Optimized Scalable Video Streaming over IEEE 802.11a/e HCCA Wireless Networks under Delay Constraints

Mihaela van der Schaar, *Senior Member*, *IEEE*, Yiannis Andreopoulos, *Member*, *IEEE*, and Zhiping Hu, *Member*, *IEEE*

Abstract—The quality-of-service (QoS) guarantees enabled by the new IEEE 802.11a/e Wireless LAN (WLAN) standard are specifically targeting the real-time transmission of multimedia content over the wireless medium. Since video data consume the largest part of the available bitrate compared to other media, optimization of video streaming for this new standard is a significant factor for the successful deployment of practical systems. Delay-constrained streaming of fully-scalable video over IEEE 802.11a/e WLANs is of great interest for many multimedia applications. The new medium access control (MAC) protocol of IEEE 802.11e is called the Hybrid Coordination Function (HCF) and, in this paper, we will specifically consider the problem of video transmission over HCF Controlled Channel Access (HCCA). A cross-layer optimization across the MAC and application layers of the OSI stack is used in order to exploit the features provided by the combination of the new HCCA standard with new versatile scalable video coding algorithms. Specifically, we propose an optimized and scalable HCCA-based admission control for delay-constrained video streaming applications that leads to a larger number of stations being simultaneously admitted (without quality reduction to any video flow). Subsequently, given the allocated transmission opportunity, each station deploys an optimized Application-MAC-PHY adaptation, scheduling, and protection strategy that is facilitated by the fine-grain layering provided by the scalable bitstream. Given that each video flow needs to always comply with the predetermined (a priori negotiated) traffic specification parameters, this cross-layer strategy enables graceful quality degradation whenever the channel conditions or the video sequence characteristics change. For instance, it is demonstrated that the proposed cross-layer protection and bitstream adaptation strategies facilitate QoS token rate adaptation under link adaptation mechanisms that utilize different physical layer transmission rates. The expected gains offered by the optimized solutions proposed in this paper are established theoretically, as well as through simulations.

Index Terms—IEEE 802.11e WLANs, delay-constrained video streaming, cross-layer optimization, QoS-enabled MAC scheduling, token rate adaptation, link adaptation.

1 Introduction

The different layers in the OSI stack have traditionally been assumed to be independent of each other. Such a design facilitated the evolution of each layer in an elegant way without influencing or conflicting with the other layers. Previously, the various layers were optimized in isolation. However, recent work [1], [2], [3], [4], [5] has demonstrated that joint optimization of the various layers improves the overall system performance for delay-constrained, loss-tolerant multimedia applications.

IEEE 802.11 WLANs [6] have emerged as a prevailing technology for the (indoor) broadband wireless access. During the last several years, an increased number of applications require the transmission of delay-sensitive and bandwidth-intense video data over WLANs. However, existing WLANs do not provide the necessary QoS to support such applications. Today, IEEE 802.11 can be considered as a wireless Ethernet, which supports only a best-effort service (not guaranteeing any service level to users/applications). For this reason, the IEEE 802.11 Working Group recently defined a new supplement (part "e") to the existing legacy Medium Access Control (MAC) sublayer of the standard in order to support QoS [7]. A new medium access method called Hybrid Coordination Function (HCF) has been proposed in the 802.11e draft [7], which combines a contention-based channel access mechanism, referred to as Enhanced Distributed Channel Access (EDCA), and a polling-based channel access mechanism, referred to as HCF Controlled Channel Access (HCCA). Both EDCA and HCCA operate simultaneously and continuously. Recent studies [8], [9], [10] have already shown that HCF enables differentiated treatment of traffic streams and can be tuned to meet QoS requirements of low latency and jitter. As such, its use for wireless multimedia streaming designs appears to be an important issue.

In order to achieve optimal transport of video over 802.11e HCCA, we need to accommodate application-layer constraints such as bandwidth variations due to variable-bitrate (VBR) coding [9], delay constraints, and selective packet retransmission [1], [3] to cope with network losses. Starting from the coding engine, nonscalable video coding algorithms do not provide graceful degradation and adaptability to a large range of wireless channel conditions

- *M. van der Schaar and Y. Andreopoulos are with the Department of Electrical Engineering, University of California Los Angeles, 66-147E Engineering IV Building, 420 Westwood Plaza, Los Angeles, CA 90095-1594. E-mail: {mihaela, yandreop}@ee.ucla.edu.*
- *Z. Hu is with Universal Electronics Inc., 6101 Gateway Drive, Cypress, CA 90630-4841. E-mail: zhu@uei.com.*

Manuscript received 18 Aug. 2005; revised 12 Nov. 2005; accepted 5 Jan. 2006; published online 17 Apr. 2006.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-0243-0805.

and power constraints. Hence, although the concepts proposed in this paper can potentially be deployed with state-of-the-art nonscalable coding with bitstream switching [11], [12] this usually entails higher complexity and smaller granularity for real-time packet prioritization and adaptive retransmissions. Consequently, in this paper, we use recently proposed scalable video coding schemes based on Motion Compensated Temporal Filtering (MCTF) [13]. MCTF-based scalable video compression is attractive for wireless streaming applications since it provides on-the-fly adaptation to channel conditions, support for a variety of wireless receivers with different resource capabilities and power constraints, and easy prioritization of various coding layers and video packets [13].

In order to accommodate delay and transmission requirements, we perform optimized *scalable* resource allocation that *jointly* considers the MAC and application layer parameters and quantify the benefits in terms of individual stations performance as well as the overall system performance. The following steps are involved in the proposed cross-layer optimization:

- Unlike in conventional wireless streaming solutions, where each video flow is admitted individually by the Admission Control Unit (ACU) collocated with the QoS-enhanced Access Point (QAP), the application-layer video flow is divided into subflows based on the delay requirements of individual video frames. As detailed in Section 3, this allows for individual bitstream components to interface with the MAC as separate flows with individual delay and transmission bandwidth requirements. We demonstrate that, without any modification to the HCCA mechanism, this allows for the MAC ACU to admit more users (stations) without any compromise in the video quality for the already admitted users.
- Based on the delay requirements of each flow, the optimal transmission scheduling strategy is established in Section 4. Our proposed solution involves low-complexity linear programming. It is shown that, under error-free transmission during the contention-free periods, the proposed optimization further increases the number of admitted stations without any compromise in the video quality.
- The inherent prioritization and graceful degradation properties of scalable coding are utilized in Section 4 in order to provide an optimized framework that defines the maximum retry limit for each MAC service data unit (MSDU) in the video subflows, given the delay constraint and distortion impact for each subflow's transmission duration. This allows an already admitted application/subflow to continue its transmission even if the channel conditions worsen, without (significantly) compromising the video quality. This graceful degradation is extremely important for real-time video applications, where a renegotiation of the TSPEC parameters would have a disrupting effect on the video quality that is unacceptable for the end user.

In order to justify the proposed algorithms and methods, simulation results are presented in Section 5. Our conclusions are drawn in Section 6.

2 Background for Video Transmission over HCCA in IEEE 802.11e

Research issues of 802.11e HCF scheduling have recently started to gain some attention. Initial contributions [8], [14] were mainly concerned with the feasibility of the EDCA and HCCA mechanisms of HCF for multimedia transmission. In an attempt to optimize scheduling for VBR video traffic, Ansel et al. [9] presented an approach for efficient scheduling on the QAP based on measured queue sizes of each traffic stream. HCCA was used, as it provides significant benefits over EDCF for applications requiring strict QoS. It is important to mention that all these approaches perform optimization either at the application layer or the MAC layer. This follows the previous legacy of application layer optimizations, e.g., rate-smoothing algorithms for QoS enabled networks [15], such as ATM networks [16]. Nevertheless, our preliminary results [17] have shown that joint application-layer and MAC-layer optimization can significantly improve the overall system performance. The next section outlines the conventional video flow scheduling and admission control in HCCA. Following that, the architecture of the deployed video coder is presented.

2.1 HCCA-Based Admission Control for Video

HCCA is used to provide a parameterized QoS service. With HCCA, there is a negotiation of QoS requirements between the QoS-enhanced wireless station (QSTA) and the Hybrid Coordinator (HC). Once a stream for a QSTA is established, the HC allocates transmission opportunities (TXOPs) via polling to the QSTA in order to guarantee its QoS requirements. The HC enjoys free access to the medium during the contention-free period and uses the highest EDCA priority during the contention period, in order to 1) send polls to allocate TXOPs and 2) send downlink parameterized traffic. It makes use of the priority interframe space (PIFS) to seize and maintain control of the medium. Once the HC has control of the medium, it starts to deliver parameterized downlink traffic to stations and issues QoS contention-free polls (QoS CF-Polls) to those stations that have requested parameterized services. The QoS CF-Polls include the TXOP duration granted to the QSTA. If the station being polled has traffic to send, it may transmit several packets for each QoS CF-poll received respecting the TXOP limit specified in the poll. In order to utilize the medium more efficiently, it is possible to piggyback both the acknowledgment (CF-Ack) and the CFPoll onto data packets. In contrast to the point coordination function of the IEEE 802.11-99 standard, HCCA operates during both the contention-free period and the contention period (see Fig. 1).

The admission control and scheduling units enable HCCA to guarantee that the QoS requirements are met once a stream has been admitted in the network. Alternatively, EDCA only provides a QoS priority differentiation via a random distributed access mechanism.

To ensure user satisfaction, it is essential that, once admitted, a video stream is guaranteed QoS for its lifetime. Thus, there is a need to control how many streams are admitted to the system and what wireless resources should

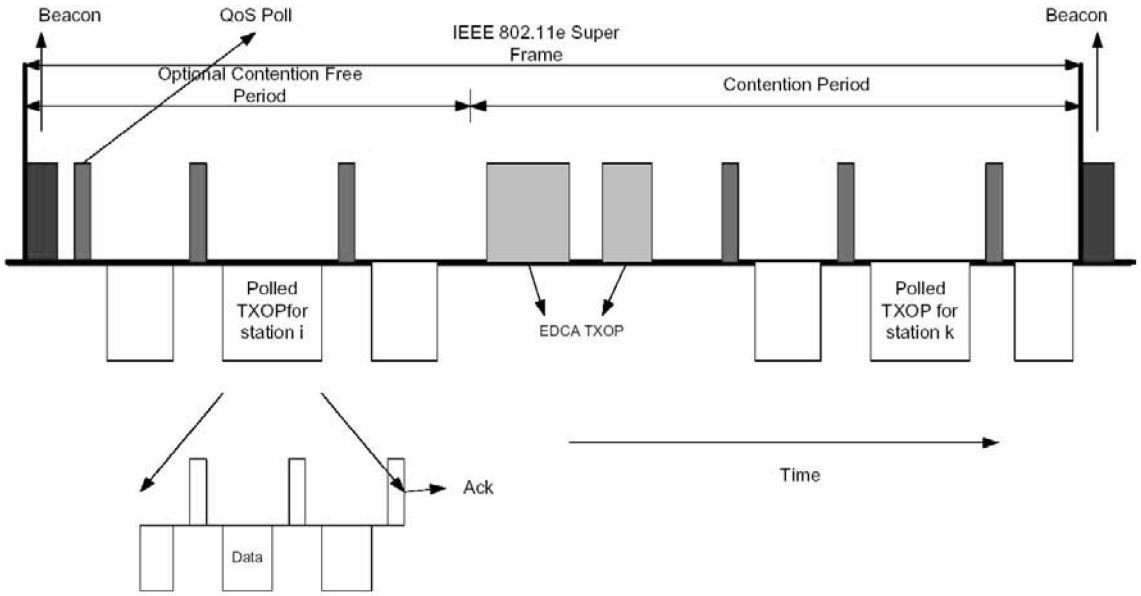

Fig. 1. Operation of the IEEE 802.11e HCF [17].

be allocated to each stream in order to obtain the optimal trade-off between a larger number of admitted stations and an acceptable video quality level for the admitted stations. In other words, a scalable admission control and adaptive protection strategy is necessary. Among the parameters defined in the traffic specification (TSPEC) element of IEEE 802.11e, we use the subset of parameters that influence the design of an efficient admission control algorithm for video applications. For each video flow i, these parameters are: the peak data rate (P_i), the mean data rate (ρ_i), the maximum burst size (σ_i), the maximum permissible delay (d_i), the nominal MSDU size (L_i), and the minimum physical-layer transmission rate (R_i).

In conventional video streaming mechanisms [8], [9], P_i, ρ_i, and σ_i are part of a twin leaky bucket mechanism [15] and are supplied to the MAC by the application layer. Based on the twin leaky bucket analysis [10], [15], the effective bandwidth for each video flow i can be computed as[1]

$$g_i = \frac{P_i}{1 + d_i(P_i - \rho_i)\sigma_i^{-1}}. \tag{1}$$

The previous bandwidth computation is "ideal" in the sense that it does not include overheads. For the transmission of each MSDU frame, there is an overhead in time based on the acknowledgment policy, the PIFS time, MAC-layer and physical-layer headers, and polling overhead. As a result, the scheduling policy has to determine and take into account these overheads, as different scheduling policies determine how many times one has to poll a QSTA in the duration of a service interval (SI), denoted as t_{SI}. Assuming that t_{SI} is known, the number of MSDUs per SI is

$$N_i = \left\lceil \frac{g_i \cdot t_{\mathrm{SI}}}{L_i} \right\rceil \tag{2}$$

and the modified guaranteed bandwidth including overheads is given by

$$g_i' = \frac{N_i(L_i + O_i)}{t_{\mathrm{SI}}}, \tag{3}$$

where O_i represents the additional bits due to overheads for the transmission of an MSDU frame corresponding to video flow i. As a result, having already $i-1$ admitted flows in the network, the admission control for the ith video flow can be expressed as

$$g_i' + \sum_{j=1}^{i-1} g_j' + g_{\mathrm{other}} \leq C, \tag{4}$$

where g_{other} represents additional bandwidth allocated to nonvideo traffic (e.g., audio or other QoS-requiring media) and C is the total guaranteed bandwidth of the wireless medium. It is important to mention that a necessary condition for nonviolation of the initially-negotiated QoS requirements is that $R_i \geq g_i'$. Based on the readjusted guaranteed bandwidth, the number of MSDUs per SI is recalculated as in (2) with g_i replaced by g_i' and, for each video flow i, we denote the modified value by N_i'. The admission control unit can now calculate the TXOP duration required to service all these MSDUs within t_{SI} as

$$t_{\mathrm{TXOP},i} = N_i' \left(\frac{L_i}{R_i} + T_{\mathrm{overhead},i} \right), \tag{5}$$

with $T_{\mathrm{overhead},i}$ the required overheads, as explained before. Similarly to (4), we can express the admission control in terms of the TXOP duration for each video flow i:

1. For the first part of our analysis presented in this section, we assume that channel or link-state analysis is used in order to determine the additional percentage that needs to be reserved for the bandwidth to cover the losses that may occur in the wireless medium. Inititally, we assume an ideal channel condition where no errors occur during the HCCA duration. The modifications imposed in the proposed admission control in order to incorporate the effects of video packet retransmission due to channel errors are presented in Section 4.

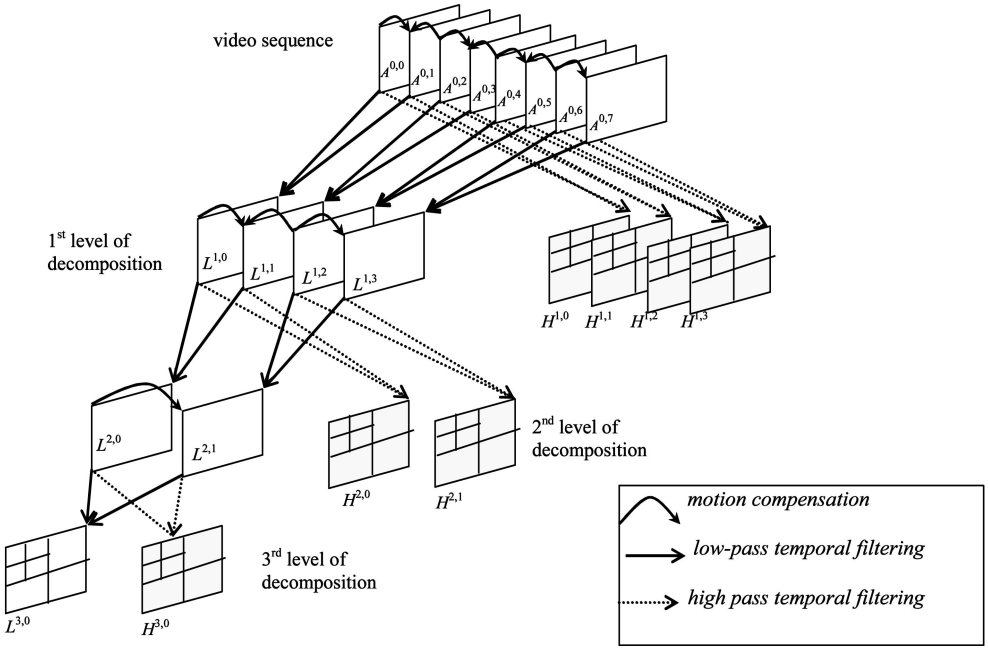

Fig. 2. MCTF decomposition.

$$\frac{t_{\mathrm{TXOP},i}}{t_{\mathrm{SI}}} + \sum_{j=1}^{i-1} \frac{t_{\mathrm{TXOP},j}}{t_{\mathrm{SI}}} + t_{\mathrm{TXOP,other}} \leq \frac{T - T_{\mathrm{CP}}}{T}, \qquad (6)$$

where $t_{\mathrm{TXOP,other}}$ indicates the TXOP allocated to nonvideo traffic, T is the beacon interval illustrated in Fig. 1, and T_{CP} is the time reserved for the contention period, i.e., for EDCA traffic. Importantly, it should be noted that the $T_{\mathrm{overhead},i}$ and $t_{\mathrm{TXOP,other}}$ have a significant impact on the number of admitted stations, as will be shown by the results of Section 5.

The admission control expressed by (6) can be used for the construction of a round-robin, standard-compliant scheduler. In particular, normative behavior set by the IEEE 802.11e draft [7] requires that the HC grants every flow i the negotiated time $t_{\mathrm{TXOP},i}$. Hence, for every video flow, the admission control described by (5) and (6) can be used. The remaining unknown parameter is t_{SI}, which is typically calculated as [8]

$$t_{\mathrm{SI}} = 0.5 \min\{d_1, \ldots, d_n\} \qquad (7)$$

for a total of n flows to be scheduled. Notice that, out of the n flows, several can be video flows, audio flows, or other delay-stringent applications. In addition, the factor 0.5 is used to accommodate the jitter constraints demanded by the particular applications [8].

2.2 Architecture of the Deployed Scalable Video Coder

In order to better understand the challenges and limitations associated with deploying the previously described HCCA admission control for video, we consider an MCTF scalable video coder,[2] whose particular architecture is outlined in this section. MCTF was shown to be the most promising technique for state-of-the-art scalable video coding schemes [13]. MCTF is aimed at removing the temporal redundancies of video sequences and, unlike predictive coding schemes, it does not employ a temporal recursive structure. Instead, it uses an open-loop, pyramidal decomposition to remove both long-term and short-term temporal dependencies in an efficient manner. During MCTF, the original video frames are filtered temporally in the direction of motion, prior to performing the spatial transformation and coding. Video frames are filtered into L (low-frequency or average) and H (high-frequency of difference) frames, as shown in Fig. 2. This filtering is performed in the direction of motion as follows. For three consecutive video frames—A (previous frame), B (current frame) and C (next frame)—an instantiation of MCTF can be written using the lifting formulation [13], which, for each pixel (x, y) of a $X \times Y$ video frame, is written as

$$\begin{aligned} H[x,y] = \frac{1}{\sqrt{2}} (B[x,y] & \\ - p_{\mathcal{F}}[x + v_x^{\mathcal{F}}, y + v_y^{\mathcal{F}}] \cdot A[x + v_x^{\mathcal{F}}, y + v_y^{\mathcal{F}}] & \\ - p_{\mathcal{B}}[x + v_x^{\mathcal{B}}, y + v_y^{\mathcal{B}}] \cdot C[x + v_x^{\mathcal{B}}, y + v_y^{\mathcal{B}}]), & \end{aligned} \qquad (8)$$

$$L[x + v_x, y + v_y] = \sqrt{2} A[x,y] + u_{\mathcal{B}}[x,y] \cdot H[x,y], \qquad (9)$$

where $(v_x^{\mathcal{F}}, v_y^{\mathcal{F}})$ and $(v_x^{\mathcal{B}}, v_y^{\mathcal{B}})$ are the forward and backward motion vectors associated with pixel (x, y) of the current frame (found by bidirectional motion estimation) and $p_{\mathcal{F}}[\cdot,\cdot]$, $p_{\mathcal{B}}[\cdot,\cdot]$, and $u_{\mathcal{B}}[\cdot,\cdot]$ are pixel weights chosen by an optimization mechanism that normalizes the information during the bidirectional (forward ($\mathcal{F}$) and backward ($\mathcal{B}$)) motion-compensated prediction of (8) and the backward ($\mathcal{B}$) inver-

2. However, it is important to emphasize that the cross-layer algorithms proposed in this paper can be deployed with any video coding scheme. The essential part of the proposed enhanced admission control mechanism is the determination of the frame dependencies of the deployed video coder (and, hence, the frame/packets delays and traffic characteristics), which can be established based on the encoding parameters and modeled by direct acyclic graphs [18].

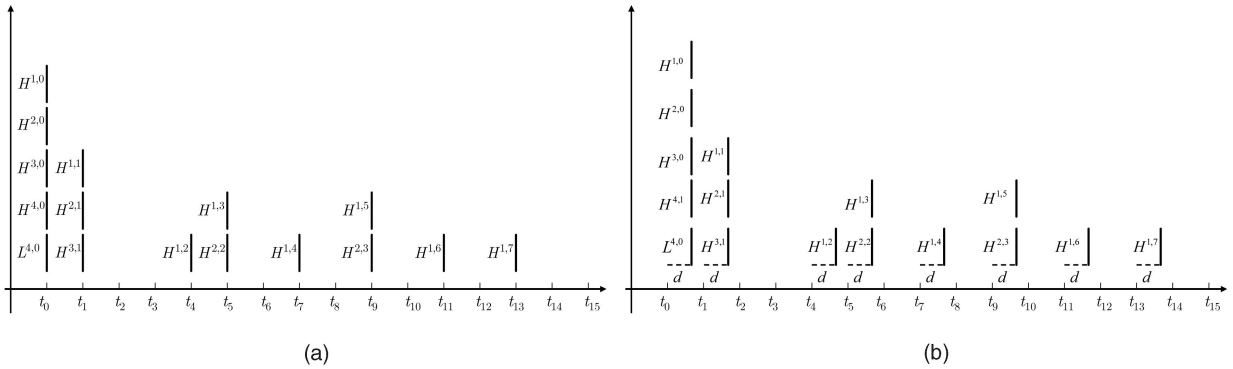

Fig. 3. (a) Playback deadlines without additional delays. (b) Playback deadlines with additional delay of d seconds.

sion of the motion information during the update step of (9). A variety of algorithms can be used to estimate these factors for various temporal decompositions in order for the filtering process to approximate an orthonormal temporal transform [13]. The process is applied initially in a group of pictures (GOP) and also to all the subsequently produced L frames, thereby forming a total of D temporal levels, as shown in Fig. 2.

We use the notation $H^{t,k}$ to indicate the $k\text{th}$ H frame of temporal level t, where $1 \leq t \leq D$ and $0 \leq k \leq 2^{D-t}$. Equivalently, the notation $L^{D,0}$ is used to indicate the remaining L frame at the last level after the completion of the temporal decomposition in the GOP.

All the produced L and H and frames are subsequently decomposed spatially by performing the discrete wavelet transform (DWT) in a number of decomposition levels, as seen in the example of Fig. 2, and embedded coding is applied to the quantized wavelet-coefficient values.

It is important to notice that the example of MCTF formulated by (8) and (9), and illustrated in Fig. 2, represents only one instantiation out of the many possible [13]. Previous research has investigated generalized forms of MCTF that use many reference frames from the past and the future for the temporal filtering [13]. In the remainder of this paper, we present the proposed algorithms based on the typical decomposition of (8), (9), shown in Fig. 2. However, more complicated MCTF schemes using longer temporal filters can easily be incorporated in our framework. Finally, it is important to mention that, although the utilized MCTF codec is based on wavelet compression (i.e., JPEG-2000 alike coding), alternative scalable coding techniques not relying on the DWT can be applied for the embedded coding of the L and H frames of the MCTF decomposition [19].

In typical MCTF-based video compression, the rate allocation for scalable bitstream extraction is performed with a maximum granularity of one GOP. This creates variable-bitrate (VBR) characteristics for the compressed video content across the frames of each GOP. In addition, each decoded frame of every GOP has its own playback deadline determined by the frame rate. Notice that, based on the MCTF structure, the decoding frame rate itself can be dyadically reduced by skipping the frames of the finer temporal levels [13]. Frame-rate scalability will be useful in our cross-layer adaptation strategy that maximizes the number of admitted stations in the wireless network.

From (8) and (9) and by inverting the temporal relations depicted in Fig. 2, we can see that frames of $L^{3,0}$, $H^{3,0}$, $H^{2,0}$, and $H^{1,0}$ should be available in the receiving buffer before reconstructing and playing back the original frame $A^{0,0}$ at the decoder. This implies that these frames in the temporal decomposition have the same deadline, called playback deadline, before which all of their video packets should arrive in the receiving buffer. Extending this calculation to four temporal levels and all the frames of the temporal decomposition of a GOP, Fig. 3 shows the playback deadlines without any extra delay and with an extra delay of d seconds. When the media server schedules and transmits a packet, it should consider the playback deadline of its corresponding video frame.

Examples of the video traffic curves of two four-GOP CIF sequences ("Foreman" and "Stefan") compressed at 2,048 kbps are shown in Fig. 4. The results were generated by an instantiation of a wavelet-based scalable video codec proposed recently [20]. The salient feature of the produced video traffic is that the traffic patterns within one GOP are similar and periodically repeated. This feature motivates us to justify our proposed scheduling algorithms within one GOP and periodically apply the scheduling across multiple GOPs.

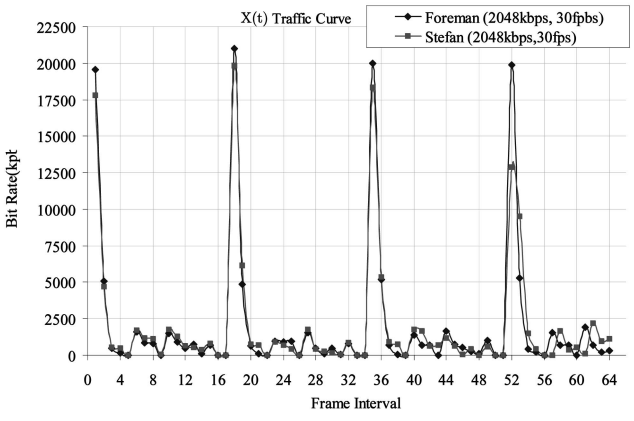

Fig. 4. Video traffic curves for 64 frames of two typical video sequences.

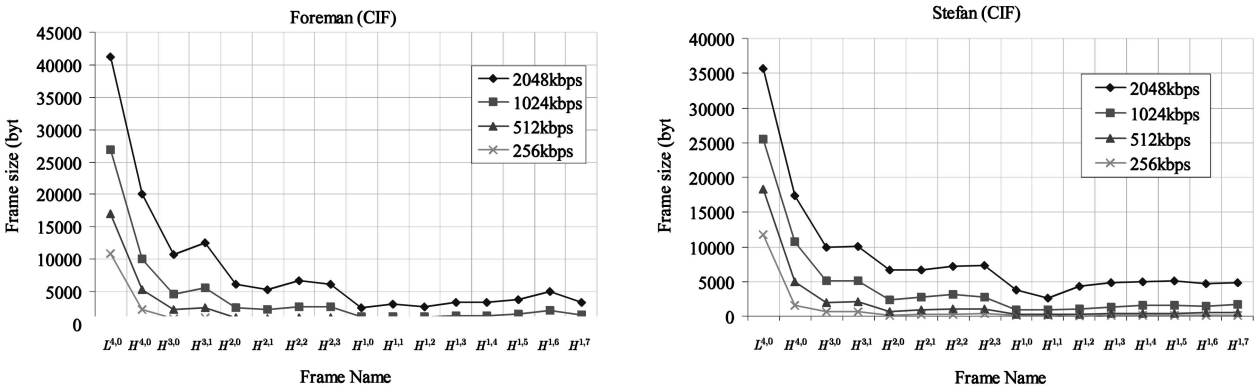

Fig. 5. Compressed frame size within one GOP of two typical CIF test sequences, extracted at different bit rates from one scalable bitstream.

Although the video traffic can be set at a constant bit rate (CBR) in a one-GOP interval, Fig. 4 demonstrates that the traffic within one GOP exhibits significant variation. These varying and bursty features result from the fact that the sizes of compressed frames vary significantly in time, as shown in the corresponding results of Fig. 5. Notice that the obtained results follow similar VBR traffic patterns of standardized video codecs; however, the hierarchical decomposition performed by the temporal filtering of Fig. 2 creates more bursty patterns in the video traffic across time than the conventional I-B-P type structure used in closed-loop coding. Thus, the problem of scalable wavelet video transmission over wireless networks under playback deadlines for each frame appears to be more challenging than conventional MPEG coding due to the complicated spatiotemporal dependencies between packets that impact the distortion and delay characteristics.

3 Enhanced Video Streaming over IEEE 802.11e—The Subflow Concept

The implementation of the simple scheduler explained in Section 2.1 is easy, but it can be quite inefficient for real-time video streaming applications. This is because video traffic varies over time and consists of frames/packets with considerably varying sizes and different delay constraints. Conventionally, the video is considered as a single stream and the TSPEC parameters are set so that the MAC of IEEE 802.11e HCCA would do the admission control and scheduling as outlined previously. To improve the overall system utilization (number of admission stations) as well as the performance of the admitted stations, we introduce the subflow concept in which a video flow (bitstream) is divided into several subflows based on their delay constraints as well as based on the relative priority in terms of the overall distortion of the decoded video. The application layer enables each subflow of the video to interface with the MAC as a separate flow. Each subflow has a different priority (determined by its distortion impact) and delay constraint. A subflow has its own TSPEC parameters and is admitted independently by ACU.

Our aim is to use the subflow mechanism to provide a joint application-MAC optimization that maximizes the number of admitted wireless stations while optimizing the video quality for each admitted station. Given the channel conditions, the ACU and the cooperating wireless stations have to determine for each application the number of subflows the application-layer can transmit, as well as their protection strategies (e.g., MAC retry limits per subflow), while maximizing the number of wireless stations in the network. In this section, we show how the global-flow traffic can be partitioned into subflows, which are then shaped by multiple token leaky buckets to determine their individual QoS token rates.

As shown in Fig. 6, frames with the same playback deadline are grouped into the same subflow and there are eight subflows within one GOP of 16 frames. The number of subflows depends on the temporal decomposition levels and the number of reference frames used for motion estimation. If we denote the number of subflows of one GOP as N_s, we have

$$N_s = 2^{D-1}, \tag{10}$$

where D is the total number of temporal decomposition levels. Each subflow is regarded as an independent traffic flow passing through a twin leaky bucket to get its own QoS guaranteed bandwidth g_i as expressed by (1), with i indicating the subflow number, $1 \le i \le N_s$, and P_i, ρ_i, σ_i, and d_i the corresponding peak data rate, mean data rate, maximum burst size, and delay constraint of subflow i, respectively. As a result, each subflow has its own TSPEC parameters and, thus, there are multiple sets of TSPEC parameters corresponding to one global video flow. A QSTA uses these multiple sets of TSPEC parameters to negotiate with the ACU.

The system performance gain that can be achieved by our proposed subflow concept can be theoretically quantified if we introduce the average transmission-opportunity duration, $\overline{t_{\mathrm{TXOP}}}$:

$$\overline{t_{\mathrm{TXOP}}} = \frac{1}{N_s}\sum_{i=1}^{N_s} t_{\mathrm{TXOP},i}. \tag{11}$$

For a global video flow i, $\overline{t_{\mathrm{TXOP}}}$ is equal to the definition given in (5). Following the admission control expressed in (6), if we assume only N_{QSTA} video flows for the HCCA transmission intervals, i.e., $t_{\mathrm{TXOP,other}} = 0$, by

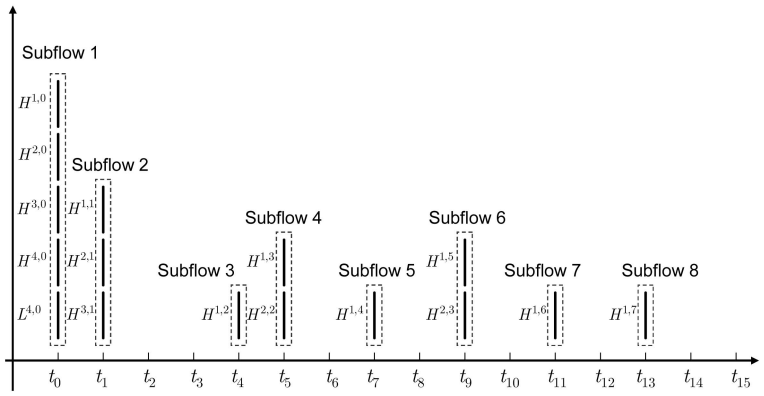

Fig. 6. Example of subflow formation.

replacing the average transmission-opportunity duration for each station by (11), we get the maximum number of admitted QSTA carrying video data as

$$N_{\mathrm{QSTA}} = \left\lfloor \frac{t_{\mathrm{SI}}(1 - T_{CP} \cdot T^{-1})}{\overline{t_{\mathrm{TXOP}}}} \right\rfloor. \tag{12}$$

In the following section, we determine the optimal allocated $t_{\mathrm{TXOP},i}$ for each subflow i (under predetermined delay constraints) such that the number of admitted stations (N_{QSTA}) is maximized.

4 Optimization of the Number of Admitted Stations under Delay Constraints

We introduce a mechanism that maximizes the number of simultaneously admitted wireless stations by optimizing the allocated transmission opportunity duration for each subflow. Our solution is based on linear programming. Given the allotted TXOP per subflow, Section 4.2 determines the maximum number of MSDU retransmissions in order to optimize the video quality under the presence of network errors and Section 4.3 presents an algorithm for dynamic adaptation of MSDU retransmissions based on this derivation. Finally, Section 4.4 explains how link adaptation can be incorporated in the proposed framework to improve the overall performance for different channel conditions.

4.1 Optimization of the Number of Admitted Stations under Delay Constraints

Although the use of subflows may increase the number of admitted stations in the HCCA traffic, if additional delay is permitted in the transmission of each subflow traffic, an optimal scheduling algorithm can yield further improvements. A visual example of such a case for one GOP of video data can be seen in Fig. 7, where each increase in the transmission duration of each subflow i, $d_{s,i}$, provides the opportunity for traffic smoothing. In order to accommodate delay requirements, we have $\max\{d_{s,1}, \ldots d_{s,2^{D-1}}\} \leq d_{\max}$ with $d_{\max}$ set by the chosen streaming scenario.

Each increase in the transmission duration of subflow i is reflected by a change in $t_{\mathrm{TXOP},i}$. Our optimization goal of maximizing N_{QSTA} given by (12) can be equivalently stated as minimizing $\overline{t_{\mathrm{TXOP}}}$ since the other parameters in (12) are unaffected by changes in the transmission duration. As a result, if we limit the optimization to the duration of one GOP (since the video-flow traffic is periodic for each GOP, as seen in Fig. 5), by combining (2)-(5) the minimization problem now becomes

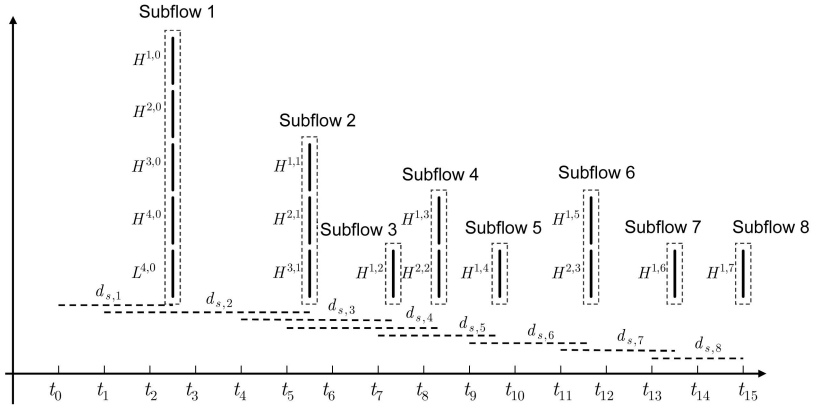

Fig. 7. Subflows with different transmission durations due to additional delay permitted. Each $d_{s,i}$, $1 \leq i \leq 2^{D-1}$ corresponds to additional transmission time for subflow i. For all cases, we assume that an upper bound for the additional delay is set, denoted by $d_{\max}$, and we have $\max\{d_{s,1}, \ldots d_{s,2^{D-1}}\} \leq d_{\max}$.

$$\textbf{Primary problem: } \{t^*_{s,1}, \ldots, t^*_{s,2^{D-1}}\} = \arg\min \sum_{i=1}^{2^{D-1}} g_{s,i} \quad (13)$$

such that $\forall i : 1 \leq i \leq 2^{D-1}$, we have

$$\sum_{j=1}^{i} t^*_{s,j} \leq \sum_{j=1}^{i} (t^*_{s,j} + d_{\max}). \quad (14)$$

In (13) and (14), $\{t^*_{s,1}, \ldots, t^*_{s,2^{D-1}}\}$ are the optimal transmission durations corresponding to subflows $1 \leq i \leq 2^{D-1}$, $g_{s,i}$ is the effective bandwidth defined by $g_{s,i} = b_{s,i}/t_{s,i}$, with $b_{s,i}$ the size (in bits) of subflow i and $t_{s,i}$ the original transmission duration of subflow i. Notice that this definition of $g_{s,i}$ corresponds to the generic definition of (1) if we assume that $P_i = \rho_i$, i.e., under the assumption of CBR transmission for the transmission duration of subflow i. In order to facilitate the optimization process, the optimization problem stated in (13), (14) can be expressed in a dual form [21] as

$$\textbf{Dual problem: } \{b^*_{s,1}, \ldots, b^*_{s,2^{D-1}}\} = \arg\min \sum_{i=1}^{2^{D-1}} \frac{b_{s,i}}{t^{\max}_{s,i}} \quad (15)$$

such that $\forall i : 1 \leq i \leq 2^{D-1}$ we have

$$\sum_{j=1}^{i} b^*_{s,j} \geq \sum_{j=1}^{i} b_{s,j}. \quad (16)$$

In (15) and (16), $\{b^*_{s,1}, \ldots, b^*_{s,2^{D-1}}\}$ are the optimal subflow sizes, and $t^{\max}_{s,1} = t_{s,1} + d_{\max}$, $t^{\max}_{s,i} = t_{s,i}$ for $2 \leq i \leq 2^{D-1}$ represent the maximum permissible transmission durations for each subflow.

Under CBR transmission for each subflow, the Primary and Dual problems stated above provide the same solution [21]. For example, by deriving the optimal subflow sizes, we can establish the optimal transmission duration corresponding to the size of each subflow as

$$t^*_{s,i} = t^{\max}_{s,i} \frac{b_{s,i}}{b^*_{s,i}}. \quad (17)$$

Nevertheless, one difference of practical significance is that the Dual problem facilitates the application of linear programming techniques, namely, the simplex minimization, for the establishment of the optimal solution. This ensures optimality with low complexity, as the algorithm converges in a number of steps proportional to the total number of subflows, N_s. The simplex optimization scans through all the vertices of the N_s-dimensional simplex in order to establish the point corresponding to the minimum of (15) [21]. It is important to mention that, in order to formulate a bounded problem for this purpose, we need to impose an upper bound to the maximum number of bits transmitted in the time interval corresponding to one GOP. Hence, we introduced an additional constraint to the problem given by

$$\sum_{j=1}^{2^{D-1}-1} b^*_{s,j} \leq \sum_{j=1}^{2^{D-1}-1} R_j \cdot t^{\max}_{s,j}, \quad (18)$$

which corresponds to the physical constraint that the maximum number of bits transmitted during the duration of one GOP together with the packetization overhead introduced at the various layers cannot exceed the mean amount of bits transmitted by the physical-layer during this time.

Finally, although the optimization problem is defined and solved for the duration of one GOP, if access to additional subflows from consecutive GOPs is possible (e.g., in the case of offline encoding), they can be included in the optimization problem of (15), (16) following the same rationale. Experimental results with real video data utilizing the proposed optimization approach are presented in Section 5.

4.2 Packet Scheduling and Retransmissions under the Proposed Admission Control

For the admitted subflows of a QSTA, the application and MAC layers can cooperate to improve the multimedia quality by adapting the retry limit. The previous studies on retry-limit adaptation [1], [3] study cross-layer strategies for 802.11 WLANs that are not HCCA enabled and, also, they do not explicitly consider the delay bound set by the application for the various packets/flows. Here, the goal of packet scheduling and prioritized MAC retransmissions is to minimize the playback distortion for a video streaming session over an 802.11a/e HCCA WLAN, under delay constraints.

Due to limits imposed by link-adaptation to different physical-layer rates [22] as well as delay constraints, the retransmission bound for the earlier-transmitted packets can be higher than the maximum retransmissions allowed for the remainder set of packets. Hence, under a scheme allowing for unequal video-packet retransmissions, a higher probability for correct reception can be provided to the first subsets of video packets. This motivates packet prioritization at the application layer depending on the video-data significance (incurred distortion due to losing the packet).

The optimal transmission duration for each subflow was already established in the previous subsection by linear programming. In this section, given the set of video packets for each subflow as well as the transmission duration, we establish which subset should be transmitted as well as the maximum permissible number of video-packet retransmissions in case of errors.

Modeling approaches have been recently proposed for the establishment of the substream significance in MCTF-based video compression [23], [24]. Most of these models use dynamic computation of the expected distortion using signal statistics or precomputed distortion metadata in conjunction with models for the error propagation across the MCTF decoding structure [23]. Although such solutions result in a model-optimized scheduling with the potential for high accuracy, they can also incur a high computational burden for online processing of many streams. In addition, if we define the number of retransmissions based only on the video-packet significance, we will not be able to take advantage of the fact that the MAC layer can provide real-time feedback concerning the correct reception of each individual MSDU frame. As a result, we have opted for the use of very low complexity (yet accurate) heuristics in the video-packet prioritization strategy:

- Packets which belong to video frames with different timestamps are ordered in timestamp order, which is derived from compressed video-frame dependency and playback deadline as discussed in the subflow creation in Section 3.
- Packets which belong to video frames with a same timestamp but different temporal levels are ordered in temporal level sequence. In particular, packets of coarser temporal levels are ordered before ones of finer temporal levels (e.g., packets belong to $H^{3,1}$ are ordered before ones belonging to $H^{2,1}$ in the example of Fig. 3). Note that the motion vector packets are ordered before the texture packets.
- Packets which belong to same video frames are ordered based on the spatial decomposition level they belong to, that is, packets in coarser spatial resolution levels ordered before ones in finer spatial resolution levels.
- Within a spatial decomposition level, packets in the low-frequency subband are ordered prior to the high-frequency subbands.
- Within a given spatial subband, packets are ordered in raster order following the JPEG-2000 convention [26].
- Packets corresponding to the same spatiotemporal subband are ordered based on the utilized-codec prioritization. For bitplane coders, this simply requires that video-packets containing information for the most-significant bitplanes are ordered prior to the ones containing the least-significant bitplanes.

Once the ordering is complete, the video packets are placed in MSDU frames and these are passed to the MAC layer in the specified order. Although these rules are simply based on the compression architecture and the proposed subflow scheduling, the layering principle of fully-scalable MCTF-based video coding ensures the optimality of such a scheduling approach. In addition, recent theoretical studies [23], [24] have shown that the expected distortion-reduction obtained by decoding each video packet is proportional to the temporal and spatial level that the packet belongs to, according to the ordering expressed in the above rules. We remark that, similar to the previous section, the scheduling algorithm operates independently for each GOP, although extensions to multiple GOPs can be envisaged following similar principles.

In this work, we assume that the transmission channel is an independent, identically distributed error channel. Thus, the channel causes errors independently in each MSDU frame and the error probability is the same for all MSDU frames with the same length at all times. Let $p_b(m)$ be the bit error probability in physical-layer mode m. Then, the error probability of an MSDU frame of size L_i (belonging to subflow i) in physical-layer mode m is a function of bit error probability $p_b(m)$ and is defined as [4]

$$p_e(m, L_i) = 1 - [1 - p_b(m)]^{L_i}. \tag{19}$$

Let $N_{\text{retry}}^{\max}(j)$ be the maximum number of retries of MSDU frame j belonging to subflow i. Notice that the value of $N_{\text{retry}}^{\max}(j)$ depends on the position of the MSDU frame in the transmission queue (derived based on the criteria outlined before), as well as on the available transmission duration for the current subflow. The probability of unsuccessful transmission after $N_{\text{retry}}^{\max}(j)$ retransmissions is

$$P_e(m, L_i, N_{\text{retry}}^{\max}(j)) = [p_e(m, L_i)]^{N_{\text{retry}}^{\max}(j)+1}. \tag{20}$$

In addition, based on $N_{\text{retry}}^{\max}(j)$, we can find the average number of transmissions for the jth MSDU frame until the packet is successfully transmitted or the retransmission limit is reached as in [3]:

$$N_{\text{average}}(j) = \frac{1 - [p_e(m, L)]^{N_{\text{retry}}^{\max}(j)+1}}{1 - p_e(m, L)}. \tag{21}$$

The corresponding average time to transmit the MSDU frame using the guaranteed channel rate g_i' for subflow i is given by

$$T_{\text{average}} = N_{\text{average}}(j)\left(\frac{L_i}{g_i'} + T_{\text{ACK}}\right), \tag{22}$$

where T_{ACK} is the overhead for the transmission of the acknowledgment frame. Assuming that the maximum time before the MSDU frame expire is $T_{\max}$, we have

$$T_{\text{average}} \le T_{\max}. \tag{23}$$

Due to CBR transmission for the duration of each subflow, the MSDU frames are evenly distributed with an interval α_i (i.e., MSDU arrival interval for subflow i). Assuming that the transmission duration for subflow i is $t_{s,i}^*$ (estimated by the optimization of Section 4.1), for the jth MSDU frame of that subflow, we have

$$T_{\max} = t_{s,i}^* - \alpha_i \sum_{k=1}^{j-1} N_{\text{retry}}^{\text{actual}}(k), \tag{24}$$

where $N_{\text{retry}}^{\text{actual}}(k)$, $0 \le N_{\text{retry}}^{\text{actual}}(k) \le N_{\text{retry}}^{\max}$, is the actual number of retries for each MSDU frame k (that precedes MSDU frame j) until an acknowledgment has been received, or the maximum number of retries has been performed. Notice that $N_{\text{retry}}^{\text{actual}}(k)$ can be determined dynamically based on feedback from the MAC layer. The last equation can be used in conjunction with (20) and (23) to establish the bound for the maximum-allowable number of retries for the current MSDU frame j:

$$N_{\text{retry}}^{\max}(j) \le \log_{p_e(m,L_i)} \left[(1 - p_e(m, L_i)) \left(\frac{L_i}{g_i'} + T_{\text{ACK}} \right)^{-1} \left(t_{s,i}^* - \alpha_i \sum_{k=1}^{j-1} N_{\text{retry}}^{\text{actual}}(k) \right) \right] - 1. \tag{25}$$

Notice that the estimated maximum number of retries determined by (25) can be negative, depending on whether we exceeded the available bandwidth for subflow i or not. In such a case, the remaining MSDU frames of the current subflow are simply discarded.

4.3 Proposed Subflow Transmission with Dynamic Adaptation

We outline the steps performed during the actual streaming process for each subflow i in Table 1. Some of the last

TABLE 1
Transmission of MSDU Frames of Each Subflow i

- **Initialization:** Establish $p_b(m)$ based on the utilized physical layer mode. Calculate $p_e(m, L_i)$ from (19).
- **For each MSDU frame j:**

1. Establish $T_{\max}$ based on (24). Calculate $N_{\text{retry}}^{\max}(j)$ based on (21)-(23). Set $\text{current\_retries} = 0$.
2. If $N_{\text{retry}}^{\max}(j) \geq 0$
   - Set $\text{current\_ACK} = \text{FALSE}$; go to Step 3.

   else
   - Discard the current packet as well as the remaining subflow packets with the same deadline.
3. While $\text{current\_ACK} = \text{FALSE}$ AND $\text{current\_retries} \leq N_{\text{retry}}^{\max}$
   - Transmit the current MSDU. Set: $\text{current\_retries} \leftarrow \text{current\_retries}+1$
   - Set current_ACK to TRUE or FALSE depending on MAC-layer feedback.
4. Set $N_{\text{retry}}^{\text{actual}}(j) = \text{current\_retries}$

MSDU frames of each subflow will not be transmitted whenever the channel condition deteriorates since the transmission duration (deadline) determined by the simplex optimization of the previous section does not take into account the retransmissions that will occur based on the algorithm of Table 1. This is checked in Step 2 of the algorithm of Table 1. Nevertheless, the use of a scalable video coding and the prioritization rules for the transmission of the video packets specified before ensure that near-optimal adaptation of the video quality will occur based on the instantaneous channel capacity since the MSDU frames with the most important video data will be transmitted first.

An alternative design can be formulated by a priori calculating the maximum number of retransmissions for each MSDU frame j based on $p_b(m)$ and using (19)-(24) with the setting of $N_{\text{retry}}^{\text{actual}}(k) = N_{\text{retry}}^{\max}(k)$ for every $1 \leq k < j$. Then, the subflow sizes can be readjusted to include the estimated number of retransmissions. This allows for the optimization algorithm of Section 4.1 to derive optimal transmission durations that include the (worst-case) expected number of retransmissions for the subflow's MSDU frames. Overall, the latter case is expected to overprovision bandwidth for each subflow while the previous case can lead to some of the least-significant video packets being dropped, depending on the channel condition.

4.4 QoS Token Rate Adaptation for Link Adaptation under the Proposed Framework

Link adaptation selects one appropriate physical-layer mode based on link conditions in order to improve the system goodput and throughput [22]. IEEE 802.11a supports eight physical-layer rates from 6 Mbit/sec to 54 Mbit/sec. QSTAs may adapt their physical-layer modulation and coding strategies depending on the link conditions [22], [27]. In particular, the physical-layer rate will be lowered dynamically when the link condition of one QSTA gets worse, i.e., when the signal to interference-noise ratio (SINR) drops. The TXOP durations calculated by (5) will not take into account the new rate when the QSTA switches its default physical-layer rate mode and, as a result, the QAP may deny the traffic stream of the QSTA whose physical rate turns out to be lower than the prenegotiated minimum rate.

In order to keep the number of admitted stations fixed and have graceful quality degradation, we can utilize the packet scheduling algorithm of Section 4.2 in order to drop MSDU frames containing less-important video data such that the precalculated TXOP duration can still guarantee the QoS when the physical-layer mode is changed. For this purpose, we need to determine the new effective bandwidth for each subflow i, g_i', under a change in the physical-layer transmission rate. If we assume that the modified rate for the duration of the subflow transmission is R_i', from (5) we have

$$N_i' = \frac{t_{TXOP}, i}{L_i \cdot (R')^{-1} + T_{\text{overhead},i}}. \tag{26}$$

Then, from (2) we get

$$g_i' = \frac{N_i \cdot L_i}{t_{\text{SI}}} \tag{27}$$

and, since CBR transmission occurs for the duration of the subflow transmission, $t_{s,i}^*$, we can calculate the modified subflow size, $b_{s,i}'$, using (26), (27) as

$$\rho_i' = \frac{b_{s,i}'}{t_{s,i}^*} = g_i' \Rightarrow b_{s,i}' = \frac{t_{s,i}^* \cdot t_{\text{TXOP},i} \cdot L_i}{\left[L_i \cdot (R')^{-1} + T_{\text{overhead},i}\right] \cdot t_{\text{SI}}}. \tag{28}$$

Notice that, in the cases where the link adaptation may change the physical layer rate more than once during the

TABLE 2
Subflow QoS Token Rates and $t_{\mathrm{TXOP},i}$ with $d_{\max} = 200$ msec for the CIF-Resolution Sequence "Foreman" (2,048 kbps, 30 fps)

Traffic Name	Components	QoS Token Rate ρ_i (kbps)	t_{TXOP} (msec)
Subflow $_1$	$H^{1,0}, H^{2,0}, H^{3,0}, H^{4,0}, L^{4,0}$	10032	13.89
Subflow $_2$	$H^{1,1}, H^{2,1}, H^{3,1}$	2840	3.96
Subflow $_3$	$H^{1,2}$	184	0.26
Subflow $_4$	$H^{1,3}, H^{2,2}$	1064	1.46
Subflow $_5$	$H^{1,4}$	264	0.37
Subflow $_6$	$H^{1,5}, H^{2,3}$	728	1.00
Subflow $_7$	$H^{1,6}$	392	0.54
Subflow $_8$	$H^{1,7}$	440	0.61

$\overline{t_{\mathrm{TXOP}}} = 2.76$ msec, $N_{\mathrm{QSTA}} = 7$

subflow transmission interval $t^*_{s,i}$, R'_i can be calculated based on the weighted sum of the different rates

$$R' = \frac{1}{t^*_{s,i}} \sum_{k=1}^{w} [R_{\mathrm{phy}}(k) \cdot t_{\mathrm{phy}}(k)], \tag{29}$$

where $R_{\mathrm{phy}}(k)$, $t_{\mathrm{phy}}(k)$ represent the rate and duration (respectively) corresponding to the kth link adaptation during time interval $t^*_{s,i}$ (out of w total adaptations).

The modified subflow size estimated by (28) may be used to restrict the number of video packets of each subflow: Depending on the adaptive retransmission scheme of Table 1, once the amount of video data sent via MSDU frames reaches $b'_{s,i}$, the remaining packets in the prioritized transmission queue are discarded. Hence, similar to the case of Section 4.2, the prioritization mechanism ensures that the most significant packets receive the highest priority under link adaptation at the physical layer. An interesting extension of the link adaptation algorithm would be to optimize the chosen MSDU frame length depending on the chosen physical layer rate [27]. This should be done having the application-layer packetization restrictions in mind in order not to affect the decoding dependencies (see our work in [28] for more details and proposed algorithms).

5 Experimental Results

The results of this section have been generated with the settings $T = 100$ msec, $T_{\mathrm{CP}} = 60$ msec, and $t_{\mathrm{SI}} = 50$ msec. First, we examine the importance of the (nonoptimized) subflow concept versus the conventional global flow scheduling. The experiment of Table 2 used a typical CIF video sequence—"Foreman"—encoded at 30 frames per second (fps), although similar results have been obtained with a variety of video content. The token rates reported in Table 2 were calculated based on a simulation with a twin leaky bucket traffic smoothing system and the delay deadline was equally extended for all subflows, such that $d_{\max} = 200$ msec. For the case of subflow scheduling, we have $\overline{t_{\mathrm{TXOP}}} = 2.76$ msec and, from (12), $N_{\mathrm{QSTA}} = 7$. Similarly, for the global flow case, we get $\overline{t_{\mathrm{TXOP}}} = 13.89$ msec and $N_{\mathrm{QSTA}} = 1$. The number of admitted stations can be increased if the optimization framework of Section 4.1 is used. This is shown by the results of Table 3, where the number of stations in the subflow case is increased to $N_{\mathrm{QSTA}} = 10$. In addition, based on the priorities shown in Table 3, we can increase the number of admitted stations if the least-significant subflows are discarded. This is illustrated in Fig. 8, where the number of admitted stations is plotted against the number of utilized subflows.

TABLE 3
Subflow QoS Token Rates and $t_{\mathrm{TXOP},i}$ with $d_{\max} = 200$ msec for the CIF-Resolution Sequence "Foreman" (2,048 kbps, 30 fps) with the Optimization Framework of Section 4.1

Traffic Name	Components	QoS Token Rate ρ_i (kbps)	Priority	t_{TXOP} (msec)
Subflow $_1$	$H^{1,0}, H^{2,0}, H^{3,0}, H^{4,0}, L^{4,0}$	4664	4	6.46
Subflow $_2$	$H^{1,1}, H^{2,1}, H^{3,1}$	2768	3	3.82
Subflow $_3$	$H^{1,2}$	216	1	0.31
Subflow $_4$	$H^{1,3}, H^{2,2}$	1376	2	1.90
Subflow $_5$	$H^{1,4}$	344	1	0.46
Subflow $_6$	$H^{1,5}, H^{2,3}$	880	2	1.24
Subflow $_7$	$H^{1,6}$	392	1	0.54
Subflow $_8$	$H^{1,7}$	440	1	0.61

$\overline{t_{\mathrm{TXOP}}} = 1.92$ msec, $N_{\mathrm{QSTA}} = 10$

The "Priority" indicates the importance (4: highest, 1:lowest) of each subflow in terms of incurred distortion at the receiver.

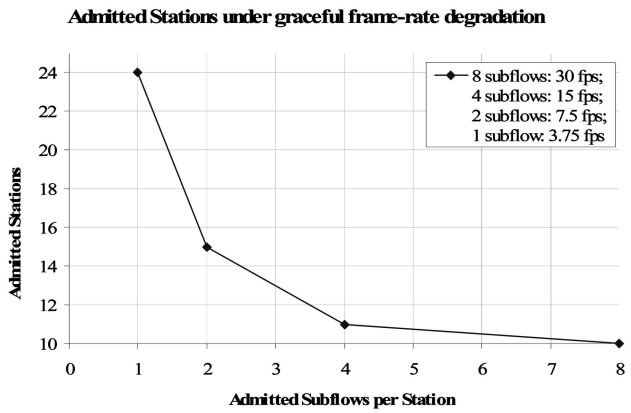

Fig. 8. A reduction of the number of admitted subflows results in dyadically reduced frame rate. However, the number of admitted stations increases. The utilized video sequences were encoded at 2,048 kbps.

The figure demonstrates that, under a progressive decrease in frame-rate, resulting from the removal (drop) of the least-significant subflows (with the significance indicated in Table 3) the number of admitted stations can be further increased. In a collaborative framework, multiple stations may opt to decrease the video frame-rate in order to allow for additional stations (or additional video flows) to utilize the wireless medium under HCCA. Also, the desired number of admitted subflows as well as how these subflows are prioritized at the application layer can be determined based on the channel resources, specific video application, and user preferences. For instance, different spatiotemporal resolutions (and corresponding subflows) should be selected for the best perceptual video quality for different channel conditions. This flexibility can be easily provided using the subflow concept.

In summary, we observe that a higher number of stations can be admitted given the same channel condition if the subflow case is used, as compared to the global flow case. Notice that the same video bitstreams are transmitted in both cases and no losses are incurred due to the use of subflows.

In order to evaluate the advantages of the proposed subflow concept under more realistic conditions, we used the ns-2 simulator package of Ansel et al. [9]. HCCA was used to stream a number of video flows generated by wavelet-based scalable video coding [20] and EDCA was used for the remaining traffic within the contention period. Our results are presented in Table 4 and Table 5 for different delay limits. The optimization algorithm of Section 4.1. was used in the case of video transmission based on subflows in order to determine the application-layer optimal transmission duration for each subflow. The results indicate that a significantly-higher number of stations could be admitted in the case of the optimized subflow-based transmission. Notice that both cases are compliant to the IEEE 802.11e definition of the standard and both transmit the same amount of video data. Moreover, even though the results obtained with the ns-2 simulation do not numerically correspond to the theoretical calculation due to overheads (i.e., Table 3 versus Table 4), the utilization of subflows offers a significant advantage.

TABLE 4
Measured TXOPs and the Total Number of Admitted Stations Based on ns-2 HCCA Simulations, with $d_{\max} = 200$ msec

	TXOP (ms)	Stations admitted
Global Flow	9.33	2
Subflow 1	9.33	5
Subflow 2	6.57	
Subflow 3	0.96	
Subflow 4	3.91	
Subflow 5	1.27	
Subflow 6	2.32	
Subflow 7	1.19	
Subflow 8	1.22	

In order to investigate the effect of retransmissions in the proposed system, we have enabled a uniform bit-error loss model in our simulations. The proposed optimized retransmission policy was compared against a conventional retransmission policy which does not set the number of retransmissions based on the expected loss rate [1], [3]. In addition, two ad hoc policies were simulated, one with an optimized packet scheduling described in Section 4.2 and one with an ad hoc scheduling that does not follow the proposed packet ordering and instead schedules the video packets in the way they are produced by the video encoder. The results are presented in Table 6. It can be seen that, across the various error rates, the proposed retransmission policy offers an average benefit of approximately 1.8 dB in PSNR versus the conventional retransmission policy, while the gains may be as high as 3.0 dB in some cases. Moreover, the scheduling policy presented in Section 4.2 provides an average benefit of 0.8 dB versus the ad-hoc scheduling.

Finally, Table 7 demonstrates the effect in video quality under the token rate adaptation for dynamic link adaptation, as discussed in Section 4.4. More details on our implementation of the link adaptation mechanism can be found in [4], [22]. The results demonstrate that, under varying SINR, the proposed adaptation mechanism can effectively reduce the transmitted subflow size by increasing the number of discarded (not transmitted) packets from each subflow. This results in a graceful degradation in the video quality. Similar results were

TABLE 5
Measured TXOPs and the Total Number of Admitted Stations Based on ns-2 HCCA Simulations, with $d_s = 400$ msec

	TXOP (ms)	Stations admitted
Global Flow	5.04	3
Subflow 1	5.04	8
Subflow 2	4.28	
Subflow 3	0.89	
Subflow 4	3.13	
Subflow 5	1.20	
Subflow 6	2.00	
Subflow 7	1.16	
Subflow 8	1.11	

TABLE 6
Average PSNR (Y Channel—Five Runs per Method/Loss Rate) for Two Typical CIF Sequences under Various Packet Loss Percentages

Foreman Packet Loss Rate (%)	Opt. Retransmission Opt., Scheduling (dB)	Fixed Retransmission, Opt. Scheduling (dB)	Fixed Retransmission, Ad-hoc Scheduling (dB)
0.00	37.93	37.93	37.93
10.00	36.11	34.96	34.59
20.00	33.73	32.25	32.07
30.00	30.41	28.47	26.40
40.00	25.27	23.18	22.79
Coastguard Packet Loss Rate (%)	Opt. Retransmission, Opt. Scheduling (dB)	Fixed Retransmission, Opt. Scheduling (dB)	Fixed Retransmission, Ad-hoc Scheduling (dB)
0.00	36.03	36.03	36.03
10.00	34.79	33.45	33.38
20.00	32.81	29.66	29.14
30.00	28.81	25.78	23.95
40.00	23.45	23.04	22.01

TABLE 7
Number of Packet Drops per Subflow and the Corresponding Average PSNR (Y Channel) Using Cross-Layer Optimization (Including Dynamic Link Adaptation)

	Number of packets discarded per Subflow for the *entire* video sequence (2560 packets)								
SINR range (dB)	Sub-flow 1	Sub-flow 2	Sub-flow 3	Sub-flow 4	Sub-flow 5	Sub-flow 6	Sub-flow 7	Sub-flow 8	PSNR (dB)
24-28	29	42	35	80	54	98	58	55	36.78
20-24	56	87	47	128	64	128	64	59	35.93

A Nominal MSDU size of 1,000 bytes was assumed and five runs for each of the two cases of SINR range were performed with the sequence "Foreman."

observed for a variety of sequences and transmission scenarios.

6 Conclusions

Improving the performance of delay-constrained video streaming over wireless networks is an important issue for various multimedia-related applications, as well as for the efficient overall utilization of the wireless medium. In this context, we have investigated cross-layer optimization strategies for HCCA-based video streaming. The proposed methods maximize the number of admitted stations by creating multiple subflows from one global video flow, each with its own traffic specification. In order to achieve an optimal scheduling policy, a low complex linear-programming solution is proposed, which effectively allocates the optimal transmission opportunity to each generated subflow in order to maximize the utilization of the wireless medium under the contention-free period. Besides the proposed method for optimization of the video traffic under HCCA transmission, the retry limit for each packet is adaptively modified in order to accommodate transmission under random packet losses. The proposed algorithm can be easily coupled with link adaptation mechanisms in order to provide efficient adaptation to dynamic network behavior. Future work should determine the impact of simultaneously scheduling different flows (such as audio, video, and data having various bit rate and delay requirements) on the admission control mechanism.

Acknowledgments

The authors would like to thank Dr. Sai Shankar N. of Qualcomm for technical discussions related to the topic of this paper. Also, the authors would like to acknowledge the kind support of US National Science Foundation Career CCF-0541867 and grants from Intel IT Research and Samsung.

References

[1] D. Majumdar, G. Sachs, I.V. Kozintsev, and K. Ramchandran, "Multicast and Unicast Real-Time Video Streaming over Wireless LANs," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 12, no. 6, pp. 524-534, June 2002.

[2] Y. Shan and A. Zakhor, "Cross Layer Techniques for Adaptive Video Streaming over Wireless Networks," *Proc. IEEE Int'l Conf. Multimedia and Expo,* vol. 1, pp. 277-280, Sept. 2002.

[3] Q. Li and M. van der Schaar, "Providing Adaptive QoS to Layered Video over Wireless Local Area Networks through Real-Time Retry Limit Adaptation," *IEEE Trans. Multimedia,* vol. 6, no. 2, pp. 278-290, Apr. 2004.

[4] M. van der Schaar, S. Krishnamachari, S. Choi, and X. Xu, "Adaptive Cross-Layer Protection Strategies for Robust Scalable Video Transmission over 802.11 WLANs," *IEEE J. Selected Areas in Comm.,* vol. 21, no. 10, Dec. 2003.

[5] M. van der Schaar and S. Shankar, "Cross-Layer Wireless Multimedia Transmission: Challenges, Principles and New Paradigms," *IEEE Wireless Comm. Magazine,* vol. 12, no. 4, Aug. 2005.

[6] IEEE Standard 802.11-1999, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999.
[7] IEEE 802.11e/D5.0, Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS), draft supplement, June 2003.
[8] A. Grilo, M. Macedo, and M. Nunes, "A Scheduling Algorithm for QoS Support in IEEE 802.11E Networks," *IEEE Wireless Comm. Magazine,* vol. 10, no. 3, pp. 36-43, June 2003.
[9] P. Ansel, Q. Ni, and T. Turletti, "An Efficient Scheduling Scheme for IEEE 802.11E," *Proc. IEEE Workshop Modeling and Optimization in Mobile, Ad-Hoc and Wireless Networking,* Mar. 2004.
[10] S. Mangold, S. Choi, G. Hiertz, O. Klein, and B. Walker, "Analysis of IEEE 802.11E for QoS Support in Wireless LANs," *IEEE Wireless Comm. Magazine,* vol. 10, no. 6, pp. 40-50, Dec. 2003.
[11] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 13, no. 7, pp. 560-576, July 2003.
[12] M. Karcewicz and R. Kurceren, "The SP- and SI-Frames Design for H.264/AVC," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 13, no. 7, pp. 637-644, July 2003.
[13] J.-R. Ohm, "Advances in Scalable Video Coding," *Proc. IEEE,* vol. 93, pp. 42-56, 2005.
[14] P. Garg, R. Doshi, R. Greene, M. Baker, M. Malek, and X. Cheng, "Using IEEE 802.11E MAC for QoS over Wireless," *Proc. IEEE Int'l Conf. Performance Computing and Comm.,* vol. 1, pp. 537-542, Apr. 2003.
[15] B.V. Patel and C.C. Bisdikian, "End-Station Performance over Leaky Bucket Traffic Shaping," *IEEE Network Magazine,* vol. 10, no. 5, pp. 40-47, Sept. 1996.
[16] A.R. Reibman and B. Haskell, "Constraints on Variable Bit-Rate Video over ATM Networks," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 2, no. 4, pp. 361-372, Dec. 1992.
[17] S. Shankar, Z. Hu, and M. van der Schaar, "Cross-Layer Optimized Transmission of Wavelet Video over IEEE 802.11a/e WLANs," *Proc. Packet Video Workshop,* Apr. 2004.
[18] P.A. Chou and Z. Miao, "Rate-Distortion Optimized Streaming of Packetized Media," Microsoft Research Technical Report MSR-TR-2001-35, Feb. 2001, also submitted to *IEEE Trans. Multimedia.*
[19] H. Schwarz, T. Hinz, H. Kirchhoffer, D. Marpe, and T. Wiegand, "Technical Description of the HHI Proposal for SVC CE1," ISO/IEC JTC1/SC29/WG11 (MPEG), m11244, Oct. 2004.
[20] D. Taubman, D. Maestroni, R. Mathew, and S. Tubaro, "SVC Core Experiment 1, Description of UNSW Contribution," ISO/IEC JTC1/SC29/WG11 (MPEG), m11441, Oct. 2004.
[21] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge Univ. Press, 2003.
[22] D. Qiao, S. Choi, and K.G. Shin, "Goodput Analysis and Link Adaptation for IEEE 802.11a Wireless LAN," *IEEE Trans. Mobile Computing,* vol. 1, no. 4, pp. 278-292, Oct. 2002.
[23] T. Rusert, K. Hanke, and J.-R. Ohm, "Transition Filtering and Optimized Quantization in Interframe Wavelet Video Coding," *Proc. SPIE Visualization Comm. and Image Processing,* vol. 1, pp. 682-693, 2003.
[24] M. Wang and M. van der Schaar, "Operational Rate-Distortion Modeling for Wavelet Video Coders," *IEEE Trans. Signal Processing,* accepted for publication.
[25] Y. Andreopoulos, R. Kelarapura, M. van der Schaar, and C.-N. Chuah, "Failure-Aware, Open-Loop, Adaptive Video Streaming with Packet-Level Optimized Redundancy," *IEEE Trans. Multimedia,* submitted.
[26] *JPEG2000: Image Compression Fundamentals, Standards and Practice,* D. Taubman and M. Marcellin, eds. Kluwer Academic, 2002.
[27] M. Schwartz, *Telecommunication Networks: Protocols, Modeling, and Analysis.* Addison Wesley, 2005.
[28] D.S. Turaga and M. van der Schaar, "Cross-Layer Aware Packetization Strategies for Optimized Wireless Multimedia Transmission," *Proc. IEEE Int'l Conf. Image Processing,* Oct. 2005.

Mihaela van der Schaar received the PhD degree from the Eindhoven University of Technology, the Netherlands, in 2001. She is now an assistant professor at the University of California Los Angeles in the Electrical Engineering Department. Prior to this, she was a senior researcher at Philips Research in the Netherlands and the United States, where she led a team of researchers working on multimedia compression, networking, communications, and architectures. In 2003, she was also an adjunct assistant professor at Columbia University. From 2003 until 2005, she was an assistant professor in the ECE Department at the University of California, Davis. She has published extensively on multimedia compression, processing, communications, networking, and architectures and holds 22 granted US patents and several more pending. Starting in 1999, she was an active participant in the ISO Motion Picture Expert Group (MPEG) standard, to which she made more than 50 contributions and for which she received three ISO recognition awards. She also chaired for three years the ad hoc group on MPEG-21 Scalable Video Coding and cochaired the MPEG ad hoc group on Multimedia Test-Bed. She was an associate editor of the *IEEE Transactions on Multimedia* and *the SPIE Electronic Imaging Journal* from 2002 to 2005. Currently, she is an associate editor of the *IEEE Transactions on Circuits and Systems for Video Technology* and an associate editor of the *IEEE Signal Processing Letters*. She received the US National Science Foundation Career Award (2004) and IBM Faculty Award (2005). She is a senior member of the IEEE.

Yiannis Andreopoulos received the electrical engineering diploma and the MSc degree in signal and image-processing systems from the University of Patras, Greece, and the PhD degree from the Vrije Universiteit Brussel, Brussels, Belgium. Currently, he is working as a postdoctoral researcher at the University of California Los Angeles (UCLA). His research interests are in the fields of transforms, fast algorithms, video coding, and video transmission through unreliable media, e.g., wireless networks and the Internet. During 2002-2003, he contributed to the ISO/IEC JTC1/SC29/WG11 (MPEG) committee (Scalable Video Coding group). He is a member of the IEEE.

Zhiping Hu received the BE degree from Central South University, Changsha, China, and the MS degree from the University of California at Davis, in 2002 and 2005, respectively, both in electrical engineering. He conducted research in areas of scalable video coding and video streaming during his graduate study and published several international conference papers. In the summer of 2004, he was with Philips Research, Briarcliff Manor, New York, working on robust multimedia streaming and resource management in home networks. He is currently working in an R&D digital platform team at Universal Electronics. He is a member of the IEEE.