti, "Towards an efficient reservation algorithm for distributed reservation protocols," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.
[22] W.-K. Kuo and C.-Y. Wu, "Supporting real-time VBR video transport on WiMedia-based wireless personal area networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1965–1971, May 2009.
[23] W.-H. Liao, Y.-C. Tseng, and K.-P. Shih, "A TDMA-based bandwidth reservation protocol for QoS routing in a wireless mobile ad hoc network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 5, Apr. 2002, pp. 3186–3190.
[24] D. Vergados, D. Vergados, C. Douligeris, and S. Tombros, "QoS-aware TDMA for end-to-end traffic scheduling in ad hoc networks," *IEEE Wireless Commun.*, vol. 13, no. 5, pp. 68–74, Oct. 2006.
[25] I. Tinnirello and P. Gallo, "Supporting a pseudo-TDMA access scheme in mesh wireless networks," in *Wireless Access Flexibility*. Berlin, Germany: Springer, 2013, pp. 80–92.
[26] Y. Khan, M. Derakhshani, S. Parsaeefard, and T. Le-Ngoc, "Self-organizing TDMA MAC protocol for effective capacity improvement in IEEE 802.11 WLANs," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.
[27] Y. Xu, K.-W. Chin, and S. Soh, "A novel distributed pseudo-TDMA channel access protocol for multi-transmit-receive wireless mesh networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2531–2542, Mar. 2018.
[28] M. Hadded, P. Muhlethaler, A. Laouiti, R. Zagrouba, and L. A. Saidane, "TDMA-based MAC protocols for vehicular ad hoc networks: A survey, qualitative analysis, and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2461–2492, Jun. 2015.
[29] X. Jiang and D. H. C. Du, "PTMAC: A prediction-based TDMA MAC protocol for reducing packet collisions in VANET," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9209–9223, Nov. 2016.
[30] V. Nguyen, T. Z. Oo, P. Chuan, and C. S. Hong, "An efficient time slot acquisition on the hybrid TDMA/CSMA multichannel MAC in VANETs," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 970–973, May 2016.
[31] (2007). *HARTCOMM, WirelessHART Specifications*. [Online]. Available: http://www.hartcomm2.org
[32] *ISA-100.11a-2009*. Accessed: 2009. [Online]. Available: http://www.isa.org
[33] *IEEE 802.15 WPAN Task Group 4e (TG4e)*. Accessed: Mar. 1, 2020. [Online]. Available: http://www.ieee802.org/15/pub/TG4e.html
[34] Y. Li, H. Zhang, Z. Huang, and M. Albert, "Optimal link scheduling for delay-constrained periodic traffic over unreliable wireless links," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2014, pp. 1465–1473.
[35] A. Ahmad and Z. Hanzálek, "Distributed real time TDMA scheduling algorithm for tree topology WSNs," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 5926–5933, Jul. 2017.
[36] A. Grilo, M. Macedo, and M. Nunes, "A scheduling algorithm for QoS support in IEEE802. 11 networks," *IEEE wireless Commun.*, vol. 10, no. 3, pp. 36–43, Jun. 2003.
[37] P. Ansel, Q. Ni, and T. Turletti, "An efficient scheduling scheme for IEEE 802.11 e," in *Proc. Model. Optim. Mobile, Ad Hoc Wireless Netw.*, Mar. 2004, pp. 24–26.
[38] D. Skyrianoglou, N. Passas, and A. Salkintzis, "ARROW: An efficient traffic scheduling algorithm for IEEE 802.11e HCCA," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3558–3567, Dec. 2006.
[39] C. Cicconetti, L. Lenzini, E. Mingozzi, and G. Stea, "Design and performance analysis of the real-time HCCA scheduler for IEEE 802.11e WLANs," *Comput. Netw.*, vol. 51, no. 9, pp. 2311–2325, Jun. 2007.
[40] M. Rashid, E. Hossain, and V. Bhargava, "Controlled channel access scheduling for guaranteed QoS in 802.11e-based WLANs," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1287–1297, Apr. 2008.
[41] K. Chandra, R. V. Prasad, and I. Niemegeers, "Performance analysis of IEEE 802.11ad MAC protocol," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1513–1516, Jul. 2017.
[42] Y. Cheng, H. Zhou, and D. Yang, "Performance evaluation of IEEE 802.11ah triggered restricted access window mode in industrial real-time applications," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2018, pp. 325–329.
[43] E. Khorov, A. Krasilov, A. Lyakhov, and D. Ostrovsky, "Dynamic resource allocation for MCCA-based streaming in Wi-Fi mesh networks," in *Wireless Access Flexibility*. Berlin, Germany: Springer, 2013, pp. 93–111.
[44] E. Shvets and A. Lyakhov, "Mathematical model of MCCA-based streaming process in mesh networks in the presence of noise," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 1887–1892.
[45] A. Ivanov, E. Khorov, and A. Lyakhov, "QoS support for bursty traffic in noisy channel via periodic reservations," in *Proc. IFIP Wireless Days (WD)*, Nov. 2014, pp. 1–6.

EVGENY KHOROV (Senior Member, IEEE) is currently the Head of the Wireless Networks Laboratory, Institute for Information Transmission Problems, Russian Academy of Sciences. He has led dozens of national and international projects sponsored by academia funds and industry. Being a voting member of the IEEE 802.11, he has contributed to the 802.11ax standard as well as to the real-time applications TIG with many proposals. He has authored over 100 articles. His main research interests include 5G and beyond wireless systems, next-generation Wi-Fi, the Internet of Things, protocol design, and QoS-aware cross-layer optimization. He was a recipient of the Russian Government Award in Science, several Best Papers Awards, and the Scopus Award Russia, in 2018. In 2015, 2017, and 2018 Huawei RRC awarded him as the Best Cooperation Project Leader. He gives tutorials and participates in panels at the significant IEEE events. He chairs TPC of the IEEE Globecom 2018 CA5GS Workshop and the IEEE BlackSeaCom 2019. He also serves as an Editor for *Ad Hoc Networks*.

ANDREY LYAKHOV (Member, IEEE) is currently a Full Professor, the Deputy Director, and the Head of the Network Protocols Research Laboratory, Institute for Information Transmission Problems, Russian Academy of Sciences. He has over 20 years of experience in Wi-Fi networks design and performance evaluation. He has authored three monographs, more than 100 articles cited in Scopus, and has ten patents. He led many joint research projects with top telecommunication companies and collaborative projects, such as FP7 ICT collaborative project FLexible Architecture for Virtualizable wireless future Internet Access (FLAVIA), from 2010 to 2012. His main research interests include design and analysis of wireless network protocols, wireless network performance evaluation methods, and stochastic modeling of wireless networks based on random multiple access. He was a member of technical and program committees of large IT conferences (ICC, MACOM, MobiHoc, Networking, and MASS) and the General Chair of the IEEE BlackSeaCom 2019 and WiFlex 2013. He was a recipient of many international and Russian awards.



ALEXANDER IVANOV received the M.Sc. degree from the Moscow Physics and Technology Institute, Moscow, Russia, in 2016. Since 2012, he has been a Researcher with the Institute for Information Transmission Problems, Russian Academy of Sciences. His research interests include analysis and developing of wireless networks protocols, and modeling of wireless networks.

IAN F. AKYILDIZ (Fellow, IEEE) is currently the Ken Byers Chair Professor of telecommunications with the School of Electrical and Computer Engineering, the Director of the Broadband Wireless Networking Laboratory, and the Chair of the Telecommunication Group, Georgia Institute of Technology, Atlanta, USA. He has been the Megagrant Research Leader with the Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, since 2018. His h-index is 121, and the total number of citations is above 115K as per Google scholar as of February 2020. His current research interests include IEEE 802.11 systems, 5G wireless systems, nanonetworks, terahertz band communications, and wireless sensor networks in challenging environments. He is a Fellow of ACM, in 1997. He received numerous awards from the IEEE and the ACM, and many other organizations.