Research Article

MAC-Layer QoS Management for Streaming Rate-Adaptive VBR Video over IEEE 802.11e HCCA WLANs

Jianfei Cai,¹ Deyun Gao,² and Jianhua Wu¹

¹ School of Computer Engineering, Nanyang Technological University, Singapore 639798

² School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

Received 18 January 2007; Revised 17 April 2007; Accepted 15 June 2007

Recommended by Mihaela Van Der Schaar

With the increasing popularity of using WLANs for Internet access, the HCCA mechanism in IEEE 802.11e WLANs has received much more attention due to its efficiency in handling time-bounded multimedia traffic. To achieve high network utilization and good end-to-end QoS in the scenario of VBR video over HCCA is a very challenging task because of the dynamics coming from both the network conditions and the video content. In this paper, we propose a cross-layer framework for efficiently delivering multiclass rate-adaptive VBR video over HCCA. The proposed framework consists of three major modules: the MAC-layer admission control, the MAC-layer resource allocation, and the application-layer video adaptation. Experimental results demonstrate the effectiveness of each individual module and the advantage of dynamic interactions among different modules.

Copyright © 2007 Jianfei Cai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

With the rapid growth of wireless communications and the advance of video coding techniques, wireless video streaming is expected to be widely deployed in the near future. Among various wireless networks, the IEEE 802.11-based wireless local area network (WLAN) is one of the most popular wireless networks and has been massively deployed in public and residential places. However, the existing 802.11 WLANs are designed for best-effort services. The two legacy medium access control (MAC) mechanisms, the distributed coordination function (DCF), and the point coordination function (PCF) [1], in the original 802.11 standard, lack quality of service (QoS) supports for multimedia applications. In order to enhance the QoS support in WLANs, a new standard called IEEE 802.11e [2] has been developed, which introduces a so-called hybrid coordination function (HCF) for medium access control. The HCF includes a contention-based mechanism named enhanced distributed channel access (EDCA) and a central-control-based mechanism named HCF controlled channel access (HCCA), which can be regarded as the extensions for the DCF and the PCF, respectively. Recently, we have seen many research studies on video over 802.11e EDCA WLANs [3, 4]. However, only a few studies investigate the HCCA such as [5]. The main reason is that distributed MAC mechanisms are much more popular than centralized

mechanisms in practice. In fact, most commercial WLAN products implement and employ DCF exclusively. However, with the increasing popularity of using WLANs for Internet access, where more and more multimedia traffic is relayed by access points as shown in Figure 1, HCCA has received more attention due to its high efficiency in handling time-bounded multimedia traffic.

For video streaming over HCCA, from the application-layer point of view, it is highly desired that video signals can be encoded in not only good average quality but also smooth video quality or less quality fluctuations among adjacent frames. However, quality-smoothed video leads to variable bit rate (VBR) bitstreams, which often exhibit significant bit-rate burstiness over multiple time scales due to the encoding frame structure and the natural variations within and between video scenes. When streaming VBR video over HCCA, the burstiness of VBR video will complicate the HCCA resource management since the resource requirements of VBR video are time-varying. On the other hand, video streaming over HCCA also faces other challenges coming from the WLAN itself. In particular, radio channels are well known for its notorious characteristics: bandwidth limited, error prone, and time varying. Under such a dynamic hostile environment, it is difficult for the WLAN to provide deterministic QoS services. In addition, wireless users could join or leave a WLAN at a random time, which further increases the

FIGURE 1: An example of video streaming relayed by access points.

dynamics of the network environment. In order to provide an end-to-end QoS, it requires not only the QoS management in the MAC layer but also the adaptation in the application layer.

Numerous solutions have been proposed for adaptive, efficient, and robust video streaming over lossy networks. Many of them are application-layer-based approaches including application-layer packetization [6], rate-distortion optimized scheduling [7], rate-reduction transcoding [8], joint source-channel coding [9], and so forth. However, the performance of these application-layer-based approaches has reached the limit. Recent research shows that carefully exploring the interactions among different layers in the network protocol stack could lead to much better performance [10]. Compared with traditional approaches, where each network layer is designed and operated independently, the cross-layer approaches jointly optimize or adjust the parameters in multiple layers.

Some cross-layer schemes for video streaming applications have been reported in literature. In [11, 12], a cross-layer protection scheme was proposed for streaming MPEG-4 FGS video over WLANs. The authors first developed an end-to-end distortion model for MPEG-4 FGS under various channel conditions and different unequal error protection strategies. Based on the developed model, the authors proposed to adaptively and jointly select the application-layer FEC, maximum MAC retransmission limit, and packet size according to the current channel conditions so that the received video quality can be maximized. In [13], Haratcherev et al. proposed a cross-layer architecture, where link adaptation is used at the MAC layer and rate control is used in the application layer. A cross-layer signaling mechanism

was proposed to convey the link-layer quality information to the video encoder. By coupling the rate control at the video encoder with the link adaptation, the proposed scheme can efficiently use the available transmission rate to achieve the best video quality. In [4], Ksentini et al. jointly considered the application, transport, and MAC layers for efficient transmission of H.264-coded video over IEEE 802.11e-based WLANs. The proposed cross-layer architecture relies on a data-partitioning (DP) technique at the application layer and an appropriate QoS mapping at the 802.11e-based MAC layer.

The major drawback of the above cross-layer strategies for wireless video streaming is that the cross-layer optimization is performed in isolation at each mobile station. In fact, the adaptation occurring in one station will affect other competing stations since wireless medium is shared among all the competing users. Therefore, although a cross-layer strategy is adopted by each individual mobile station, it should not be optimized in isolation. Instead, it should be considered from the entire network perspective, so that the overall system utility can be maximized. Similar ideas have been presented in [14, 15]. In particular, the authors in [14] studied efficient bandwidth resource allocation for streaming multiple MPEG-4 FGS video streams to multiple users, where the variations in the scene complexity of different video streams are explored and the system resources are dynamically and jointly distributed among users. In [15], Weber and Veciana proposed both optimal and practical mechanisms to maximize the customer average QoS defined in terms of received normalized time-average rate.

In this paper, we study rate-adaptive VBR video over HCCA using cross-layer design. We jointly consider the MAC-layer QoS management with the application-layer video adaptation in order to achieve not only good end-to-end QoS but also high network utilization. In particular, we apply the existing statistical multiplexing technique to the admission control problem to exploit the multiplexing gain among multiple VBR traffic. Unlike our previous admission control work in [16], which only considers one class of VBR traffic, in this paper we extend it to multiple classes of traffic flows. In addition to admission control, we also propose a dynamic network resource allocation scheme, where we take into account not only the average bit rates but also the burstiness of traffic flows. Experimental results demonstrate the effectiveness of each individual module, and the advantage of dynamic interactions among different modules.

This paper is organized as follows. Section 2 gives an overview of the HCCA mechanism. Section 3 describes the overall cross-layer architecture. Section 4 introduces the extended admission control scheme and the proposed dynamic resource allocation in the MAC layer. Section 5 shows the simulation results. Finally, Section 6 concludes this paper.

2. OVERVIEW OF HCF CONTROLLED CHANNEL ACCESS

Compared with the legacy PCF scheme in 802.11, HCCA also provides contention-free access to the wireless medium

FIGURE 2: The CAP/CFP/CP periods.

through polling stations but with some difference. In particular, HCCA uses a QoS-aware hybrid coordinator (HC), which is typically located at the QoS access point (QAP) in infrastructure WLANs. HC uses point interframe space (PIFS) to gain control of the channel and then allocates transmission opportunities (TXOPs) to QoS stations (QSTAs), which are referred as HCCA TXOPs or polled TXOPs. Unlike PCF, HCCA can poll the QSTAs during not only contention-free periods (CFPs) but also contention periods (CPs), and HCCA takes into account QSTAs' specific flow requirements in packet scheduling. Figure 2 illustrates the different periods under HCCA. Note that the CAP (controlled access phase) is defined as the time period when HC maintains the control of the medium. It can be seen that CAPs can be generated (or allocated by the HC) during CFPs or CPs.

After grabbing the channel, the HC polls QSTAs in turn according to its polling list. In order to be included in the polling list of the HC, a QSTA must send a QoS reservation request using the special QoS management frame that carries the traffic specification (TSPEC) parameters, and each individual flow needs one particular reservation request. The definitions of the TSPEC parameters can be found in the 802.11e standard [2], where the major TSPEC parameters include the following:

- (i) *peak data rate (P)*: the maximum bit rate allowed for packet transfer, in bits per second (bps);
- (ii) *mean data rate (A)*: the average bit rate for packet transmission, in bps;
- (iii) *maximum burst size (M)*: the maximum size of a data burst that can be transmitted at the peak data rate, in bytes;
- (iv) *delay bound (T^D)*: the maximum delay allowed to transport a packet across the wireless interface (including queuing delay), in milliseconds;
- (v) *maximum service interval (SI_{max})*: the maximum time allowed between neighbor TXOPs allocated to the same station, in microseconds;
- (vi) *nominal MSDU size (L_p)*: the nominal size of a packet, in bytes;
- (vii) *minimum PHY rate (A_{min}^{PHY})*: the minimum physical bit-rate assumed by the scheduler for calculating transmission time, in bps.

3. CROSS-LAYER FRAMEWORK

The key of the cross-layer optimization between the application layer and the MAC layer is to define interface parameters, based on which the two layers can talk and affect each other. It is intuitive that the interface parameters should come from traffic rate statistics. This is because traffic rate statistics can be easily understood by the two layers and they directly affect both the end user quality and the network resource utilization. However, to generate accurate traffic rate statistics requires a general model that can describe the characteristics of a VBR video. This is a nontrivial task since the rate distributions of a VBR video are time-varying and non-stationary. Traditional network resource management studies typically assume a traffic flow can be modeled as an ideal Poisson process, which is not true for VBR videos. In this paper, we bypass the video traffic modelling problem and directly work on the three TSPEC rate parameters: mean data rate A , peak data rate P , and maximum burst size M . Although (A, P, M) cannot fully describe the rate characteristics of a VBR video, it is sufficient to depict the traffic rate envelop, based on which a certain degree of optimization can be performed.

3.1. Traffic characteristics

In order to guarantee each traffic flow will conform to its claimed traffic parameters (A, P, M) , similar to the work in [17], we adopt the dual token bucket (DTB) as the traffic shaper to shape each traffic flow before entering the network. In particular, a DTB consists of two token buckets, where the first bucket is used to constrain the traffic flow with peak data rate P , and the other is used for maintaining the traffic flow with mean data rate A . Here, we use the second bucket as an example to explain how it works. Basically, each packet needs the same amount of tokens to be admitted. Tokens arrive at the token buffer at the rate A . If the total number of tokens in the bucket reaches the bucket depth B , a newly generated token will simply be discarded. When a packet arrives at the token bucket, it will be sent down to the MAC layer immediately if there are sufficient tokens available, and the corresponding tokens are removed from the token bucket. On the other hand, if there are not enough tokens available, the packet is either discarded directly or buffered if there is an incoming buffer in front of the token bucket. When a burst

FIGURE 3: The traffic envelop for a traffic flow with the DTB shaper.

of packets arrives, it is allowed to pass if enough tokens have been accumulated in the token bucket.

Let $a(t, t + \tau)$ denote the total number of arrived data of a flow in the interval $(t, t + \tau)$. Clearly, after passing through the DTB shaper, $a(t, t + \tau)$ is deterministically bounded by a function $b(\tau)$, which is called rate envelop and is defined as

$$b(\tau) = \begin{cases} P\tau, & 0 < \tau < \eta, \\ B + A\tau, & \tau \geq \eta. \end{cases} \quad (1)$$

Considering the three TSPEC parameters (A, P, M) , we can derive $B = M(1 - A/P)$ and $\eta = M/P$, as shown in Figure 3. In this way, we can guarantee that, after shaping, any random process $a(t, t + \tau)$ is fully conforming to the three TSPEC parameters.

3.2. Cross-layer architecture

Figure 4 shows the overall system architecture. Basically, we consider multiple adaptable VBR video transmitted over reliable wired channels to an AP, and the AP uses the centralized HCCA mechanism to deliver video traffic to multiple mobile stations over unreliable wireless channels. We assume the wired channels between the video source and the AP are perfect, and the bottleneck for end-to-end QoS lies in the wireless channels.

In particular, multiple adaptable video sources could be physically generated at one video server or multiple video servers or multiple endpoints. The application layer of an adaptable video source contains two major modules: video adaptation and traffic shaper. The video adaptation module is to approximately adapt the video flow to the allocated traffic rate parameters $\Omega = \{A, P, M\}$ while the traffic shaper is to guarantee the video traffic conforms to Ω . For practical applications, it is highly desired that video quality can be controlled in a certain range. Specifically, users expect received video quality should not be below an acceptable quality, while achieving an extremely high quality is not necessary. Thus, in this paper, we simply use the common PSNR

(peak signal to noise ratio) metric to define two video quality thresholds, $U^{\min}$ and $U^{\max}$, which correspond to the acceptable video quality and the highest video quality specified by users. Note that these MSE-based thresholds could be determined according to human visual systems (HVS). The adaptation between $U^{\min}$ and $U^{\max}$ can be implemented through many video adaptation techniques such as layered video coding, scalable video coding, or bitstream switching. Since we consider stored video, the corresponding traffic statistics including $\Omega^{\min}$ and $\Omega^{\max}$ for each adaptation level can be pre-generated.

At the AP side, there are three major modules: admission control, dynamic bandwidth allocation, and physical rate adjustment. The physical rate adjustment is to adaptively adjust the transmission rates from the AP to QSTAs according to the feedback information from QSTAs so that the physical-layer bit errors can be effectively reduced. Many physical rate adaptation schemes [13, 18] have been proposed in literature. Some are based on the statistics of the performance parameters such as throughput, frame error rate, or frame retransmission rate. Others are according to the receiver SNR which directly determines the decoding error rate. In this research, although we do not study the mechanisms for physical rate adjustment, it could be easily incorporated into our proposed cross-layer framework. As for the admission control module, the purpose is to limit the amount of traffic admitted into the WLAN communications so that the QoS of the existing flows will not be degraded while at the same time the wireless medium resources can be maximally utilized. The dynamic bandwidth allocation module is to reallocate the bandwidth if the network conditions or the traffic conditions are changed. The network condition change could be due to three reasons: (1) one new traffic flow is admitted; (2) one of the existing flows is finished; (3) some QSTAs' physical-layer rates have been changed either due to their movement or wireless channel variation. The traffic-condition change is due to the variation of video content such as scene changes.

4. MAC-LAYER QoS MANAGEMENT

A simple admission control and resource allocation scheme for HCCA has been developed as a reference in the 802.11e standard [2], where the mean data rate and the mean packet size are used to calculate the resource needed by a flow. This reference scheme works fine for CBR (constant bit rate) traffic which strictly comply with their QoS requirements. However, it is not suitable for VBR traffic, where the instantaneous sending rate and packet size are usually quite different from the corresponding mean values. Recently, we have seen some admission control and resource allocation algorithms [17, 19–23] being proposed for delivering VBR traffic over HCCA. In [19], the authors adopted the reference scheme for admission control and proposed to consider the application deadline at the time of allocating a TXOP. In [20], the authors proposed a dynamic bandwidth allocation algorithm, where the classic feedback control theory is applied to take into account the queue levels in the QoS stations (QSTAs). In [21], two types of schedulers are proposed: QoS access point

FIGURE 4: The cross-layer architecture for video streaming over HCCA.

(QAP) scheduler and node scheduler. The QAP scheduler estimates the queue length of each QSTA and adapts the TXOPs allocation accordingly. The node scheduler of a QSTA is to redistribute the unused time among its own multiple traffic flows. In [22], the authors proposed to estimate the application's mean data rate through the queue length and then allocate the resource accordingly.

Although the above methods improve the efficiency of resource allocation, all of them still use the reference scheme for admission control, which might wrongly admit or reject new flows since it does not consider the characteristics of VBR traffic. In our previous work [23], we proposed an effective TXOP-based admission control scheme for VBR over HCCA. The basic idea is to use the effective TXOP to statistically guarantee a certain packet loss ratio. Recently, a guaranteed-rate-based admission control was proposed in [17], where the DTB is used as the traffic shaper to shape each traffic flow. Based on the characteristics of shaped traffic flows, the authors derived guaranteed rates for each flow. Although these two admission control schemes indeed take the VBR characteristics into consideration, they are still not efficient because both schemes consider each traffic flow individually and the multiplexing gain among multiple VBR flows has not been explored at all.

4.1. Admission control

Since the purpose of admission control is to admit as many flows as possible under the constraints of satisfying the minimum QoS requirement of all the flows, it is obvious that the decision of admission control should be based on the minimum traffic rate statistics $\Omega^{\min} = \{A^{\min}, P^{\min}, M^{\min}\}$. We can summarize the admission control problem as follows: given the QoS requirement of a new flow, including the minimum traffic rate statistics $\Omega^{\min}$, the delay bound T^D , and the tolerable packet loss rate ϵ , how to decide whether this flow should be admitted or not? It is clear that the packet delay consists of the transmission delay in the PHY layer and the queuing delay in the MAC layer. The transmission delay can

be neglected because of the short distance between the AP and mobile stations in WLANs. The MAC-layer queuing delay is determined by the queue scheduling algorithms in the MAC layer. On the other hand, the packet loss could cause by wireless channel errors, the DTB shaper, and the delay bound violation, where we consider a packet delayed long than T^D as a lost packet. Since the physical rate adjustment is used in each station, which automatically adjusts the physical transmission rate according to wireless channel conditions, the MAC-layer frame loss rate due to wireless channel errors can be greatly reduced and typically the frame loss rate is less than 2.5% [24]. Further considering the use of large retry limit (e.g., the default value of 7), we can neglect the packet loss due to wireless channel errors. The packet loss caused by the DTB shaper can also be neglected because of the application-layer adaptation and buffer control. In this way, we can deem that the packet loss is primarily caused by the delay bound violation and thus the packet loss threshold ϵ becomes the same as the delay bound violation threshold, that is, $P\{d > T^D\} \leq \epsilon$.

In this research, we consider multiple classes of VBR video flows. Let N denote the total number of video classes and let K_i denote the number of video flows in the i th class. Suppose all the video flows in a class i have the same QoS requirement (T_i^D, ϵ_i) and the QoS requirements in different classes are different. We employ the popular weighted fair queueing (WFQ) to provide the service differentiation among multiple classes. Although other types of scheduling algorithms such as the earliest deadline first (EDF) algorithm [25], which schedules packets in ascending order according to their deadlines, can utilize the resources well, they are too complicated to be implemented in AP. On the contrary, implementing the WFQ is very easy. The WFQ scheduling algorithm simply separates packets into different queues according to their QoS requirements. The first-come-first-served (FCFS) principle is used in each queue, and the resource is dynamically allocated among different queues by adjusting the weights, which are determined by the resource allocation algorithm.

Let C_i^g , ($i = 1, \dots, N$) denote the minimum bandwidth needed to guarantee the QoS requirements, $P\{d_i > T_i^D\} \leq \epsilon_i$, for each class. Clearly, C_i^g depends on the aggregated traffic rate statistics of the i th class, and it needs to be carefully selected. If we simply choose C_i^g according to the aggregated peak rates ($\sum_{j=1}^{K_i} P_{ij}^{\min}$) of all the flows in the i th class, no packet loss will occur but a substantial amount of bandwidth will be wasted at most of the time. On the other hand, if we choose C_i^g according to a data rate much lower than the aggregated peak rate (e.g., the aggregated mean data rate $\sum_{j=1}^{K_i} A_{ij}^{\min}$), we might experience large delay and excessive packet loss since the instantaneous sending rates of VBR traffic are usually quite different from the corresponding mean values. Therefore, it is a challenging task to obtain optimal C_i^g values that achieve the best tradeoff between network utilization and service quality for VBR traffic over HCCA.

Fortunately, the relationship between the probability of queuing delay and the aggregated traffic rate statistics has been derived in [26, 27], where the delay probability is modeled as a Gaussian-like distribution. Applying the finding to our case, we obtain the delay-bound violation probability for the i th class as

$$P\{d_i > T_i^D\} \approx \max_{0 \leq \tau \leq \beta_i} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(C_i^g(\tau + T_i^D) - \mu_i)^2}{2\sigma_i^2}\right) \quad (2)$$

with

$$\begin{aligned} \mu_i &= \sum_j \tau \phi_{ij}, \\ \sigma_i^2 &= \sum_j \tau^2 RV_{ij}(\tau), \end{aligned} \quad (3)$$

where β_i is the busy period bound, μ_i is the aggregate mean traffic rate, σ_i^2 is the aggregate rate variance, and ϕ_{ij} and RV_{ij} are the long-term average rate and the rate-variance envelop for the j th flow of the i th class, respectively. If a flow j is stationary and its arrival $a_{ij}[t, t + \tau]$ is upper bounded, that is, $a_{ij}[t, t + \tau] \leq q_{ij}(\tau)$ for all $t, \tau > 0$, according to [27], its rate-variance envelop should be upper bounded as

$$RV_{ij}(\tau) \leq \frac{\phi_{ij} q_{ij}(\tau)}{\tau} - \phi_{ij}^2, \quad (4)$$

$$\phi_{ij} = \lim_{\tau \rightarrow \infty} \frac{q_{ij}(\tau)}{\tau}, \quad (5)$$

where q_{ij} is the PHY rate envelop. Considering the worst case, we let

$$RV_{ij}(\tau) = \frac{\phi_{ij} q_{ij}(\tau)}{\tau} - \phi_{ij}^2. \quad (6)$$

In our system, since each video flow a_{ij} matches the DTB shaper, we know it is upper bounded by b_{ij} defined in (1). However, b_{ij} is not the same as the physical rate envelop q_{ij} since we need to consider the protocol headers and overhead when packets pass through different network layers. Thus,

we introduce a new variable, network resource utilization ratio r_{ij} , which is defined as the ratio between the network resource used by the arrived traffic and the total network resource used to successfully deliver the traffic. Combining b_{ij} and r_{ij} , we express $q_{ij}(\tau)$ as

$$q_{ij}(\tau) = \min(P_{ij}^{\min} \tau r_{ij}, (B_{ij}^{\min} + A_{ij}^{\min} \tau) r_{ij}). \quad (7)$$

This is the rate envelop bound from the PHY-layer point of view. According to (5) and (7), we obtain

$$\phi_{ij} = A_{ij}^{\min} r_{ij}. \quad (8)$$

After that, substituting ϕ_{ij} and $q_{ij}(\tau)$ back to (6), we derive

$$RV_{ij}(\tau) = \begin{cases} A_{ij}^{\min} (P_{ij}^{\min} - A_{ij}^{\min}) (r_{ij})^2, & 0 \leq \tau \leq \eta_{ij}, \\ \frac{A_{ij}^{\min} B_{ij} (r_{ij})^2}{\tau}, & \eta_{ij} \leq \tau \leq \beta_i, \end{cases} \quad (9)$$

where $\eta_{ij} = B_{ij} / (P_{ij}^{\min} - A_{ij}^{\min})$ and $B_{ij} = M_{ij}^{\min} (1 - A_{ij}^{\min} / P_{ij}^{\min})$.

The busy period bound β can be calculated as [28]

$$\beta_i = \min \left\{ \tau > 0 \mid \sum_j q_{ij}(\tau) \leq C_i^g \tau \right\}. \quad (10)$$

We can see that β_i is actually the minimum time that the network needs to accommodate the aggregated VBR burst. Clearly, if we use an upper bound to replace $q_{ij}(\tau)$ in (10), it will only result in a larger value of β_i , which will not affect the solution of $P\{d_i > T_i^D\}$. Thus, we use $(B_{ij} + A_{ij}^{\min} \tau) r_{ij}$ to replace $q_{ij}(\tau)$ in (10) since $q_{ij}(\tau) \leq (B_{ij} + A_{ij}^{\min} \tau) r_{ij}$ and we obtain

$$\begin{aligned} \beta_i &= \min \left\{ \tau > 0 \mid \sum_{i=1}^{K_i} (B_{ij} + A_{ij}^{\min} \tau) r_{ij} \leq C_i^g \tau \right\} \\ &= \frac{\sum_{j=1}^{K_i} B_{ij} r_{ij}}{C_i^g - \sum_{j=1}^{K_i} A_{ij}^{\min} r_{ij}}. \end{aligned} \quad (11)$$

In this way, we have derived all the parameters except the bandwidth C_i^g for calculating the delay-bound violation probability $P\{d_i > T_i^D\}$ defined in (2). In other words, given the traffic rate statistics $\Omega_{ij}^{\min}$ and the allocated bandwidth C_i^g for the i th service class, we are able to derive the delay-bound violation probability. In reverse, given the traffic rate statistics $\Omega_{ij}^{\min}$ and the QoS requirement $P\{d_i > T_i^D\} \leq \epsilon_i$, we can also derive the minimum bandwidth needed for the i th class, that is, C_i^g .

Based on the above discussion, the admission control algorithm can be simply designed as follows. When a new flow arrives, we first classify it into a service class i according to its QoS requirement. Then, we calculate the needed minimum bandwidth C_i^g if this new flow is admitted. After that, we add the minimum bandwidth for all the classes together, that is, $C^g = \sum_{i=1}^N C_i^g$, and compare C^g with the link capacity $C^{\text{PHY}} \delta$, where C^{PHY} is the physical bandwidth of a WLAN and δ is the percentage of polling-based transmission specified in HCF. If $C^g \leq C^{\text{PHY}} \delta$, we accept the new flow. Otherwise, the new flow should be rejected.

4.2. Dynamic resource allocation

After a new flow is being accepted by the admission control algorithm, the next task we need to solve is how to allocate the network resource to the new flow and all the previously existing flows. As mentioned in Section 3.2, such a bandwidth allocation task also exists in other scenarios including the network variation and also the traffic variation caused by the change of video content. The objective of the resource allocation is to maximize the overall utility (the same as video quality), which is the utility sum of all the video flows,

$$U = \sum_{i=1}^N \sum_{j=1}^{K_i} U_{ij}. \quad (12)$$

The network resource we need to allocate is the remaining capacity defined as the difference between the total capacity for polling-based transmission and the total minimum bandwidth needed for all the classes, that is,

$$C^{\text{rm}} = C^{\text{PHY}} \delta - C^g. \quad (13)$$

The resource allocation problem can be summarized as follows. Given the ranges of the traffic rate statistics of all the flows $(\Omega_{ij}^{\min}, \Omega_{ij}^{\max})$, how to distribute the remaining capacity C^{rm} through selecting the optimal traffic rate statistics Ω_{ij} for each flow so that the overall utility U can be maximized?

In order to solve this overall optimization problem, a general model that can characterize the relationship between U_{ij} and Ω_{ij} is needed. However, it is very hard to develop such a model since the R-D behavior of video coding is very complicated, and moreover there are three parameters included in the traffic rate statistics Ω_{ij} . In this research, for simplicity we assume that the utility U_{ij} only depends on the average bit rate A_{ij} with a linear relationship between them, that is,

$$U_{ij} = \frac{U^{\max} - U^{\min}}{A_{ij}^{\max} - A_{ij}^{\min}} (A_{ij} - A_{ij}^{\min}) + U^{\min}, \quad (14)$$

where the two constants $U^{\max}$ and $U^{\min}$, as mentioned in Section 3.2, are the highest video quality needed and the acceptable video quality, respectively, and $A_{ij}^{\max}$ and $A_{ij}^{\min}$ are the corresponding average bit rates. Under such a linear relationship, it seems that the bandwidth allocation would be straightforward, that is, allocating more bandwidth to video flows with larger slope values $1/(A_{ij}^{\max} - A_{ij}^{\min})$ since they achieve higher utility increase for the same average bit rate increase. However, the same average bit rate increase does not mean the same network resource consumption since different video flows have different traffic burst characteristics. A highly bursty video flow with a lower average bit rate might require more network resource than a less bursty video flow but with a higher average bit rate.

In this paper, we divide this bandwidth allocation problem into two tasks: the first task is to distribute the remaining bandwidth C^{rm} among different classes, and the second task is to allocate the bandwidth among different video flows within one class. For the first task, we propose to proportionally allocate the remaining bandwidth to each class according

to the needed minimum network resource C_i^g , which we have computed in admission control. This is reasonable since the class with higher minimum bandwidth requirement should be allocated more bandwidth. Thus, we calculate the weights ω_i for the WFQ scheduler as

$$\omega_i = \frac{C_i^g}{\sum_{i=1}^N C_i^g} \quad (15)$$

and the total bandwidth C_i for the i th class becomes

$$C_i = \omega_i C^{\text{rm}} + C_i^g. \quad (16)$$

Although we are able to allocate the bandwidth among different classes, we still face the problem of allocating bandwidth among different video flows within one class. Considering the key term $f = ((C_i(\tau + T_i^D) - \mu_i)^2 / 2\sigma_i^2)$ ($C_i = C_i^g$ when we consider the minimum traffic rate statistics $\Omega_{ij}^{\min}$) in (2), when the traffic rate statistics of a video flow is changed from $\Omega_{ij}^{\min}$ to $\Omega_{ij}^{\max}$, μ_i and σ_i^2 will correspondingly change with the increments of $\Delta\mu_i$ and $\Delta\sigma_i^2$. If there is also an increment ΔC_i for C_i that can make f remain unchanged, we can deem this ΔC_i is the corresponding network resource increment in order to accommodate the increment in traffic rate statistics. Using Taylor's expansion for $f(C_i + \Delta C_i, \mu_i + \Delta\mu_i, \sigma_i^2 + \Delta\sigma_i^2)$, we derive

$$\frac{\partial f}{\partial C_i} \Delta C_i + \frac{\partial f}{\partial \mu_i} \Delta\mu_i + \frac{\partial f}{\partial \sigma_i^2} \Delta\sigma_i^2 = 0. \quad (17)$$

Solving the equation above, we obtain

$$\Delta C_i = \frac{C_i(\tau + T_i^D) - \mu_i}{2(\tau + T_i^D)\sigma_i^2} \cdot \Delta\sigma_i^2 + \frac{1}{(\tau + T_i^D)} \cdot \Delta\mu_i. \quad (18)$$

By assuming $\tau/(\tau + T_i^D) = 1$, $RV_{ij} = (A_{ij}r_{ij})(B_{ij} + A_{ij}\tau)r_{ij}/\tau - (A_{ij}r_{ij})^2$ and $B_{ij}/\tau = P_{ij}$, we approximate ΔC_i as

$$\begin{aligned} \Delta C_{ij} = & \frac{C_i^g - \sum_j A_{ij}^{\min} r_{ij}}{2 \sum_j A_{ij}^{\min} P_{ij}^{\min}} \cdot (A_{ij}^{\max} P_{ij}^{\max} - A_{ij}^{\min} P_{ij}^{\min}) \\ & + (A_{ij}^{\max} - A_{ij}^{\min}) r_{ij}, \end{aligned} \quad (19)$$

where the first term is the network resource needed to accommodate the increment in traffic burst and the second term is to accommodate the increment in traffic average bit rate. Note that the reason we made many assumptions for the approximation in (19) is that we are not aiming to obtain accurate values of ΔC_{ij} . Instead, we try to obtain some quantitative values which can relatively reflect more or less network resource being consumed for each flow for achieving a utility increase $(U^{\max} - U^{\min})$. Experimental results presented later show that the bandwidth allocation based on ΔC_{ij} in (19) outperforms the approach based on $(A_{ij}^{\max} - A_{ij}^{\min})$.

After obtaining ΔC_{ij} , we put all the video flows in the i th class into one queue at an increasing order of ΔC_{ij} . Clearly, as long as we still have unallocated network resource, we will increase the traffic rate of the first video flow in the queue

FIGURE 5: The number of admitted flows using different admission control schemes. (a): using the rate guarantee-based admission control. (b): our proposed rate-variance-based admission control.

until it reaches the maximal traffic rate statistics. Then, the next video flow will move to the queue head and the process repeats. After we allocate all the remaining network bandwidth, we can imagine that all the video flows except the middle one will be either allocated maximum traffic rate statistics or minimum traffic rate statistics. Therefore, by forcing the middle one also taking its minimum traffic rate statistics, our obtained bandwidth allocation can be expressed as

$$\Omega_{ij} = \begin{cases} \Omega_{ij}^{\max}, & j = 1, 2, \dots, k-1, \\ \Omega_{ij}^{\min}, & j = k, \dots, K_i, \end{cases} \quad (20)$$

where the position of k can be determined through search under the constraints of $P\{d_i > T_i^D\} \leq \epsilon_i$ and C_i . Note that this kind of bandwidth allocation only requires the application layer to provide two adaptation levels, $\Omega_{ij}^{\max}$ and $\Omega_{ij}^{\min}$, and it can achieve good overall system utility although the quality of individual video flows might change sharply, that is, jumping between $U^{\max}$ and $U^{\min}$. We would also like to point out the previous discussion is for the case of $C^{\text{rm}} > 0$. If $C^{\text{rm}} < 0$ (e.g., due to the channel deterioration), we have to reject some of the existing video flows. It can be conducted based on the arrival time of video flows, that is, keep deleting the latest video flow until C^{rm} becomes not less than zero.

5. SIMULATION RESULTS

5.1. Results of MAC-layer QoS management

We first evaluate the efficiency of our proposed rate-variance envelop-based admission control and compare it with the guaranteed-rate-based admission control (GRAC) in [17]. Since IEEE 802.11e only specifies the MAC-layer mechanisms, we use the parameters of the IEEE 802.11a physical layer in the experiments. In particular, the physical transmission rates of all the nodes are set to 54 Mbps and 24 Mbps

for data frames and control frames, respectively. We consider two classes of traffic flows. For the first class, the average bit rate is randomly chosen from the range of [50, 100] kbps, and the peak bit rate is randomly chosen from [5, 10] times of the average bit rate. For the second class, the average bit rate is randomly chosen from the range of [100, 150] kbps, and the peak bit rate is randomly chosen from [10, 15] times of the average bit rate. The burst sizes for both classes are set to 0.2 second peak rate. We test the admission control performance under different delay bounds ranging from 0.01 second to 0.15 second and from 0.16 second to 0.30 second for the two classes, respectively. The delay bound violation probabilities are set to 10^{-6} and 10^{-5} for the two classes, respectively.

The number of admitted flows is one of the important criteria to measure the performance of admission control in terms of network utilization. The larger the number of admitted flows is, the better network utilization the admission control achieves. Figure 5 shows the numerical results of the number of admitted flows under different delay bounds. It can be seen that our proposed admission control scheme always outperforms the GRAC scheme in terms of admitting much more traffic flows. For example, in the case that the delay bounds of class 1 and class 2 are set to 0.15 second and 0.30 second, respectively, GRAC admits 19 and 18 flows in each class while our proposed admission control admits 40 and 39 flows for class 1 and class 2, respectively. We have also used the same traffic parameters in NS-2 simulations. We find that the average delay is very small and delay bound violation is nearly zero. For example, in the case that the delay bounds of class 1 and class 2 are set to 0.15 second and 0.30 second, the NS-2 simulation shows that the average delay and the maximum delay for class 1 are 2.985 milliseconds and 14.903 milliseconds, respectively, which means that no packet will be dropped due to delay bound violation and the QoS performances of video flows are still satisfied. The results for class 2 are similar.

FIGURE 6: The number of flows admitted at $\Omega^{\max}$ and $\Omega^{\min}$ using different resource allocation schemes.

Next, we compare two different dynamic resource allocation schemes. As described in Section 4.2, one is purely based on the average bit rate and the other is our proposed algorithm that considers both the average bit rate and the traffic burstiness. For illustration purpose, we only consider one traffic class with four different types of VBR flows, where $A^{\min} = 100$ kbps, $p^{\min} = 8 \cdot A^{\min}$, $M^{\min} = 0.2 \cdot p^{\min}$, $A^{\max} = 200$ kbps, $M^{\max} = 0.2 \cdot p^{\max}$, and $p^{\max}$ is set to $\{6, 5, 4, 3\}$ times of the average bit rate for different types. We first choose a delay bound of 0.05 second and a delay violation probability of 10^{-6} . The four types of VBR flows are added into the network in turn until no new flow can be accepted. We find the total number of admitted flows is 100 and all the flows are allocated with their corresponding $\Omega^{\min}$ as we expect. Then, we fix the total number of flows to 100 and increase the delay bound from 0.05 second to 0.25 second. Figure 6 shows the number of flows allocated with either $\Omega^{\max}$ or $\Omega^{\min}$ using the two different resource allocation schemes under different delay bounds. It can be seen that our proposed scheme allows much more flows to use their maximum traffic parameters and thus achieves higher overall system utility.

5.2. Results of video over HCCA

In this section, we evaluate the performance of transmitting VBR videos over our proposed MAC-layer QoS management system. The 300-frame QCIF Foreman and QCIF Akiyo video are used as the test video sequences, where the Foreman sequence is considered as a high-motion sequence and the Akiyo is a low-motion one. H.263 is applied to code the video sequences at both 10 fps and 30 fps. The quality thresholds, $U^{\max}$ and $U^{\min}$, are set to 38 dB and 32 dB, respectively. For VBR video encoding, we adopt the encoder-based rate smoothing approach proposed in [29]. Let U_T be

the PSNR value of the target picture quality, and let $U(n)$ and $R(n)$ be the actual PSNR value and bit rate of the n th frame. The basic idea of the encoder-based rate smoothing scheme is to let $U(n)$ vary within a small range $[U_T - \delta, U_T + \delta]$, and try to make $R(n)$ as close to $R(n-1)$ as possible. In this experiment, δ is set to 1 dB. There is one I-frame every two seconds and the rest of the video frames are encoded as P-frames. Table 1 shows the generated four types of adaptable VBR video traffic. For each adaptable VBR video traffic, the bitstream switch technique is employed to adapt the video traffic between $\Omega^{\max}$ and $\Omega^{\min}$. Note that although we use encoder-based rate smoothing scheme for VBR video encoding, any other VBR encoding scheme can be adopted in our cross-layer framework.

For simplicity, we only consider one traffic class, and the delay bound and the delay bound violation probability are set to 0.2 second and 10^{-6} . We send the four different flows, that is, the two video sequences with two different frame rates, to the network in turn until the number of admitted flows reaches 80. Then, we keep sending the Foreman with 30 fps to the network until the total number of admitted flows reaches 90. Every 0.5 second, one flow is added to the network, and the total simulation time is 80 seconds. Every 1 second, an interaction between the application layer and the MAC layer is performed. In addition to the network dynamics, we purposely make one scene change for each of the last ten flows, that is, the 10 Foreman sequences with 30 fps are changed to Akiyo with 30 fps.

Figure 7(a) shows the average PSNR performance, where we compare the static strategy and the dynamic strategy. Both of them use the same setting, that is, our proposed admission control and dynamic bandwidth allocation, except that under the static strategy, the application layer only sends the traffic parameters once to the MAC-layer. It can be seen that the average PSNR of the dynamic strategy is better since it dynamically reacts to the scene changes occurring after 55 seconds. Note that the average PSNR gain will be more significant if there exists larger number of scene changes. Figure 7(b), we use a particular flow, the 46th flow, as an example to show the PSNR result of a flow. The 46th flow is admitted at 23 seconds with the assigned traffic rate statistics $\Omega^{\max}$, which lead to an average PSNR of 38 dB. Starting at 43 seconds, the assigned traffic rate statistics for the flow are changed to $\Omega^{\min}$ due to more video flows added to the network. Thus, the average PSNR value is dropped to 32 dB. However, at 46 seconds, the flow is being assigned with $\Omega^{\max}$ again. This is because the scene changes occurring at the last ten flows, which requires less network resource, cause the network to dynamically change the 46th flow's traffic rate statistics from $\Omega^{\min}$ to $\Omega^{\max}$.

6. CONCLUSION

In this paper, we have extended the existing statistical multiplexing technique to the admission control of multiclass multitype VBR flows in HCCA. Experimental results show that the number of admitted VBR flows using our approach is two times of that using the existing admission control

TABLE 1: The traffic parameters for Akiyo and Foreman at different frame rates and different target PSNR values.

Seq. (frame rate, $U^{\min}/U^{\max}$)	($A^{\min}/A^{\max}$, $P^{\min}/P^{\max}$, $M^{\min}/M^{\max}$)
Akiyo (10 fps, 32 dB/38 dB)	(14.540/37.680 kbps, 125.440/267.920 kbps, 20.731/48.837 kbits)
Foreman (10 fps, 32 dB/38 dB)	(69.710/212.890 kbps, 236.40/551.200 kbps, 126.080/339.253 kbits)
Akiyo (30 fps, 32 dB/38 dB)	(16.890/55.680 kbps, 376.320/803.760 kbps, 20.799/61.346 kbits)
Foreman (30 fps, 32 dB/38 dB)	(104.430/393.360 kbps, 709.200/1653.600 kbps, 139.649/454.271 kbits)

(a)

(b)

FIGURE 7: (a): the average PSNR. (b): the PSNR result of the 46th video flow.

scheme for HCCA. Moreover, we have also proposed the optimized resource allocation scheme that considers not only the average traffic rate but also the burstiness of VBR flows. Experimental results show that compared with the scheme only based on the average bit rate, our proposed resource allocation utilizes the network resource very well, up to 50% more flows being allocated maximal traffic rate statistics. In addition, we have integrated the MAC-layer QoS management modules with the application-layer video adaptation to have a cross-layer design. Experimental results show that the cross-layer framework is able to handle the dynamics of video over WLANs.

ACKNOWLEDGMENT

This research is supported by Singapore A*STAR SERC Grant (032 101 0006).

REFERENCES

- [1] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std. 802.11, 1999.
- [2] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, IEEE Std. 802.11e-2005, 2005.
- [3] Y. Xiao, H. Li, and S. Choi, "Protection and guarantee for voice and video traffic in IEEE 802.11e wireless LANs," in *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '04)*, vol. 3, pp. 2152–2162, Hongkong, March 2004.
- [4] A. Ksentini, M. Naimi, and A. Guroui, "Toward an improvement of H.264 video transmission over IEEE 802.11e through a cross-layer architecture," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 107–114, 2006.
- [5] M. van der Schaar, Y. Andreopoulos, and Z. Hu, "Optimized scalable video streaming over IEEE 802.11a/e HCCA wireless networks under delay constraints," *IEEE Transactions on Mobile Computing*, vol. 5, no. 6, pp. 755–768, 2006.
- [6] S. Dumitrescu, X. Wu, and Z. Wang, "Globally optimal uneven error-protected packetization of scalable code streams," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 230–239, 2004.
- [7] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," Technical Report MSR-TR-2001-35, Microsoft Research, Redmond, Wash, USA, February 2001.
- [8] A. Vetro, J. Cai, and C. W. Chen, "Rate-reduction transcoding design for wireless video streaming," *Journal of Wireless Communications and Mobile Computing*, vol. 2, no. 6, pp. 549–552, 2002.
- [9] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 511–523, 2002.

- [10] M. van der Schaar and S. Shankar N, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 50–58, 2005.
- [11] 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 Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1752–1763, 2003.
- [12] 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 Transactions on Multimedia*, vol. 6, no. 2, pp. 278–290, 2004.
- [13] I. Haratcherev, J. Taal, K. Langendoen, R. Lagendijk, and H. Sips, "Optimized video streaming over 802.11 by cross-layer signaling," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 115–121, 2006.
- [14] G.-M. Su and M. Wu, "Efficient bandwidth resource allocation for low-delay multiuser video streaming," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 9, pp. 1124–1137, 2005.
- [15] S. Weber and G. Veciana, "Rate adaptive multimedia streams: optimization and admission control," *IEEE/ACM Transactions on Networking*, vol. 13, no. 6, pp. 1275–1288, 2005.
- [16] D. Gao, J. Cai, and C. W. Chen, "Admission control with traffic shaping for variable bit rate traffic in IEEE 802.11e WLANs," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '06)*, pp. 1–5, San Francisco, Calif, USA, November 2006.
- [17] C.-T. Chou, S. Shankar, and K. G. Shin, "Achieving per-stream QoS with distributed airtime allocation and admission control in IEEE 802.11e wireless LANs," in *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '05)*, vol. 3, pp. 1584–1595, Miami, Fla, USA, May 2005.
- [18] J. Kim, S. Kim, S. Choi, and D. Qiao, "CARA: collision-aware rate adaptation for IEEE 802.11 WLANs," in *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM '06)*, pp. 1–11, Barcelona, Spain, April 2006.
- [19] A. Grilo, M. Macedo, and M. Nunes, "A scheduling algorithm for QoS support in IEEE802.11E networks," *IEEE Wireless Communications*, vol. 10, no. 3, pp. 36–43, 2003.
- [20] G. Boggia, P. Camarda, L. A. Grieco, and S. Mascolo, "Feedback-based bandwidth allocation with call admission control for providing delay guarantees in IEEE 802.11e networks," *Computer Communications*, vol. 28, no. 3, pp. 325–337, 2005.
- [21] P. Ansel, Q. Ni, and T. Turletti, "An efficient scheduling scheme for IEEE 802.11e," in *Proceedings of IEEE Workshop on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt '04)*, Cambridge, UK, March 2004.
- [22] L. Yang, "Enhanced HCCA for real-time traffic with QoS in IEEE 802.11e based networks," in *IEEE International Workshop on Intelligent Environments (IE '05)*, pp. 203–209, Colchester, UK, June 2005.
- [23] W. F. Fan, D. Gao, D. H. K. Tsang, and B. Bensaou, "Admission control for variable bit rate traffic in IEEE 802.11e WLANs," in *Proceedings of the 13th IEEE Workshop on Local and Metropolitan Area Networks, (LANMAN '04)*, pp. 61–66, Mill Valley, Calif, USA, April 2004.
- [24] M. Hassan and R. Jain, Eds., *High Performance TCP/IP Networking*, Prentice-Hall, Upper Saddle River, NJ, USA, 2003.
- [25] A. Grilo, M. Macedo, and M. Nunes, "A scheduling algorithm for QoS support in IEEE802.11E networks," *IEEE Wireless Communications*, vol. 10, no. 3, pp. 36–43, 2003.
- [26] E. W. Knightly, "Second moment resource allocation in multi-service networks," in *Proceedings of the ACM Sigmetrics International Conference on Measurement and Modeling of Computer Systems*, pp. 181–191, Seattle, Wash, USA, June 1997.
- [27] E. W. Knightly, "Enforceable quality of service guarantees for bursty traffic streams," in *Proceedings of IEEE 17th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '98)*, vol. 2, pp. 635–642, San Francisco, Calif, USA, March-April 1998.
- [28] R. L. Cruz, "A calculus for network delay—I: network elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 114–131, 1991.
- [29] Z. He, " ρ -domain rate-distortion analysis and rate control for visual coding and communication," Ph.D. dissertation, ECE of University of California, Santa Barbara, Calif, USA, 2001.