CROSS-LAYER RESOURCE ALLOCATION FOR DELAY-CONSTRAINED WIRELESS VIDEO TRANSMISSION

Anna Scaglione[1] *& Mihaela van der Schaar*[2]

[1] Cornell University, [2] UC Davis; School of Electrical and Computer Eng.

ABSTRACT

We consider a set of wireless stations that transmit and receive real-time video over a wireless channel. In our analysis, each video packet experiences a single flat fading state (block fading model) and additive Gaussian noise. We assume that the flat fading parameters are known at both transmitter and receiver sides and that each packet is sufficiently long such that, through error correction, it it possible to recover it correctly at the receiver end. Under an analogous model optimal rate allocation policies have been developed in an information theoretic framework. These policies maximize aggregate throughput or minimize queueing delays, but do not explicitly consider the application layer parameters in the rate allocation. We are interested in developing the framework to derive the optimum rate adaptation for video sources and evaluating the aggregate PSNR achievable with an optimal cross-layer design combining Application-MAC-PHY layers.

1. INTRODUCTION

In layered architectures multiple access consists of: the *physical layer* task of creating bit pipes (multiplexing, coding) and the *data-link layer* task of scheduling their use. To avoid cross-layer interactions the bit pipes are equalized in cost and performance. However, in a shared wireless medium the quality that can be offered depends on the traffic of signals in the medium itself and on the channel time-varying gain (fading). Equalizing the quality of the bit pipes over all possible traffic and fading conditions leads to spectrally inefficient multiplexing and modulation strategies.

Not surprisingly, at the lower layers (PHY, MAC), significant gains have been reported by adopting cross-layer optimization: we will not survey the extensive literature dedicated to the cross-layer optimization of the MAC and PHY, but rather we will focus on the criteria that stem from Shannon theory and in particular the works [4] and [1]. With a combined PHY-MAC, more users and fastest fading variations lead to greater aggregate rates and bandwidth efficiency (this phenomenon is referred to as *multi-user diversity*). The contributions in [4] and [1] are aimed at maximizing the throughput and minimizing queueing delays respectively for a given power budget, without taking into consideration the multimedia content and traffic characteristics, delay, and relative importance and dependencies among the various packets.

The advances achieved in cross-layer design at the lower layer algorithms can be further enhanced by taking into account multimedia characteristics and requirements allowing existing wireless networks to provide optimal time-varying Quality of Service (QoS) for the delay-sensitive, bandwidth-intense and loss-tolerant multimedia applications. The variability of wireless resources has considerable consequences for multimedia applications and often leads to unsatisfactory user experience due to their following characteristics: "High bandwidths - many consumer applications, e.g. High Definition TV (HDTV), require transmission bit-rates of several Mbps;" Very stringent delay constraints - delays of less than 200 milliseconds are required for interactive applications, such as video-conferencing, surveillance etc., while for multimedia streaming applications delays of 1-5s are tolerable. Packets that arrive after their display time are discarded at the receiver side or, at best, can be used for concealing subsequently received multimedia packets. Fortunately, multimedia applications can cope with a certain amount of packet losses depending on the sequence characteristics and error concealment strategies available at the receiver (e.g. packet losses up to 5% or more can be tolerated at times). Consequently, unlike file transfers, real-time multimedia applications do not require a complete insulation from packet losses, but rather require the application layer to cooperate with the lower layers to select the optimal wireless transmission strategy that maximizes the multimedia performance. For instance, as discussed in this paper, the scheduling of different video packets will be determined based on the channel condition, but also on the application layer delay constraints.

The goal of this paper is to provide a framework to determine the optimum rate-adaptation within the MAC-PHY Capacity region that maximizes the multimedia quality.

Note that the rate allocation policies we consider in can change the rates continuously within the MAC capacity region and inside of the region the communication is error free. The aim of this idealistic information theoretic analysis is to find bounds and solutions that can provide intuition on how to setup optimized practical policies, with discrete

rate vectors and finite error probabilities.

Notation: Boldface letters are vectors if lower case and matrices if capital. S denotes a set of indices and, given a vector $\mathbf{a}$, $\mathbf{a}(S)$ contains the entries of $\mathbf{a}$ corresponding the indices in S. $\mathcal{CN}(\mu, \sigma^2)$ is a complex, circularly symmetric, Gaussian distribution with mean μ and variance σ^2. The letter π indicates a permutation of a set and π_i is the ith element in the permutation. $\mathbf{0}$ and $\mathbf{1}$ are vectors with entries all equal to 0 and 1 respectively.

2. OPTIMAL MULTIPLE ACCESS

We consider a wireless multiple access noisy with I users affected by random flat fading. The Nyquist samples $y[n]$ of the complex envelope of the received signal are:

$$y[n] = \sum_{i=1}^{I} \sqrt{H_i P_i} e^{j\phi_i} s_i[n] + w_i[n] \qquad (1)$$

where H_i denotes the random flat fading power gain (which in the following will be assumed to have p.d.f. $p_{H_i}(h) = e^{-h} \;\; h > 0$ (Rayleigh fading), $s_i[n]$ is the user code, such that $E\{s_i^2[n]\} = 1$, ϕ_i is the carrier phase offset, P_i is the transmit power and, denoting by W the signal bandwidth, $w_i[n] \sim \mathcal{CN}(0, N_0 W)$ is the complex Gaussian circularly symmetric noise sample. The fading H_i varies over time due to mobility but its variations are quite slow compared to the transmission rate. Thus, H_i can be considered a constant over a long block of data, whose length is approximately equal to W/f_d, where f_d is the Doppler frequency[1].

The rate assignment can ideally be updated for *every channel use*, i.e. the service time service is $T = 1/W$ equal to the one symbol period $s_i[n]$ (the Nyquist rate is W). However, since channel fading coefficients $\sqrt{H_i}$ are highly correlated during the coherence time $1/f_d$ we will assume that the service time is $T = 1/f_d$, updating the optimum vector of rates $\mathbf{r}$ much less frequently. Also the channel states can be assumed to be nearly independent from one service time to the next. The bits transmitted during the service time are $w_i = T r_i$ (*MAC packet* size) for the ith user where $r_i = \{\mathbf{r}\}_i$ is the rate allocated for that user. For channel models such as the one in (1) the MAC Capacity region has been derived at under the classical assumption of infinitely backlogged systems, where users continue to produce data at a constant rate. Specifically, Tse and Hanly in [4] studied the capacity region of wireless channels such as the one in (1) when the channel parameters are perfectly known at both transmitter and receiver sides.

[1]The Doppler frequency $f_d = v/\lambda$ the wavelength divided by the maximum velocity of the mobile (the transmitter/receiver or any scatterer in the environment)

2.1. MAC with optimum aggregate rate

Under individual power constraints for the users [4] indicates how to find the power control and rate allocation policies that maximize the weighted sum of the users rates, by exploiting the *polymatroid structure of the capacity region*. Specifically, for a given set of channel states H_i and powers $P_i \; i = 1, \ldots, I$, the capacity region is:

$$\mathcal{C}_g(\mathbf{h}, \mathbf{p}) = \left\{ \mathbf{r} : \mathbf{r}(S) \leq W \log\left(1 + \frac{\sum_{i \in S} H_i P_i}{N_0 W}\right); \forall S \in \{1, \ldots, I\} \right\}, \qquad (2)$$

which is characterized by $2^I - 1$ constraints, i.e. all non empty subsets S. It is also known that in the boundary of the region $\mathcal{C}_g(\mathbf{h}, \mathbf{p})$ there are $I!$ vertices in the positive quadrant. Each of the vertexes is achievable by successive decoding at the receiver using one of the $I!$ possible orderings π of the user indexes $\{1, \ldots, I\}$; i.e. the rate assignment in the vertex corresponding to a specific permutation π is:

$$\begin{aligned} r_{\pi_1} &\leq W \log(1 + H_{\pi_1} P_{\pi_1}) \qquad (3) \\ &\ldots \\ r_{\pi_I} &\leq W \log\left(1 + \frac{H_{\pi_I} P_{\pi_I}}{\sum_{i=1}^{I-1} H_{\pi_i} P_{\pi_i} + N_0 W}\right) \end{aligned}$$

Because $\mathcal{C}_g(\mathbf{h}, \mathbf{p})$ in (2) is a polymatroid [4], the solution of the linear programming problem:

$$\max_{\mathbf{r}} \boldsymbol{\lambda}\mathbf{r} \;\; \text{subject to} \;\; \mathbf{r} \in \mathcal{C}_g(\mathbf{h}, \mathbf{p}) \qquad (4)$$

is given by the vertex of $\mathcal{C}_g(\mathbf{h}, \mathbf{p})$ among the $I!$ that corresponds to the same $\pi : \lambda_{\pi_1} \geq \lambda_{\pi_2} \ldots \geq \lambda_{\pi_I}$. In words, the optimum rate assignment for a given set of weights $\boldsymbol{\lambda}$ is equal to a *greedy* iterative solution obtained by sorting all the rates in the same order as the descending order of the weights $\boldsymbol{\lambda}$, setting them initially to zero and increasing them one by one in order until the constraint $\mathcal{C}_g(\mathbf{h}, \mathbf{p})$ becomes tight. The greedy algorithm leads to place $\mathbf{r}$ in the vertex in (3). The optimum rate allocation and power control policy can be found exploiting the same result:

$$\max_{\mathbf{r},\mathbf{p}} \;\; \boldsymbol{\lambda}\mathbf{r} - \boldsymbol{\mu}\mathbf{p} \;\; \text{subject to} \;\; \mathbf{r} \in \mathcal{C}_g(\mathbf{h}, \mathbf{p}) \qquad (5)$$

$$\max_{\mathbf{p}} \;\; -\sum_{i=1}^{I} \mu_{\pi_i} P_{\pi_i} + \lambda_{\pi_1} W \log(1 + H_{\pi_1} P_{\pi_1}) \qquad (6)$$

$$+ \;\; \sum_{i=2}^{I} \lambda_{\pi_i} W \log\left(1 + \frac{H_{\pi_i} P_{\pi_i}}{\sum_{k=1}^{i-1} H_{\pi_k} P_{\pi_k} + N_0 W}\right)$$

To identify the optimum power control policy, [4] introduced a marginal user utility function:

$$u_i(z) = \frac{\lambda_i}{z + N_0 W} - \mu_i; \qquad (7)$$

by noting that:

$$\lambda \log\left(1+\frac{a}{\sigma^2+b}\right)-\mu a=\int_b^{a+b}\left(\frac{1}{z+\sigma^2}-\mu\right)dz \quad (8)$$

hence the optimization above can be seen as follows

$$\max_{\mathbf{p}}\int_0^{q_{\pi_1}} u_{\pi_1}(z)dz+\sum_{i=2}^{I}\int_{q_{\pi_{i-1}}}^{q_{\pi_i}} u_{\pi_i}(z)dz \quad (9)$$

where, for $i=1,\ldots,I$:

$$q_{\pi_1}=H_{\pi_1}P_{\pi_1};\;\; q_{\pi_i}=q_{\pi_{i-1}}+H_{\pi_i}P_{\pi_i} \quad (10)$$

Hence, if there exist a z_1 such that $u_{\pi_2}(z)\geq u_{\pi_1}(z)$ for $z\geq z_1$ then $q_{\pi_1}=z_1$, if there exist a z_2 such that $u_{\pi_3}(z)\geq u_{\pi_2}(z)$, then $q_{\pi_2}=z_2-q_{\pi_1}$ and so on, always selecting the edges of the intervals such that the $\max_i[u_i(z)]$ is the function integrated. This framework includes and explains, as a special case, the result shown in [2] that is, for wireless users experiencing fading H_i with equally distributed statistics over time and equal power constraint, the maximum aggregate capacity (which corresponds to $\boldsymbol{\lambda}=\mathbf{1}$) is achieved by *time sharing* among the users that have the best instantaneous channel condition. The reason is due to the fact that all $u_i(z)$ are equal in this case and there is no incentive in changing user if the users are ordered such that $H_{\pi_1}\geq H_{\pi_2}\geq\ldots\geq H_{\pi_I}$. This relatively simple optimum time sharing solution has been named Opportunistic Multiple Access [2]. The maximization of the aggregate rate in wireless channels leads to the so called *multi-user diversity* gain. In a large pool of users whose links are independently faded it becomes statistically more and more likely to have users with excellent channel conditions. As a result, the aggregate rate scales positively with the number of users. Needless to say, it leads to increasing latency for the users and respective application layers.

The question we ask is: 1) how effective are these protocols when users packet's queues are not equally backlogged with jobs and 2) what type of joint application layer adaptation, queueing discipline and MAC should be applied such that the multimedia quality is maximized?

2.2. Minimum delay MAC in packet switched networks

The first question posed at the end of the previous section was considered by [3] and, more recently in [1]. These works found an optimum policy whose name is self-explanatory: Longest Queue Highest Possible Rate (LQHPR).

Let $q_i[n]$, $i=1,\ldots,I$ be the number of bits in left in the ith user queue at time n and let $\mathbf{q}_n$ be the vector of such queue states (the vector whose entries are $q_i[n]$, $i=1,\ldots,I$). Let $a_i[n]$ be the number of packets that arrives during the nth service time S. We can write:

$$\mathbf{q}_{n+1}=\mathbf{a}_n+(\mathbf{q}_n-T\mathbf{r}_n)^+. \quad (11)$$

where $(a)^+$ stands for $\max(a,0)$ (if a is a vector or matrix the definition is applied to each of its entries). The main result in [1] is that the average length of the queues $E\{\sum_{i=1}^{I} q_i[n]\}$ is minimized if

$$\max_{\mathbf{r},\mathbf{p}}(\mathbf{q}_n^T\mathbf{r}-\boldsymbol{\mu}\mathbf{p}) \;\text{ subject to }\; \mathbf{r}\in\mathcal{C}(\mathbf{h},\mathbf{p}) \quad (12)$$

i.e. if the weights $\boldsymbol{\lambda}$ in (5) are chosen to be equal to the user's queue states, i.e. $\boldsymbol{\lambda}=\mathbf{q}_n$. The assumptions are that the users channel states have statistics that are symmetric or interchangeable, that the packet arrival processes are independent identical Poisson processes and that the random fading coefficients vary also as a Markov process. According to [1], the LQHPR policy leads the system to have the highest stable throughput (i.e. the queues do not blow in size) and the shortest average queue length per user. By mean of Little's law, [1] argues that the LQHPR policy leads also to the shortest average delay.

2.3. Cross-layer video rate-adaptation

To enable the optimized adaptation of multimedia transmission for the various users, we need to quantify the rate-distortion function for all the video users in a given service time. To do this, it is helpful to perform a differential treatment of the user packets by introducing the concept of *sub-flow*: a video flow (bitstream) is divided into N sub-flows to which a fraction $0\leq\alpha_{i,n}\leq 1$ of the total user rate r_i is assigned (i.e. $\sum_{n=1}^{N}\alpha_{i,n}=1$). The subdivision in sub-flows can be done using a state-of-the-art wavelet video coding scheme for wireless transmission as well as any other codec, so that each sub-flow groups frame-bits according to their delay constraints and relative contribution to the overall distortion reduction of the decoded video. Specifically, we use two distinct sub-flow definitions: 1) The first definition groups frames based on their impact on the distortion, e.g. the motion compensated wavelet decomposition frames are divided in independent sub-flows grouping the low pass temporal frames and the various levels of the high pass temporal frames; 2) The second definition groups together in a sub-flow frames with similar delay constraints τ_i.

Each sub-flow has a rate-distortion $R_i(D_i)$ value associated with it and the contribution to the overall distortion by the various sub-flows can be determined for instance, using the operational Rate-Distortion models proposed in [5] for MPEG-like coders or proposed by us in [6] for motion-compensated wavelet video coders. Alternatively, the Rate-Distortion values for the various sub-flows can be computed in real-time. Delay constraints can also be accounted for in the distortion model, associating an amount of incremental distortion equal to the zero rate distortion value $D_i(0)$ when the sub-flow delay constraint is not met. By differentiating the delay constraints, a larger number of users will be admitted to the service.

Given a certain rate allocated for the user r_i there exist an optimum partition of the rate $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,N})$ among sub-flows, such that the ith user distortion:

$$D_i(r_i) = \min_{\boldsymbol{\alpha}_i} \sum_{n=1}^{N} D_{i,n}(\alpha_{i,n} r_i) \qquad (13)$$

The objective of the cross layer controller is to minimize the total distortion over the users:

$$\min_{\mathbf{r},\mathbf{p}} \sum_{i=1}^{I} D_i(r_i) + \boldsymbol{\mu}\mathbf{p} \;\text{ subject to }\; \mathbf{r} \in \mathcal{C}(\mathbf{h}, \mathbf{p}). \qquad (14)$$

The objective of the admission control is instead to determine the maximum number of stations that can transmit their data while meeting the quality demanded by the application.

Below (Fig. 1) is a graphical example that illustrate the numerical optimization of (14) for $I = 2$ video users. In the figure it is assumed that the channel gain for user 2 is greater than the one for user 1 and that, in contrast, the total number of bits in all sub-flows produced by the video encoder of user 1 exceeds that of user 2. The rate vectors $\mathbf{r} = (r_1, r_2)$ chosen by the criterion maximizing the aggregate throughput (opportunistic MAC) and minimizing the delay (LQHPR) are the red and green vectors respectively. The blue lines represent the curves of constant aggregate distortion for the users, to be calculated solving (13). The optimum rate adaptation for video quality is the blue vector, which is at the intersection between the lowest aggregate distortion curve and the edge of the Capacity region. The

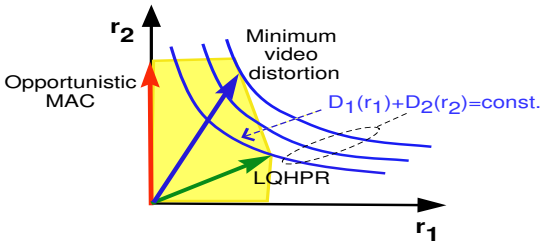

Fig. 1. A graphical example of the rate allocation for $I = 2$, $H_2 > H_1$ and $q_1[n] > q_2[n]$.

joint distortion curves depend on the type of video sent. To illustrate the difference in PSNR (distortion) among the various video sequences at different rates, we illustrate in Figure 2 (a) and (b) our PSNR results obtained using a state-of-the-art motion-compensated wavelet video coder. The sequences are at CIF resolution, 30 Hz. From the figures it is clear that if the two users would like to operate at the same video quality level (e.g. 35dB), they will need very different rates. For instance, the Foreman sequences requires only 350kbps, while Football will require almost 2000kbps.

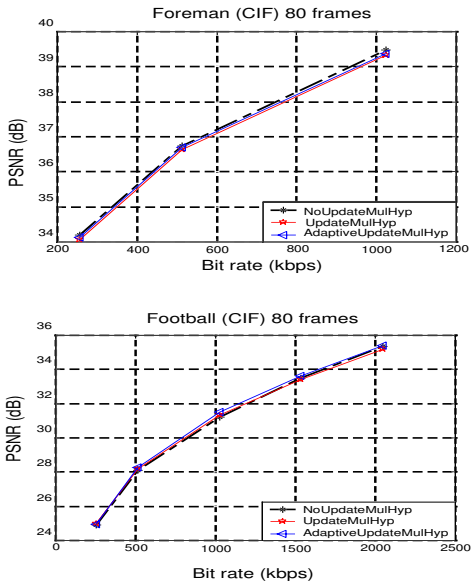

Fig. 2. PSNR vs.rate: Foreman and Football videos.

3. CONCLUSIONS

Our paper shows that information theoretic solutions for wireless MAC such as opportunistic MAC and LQHPR are sub-optimal for video performance. The optimal solution is achieved by transmitting an incrementally larger number of video sub-flows, leading to an increasing PSNR, until the capacity region is met.

4. REFERENCES

[1] Edmund M. Yeh and Aaron S. Cohen, "Delay Optimal Rate Allocation in Multiaccess Fading Communications," Proc. of the Allerton Conference. Monticello, IL, 2004.

[2] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," Proc. of ICC, (Seattle, WA), 1995.

[3] I. E. Telatar and R. Gallager, "Combining queueing theory with information theory for multiaccess," IEEE JSAC, vol. 13, no. 6, pp. 963969, 1995.

[4] D. Tse and S. Hanly, "Multi-access fading channels: Part I: Polymatroid structure, optimal resource allocation and throughput capacities," IEEE Trans. on Inf. Theory, vol. 44, no. 7, pp. 27962815, 1998.

[5] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," IEEE J. Select. Areas Commun., vol. 18, pp.1012-1032, June, 2000.

[6] M. Wang, M. van der Schaar, "Rate-Distortion Modeling for Wavelet Video Coders", to appear proc. of ICASSP 2005.