https://doi.org/10.1016/j.comcom.2017.04.010

Koutsakis, P. (2017) Scheduling for telemedicine traffic transmission over WLANs. Computer Communications, 108 . pp. 17-26.

Scheduling for Telemedicine Traffic Transmission over WLANs

Polychronis Koutsakis

PII: S0140-3664(17)30490-5
DOI: 10.1016/j.comcom.2017.04.010
Reference: COMCOM 5494

To appear in: *Computer Communications*

Received date: 5 July 2016
Revised date: 13 March 2017

Scheduling for Telemedicine Traffic Transmission over WLANs

[a] **Polychronis Koutsakis**

School of Engineering and Information Technology

Murdoch University, Australia

email: p.koutsakis@murdoch.edu.au

Abstract

The major disadvantage of the Enhanced Distributed Channel Access (EDCA, the contention-based channel access function of 802.11e) is that it is unable to guarantee priority access to higher priority traffic in the presence of significant traffic loads from low priority users. This problem is enhanced by the continuously growing number of multimedia applications and the popularity of Wireless Local Area Networks (WLANs). Hence, solutions in scheduling multimedia traffic transmissions need to take into account both the Quality of Service (QoS) requirements and the Quality of Experience (QoE) associated with each application, especially those of urgent traffic, like telemedicine, which carries critical information regarding the patients' condition. In this work, we propose an easy-to-implement token-based and self policing-based scheduling scheme combined with a mechanism designed to mitigate congestion. Our approach is shown to guarantee priority access to telemedicine traffic, to satisfy its QoS requirements (delay, packet dropping) and to offer high telemedicine video QoE while preventing bursty video nodes from over-using the medium.

1. Introduction

The IEEE 802.11e [1] ensures service differentiation via the use of four access categories (ACs), namely Background (BK), Best Effort (BE), Video and Voice, in increasing priority order. The ACs are differentiated via the use of AIFS (Arbitration Interframe Space), CW (Contention Window), and TXOP (Transmission Opportunity). AIFS is the minimum time interval that a station needs to sense that the medium is idle before transmitting; CW values are smaller for higher priority traffic, hence lower priority users wait longer to retransmit after a transmission failure; TXOP is a bounded time interval indicating the maximum amount of time for which a terminal can initiate transmissions. A TXOP equal to zero means that the terminal can only transmit a single frame.

It is clear, from the definition of the four ACs, that none of them are applicable to urgent traffic; this type of traffic may need absolute priority but its sparse nature does not justify the use of a dedicated network, not shared by other traffic, for its transmission (unless the WLAN is set up, for example, in a hospital). A preeminent example of such urgent traffic is telemedicine traffic, on which we focus in this work. Telemedicine is already being employed in many areas of healthcare (intensive neonatology, critical surgery, pharmacy, patient education). In addition to ambulance vehicles, it is also of critical importance for the provision of health care services at understaffed areas like ships, trains, airplanes, as well as home monitoring [2]. Mobile healthcare (M-health) is a new paradigm that brings together the evolution of emerging wireless communications and network technologies with the concept of "connected healthcare" anytime and anywhere, for a large variety of medical purposes [3-7]. To the best of our knowledge, the problem of guaranteeing telemedicine QoS over WLANs has rarely been addressed in the relevant literature; exceptions include some works which are discussed in Section 2.

Diagnostic Losslessness (as opposed to Mathematical Losslessness, implying no loss of any digital information) is required for the transmission of telemedicine video, i.e., lossy compression may exist, as long as it does not compromise visual medical assessment in any way and hence suffices for making a diagnostic evaluation [8]. Telemedicine data, on the other hand, needs to be transmitted with Mathematical Losslessness, therefore it cannot be transmitted as BK, nor as BE traffic. Additionally, telemedicine video needs to be separated from regular video traffic, in order to be transmitted quickly and without contention with the lower priority regular videos. Although EDCA is able to prioritize traffic, it still provides only *statistical* priority access; it cannot guarantee priority access to high priority traffic, especially in the presence of significant traffic loads from low priority users. Hybrid Coordination function Controlled channel Access (HCCA), the polling-based channel access function of

802.11e, does provide guaranteed services and therefore outperforms EDCA when centralized access is possible. However, not only does it incur overhead and has complicated software architecture, but also recent work [9] shows that the performance of EDCA can be clearly better than that of HCCA for Variable Bit Rate (VBR) video streams, particularly in multi-collision domains where APs in neighboring basic service sets poll the stations in the overlapping area, resulting in collisions.

For these reasons, this paper focuses on the improvement of EDCA for the transmission of various types of telemedicine traffic, with QoS guarantees. A comparison with HCCA will, however, be presented in order to discuss the efficiency of our approach.

We specifically focus on the transmission of four different and crucial, in nature, types of telemedicine traffic. We propose the addition of a new Access Category in EDCA, namely URgent traffic (UR), which acquires its own AIFS and TXOP values, where AIFS is constant and TXOP is defined via self-policing. All types of telemedicine traffic are transmitted as UR traffic in our scheme. In our new scheduling scheme for telemedicine and regular multimedia traffic, we propose the mitigation of congestion by:

a) utilizing the H.264 or H.265 video decoding procedure, in order to selectively drop packets of lesser importance for regular video users.

b) defining a new QoS metric to be used by each video node in order to estimate the current congestion level, and discussing another metric as a possible candidate. The basic idea of our scheme is to force regular video users to have the *worst possible acceptable* QoS. This will allow urgent traffic the necessary "room" to be transmitted without having its QoS requirements violated.

Our scheme is compared with EDCA and shown to significantly improve the channel utilization, under both light and heavy traffic loads. It is also shown to provide significantly better QoE than EDCA. An important advantage of our scheme is that the proposed modifications to EDCA, although major in essence, are easy to implement. Our scheme is also compared with HCCA and shown to provide at least comparable and often better results.

2. Related Work

Although there are several studies in the literature, especially recently, on medical video streaming, few of them have focused on scheduling for WLANs.

Recent research efforts on different types of networks include [10-13], which focus on the problem of medical video streaming over 4G cellular networks, as well as [14], which proposes a framework for transmitting medical multimedia over cognitive radio networks.

The majority of the existing work on Medium Access Control (MAC) protocols for WLANs has focused on the transmission of integrated voice and data traffic. The problem of transmitting video traffic along with voice and data has received attention in the past decade [15-17]. The authors in [15, 17] proposed alternate approaches for defining the length of the TXOP. In [15], TXOP is calculated based on the number of MAC service data units (MSDUs) in the current queue of each station. In [17], the authors use a window *w* of already known real queue length measurements to tune their estimation of the TXOP. However, both of these approaches are insufficient for bursty video traffic. The reasons are that: a) the current queue length may be irrelevant to the size of the next video frame, leading to a quite false estimate, and b) history information does not provide an adequate estimation on the future behavior of the video source, especially for short video sequences.

In [18] the authors proposed a cross-layer design technique for transmitting telemedicine over WLANs. However, their scheme does not consider the transmission of telemedicine data and assumes that the only video traffic transmitted in the network is telemedicine video. Their proposal also uses a complex three-layer approach in order to guarantee bandwidth to telemedicine applications; in a scenario with multiple types of regular and telemedicine traffic this approach can be both time-consuming and lead the system to bandwidth starvation.

In [19] the authors proposed the use of two channels, the primary one for information transmission and the secondary (slim) one for the telemedicine application devices to send short alert messages to the Access Point. Then, the Access Point schedules to broadcast a beacon to all applications in the network, indicating the reservation of resources for the emergency device. The Access Point itself can gain access to the primary channel via HCCA. The underlying assumption of the proposed mechanism, however, is that telemedicine data will be extremely sparse. In the case of periodic telemedicine traffic transmitted from a multitude of users (e.g., vital signs and physiological measurements) such an approach would not be efficient as the secondary channel would not suffice.

The works in [20-21] both focus only on the case of a WLAN serving patients specifically within a medical facility, i.e., they do not consider the case of a patient using a WLAN to transmit telemedicine-related video or data outside a hospital. Also, [21] proposes the use of medical data transmission only when video is not being transmitted, or alternatively the use of extra, special subcarriers for the transmission of medical data. Our work considers non-dedicated WLANs and does not require extra subcarriers for parts of the transmitted traffic.

The work in [22] considers the utilization of Scalable Video Coding (SVC) for improved scheduling for medical applications, and only uses Advanced Video Coding (AVC) for the vital signals. The authors point out, however, that AVC (which is used in our work) provides better picture quality. The work in [12], which focuses on cellular networks, also uses SVC.

The idea of a token-based scheduling scheme, which practically eliminates collisions and hence increases channel utilization, was first proposed in [23], for the transmission of voice and data traffic in a fully-connected WLAN where all the nodes can hear each other. In [24] we proposed an extension to the work in [23] to include video traffic. More specifically, we proposed a change in EDCA, via the use of tokens for all types of nodes and self-policing for video nodes, which was shown to improve channel utilization. However, as we explained in [24], one feature missing from that scheme was the ability to handle applications generating urgent data; the ideas presented in that work are not sufficient to provide any guarantees for telemedicine QoS. Also, [24] did not study the problem of QoE satisfaction for any type of traffic.

In this work we extend the work presented in [23, 24], in order to integrate various types of telemedicine traffic and bursty video traffic with voice and data traffic over WLANs, while achieving high QoS and QoE for all telemedicine users in our system.

3. The Proposed Scheme

We consider only the traffic from the wireless devices to the Access Point. Our scheme incorporates two of the three EDCA enhancements of DCF, i.e., AIFS and TXOP. We propose the following additions/changes to EDCA and the work in [23], and we "translate" some of the ideas presented in [23] in the context of EDCA, as [23] was compared against DCF and did not include the EDCA enhancements. In our target application scenario, telemedicine and non-telemedicine nodes transmit over a single information channel. Regular (non-telemedicine) video traffic can be transmitted for various purposes, including video calls (1-to-1 videoconference, multi-party collaborative calls, immersive video), live transmission of events through mobile devices, video surveillance, screen mirroring. Our proposed modifications can be implemented at the firmware level, similarly to [48].

3.1. Self-Policing for Telemedicine and Regular Video Users

There are four tokens in the system in our proposed scheme, as opposed to two in [23], since that work does not consider telemedicine or video traffic. These four tokens will be named "*permission tokens*" for the rest of the paper. The first permission token is circulated among telemedicine nodes, due to the urgency of this type of traffic. The second permission token is circulated among voice nodes, the third among video nodes and the fourth among data nodes. When a node holds the token, it will transmit its packet(s), when the channel is available. The portion of channel time unused by telemedicine nodes is shared in our scheme by voice nodes, first, and then by video and by data nodes (BE and BK, respectively).

A voice node transmits all its backlogged packets after obtaining the voice token. A data node is assigned a maximum channel occupancy time, equal for all data nodes. During this time, the data node can transmit one or multiple packets depending on its packet size and transmission rate. We assign a TXOP equal to zero for BK and BE traffic and a TXOP equal to 1504 μs to voice traffic. The proposed scheme works in a distributed manner; there is no central controller passing the tokens to others. The current token holder decides which will be the next token holder. When a backlogged node holds the token, it piggybacks the token in its voice/data packet transmission and passes it to the next node. When a data token holder has no packet to transmit or a voice token holder changes from the *on* state to the *off* state, the node passes the token directly to the next holder.

However, the above ideas proposed in [23] cannot be used for telemedicine nodes and regular video nodes, as they do not take into account the urgency and the burstiness of each type of traffic, respectively. If a video node was allowed to transmit all its backlogged packets, it could greedily occupy the channel for a significant amount of time, in case of a burst (which, in the case of the H.264 traces used in this study, can be up to 25 times the mean and in the case of the H.265 traces used can be up to 83 times the mean). The same is true in the case of the transmission of a large telemedicine image or X-Ray file, and in the case of telemedicine video. On the other hand, if a strict TXOP value was defined for all telemedicine and regular video nodes, this could lead to unfairness for certain nodes. For example, this could occur in the case of a regular video node transmitting at a lower rate than its declared mean for a while, and now needing to transmit a significantly larger video frame, e.g., a new *I* frame denoting a significant scene change. Similarly, telemedicine data nodes can not be assigned a

TXOP value equal to zero (similarly to data nodes), as this could jeopardize the timely transmission of telemedicine traffic.

To solve these problems, our scheme works as follows. When a telemedicine or a regular video node obtains the respective permission token, it does *not* transmit all its backlogged packets before sending the permission token to the next node. Instead, assuming non-selfish nodes, we propose the use of *self policing* in each node, based on the accurate video traffic model presented in our work in [25] and assuming a similarly accurate model for telemedicine data traffic, like the ones taken from the literature and presented in Section 4.

More specifically, each telemedicine/regular video node runs a jumping window policer, which is shown in [25] to be the most lenient traffic policing mechanism among other mechanisms studied. The Jumping Window (JW) mechanism uses windows of a fixed length T side by side through time. A new window starts immediately after the conclusion of the previous one. During a window, only K bytes (or packets) can be submitted by the source to the network. In the case that a source attempts to transmit more than K bytes, the excessive traffic is dropped, or marked as nonconforming, as in the case of the Token Bucket. The mechanism is implemented with the use of a transmission token counter. The number of *transmission* tokens is the equivalent of the number of packets that the user is allowed to transmit and should not be confused with the *permission* tokens, representing the turn of the user to transmit. The dynamic approach proposed in [25] was shown to outperform static traffic policing, for H.263 video traffic. Here, it is implemented on H.264 and H.265 regular video traffic, on H.264 telemedicine video traffic and on telemedicine data traffic. Our work in [25] focused only on traffic policing (not on scheduling as this paper does) and focused on cellular networks, not WLANs.

It needs to be emphasized that the JW mechanism is not used in our scheme in order to drop or mark excessive traffic, but to control each user's TXOP. Therefore, each user's TXOP is equal to the time needed for the transmission of the number of packets that the user is "allowed" by its policer to send. This means that in our proposed scheme, contrary to the approach of EDCA (for all types of traffic) and

[23] (for voice/data traffic), *the TXOP is not the same for all telemedicine nodes, nor for all video nodes*. To the best of our knowledge, the idea of using variable TXOPs per user, the value of which is controlled via a traffic policing-like mechanism, has only been proposed once in the relevant literature (it was introduced in our work in [24], but not for telemedicine traffic). This approach solves, as will be shown from our simulation results, the aforementioned problem of how to define TXOP for telemedicine and video nodes. It is combined in our work with the idea of token passing which solves the well-known problem of EDCA where a large number of stations from the same AC (hence, same AIFS, TXOP) can lead to a high collision probability and lower channel utilization.

The use of self-policing for telemedicine traffic as well is required, despite its urgent nature. The reason is that many different types of telemedicine traffic may need to be transmitted and, respectively, a significant number of patients/physicians may need to communicate at a given time; hence, even a telemedicine node cannot be allowed to dominate the channel.

It should also be mentioned that there is no "optimal" solution when our approach is used, in the sense that there is no optimal traffic policer. We have opted for using the lenient JW policer since our goal, as explained above, is to control each user's TXOP, not to drop or mark excessive traffic.

3.2. Access Priority and Dynamic Token Passing

EDCA assigns the same AIFS, equal to 2, for the video and voice categories, while the values for best effort and background traffic are 3 and 7, respectively. Therefore, no AIFS value is available for urgent traffic of any type. The fact that voice traffic is considered of higher priority than video traffic is expressed via the values of CW_{min} and CW_{max} for each type of traffic (smaller values for voice nodes). Since our proposed scheme does not use contention windows, as it practically eliminates contention with the use of tokens, a different mechanism needs to be implemented in order to enforce voice priority. *For this reason, we change the default AIFS values of AC_VI and AC_BE, and we introduce one* *more AIFS value, AC_UR, for urgent traffic (in our case, telemedicine traffic),which is lower than that for voice traffic.*

More specifically, the AIFS values for {BK, BE, VI, VO, UR} respectively are {7, 5, 4, 3, 2}.

Regarding the basic rate and channel rate, we need to mention that the basic rate set is the set of rates that all devices that want to associate with a given access point must support. It is a subset of the transmit rate set, which is a set of the fastest rates that an AP or wireless router will send data. This information is transmitted ("advertised") by an access point to the network. All control, multicast, and broadcast packets are transmitted using one of the basic rates. In our work, we assume that the basic rate is fixed at 2 Mbps and that the transmit rate is 11 Mbps.

When a *new* telemedicine or regular video user enters the network it waits for the channel to be idle for T_{NEWUR} < AIFS(AC_UR) or T_{NEWVID}<AIFS(AC_VI), respectively, and transmits. Nodes broadcast a JOIN message and a LEAVE message when they arrive and depart from the network, respectively.

In the case that a telemedicine or video node leaves the WLAN or ends its transmission, the node sends a message to announce this and passes the token to the next node of the same type. The previous token holder makes note of this, so that it will not send the token to the departing node again in the future. Also, the same token initialization procedure is followed if a node which has already transmitted “crashes” and does not send the LEAVE message. The rest of the token passing procedure is the same as that in [24].

3.3. Using Packet Importance to Mitigate Congestion

In the case of multimedia traffic being transmitted over a WLAN, video traffic can easily become responsible for congestion, due to the much higher loads it carries. The burstiness of a video trace can be somewhat restricted through the use of self-policing to define the nodes’ TXOP, as explained in Section 3.1. However, this restriction is not enough to prevent congestion. There are two possible causes for congestion: a) the total video traffic load is too high for a period of time, b) a video user may have been accumulating transmission tokens during a low rate transmission period, and then the user is allowed to send a burst through the network. Our intention is to use a simple but efficient mechanism, to mitigate congestion. We believe that the video decoding procedure offers a significant opportunity towards this goal.

The decoding of an I frame in a typical H.264/AVC or H.265 HEVC single layer trace is independent of other video frames. The decoding of P frames depends on the successful decoding of the I frame. The decoding of B frames depends on the successful decoding of I and P frames. As noted in [26], an I-frame packet loss results in a visible artifact duration of GoP length or even longer. The artifact duration caused by P frame packet loss is smaller, although it can be significant in specific cases. The artifacts caused by a B frame packet loss, if noticeable, look like an instant glitch, because there is no error propagation from B frames. Similar conclusions were reported in [27], where the loss of P and B frames was shown to have minor difference in perceptual quality for low motion clips (such as the ones considered in our work), whereas the loss of P frames was more important for high motion clips.

Therefore, *we use the following procedure* to help the system dynamically react and mitigate congestion, and hence decrease packet transmission delays for all users: firstly, a QoS threshold, based on a respective QoS metric, needs to be defined, beyond which the nodes consider the network to be congested. If the network is not congested, the node transmits all the frames of the video trace. If the network is congested, the node moves on to a lower quality mode and transmits all I and P video frames (B frames are dropped at the transmitter). If the network remains congested, the node moves on to an even lower quality and transmits only the I frames, i.e., only the basic information of each Group of Pictures (GoP); P and B frames are dropped in this case. This procedure of moving to a lower quality is simple to implement, as the node simply needs to decide whether to transmit or not a frame, depending on the frame type, which is known from the video header.

The important problem is to correctly define the QoS metric to be used, in order to mitigate congestion without compromising more of the video quality than needed. A good metric to use as an indicator for video QoS of the system is the video packet transmission delay; this metric can be used at each node for QoS evaluation. However, by using the mean video packet transmission delay at each node as an indicator of congestion, a node may be led to decrease the quality of its video transmissions unnecessarily. The reason is that an increase in the mean video packet transmission delay experienced by a video node could be owed to a large TXOP allowed temporarily to other video users at a given point in time. Similarly, the mean packet transmission rate (mean throughput) of a node is not the proper choice, as the node may be transmitting smaller traffic loads because of the actual current content of its transmission. For this reason, *we propose the use, as a QoS metric, of the delivery ratio*

R= (packets_transmitted)/(packets_generated)

This metric is computed by the node, and if it is found to have decreased in two consecutive intervals of length *delta*, then the node moves on to a lower video quality using the procedure described in the previous paragraph. The value for the parameter *delta* is discussed in Section 5. The inverse is also implemented: if a node has earlier decreased its quality and the ratio is found to have increased in two consecutive intervals, the node moves on to a higher video quality. This metric also has the advantage that it is very easy for the node to compute (as it has knowledge of how many packets it has generated and transmitted) therefore it does not increase the computational complexity of the scheme.

There is another metric which we found equally useful, in terms of results, as the above mentioned ratio. That metric is the coefficient of variation of the video packet transmission rate per video frame type at the node. In all the cases we studied, an increase of the coefficient of variation of the rate (calculated per GoP) was due to a significant increase in the standard deviation in comparison to the mean, because of congestion. However, since this calculation needs to be done for each different type of frame and for every GoP, the use of this metric leads to unnecessary complexity; for this reason we only adopt the delivery ratio R for use in our scheme.

We need to emphasize that *our solution is only enforced on regular video traffic*; telemedicine video is transmitted without dropping frames at the transmitter, due to the crucial nature of its data.

4. QoS and QoE Metrics and Parameters

Telemedicine arrivals and regular video arrivals are Poisson distributed [28]. We consider four types of telemedicine traffic: a) Occasional measurements (snapshot mode) for four vital parameters (blood pressure, heart pulse, pulse oxymetry and electrocardiogram (ECG) signals), b) X-Rays, c) Medical still images (which may represent various types of medical files such as MRI or Ultrasound), and d) Telemedicine video. We consider all four types of telemedicine traffic to be of the same priority (i.e., all are characterized as urgent traffic). Transmission of video is important for remote diagnosis; seeing the patient gives the physician access to useful indicators such as the patient posture, skin coloration and perspiration, skeletal and tissue deformations. For this reason, the requirements for transmitting diagnosis-grade video [29] are different from those of video conferencing (e.g., in terms of colour faithfulness). The work in [8] studied the relation between the network bandwidth dedicated to the video transfer and the diagnosis quality, and compared uniform and Region-of-Interest (ROI) video encoding. In the case of ROI, a subset of the picture (the region of interest) is encoded at higher quality compared to the video background. ROI encoding is shown in [8] to save 53% of the bandwidth (the required mean bit rate of telemedicine video dropped from 1.077 Mbps to 500 Kbps). Hence, we use the video traffic model from [25] in order to simulate the traffic of a H.264 trace with mean equal to 500 Kbps and standard deviation equal to 600 Kbps, assuming that at first the region of interest might be unclear but soon it will be defined by the remote physician. We choose to use the model for H.264, because H.264 has been shown to perform much better than MPEG-4 for telemedicine applications

[30]. We set an upper bound of 0.01% video packet dropping for telemedicine video. Packets are dropped if they are not transmitted within 100 ms [31].

In an actual system, snapshot mode for the four vital parameters is done with the use of a 3-lead ECG. A recording session of 10 secs weighs about 90 KB of data which is transferred with a target delivery delay of 30 secs. Taking into account the overhead incurred by encapsulation into packets, it requires a bit rate of 25 Kbps [29]. We consider that a typical X-Ray file size is 200 Kbytes [6], with a standard deviation of 20 Kbytes, and that the aggregate X-Ray file arrivals are Poisson distributed with mean λ_X files/frame. Medical image files have sizes ranging between 15 and 20 Kbytes/image and their arrivals are Poisson distributed with mean λ_I files/frame. The upper bound for the transmission delay of an X-Ray file is set to 1 minute (the average download time needed in [6]), and the upper bound for the transmission delay of an image is set to 5 seconds. A telemedicine node arriving in the network chooses one of the four types of telemedicine traffic with equal probability. The QoS requirements for all telemedicine traffic types are rather strict, because of the urgency of the applications. This strictness has been discussed in [32], which considered the problem of transmitting telemedicine traffic over wireless cellular networks.

Voice traffic is represented by a 2-state (on/off) Markov model. In line with the traffic characteristics digitized with the G.711 coding standard, the voice packet inter-arrival period is 20 ms and the packet size is 160 bytes. The inter-arrival time for BE data traffic is 7.5 ms and the packet size is 1000 bytes. Hence, the voice rate is 64 Kbps and the best effort data rate is 1.07 Mbps, respectively [15].

For BK traffic we adopt the http traffic model from [33].

For regular video, we have used four sequences of H.264/AVC VBR encoded videos in our study, from [34] and two sequences of H.265/Single Layer HEVC from [46]. We chose to use sequences encoded with both standards given that H.265 (the newest video coding standard) generally offers a better visual quality for the same bit rate (it has also been used in [43, 44] for telemedicine), however it is not yet as popular as H.264.

The interframe period is 33.3 ms for the H.264 traces and 41.67 ms for the H.265 traces. The packet size is 1280 bytes [15]. The trace statistics are presented in Table I. A video node arriving in the network chooses one of the six traces with equal probability.

Video Trace	**Encoding**	**[G, B, F]**	**Mean (Kbps)**	**Standard Deviation (Mbps)**	**Peak (Mbps)**
NBC News	H.264	[16, 3, 28]	439	5.2	5.5
NBC News	H.264	[16, 7, 38]	121	1.9	2.4
Sony Demo	H.264	[16, 3, 28]	384	6.7	6.7
Sony Demo	H.264	[16, 7, 38]	105	2.3	2.6
Blue Planet	H.265	[24, 7, 45]	85	0.4	7.1
Finding Neverland	H.265	[24, 7, 45]	90	0.2	3.9

Table I. Video Trace Statistics (G, GoP size; B, number of successive B frames; F, quantization Parameters)

Parameter	**Value**
Slot	20 μs
SIFS	10 μs

PHY preamble	192 µs
RTS frame size	20 bytes
CTS frame size	14 bytes
Token frame size	36 bytes
Channel Rate	11 Mbps
Basic Rate	2 Mbps
AIFS (AC_BK)	7
AIFS (AC_BE)	5
AIFS (AC_VI)	4
AIFS (AC_VO)	3
AIFS (AC_UR)	2
T_{NEWUR}	30µs
TXOP (AC_BK)	0
TXOP (AC_BE)	0
TXOP (AC_VI)	Variable per node
TXOP (AC_VO)	1504 µs
TXOP (AC_UR)	Variable per node
Error Rates	Various values between 10^{-1} and 10^{-5}

Table II. Simulation Parameters

Table II contains an overview of the basic simulation parameters.

Each point of the event-driven simulation is derived as the average of 10 independent Matlab runs (Monte-Carlo simulation), each simulating 180 seconds of channel time. All our results have been derived for 95% t-confidence intervals (constructed in the usual way [35]). For each new regular video node arrival, a 2-minute sequence of the chosen video trace is used at random. For comparison with EDCA, the TXOP for AC_VI is 3008 μs in our EDCA simulations, and telemedicine traffic in EDCA is treated as BE. We consider a 2-state (Gilbert-Elliott) channel error model. We have studied our system for various values of packet error rates, ranging between 10^{-1} and 10^{-5} [36, 37].

In [38], the authors expand the work first presented in [39] on a Mean Opinion Score (MOS) prediction process flow. MOS provides a numerical representation of the perceived quality of received media after compression and/or transmission. The work in [39] presented a function for accurately predicting MOS for video sequences using as input the total number of dropped and repeated video frames. Values of MOS ranging from 1 to 5 correspond, respectively, to bad and excellent video quality. If *l* is the total number of dropped frames, *r* the total number of repeated frames, and *f* the number of frames used for the prediction, then the predicted value of MOS at time t, Q(t) is computed as:

$$e=l+r$$

$$\hat{e}=240*e/f \qquad (1)$$

$$Q(t)=-0.571*\ln(\hat{e})+4.6836$$

Due to the fact that the authors in [39] provide evidence for the model accuracy only for frame sequences of 240 frames but not for arbitrarily long sequences of f frames, [38] proposed to use an exponentially weighted average of a number of qualities estimated in the past over continuous short-term sequences, via (1). Hence, if Q(t) is the short-term MOS of the past τ-second interval at time t, then the estimated overall quality $Q_E(t)$ of a long interval of T seconds considered at time t is computed as:

$$Q_E(t)=\beta*Q(t)+(1-\beta)*Q_E(t-\tau) \qquad (2)$$

where $\beta \in [0,1]$ is a weighting factor which, when close to 1, gives more weight to more recent samples.

In our work, we use Equations (1) and (2) assuming that there are no repeated (frozen) frames, i.e., we only take dropped video frames into account, hence e=l.

In our experiments we used two values of β, namely 0.6 (which was used in [38]) and 0.9, in order to evaluate how the difference in the value of β influences our results.

5. Results and Discussion

The results presented in Figures 1-3 have been derived for a channel with a low packet error rate of 10^{-5}. Figure 1 presents the average video packet delay with 5 video nodes and 3 telemedicine nodes present in the network, in the absence of voice traffic and for an increasing number of data nodes. It also presents the average voice packet delay with 25 voice nodes present in the network (case of a WLAN for a health institution, or a cellular/WLAN deployment), for an increasing number of data nodes and a constant number of video and telemedicine nodes, equal to 2 for each node type. Our results show that with the use of our scheme, which provides guaranteed priority to voice and video nodes over data nodes, the delays remain very low, below 2 ms for voice packets and below 12 ms for video packets, even in the case where no congestion mitigation is enforced. Figure 1 also shows that when we enforce our mechanism for congestion mitigation (regular video nodes transmitting at a lower video quality), the video packet delay achieved with our scheme drops even further (by 7.2% on average). For a large number of data nodes, our scheme becomes comparable or outperforms even HCCA, which shares with our scheme the concept of eliminating collisions (our scheme through token passing, HCCA through polling). These results have been acquired by setting the value of the interval *delta* in our scheme to be equal to 0.5 ms, but we have also experimented with other values, between 0.2 ms and 1 ms, and the results were very close to those presented in Figure 1. We also computed the jitter (variance of delay) results. The jitter, which did not exceed 8 ms even for the highest loads examined in our simulations, was marginally lower in our scheme for high loads in comparison to HCCA. There are two reasons for this: a) when a poll is lost in HCCA, there is a certain delay until the same station is polled again, b) the selective transmission of just I and P frames or just I frames from nodes which implement our congestion mitigation mechanism.

The QoS requirements for all types of telemedicine traffic are met with the use of our scheme, even for 50 data nodes, while EDCA fails to satisfy them even when 10 data nodes are present in the system.

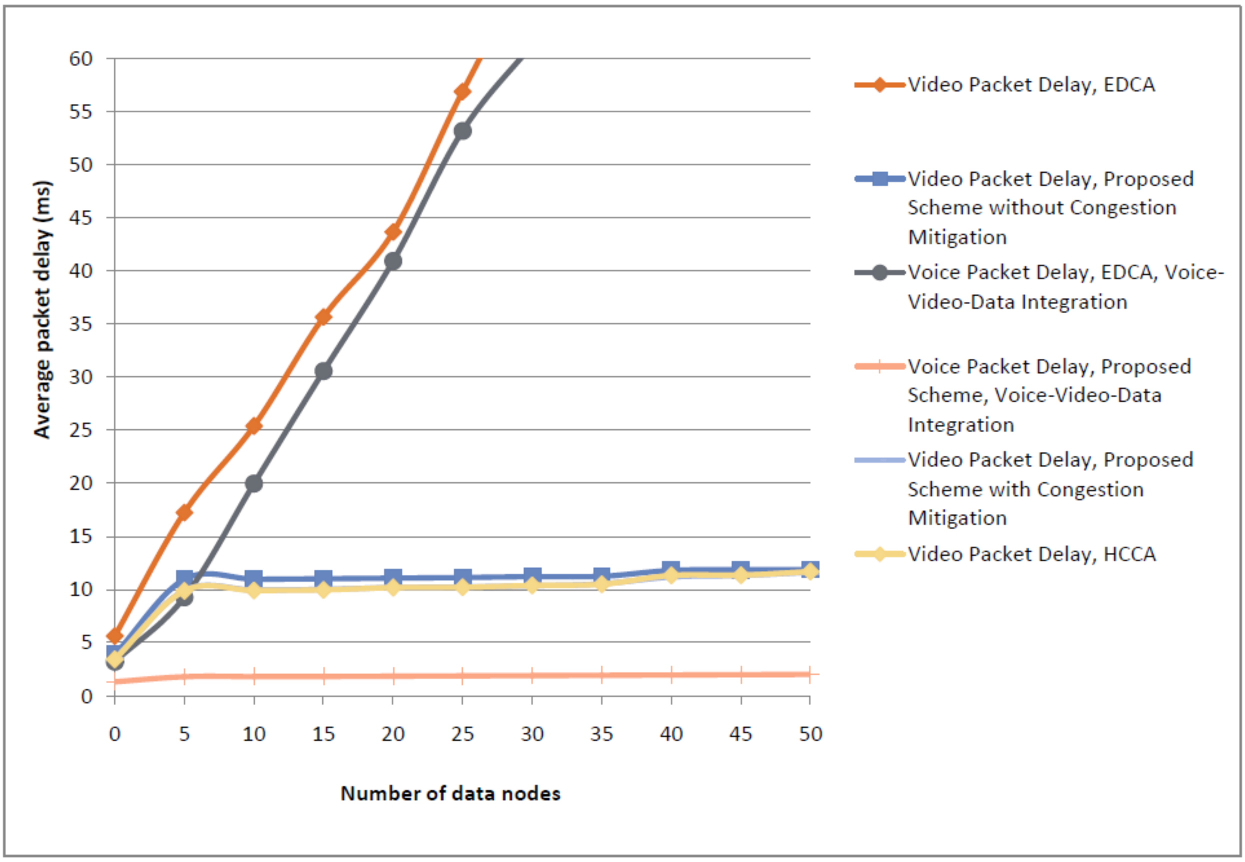

Figure 1. Average Voice and Video Packet Delays vs. the number of data nodes.

In the following discussion, by “Proposed Scheme” we are referring to the implementation of our scheme with the inclusion of the congestion mitigation mechanism.

In Figure 2 we compare the video packet delay results in the case where only token passing is used and TXOP is *fixed*. The values of TXOP used for our scheme are those corresponding to: a) the time needed for a video user to transmit its mean video frame size, b) the time needed for a video user to transmit its peak video frame size. The respective fixed values of TXOP for all video users in EDCA were those corresponding to the transmission of: a) the mean of the average video frame sizes, b) the mean of the peak video frame sizes.

In both cases, our scheme again outperforms EDCA, *up to the point that the performance of our scheme for a TXOP equal to the peak (i.e., constant overallocation) is comparable to EDCA's performance for a TXOP equal to the mean*. Still, the results are clearly worse than those achieved when we use variable TXOP in our scheme. The reason is that, with the use of our accurate video traffic model, the choice of dynamic TXOP is close to optimal, whereas with a fixed TXOP value there is always the problem of overallocating or underallocating time to a video user. Our results agree conceptually with those in [40], where the authors concluded that allocating TXOP limit based on the burst size distribution can improve the network performance under bursty traffic.

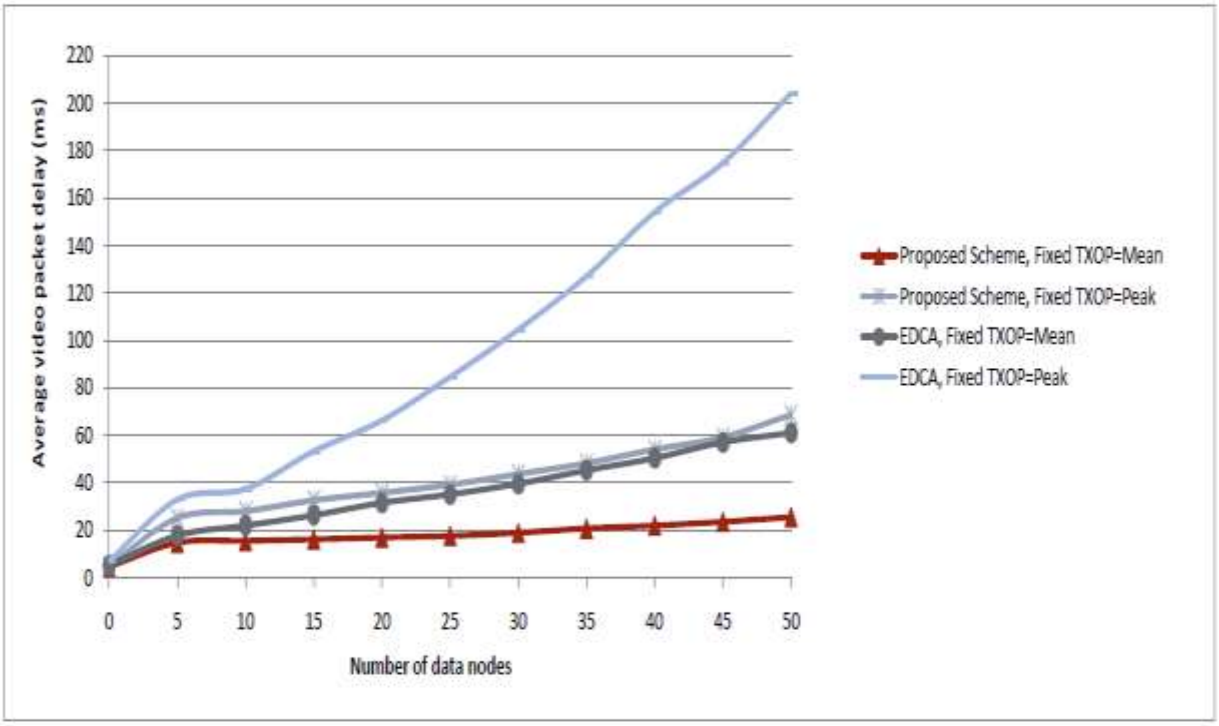

Figure 2. Average Video Packet Delays vs. the number of data nodes, for fixed TXOPs.

Figure 3 shows that the increase in the number of video nodes does not affect the voice packet delay with the use of our proposed scheme, as the delay for 25 voice calls does not surpass 2 ms. However, it has a significant impact when EDCA is used. A constant number of 10 data nodes and 2 telemedicine nodes was used in these simulations. By transmitting at a lower video quality, regular video users give the system room to "breathe" and hence not only help decrease their own (video packet) delay but also the delays for all other types of traffic, including voice. The use of congestion mitigation decreases voice packet delay on average by 22% (the improvement ranges from 10-30% as shown in the Figure).

Regarding voice traffic, as pointed out in [41], acceptable one-way delays are of the order of 150 ms, which means that our scheme can provide excellent QoS to VoIP users, by offering the second highest priority, after urgent traffic, to voice.

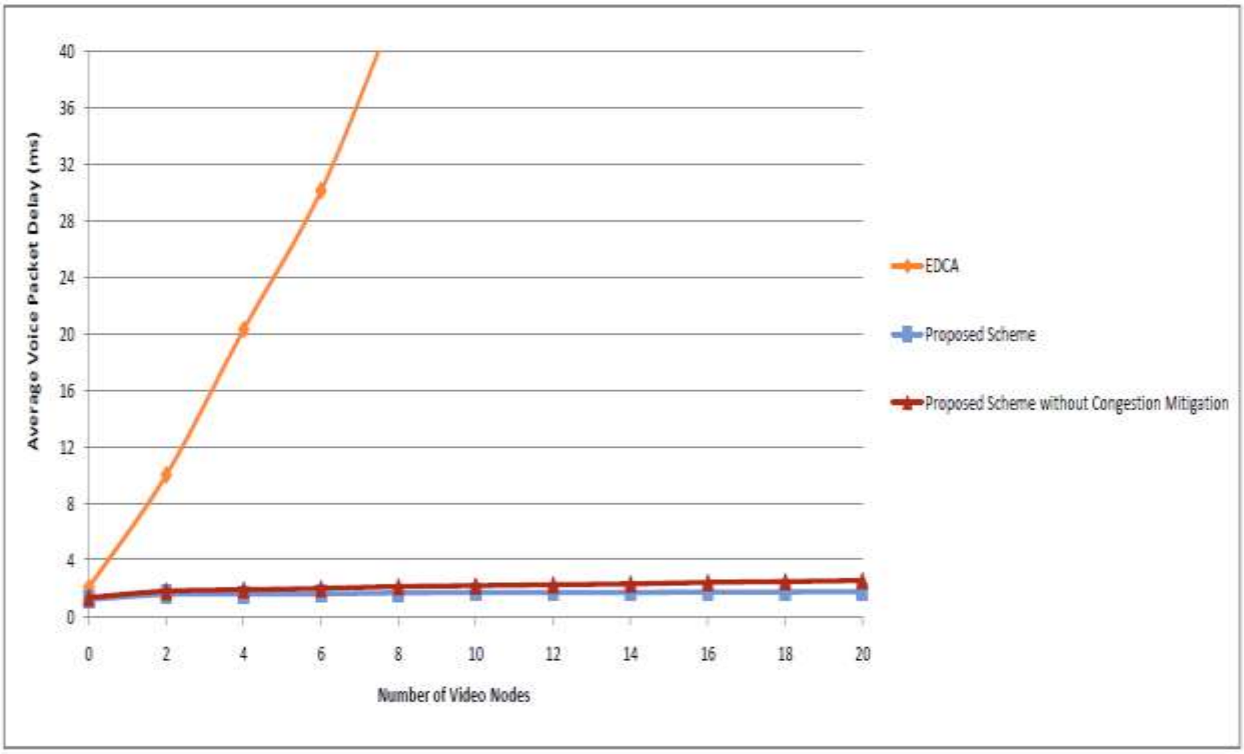

Figure 3. Average Voice Packet Delay vs. the number of video nodes.

Figure 4 presents our results on the channel utilization achieved by our proposed scheme and EDCA, respectively. As shown in the Figure, even for a high packet error rate of 10^{-1} our scheme provides higher channel utilization than EDCA does for a channel with a low packet error rate of 10^{-5}. These results were derived by running ten different simulation scenarios, where the system traffic load was generated with a specific mixture of telemedicine, voice, video, BE and BK traffic. The telemedicine traffic load ranged, in all the scenarios, between 5% and 10%. The voice traffic load ranged between 10% and 60%. The video traffic, BE and BK loads ranged between 10% and 70%. HCCA is shown to provide only marginally better results in the case of high traffic loads. The results for HCCA do not take into consideration the problems of HCCA's complicated software architecture and performance in multi-collision domains.

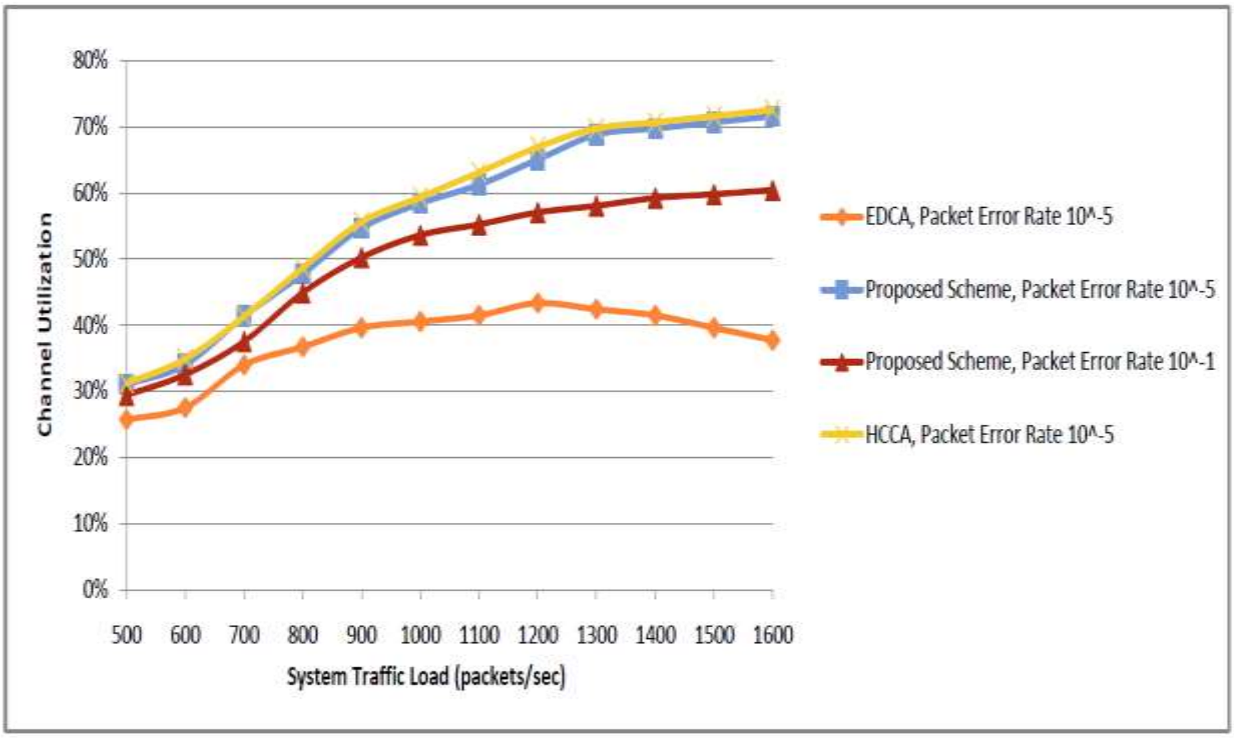

Figure 4. Channel Utilization vs. the system traffic load

The results presented in Figure 5 focus on fairness, based on the video packet delay encountered by individual video streams. The use of Jain's fairness index [42] shows once again that the idea of using variable TXOP for video users, based on an accurate video traffic model, clearly outperforms EDCA, as well as the implementations of the proposed scheme with fixed TXOP. Still, for high loads, our scheme is shown to provide worse fairness results than its implementation without the congestion mitigation mechanism. The reason is that, under high traffic loads, certain regular video nodes will transmit at a lower quality, while others will retain their higher quality (generally, a few nodes transmitting at a lower quality is enough to alleviate congestion). This leads to inequalities among the regular video nodes in terms of their allocated TXOPs and in terms of the waiting times for each video to transmit, which in turn can lead to smaller fairness, as shown in the Figure.

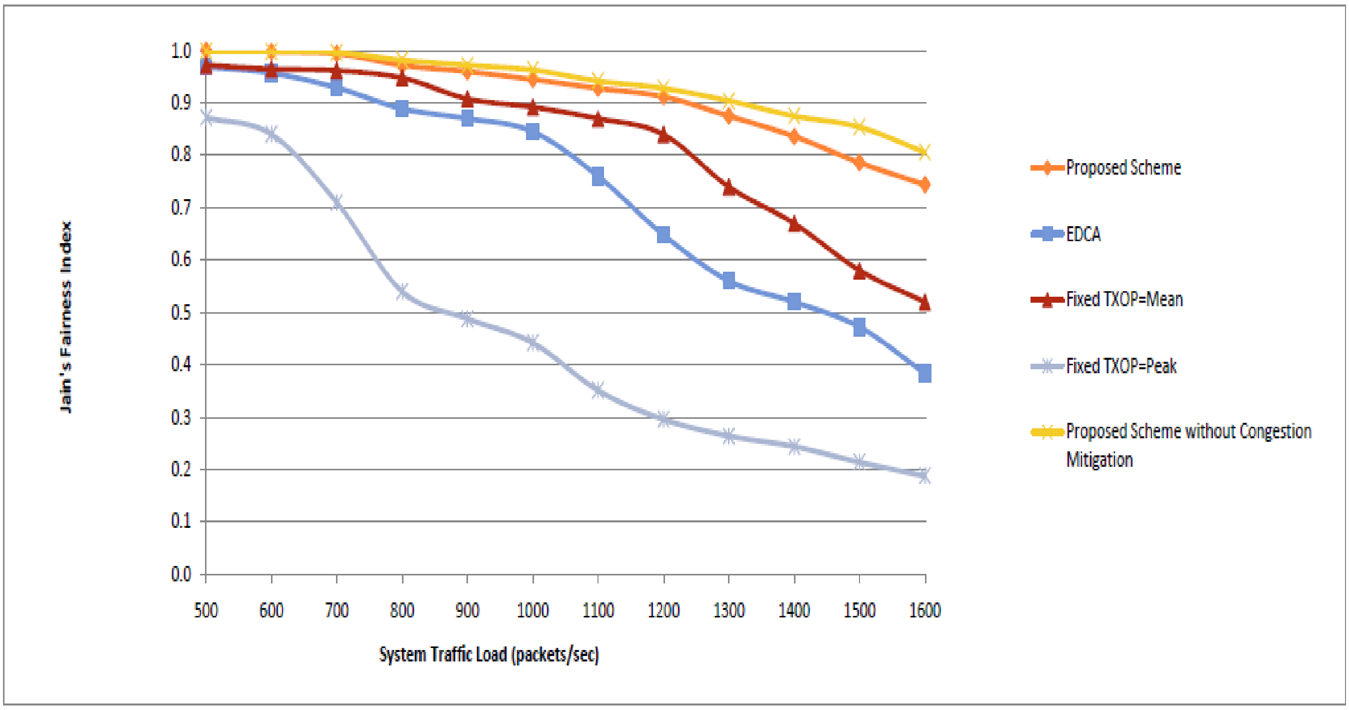

Figure 5. Fairness Index vs. system traffic load

Regarding the Quality of Experience for telemedicine video users, the use of Equations (1) and (2) for telemedicine video users gives predicted MOS values ranging from 3.5 to 4.2 for all of our simulations which were conducted with f equal to 60 and 120, while the respective values for EDCA range from 2.8 to 3.2. The difference in the value of β is shown to have small effect in our results, with the 0.9 value leading to slightly smaller estimated MOS than the 0.6 value, as it tends to over-weigh any recent frame loss.

We also need to point out that, the larger the number of telemedicine nodes in the system (for the same total traffic load) the worse are the results for voice, video and data (BK and BE) packet delays. This result is expected, as telemedicine acquires top priority in the system, but again it is not a result that should cause concern, because telemedicine traffic is sparse by nature, and can only become dominant in the channel and create significant delays for the other types of traffic in a mass medical crisis situation. And in that case, its dominancy would be welcome.

In our study, we have assumed that the transmission of all types of telemedicine traffic that we consider is urgent, and for this reason we have treated it as such. We also wanted to show that with the use of our scheme, even in the case of significant volumes of urgent traffic the required QoS is satisfied. In reality, the results of, e.g, an X-Ray, depending on a patient’s condition, may be urgent for the doctors to see or not. In the case that part of the telemedicine traffic is not urgent, it could be treated as Best Effort traffic (the sender could have the option of indicating the urgency of the transmission), with worse QoS.

A few comments are necessary on the limitations of this study. We assume a fully-connected WLAN where all the nodes can hear each other. Extending to a partially connected WLAN, as also pointed out in [23], is technically challenging. In general, as explained in [47] which proposed a token-passing protocol for wireless industrial networks, bringing token-passing to the wireless domain contains challenges. Instability in the network can be caused by token transmission errors or in the cases when a station holding the token fails. In wireless networks, error bursts caused by fading are common, and mobility/propagation effects can cause correlated failures in multiple wireless links simultaneously. Still, as shown in [47] these problems can be significantly mitigated through cooperative ARQ (Automatic Repeat Request).

Also, in order to achieve the superior performance supported by the proposed scheme, the scheme needs to be implemented in all the nodes of the fully-connected WLAN. In the case when standard 802.11e devices or legacy devices using DCF coexist in the network, the only way for the scheme to be backward compatible is for the "evolved" nodes to work in 802.11e mode, which would lead to the significantly worse EDCA results in regards to the QoS and QoE of telemedicine traffic.

Finally, depending on the profile employed in H.264/AVC, B-frames may not be used in the encoding structure; in these cases our proposed scheme will resort to dropping only P-frames, if needed. Furthermore, B-frames can at times be used as reference frames in H.264 (this is called B-frame pyramiding); this is not common, because there are players who do not properly decode B-frames as reference frames, however in such cases our scheme would not be useful, as the dropping of B-frames would result in a significant deterioration of video quality.

6. Conclusions

We have proposed a new scheduling scheme using token-passing and self-policing for the integration of telemedicine traffic with voice, video and data traffic over WLANs. Our scheme introduces significant in essence but not difficult to implement modifications of the EDCA, to ensure the proper prioritization among different Access Categories which leads to high telemedicine video QoE and to the satisfaction of the QoS requirements of four different types of telemedicine traffic. Our proposal practically eliminates contention and TXOP idle time, and dynamically reacts to cases of congestion, by forcing video nodes to transmit at a lower quality; hence, it leads to a significant increase in channel utilization when compared with EDCA.

A distinctive characteristic of our proposed scheme is its decentralized nature (congestion mitigation is done *locally, at the nodes*) and the fact that it actually helps higher layers by reducing congestion so that the transport layer will need to solve congestion control problems more rarely. Given that:

a) there is no need for changes in the architecture of the healthcare environment in order to implement our scheme, and

b) the changes that we suggest do not affect any standard besides the small changes in EDCA,

we believe that our scheme can be of significant practical value.

In future work, we will address the well-known hidden-node problem which could affect our scheme, since it uses token-passing (still, solutions have been proposed in the literature to resolve it, such as the increase in transmission power from the nodes and the use of omnidirectional antennas) and we will focus on the problem of providing guaranteed QoS to all types of multimedia traffic, not just telemedicine. We intend to gather a large amount of measurement of actual telemedicine data from major hospitals and to conduct experiments in the hospital environment, in order to use it for the further evaluation of our scheme in practice (e.g., evaluating in practice the delays associated with self-policing in each node). We will use PSNR among the QoE metrics for telemedicine video and we will associate it with the MOS of medical experts clinically evaluating the transmitted telemedicine video streams, similarly to the work in [45], which used the HIPERMED platform for its experiments, and [43], which evaluated state-of-the-art video quality metrics with respect to compressed medical ultrasound video sequences.

Finally, we will study how the proposed scheme compares and how it can be possibly integrated with other standards' (such as 802.11ac, 802.16) QoS guarantees.

References

1. IEEE Std. 802.11-2012, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.

2. H F. Rashvand et al., "Ubiquitous Wireless Telemedicine", *IET Communications*, Vol. 2, No. 2, 2008.

3. M. Singh et. al., “Application of Handheld Tele-ECG for Health Care Delivery in Rural India”, *International Journal of Telemedicine and Applications*, Vol. 2014, Article ID 981806, 2014.

4. S. Garawi et al.,“3G Wireless Communications for Mobile Robotic Tele-Ultrasonography Systems”, *IEEE Communications Magazine*, Vol. 44, No. 4, 2006.

5. A. Zvikhachevskaya et al., “Quality of Service Consideration for the Wireless Telemedicine and E-Health Services”, *in Proc. of the IEEE WCNC* 2009.

6. M. G. Hadjinicolaou, et al., “Emergency TeleOrthoPaedics M-health System for Wireless Communication Links”, *IET Communications*, Vol. 3, No. 8, 2009.

7. K. Halje et. al., “Towards mHealth Systems for Support of Psychotherapeutic Practice: A Qualitative Study of Researcher-Clinician Collaboration in System Design and Evaluation”, *International Journal of Telemedicine and Applications*, Vol. 2016, Article ID 5151793, 2016.

8. S. P. Rao et al., "Delivering Diagnostic Quality Video over Mobile Wireless Networks for Telemedicine", *International Journal of Telemedicine and Applications*, Vol. 2009, Article ID 406753, 2009.

9. Y. He et al., “A Reservation Based Backoff Method for Video Streaming in 802.11 Home Networks”, *IEEE Journal on Selected Areas in Communications,* Vol. 28, No. 3, 2010.

10. I. U. Rehman, N. Y. Philip and R. S. H. Istepanian, “Performance analysis of medical video streaming over 4G and beyond small cells for indoor and moving vehicle (ambulance) scenarios”, *in Proc. of the 4th International Conference on Wireless Mobile Communication and Healthcare (MobiHealth) 2014*.

11. S. M. S. Al-Majeed, S. K. Askar and M. Fleury, “H.265 Codec over 4G Networks for Telemedicine System Application”, *in Proc. of the 16th International Conference on Computer Modelling and Simulation 2014*.

12. S. Cicalo et al., “Multiple Video Delivery in m-Health Emergency Applications”, *IEEE Transactions on Multimedia*, Vol.18, No. 10, 2016, pp.1988-2001.

13. I. U. Rehman and N. Y. Philip, “M-QoE Driven Context, Content and Network Aware Medical Video Streaming based on Fuzzy Logic System over 4G and beyond Small Cells”, *in Proc. of EUROCON 2015*.

14. D. Ouattara et. al, “A QoS-control framework for medical multimedia data transmission in CRN environment”, *in Proc. of the IEEE Symposium on Computers and Communication (ISCC) 2014*.

15. H. Zen et al.,“Adaptive Segregation-Based MAC Protocol for Real-Time Multimedia Traffic in WLANs”, *in Proc. of the IEEE ICON 2007.*

16. R. Haywood et al.,"Investigation of H.264 Video Streaming over an IEEE 802.11e EDCA Wireless Testbed", *in Proc. of the IEEE ICC 2009.*

17. P. Ansel et al.,"FHCF: A Simple and Efficient Scheduling Scheme for IEEE 802.11e Wireless LAN", *Mobile Networks and Applications Journal*, Vol. 11, No. 3, 2006.

18. E. Supriyanto et al., "Cross Layer Design of Wireless LAN for Telemedicine Application Considering QoS Provision", in "Advances in Telemedicine: Technologies, Enabling Factors and Scenarios", InTech publishers, G. Graschew and T. A. Roelofs (editors), 2011.

19. C. Chigan and V. Oberoi, "Providing QoS in Ubiquitous Telemedicine Networks", *in Proc. of the 4th International Conference on Pervasive Computing and Communications Workshops (PERCOMW 2006)*.

20. A. Panayides, I. Eleftheriou and M. Pantziaris, "Open-Source Telemedicine Platform for Wireless Medical Video Communication", *International Journal of Telemedicine and Applications*, Vol. 2013, Article ID 457491, 2013.

21. D. Lin and F. Labeau, "An Algorithm that Predicts CSI to Allocate Bandwidth for Healthcare Monitoring in Hospital's Waiting Rooms", *International Journal of Telemedicine and Applications*, Vol. 2012, Article ID 843527, 2012.

22. T. Ojanpera, M. Uitto and J. Vehkapera, "QoE-based Management of Medical Video Transmission in Wireless Networks", *in Proc. of the Network Operations and Management Symposium (NOMS) 2014*.

23. P. Wang and W. Zhuang, "A Token-Based Scheduling Scheme for WLANs Supporting Voice/Data Traffic and its Performance Analysis", *IEEE Transactions on Wireless Communications*, Vol. 7, No. 5 (1), 2008.

24. P. Koutsakis, "Token- and Self-Policing-Based Scheduling for Multimedia Traffic Transmission over WLANs", *IEEE Transactions on Vehicular Technology*, Vol. 60, No. 9, 2011.

25. P. Koutsakis, "Dynamic versus Static Traffic Policing: A New Approach for Videoconference Traffic over Wireless Cellular Networks", *IEEE Transactions on Mobile Computing*, Vol. 8, No. 9, 2009.

26. N. Liao and Z. Chen, "A Packet-Layer Video Quality Assessment Model with Spatiotemporal Complexity Estimation", *EURASIP Journal on Image and Video Processing*, 2011:5, pp. 1-13.

27. M. Venkataraman and M. Chatterjee, "Effects of Internet Path Selection on Video-QoE", *in Proc. of the 2nd ACM Conference on Multimedia Systems (MMSys 2011)*.

28. K. Molnar et al., "Optimization of Link Capacity for Telemedicine Applications", *in Proc. of the 6th International Conference on Systems and Networks Communications (ICSNC 2011)*.

29. L. Franck et al., "Modeling and Evaluation of an Aeronautical Telemedicine Service over Satellite", *in Proc. of the 28th AIAA International Communications Satellite Systems Conference (ICSSC)*, Anaheim, USA, 2010.

30. P. Malindi, "QoS in Telemedicine", Telemedicine Techniques and Applications, G. Graschew and S. Rakowsky (editors), InTech, 2011.

31. M. Clarke et al., “Optimum Delivery of Telemedicine over Low Bandwidth Satellite Links”, *in Proc. of the 23rd International Conference of the IEEE Engineering in Medicine and Biology Society 2001*.

32. L. Qiao and P. Koutsakis, “Adaptive Bandwidth Reservation and Scheduling for Efficient Wireless Telemedicine Traffic Transmission”, *IEEE Transactions on Vehicular Technology*, Vol. 60, No. 2, 2011.

33. P. Tran-Gia et al,“Source Traffic Modeling of Wireless Applications”, *International Journal of Electronics and Communications,* Vol. 55, No. 1, 2001.

34. G. Van der Auwera, P. T. David, and M. Reisslein, “Traffic and Quality Characterization of Single-Layer Video Streams Encoded with H.264/MPEG-4 Advanced Video Coding Standard and Scalable Video Coding Extension”, *IEEE Transactions on Broadcasting*, Vol. 54, No. 3, 2008.

35. A. M. Law and W. D. Kelton, “Simulation Modeling & Analysis”, 2nd Ed., McGraw Hill Inc., 1991.

36. G. Min et al., “Performance Analysis of the TXOP Scheme in IEEE 802.11e WLANs with Bursty Error Channels”, *in Proc. of the IEEE WCNC 2009*.

37. S. G. Sitharaman and K. M. Anantharaman, “Impact of Retransmission Delays on Multilayer Video Streaming over IEEE 802.11e Wireless Networks”, *in Proc. of COMSWARE 2007.*

38. C. Vassilakis and I. Stavrakakis, “Minimizing Node Churn in Peer-to-Peer Streaming”, *Computer Communications*, Vol. 33, No. 2010, pp. 1598-1614.

39. A. Younkin et al., “Predicting an average end-user’s experience of video playback”, *in Proc. of the 3rd International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2007).*

40. S. Rashwand and J. Misic, “IEEE 802.11e EDCA under Bursty Traffic – How much TXOP Can Improve Performance”, *IEEE Transactions on Vehicular Technology*, Vol. 60, No. 3, 2011.

41.[Online]:http://www.cisco.com/en/US/tech/tk652/tk698/technologies_white_paper09186a00800a8993.shtml

42. R. Jain, “The Art of Computer Systems Performance Analysis”, John Wiley&Sons, 1991.

43. M. Razaak, M. G. Martini and K. Savino, “A Study on Quality Assessment for Medical Ultrasound Video Compressed via HEVC”, *IEEE Journal of Biomedical and Health Informatics*, Vol.18, No. 5, 2014, pp. 1552-1559.

44. Z. Ul-Abdin, M. Shafique and M. A. Qadir, “Evaluating Video Codecs for Telemedicine Under Very-Low Bitrates”, *in Proc. of the 8th International Congress on Image and Signal Processing (CISP 2015).*

45. A. Chaabouni et. al, “Subjective and objective quality assessment for H264 compressed medical video sequences”, *in Proc. of the 4th International Conference on Image Processing Theory, Tools and Applications (IPTA) 2014*.

46. [Online]: http://trace.eas.asu.edu/videotraces2/h265/

47. C. Dombrowski and J. Gross, “EchoRing: A Low-Latency, Reliable Token-Passing MAC Protocol for Wireless Industrial Networks”, *in Proc. of the European Wireless Conference 2015.*

48. I. Tinnirello et al., “Wireless MAC processors: Programming MAC protocols on Commodity Hardware”, *in Proc. of the IEEE INFOCOM 2012.*