Cross Layer Design and Analysis of Multiuser Wireless Video Streaming Over 802.11e EDCA

Ju-Lan Hsu and Mihaela van der Schaar, *Senior Member, IEEE*

***Abstract*—We propose a novel cross layer solution for optimizing the performance of multiple users, which are concurrently transmitting their delay-sensitive video streams over the same shared contention-based wireless LAN. To improve the performance of their applications, they dynamically adapt their medium access control (MAC) strategies (i.e., retry limits). Higher retry limits will benefit the user, but they will also lead to an increased congestion for the other users. Hence, we explicitly consider how the cross layer strategy adopted by one video user will impact the strategies of the competing users. Specifically, we analytically model how the cross layer transmission strategies adopted by each wireless user impact their own incurred distortion and delay but also that of their competing wireless stations. We also propose a distributed, low-complexity, and highly scalable optimization algorithm which can maximize the utility of the delay-sensitive video application. We compare our scheme to existing retry limit-based solutions, and the simulation results show that the proposed solution can outperform these solutions and maximize the users' utilities in a highly congested multiuser wireless environment.**

***Index Terms*—Cross layer design, multimedia communication.**

I. Introduction

With the proliferation of wireless LANs (WLAN) technologies, wireless multimedia streaming has started to emerge in homes, campuses, meeting venues, training facilities, etc. Contention-based WLAN solutions such as CSMA/CA mechanisms used in 802.11a/b/g and 802.11e Enhanced Distributed Channel Access (EDCA) MAC are especially attractive for multiuser multimedia streaming. However, transmitting multimedia in real-time over contention-based WLANs is challenging since the video applications have no time guarantee for accessing the channel [6], [7]. In this letter, we focus on how the cross layer MAC and application strategies can be optimized when multiple users are concurrently deploying real-time prioritized video streaming applications over a contention-based CSMA/CA (single hop) wireless network using the 802.11e EDCA WLAN standard.

IEEE 802.11e is a QoS-supported extension to the legacy 802.11 MAC protocols [3]. Past research works on multiuser multimedia transmissions over WLANs have mostly focused on contentionless MAC protocols. To enable cross layer optimization in contention-based EDCA MAC, accurate analytical models need to be constructed. In [2], the performance of EDCA is analyzed under saturated scenarios, yielding infinite packet delays. In [9], upper and lower bounds of the EDCA queueing delay are derived. In [7], the queueing delay is calculated for the case when each user carries traffic in only one access category (AC) and when the differentiated channel accesses of users of different ACs are neglected. In [6], various MAC adaptation schemes are proposed for real-time applications. In [1], the presented error protection method provides QoS for layered coded video by retry limit adaptation. Summarizing, no existing analysis can predict when congestion will occur in EDCA when prioritized video is transmitted. We perform such an analysis and, based on it, propose a cross layer mechanism for optimizing delay-sensitive multiuser streaming over contention-based WLANs. Our solution focuses on the transmission of prioritized video streams. In Section II, we formalize the transmission problem, while in Section III, we present the corresponding cross layer analysis and propose a distributed cross layer solution. Simulation results are presented in Section IV, while conclusions are drawn in Section V.

Manuscript received August 08, 2008; revised November 05, 2008. Current version published February 17, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lap-Pui Chau.

The authors are with the Electrical Engineering Department, University of California Los Angeles, Los Angeles, CA 90095 USA (e-mail: jlhsu@ee.ucla.edu; mihaela@ee.ucla.edu).

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

Digital Object Identifier 10.1109/LSP.2008.2010821

II. Multiuser Video Transmission in EDCA Networks

A. System Considerations

We consider a set of distinct network users V, each user v carrying a layered video streaming application, potentially along with other traffic such as voice and data. The layered coding techniques produce a number of quality layers, or sub-flows [8]. We denote the number of sub-flows by N and the kth sub-flow of v by f_{vk}. Each sub-flow is characterized by the source rate, the delay deadline d_{vk}, and the priority level k. We label the sub-flows in ascending priority order. We assume that the packets within each sub-flow have the same delay deadline. The generated video is queued per sub-flow and labeled with their delay and priority information. The packets are then passed to the EDCA and mapped to one of the four ACs. EDCA [3] allows traffic prioritization that can be used to provide graceful degradation in a congested multiuser setting. It specifies several differentiated MAC parameters for each AC: the contention window, retry limit, and inter-frame spacing.

We use a nonpreemptive priority queuing structure to model the EDCA packet scheduler within each video sender. The service time accounts for the medium access time of a head-of-line (HOL) packet. As in [1], for video user v, we assume that the video packet arrivals at layer k are characterized by an i.i.d. inter-arrival time distribution $a_{vk}(t)$, where t is in seconds. We assume that $a_{vk}(t)$ is Poisson distributed with intensity equal to R_{vk}/h_{vk}, where R_{vk} is the video source rate and h_{vk} is the average packet length. The Poisson traffic model has been used in [1], among many other papers, to model the multimedia traffic. We denote the service time distribution by $b_{vk}(t)$ and its ith moment by $\beta_{vk}^{(i)}$. Denote the ith moment of $a_{vk}(t)$ by $\alpha_{vk}^{(i)}$. The M/G/1 queuing system of video user v has a utilization factor ρ_v given by $\rho_v = \sum_{k=1}^{N} {\beta_{vk}}^{(1)}/{\alpha_{vk}}^{(1)}$. Note that ρ_v must be

less than 1 for the system to be unsaturated. If ρ_v is larger than 1, there exists a layer q_v such that

$$q_v = \sup\left\{ p : \sum_{k=p}^{N} \frac{\beta_{vk}{}^{(1)}}{\alpha_{vk}{}^{(1)}} \geq 1 \right\}. \quad (1)$$

In this case, the packets of layers $1, \ldots, \max(q_v - 1, 1)$ will experience infinite delay and packets of layer q are partially served. In fact, the transmissions made by the lower layers $1, \ldots, q_v$ will further worsen the video quality by increasing the experienced delay of the higher priority layers. Thus, the transmission rates should not exceed the network capacity.

B. Problem Formulation

We now formulate the cross layer optimization problem. The probability that a sub-flow f_{vk} packet is not received successfully at the video receiver is identified by P_{vk}, which will be computed in the next section. The objective of each video user v is to maximize its video quality utility U, which is a nondecreasing function of its effectively received video rate vector. This rate can be computed as the element wise product of the transmission source rate $\boldsymbol{R_v} = [R_{vk}]_{k=1,\ldots,N}$ and the probability of successfully receiving the packets $(\mathbf{1} - \boldsymbol{P_v})$. We formulate the cross layer optimization problem for user v as follows:

$$\max_{L_v} U(\boldsymbol{R_v}(\mathbf{1} - \boldsymbol{P_v}(\boldsymbol{R_v}, \boldsymbol{L_v})), \ \boldsymbol{P_v} = [P_{vk}]_{k=1,\ldots,N} \quad (2)$$

where $\boldsymbol{P_v}(\cdot)$ denotes the functional dependence and $\boldsymbol{L_v} = [L_{vk}]_{k=1,\ldots,N}$ denotes the retry limit at the different ACs employed by user v. The dependence of $\boldsymbol{P_v}(\cdot)$ on other users' activities is not expressed in (2) for writing simplicity. Intuitively, adopting a large retry limit increases the packet transmission. However, the packet service time increases under highly loaded network scenario, thereby leading to an increased experienced delay, eventually causing the queue to overflow. There exists an optimal vector $\boldsymbol{L_v}$ that produces a low $\boldsymbol{P_v}$ and a maximum received video quality.

III. Cross Layer Analysis and Adaptive Retry Mechanism

A. Cross Layer Analysis

Assume that V video users contend for the channel. Assuming stationary, the probability of user v starting to transmit at an arbitrary slot is denoted by τ_v and its probability of observing a busy channel by p_v, $v = 1, \ldots, V$. We obtain

$$p_v = 1 - \prod_{i=1, i\neq v}^{V} (1 - \tau_i) \quad (3)$$

where τ_v represents the combinatory effect of multiple ACs' transmission probabilities at a backoff slot (denoted by τ_{vk}). We approximate τ_v by the weighted sum expression in (4) at the bottom of the page, where q_v is defined in (1). After serving a packet, if v's queue is not empty, the HOL packet (of the highest priority non-empty queue) proceeds to backoff. The backoff process can be modeled based on a two-dimensional Markov chain, where the state transitions characterize the behavior of the backoff instance perceived by the shared channel. Such a backoff Markov chain can be identified for each class-k flow of each user v. A state (j, n) indicates the state in which the backoff instance is in its jth backoff stage and has a backoff counter equal to n. However, $AIFS$ differentiation causes the shared channel to be perceived differently by different ACs, and thus, the backoff processes are not the same for each AC. To account for the higher channel blockage probability induced by the $AIFSs$ for the lower priority class packets, a probability p^*_{vk} is introduced to approximate the equivalent "blocked-from-backoff" probability seen by the backoff instance, at a generic backoff slot. The latter describes the effect of the virtual busy channel period [i.e., the extra guarding slots $(A_k - A_N)$ following the actual busy channel period and the shortest spacing $AIFS_N$, where $AIFS_k = SIFS + A_k T_e$, and A_k is an integer). At each state (whose backoff counter is not equal to 0] of the backoff Markov chain, the transition probability of staying at the same state is equal to p^*_{vk} [9]. Its value is approximated by us to be given by $p^*_{vk} = \min[1, (A_k - A_N)p_v]$.

The next step is to obtain the steady-state probabilities s_{jn} of state (j, n). The calculations are omitted here ,and only the results are presented. After the steady-state probabilities are obtained, the probability of a class-k packet of user v transmitting in a slot during its backoff session (τ_{vk}) can be represented as the summation of the $(j, 0)$ steady-state probabilities, $j = 0, \ldots, L_{vk}$. Let cw_{kj} denote the contention window assumed by a sub-flow f_{vk} packet in the jth backoff stage. The state probabilities and the resulting transmit probabilities can be determined as

$$s_{0,0}{}^{-1} = \sum_{j=0}^{L_{vk}} \left(1 + \frac{(1 - p^*_{vk})^{-1} CW_{kj}(CW_{kj} - 1)}{2} \right) \cdot {p_v}^j$$

$$\tau_{vk} = \sum_{j=0}^{L_{vk}} s_{j,0} = \frac{s_{0,0}\left(1 - {p_v}^{L_{vk}+1}\right)}{(1 - p_v)}. \quad (5)$$

The values of τ_{vk}, τ_v, and p_v can then be jointly solved by (3)–(5), provided that the mean inter-arrival time $\alpha^{(1)}_{vk}$ and service time $\beta^{(1)}_{vk}$ in (4) are known. The former information can be acquired either at the application (video source) or at the MAC control unit by monitoring such statistics. The first two moments $\beta^{(1)}_{vk}$ and $\beta^{(2)}_{vk}$ are obtained in the Appendix .

Having obtained the service time, the medium contention, and the packet collision probabilities from the previous section, we now calculate the waiting time and delay characteristics.

$$\tau_v = \begin{cases} \sum_{k=1}^{N} \left(\frac{\beta_{vk}{}^{(1)}}{\alpha_{vk}{}^{(1)}} \right) \tau_{vk}, & \text{if } \rho_v < 1 \\ \sum_{k=q_v+1}^{N} \left(\frac{\beta_{vk}{}^{(1)}}{\alpha_{vk}{}^{(1)}} \right) \tau_{vk} + \left(1 - \sum_{k=q_v+1}^{N} \frac{\beta_{vk}{}^{(1)}}{\alpha_{vk}{}^{(1)}} \right) \tau_{q_v}, & \text{if } \rho_v \geq 1 \end{cases} \quad (4)$$

The mean waiting time can be calculated using M/G/1 priority queuing [4], given by (6) at the bottom of the page. When ρ_v is larger than or equal to 1, the mean waiting time is finite only for the higher ACs.

When the system is saturated ($\rho_v \geq 1$), the waiting time for the nonsaturated higher priority layer packets contains an additional term that is dependent on the service time of the partially served layer q. Hence, a user should prevent layer $1, \ldots, q$ from being transmitted to improve waiting time performance. The tail probability of the waiting time can then be approximated to be exponential [5] as follows:

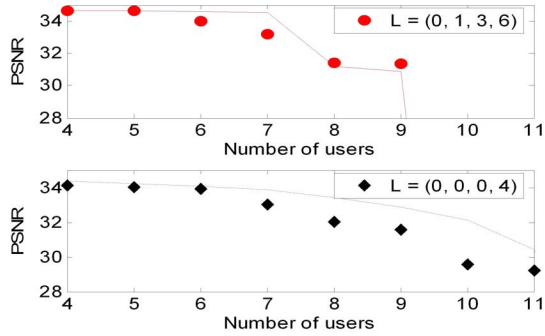
$$P(W_{vk} > t) \approx \rho_v \exp\left(-\frac{\rho_v}{E[W_{vk}]}t\right). \quad (7)$$

The packet loss probability P_{vk} can be expressed as function of two terms: the probability P_{vk}^{drop} that the packet is dropped at the MAC because its retry limit is exceeded, and the probability P_{vk}^{late} that the packet is received successfully, but it misses its delay deadline. Its value is computed as follows:

$$P_{vk} = 1 - (1 - P_{vk}^{late})(1 - P_{vk}^{drop})$$
$$P_{vk}^{late} = P(\beta_{vk}^{(1)} + W_{vk} > d_{vk}) \text{ with } P_{vk}^{drop} = {p_v}^{L_{vk}+1}. \quad (8)$$

B. Verification of the Analytical Results

We use iterative numerical techniques to jointly solve the system of (3)–(5) and (11) (given in the Appendix). The existence of a fixed point is guaranteed following the Schauder's Fixed Point Theorem. For illustration, we use the following traffic setup in all tested cases. Each network user carries 3 G711 coded voice streams (200-byte packets every 20 ms) in the highest AC. In addition, a scalable coded video stream (the Forman sequence) is mapped to the middle two ACs, at 30 video packets per layer per second. The delay deadline is set to 0.533 s for both sub-flows, which is the interval of one GOP (16 frames per GOP at 30-Hz frame rate). Each user also carries 80 kbps of data at the lowest AC. Both the packet sizes of video and data are set equal to 1000 bytes. The 802.11b physical layer of 11 Mbps bit rate is employed. In Fig. 1(a), the PSNR of the video stream as the number of competing users increases is depicted for two selected retry vector configurations. The lines represent the predicted analytical results, while the markers indicate the simulation results. We note that the analytical results based on Poisson traffic assumption are able to predict the achieved PSNRs. Voice traffic is served using different retry limits, depending on the network congestion, and the number of users in the network. Data traffic is not protected in both cases, since the retry limit for the lowest AC is 0. More results on voice and data traffic are reported in next section. Fig. 1(a) also indicates that one can allow



(a)

number of users	5	8	11
optimal retransmission vector	(7, 7, 7, 7)	(0, 1, 3, 6)	(0, 0, 0, 4)

(b)

Fig. 1. (a) PSNR versus number of users. Lines and markers represent analytical and simulation results, respectively. (b) Optimal retransmission vector under selected scenarios.

graceful performance degradation by adjusting the retry limit vector dynamically and adaptively. In Fig. 1(b), the cross layer optimized retry vector is calculated under 5, 8, and 11 users. As the network contention increases, the optimal retry limits of the lower priority ACs begin to degrade. Under high contention (11 users), all retry limits except for the highest AC drop to 0. In addition, the retry limit for the highest AC degrades to 4, since a too high limit will cause unacceptable delays, and there is no point to transmit late packets.

IV. Simulation Results

We evaluate the cross layer mechanism by conducting simulation-based performance comparisons using Qualnet 4.0. We assume that the delay deadline for voice is 0.1 s and there is no delay deadline for data. For comparison, we use three other retry limit solutions: 1) Fixed short retry: the 802.11 short retry limit (= 7) is always used; 2) Fixed long retry: the 802.11 long retry limit (= 4) is always used; 3) Adaptive retry: a time stamp is recorded as a packet enters HOL of its AC. A packet is retransmitted until either the transmission is successful or its HOL delay exceeds the delay deadline.

In Fig. 2(a), the effective voice (that satisfies the delay target) and data throughput obtained per user is plotted against the number of users in the system; in Fig. 2(b), the obtained PSNR per video stream user is plotted. As the contention grows and the network becomes overloaded, by using our cross layer mechanism, the data streams are first to starve. The video quality layers

$$E[W_{vk}] = \begin{cases} W_{0k} \cdot \sum_{i=1}^{N} \frac{\beta_{vi}^{(2)}}{\alpha_{vi}^{(1)}}, & k = 1, \ldots N, \quad \rho_v < 1 \\ W_{0k} \cdot \left[\sum_{i=q+1}^{N} \frac{\beta_{vi}^{(2)}}{\alpha_{vi}^{(1)}} + \left(1 - \sum_{i=q+1}^{N} \frac{\beta_{vi}^{(1)}}{\alpha_{vi}^{(1)}}\right) \frac{\beta_{vq}{}^{(2)}}{\beta_{vq}{}^{(1)}} \right], & k = q+1, \ldots, N, \quad \rho_v \geq 1 \end{cases}$$
$$\text{with } W_{0k} = \left(2\left(1 - \sum_{i=k}^{N} \frac{\beta_{vi}{}^{(1)}}{\alpha_{vi}{}^{(1)}}\right)\left(1 - \sum_{i=k+1}^{N} \frac{\beta_{vi}{}^{(1)}}{\alpha_{vi}{}^{(1)}}\right)\right)^{-1} \text{ and } q \text{ defined in (1)} \quad (6)$$

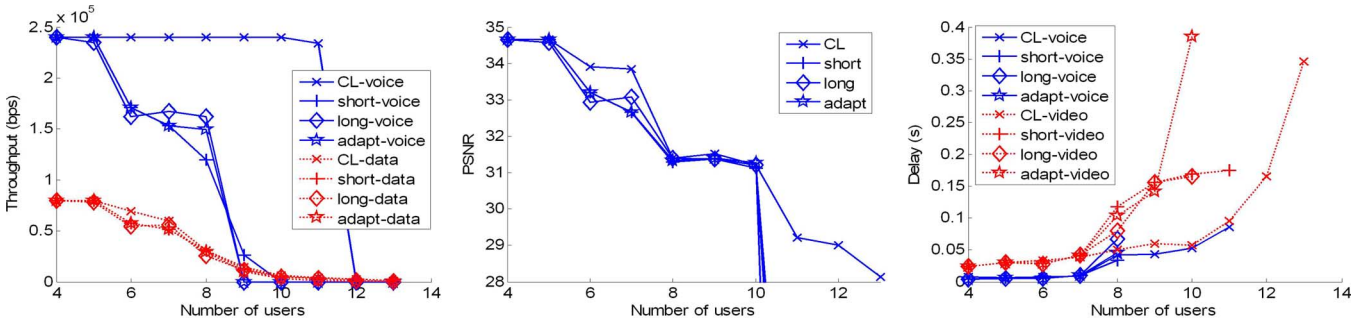


Fig. 2. Simulation results. (a) Effective throughput versus number of users for voice and data. (b) PNSR of video streaming versus number of users. (c) Delay versus number of users for voice and video traffics.

degrade gracefully as the contention continues to grow. The cross layer mechanism supports voice for as high as 11 simultaneous users. On the other hand, by using the other three policies, voice traffics fail to meet the QoS requirement as soon as increasing congestion causes data traffics to stop from transmitting. The other three policies are able to support up to ten video traffics with acceptable PSNRs, while the cross layer scheme supports up to 13 video traffics. We find that the adaptive policy, despite that it considers delay deadlines, performs similar to the nonadaptive fixed policies. The explanation is that without an effective analysis, the adaptive policy fails to consider the overall network congestion and to react in advance, and it can only adapt passively. In Fig. 2(c), the total delay performance of voice and video is plotted. The cross layer mechanism significantly enhances the performance behavior of the received video quality.

V. Conclusion

We propose a cross layer solution that allows each user to optimize its quality by adapting the 802.11e EDCA parameters. We design a mechanism that is dynamically adaptive, distributed, low-complexity, and compatible to the EDCA standard. Given network activity states and statistics, the setting of the retry limit parameter effectively balances link failure and queueing delay to optimize overall received video quality. Our simulation results show that the proposed cross layer strategy is able to maximize the utility of delay-sensitive applications in heavily congested networks.

Appendix
Computation of Service Time Distribution

The packet service time consists of two parts: the backoff and the transmission time duration. First, we define the random variable S that denotes the number of transmissions of a packet until the MAC server moves on to the next packet. It is modeled by a geometric distribution as follows:

$$\Pr\{S_{vk}=r\}=\begin{cases}(1-p_v){p_v}^{r}, & r=0,\dots L_{vk}-1\\ {p_v}^{L_{vk}}, & r=L_{vk}.\end{cases} \tag{9}$$

Given the value of S_{vk}, we set a random vector $\boldsymbol{B}=[B_i]_{i=0,\dots,S_{vk}}$, whose ith entry (B_i) represents the number of backoff slots chosen in the ith backoff stage. B_i is uniformly distributed in $[0, cw_{ji}-1]$. Let $T_m^{(i,b)}$ denote the random time duration of the bth slot in the ith backoff stage, $i=0,\dots,S_{vk}, b=1,\dots,B_i$. The random matrix $[T_m^{(i,b)}]$ is assumed to consist of i.i.d. random variables. Their outcomes can be T_e, T_c, or T_s, which, respectively, represent the slot time duration, the time period engaging a successful transmission, and that consumed by an unsuccessful transmission. Each outcome is associated with a probability equal to $(1-p_v)$, (p_v-p_s), and p_s, respectively, where p_s is given by

$$p_s=\sum_{n\neq v}\tau_n\prod_{t\neq n}(1-\tau_t). \tag{10}$$

The first two moments of the service time are given by

$$\begin{aligned}\beta_{vk}^{(1)}=&\left[T_e\cdot(1-p_v)+T_s\cdot p_s+T_c\cdot(p_v-p_s)\right]\\ &\cdot\sum_{j=0}^{L_{vk}}\frac{CW_{kj}-1}{2}{p_v}^{i}+\left[T_c\cdot p_v+T_s\cdot(1-p_v)\right]\\ &\cdot\frac{1-{p_v}^{L_{vk}}}{1-p_v}\end{aligned} \tag{11}$$

$$\begin{aligned}\beta_{vk}^{(2)}=&E[{T_m}^2]\cdot E\left[\sum_{i=0}^{S}B_i\right]+E\left[\left(\sum_{i=0}^{S}B_i\right)^2-\sum_{i=0}^{S}B_i\right]\\ &\cdot E[T_m]^2+E\left[\left(S_{vk}T_c+T_s\right)^2\right]+2E\left[\left(S_{vk}T_c+T_s\right)\right]\\ &\cdot E[T_m]E\left[\sum_{i=0}^{S}B_i\right].\end{aligned} \tag{12}$$

References

[1] 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.

[2] J. Robinson and T. Randhawa, "Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function," *IEEE J. Select. Areas Commun.*, vol. 22, no. 5, pp. 917–928, Jun. 2004.

[3] Draft Supplement to Part 11: Wireless Medium Access Control and Physical Layer Specifications: Medium Access Control Enhancements for Quality of Service, 2005, IEEE 802.11e/D13.0.

[4] Kleinrock, *Queueing Systems*. New York: Wiley, 1975, vol. 1–2.

[5] T. Jiang, C. K. Tham, and C. C. Ko, "An approximation for waiting time tail probabilities in multiclass systems," *IEEE Commun. Lett.*, vol. 5, no. 4, pp. 175–177, Apr. 2001.

[6] Y. Xiao, "QoS guarantee and provisioning at the contention-based wireless MAC layer in the IEEE 802.11e wireless LANs," *IEEE Wireless Commun.*, vol. 13, no. 1, pp. 14–21, Feb. 2006.

[7] X. Chen, H. Zhai, X. Tian, and Y. Fang, "Supporting QoS in IEEE 802.11e wireless LANs," *IEEE Trans. Wireless Commun.*, vol. 5, no. 8, pp. 2217–2227, Aug. 2006.

[8] *Multimedia Over IP and Wireless Networks*, M. van der Schaar and P. Chou, Eds. New York: Elsevier, 2007.

[9] P. E. Engelstad and O. N. Østerbø, "Queueing delay analysis of IEEE 802.11e EDCA," in *Proc. Wireless On Demand Network Systems and Services (WONS'06)*, 2006.